A Review of Selected Techniques in Inverse Problem Nonparametric Probability Distribution Estimation

H.T. Banks, Zackary R. Kenz, and W. Clayton Thompson
Center for Research in Scientific Computation
Center for Quantitative Sciences in Biomedicine
N.C. State University
Raleigh, NC

May 24, 2012

Abstract

Techniques for the nonparametric estimation of probability distributions are reviewed. Methods are divided into two categories for estimation problems: those applied to situations in which individual data is available, and those where only aggregate data is available. For each technique, the general ideas, strengths, and weaknesses of the corresponding methodology are discussed. In addition, the generation of estimates of not only the parameter distributions but also any structural parameters that are constant across the population is outlined. When discussing various techniques, particular consideration is given to the theory behind finite dimensional approximations (since the space of measures on a viable parameter space Θ is an infinite dimensional space) in the development of computational methodologies. Additional issues, such as consistency of the estimation procedure, are considered when appropriate. Various sample problems on which the methods can be applied are referenced throughout the review.

Key words: Inverse problems, nonparametric probability distributions, least squares, maximum likelihood, finite dimensional approximations, computational methodologies.

1 Motivation

In the mathematical modeling of physical and biological systems, the case often arises where some facet of the underlying dynamics (in the form of a parameter) is not constant but rather is distributed probabilistically within the system or across the population under study. While traditional inverse problems involve the estimation, given a set of data/observations, of a fixed parameter contained within some admissible set, models with distributed parameters require the estimation of a probability measure or distribution over the set of admissible parameters. This problem is well-known to both applied mathematicians and statisticians and both schools have developed their own set of tools for the estimation of a measure or distribution over a set. Not surprisingly, the methods developed by the two schools are best suited to the particular types of problems most frequently encountered in their respective fields. As such, the techniques for measure estimation (along with their theoretical justifications) are widely scattered through the literature in applied mathematics and statistics, often with few cross references to related ideas.

In this report we provide a survey of some the techniques from both fields, paying particular attention to the theoretical underpinnings of the various methods as well as to computational issues which result from the theory. Throughout this report, we will focus on the *nonparametric* estimation of a probability measure. That is, we want to establish a meaningful estimation problem while simultaneously placing a minimal set of restrictions on the underlying measure. The fully parametric case (that is, when the form of the underlying measure is assumed to be known and determined uniquely by a small number of parameters, for example, its mean μ and covariance matrix Σ) can be considered as a standard parameter estimation problem over Euclidean space for which computational strategies are well-known and readily available.

Not surprisingly, the applicability of various techniques for the nonparametric estimation of a probability measure is heavily dependent upon the type of data available for use with a given model in an inverse problem. In some studies, data is available at particular times for each individual in the population under study. Such might be the case, for example, in a pharmacokinetics study in which one desires to estimate drug absorption and excretion rates from blood samples taken serially in time from individuals in the population. For example, consider a dynamic pharmacokinetics model for an individual's blood concentration of the drug theophylline with solution

$$C(t, F, D, k_a, k_e, V) = \frac{k_a F D}{V(k_a - k_e)} (e^{-k_e t} - e^{-k_a t}), \quad (1)$$

where V is blood volume, D is the dose received, F is the uptake fraction, and k_a and k_e are rates of drug absorption and excretion, respectively. For sufficient data collected from a single individual, it is a standard problem in statistics (e.g., [29, Ch. 2]) and mathematics (e.g., [21, Ch. 3]) to estimate, for example, the parameters F , k_a , and k_e for V and D known. Of course, it is highly likely that individual parameters (say, k_a and/or k_e) might vary from one individual to the next within the sampled population. Thus, our goal here is to use the entire sample of individuals to estimate the probability measure describing the distribution of certain parameters in the full population.

In other investigations, a situation might arise in which one is only able to collect aggregate data. Such might be the case, for example, in catch and release experiments in which one cannot be certain of measuring the same individual multiple times. Consider the Sinko-Streifer [21, 47] model for a size-structured population given by

$$\begin{aligned} \frac{\partial n(t, x)}{\partial t} + \frac{\partial [g(t, x)n(t, x)]}{\partial x} &= -\mu(t, x)n(t, x) \\ n(0, x) &= n_0(x) \\ g(t, x_0)n(t, x_0) &= \int_{x_0}^{x_1} k(t, \xi)n(t, \xi)d\xi \\ g(t, x_1) &= 0, \end{aligned} \quad (2)$$

which describes the size-structured density (number/length) $n(t, x)$ at time t and size x with death rate $\mu(t, x)$ and fecundity kernel $k(t, x)$. In this model, all individuals with size x at time t grow with individual dynamic growth rate $\frac{dx}{dt} = g(t, x)$. For example, one might take $g(t, x) = r(\tilde{x} - x)$, with $\tilde{x}$ a maximum size and r an intrinsic rate of growth. Then the parameters r and $\tilde{x}$ are expected to vary across individuals in the population.

Unlike the pharmacokinetics example, one does not have individual growth data, but rather histograms showing the aggregate number of individuals sampled from the population at a given time and having a given size ([6] and [21, Chap 9]). The goal is to estimate the probability distributions describing the variability of the parameters r and $\tilde{x}$ across the population.

As will be shown, these two situations lead to fundamentally different estimation problems. In the first case, one attempts to use a mathematical model describing a single individual along with data collected from multiple individuals (but data in which subsets of longitudinal data points could be identified with a specific individual) in order to determine population level characteristics. In the second case, while one again has an individual model, the data collected cannot be identified with individuals and is considered to be sampled from the population. It is worth noting that special care should be taken in this latter case to identify the model (2) as an individual model in the sense that it describes an individual population—that is, all ‘individuals’ (i.e., mosquitofish [6]) described by the model share a common growth rate function $g(t, x)$. The confusing terminology is largely the result of the disparate fields in which such problems are commonly found. In pharmacokinetics, one typically discusses ‘individual’ patients which together constitute a ‘population’. In ecology, one often works with an ‘aggregate’ population which itself consists of smaller ‘individual’ subpopulations. Mathematically, here we define the ‘individual’ in terms of the underlying parameters, using ‘individual’ to describe the unit characterized by a single parameter set (excretion rate, growth rate, etc.).

In this report, we consider two generic estimation problems (for which the examples above are specific instances). First, we consider the case that one has a mathematical model for individual dynamics along with individual data (called the *individual dynamics/individual data* problem). This is the case with the pharmacokinetics example above. Second, we consider the case as in the structured density model above that one has a mathematical model for individual dynamics but only aggregate data (the *individual dynamics/aggregate data* problem). There is a third possibility—that one has only an aggregate model (i.e., the dynamics depend explicitly on a distribution of parameters across the population) with aggregate data. Such examples arise in electromagnetic models with a distribution of polarization relaxation times for molecules (e.g., [15, 16]); in biology with HIV cellular models [3, 4, 5]; and in wave propagation in viscoelastic materials such as biotissue [20]. The measure estimation problem for such examples is sufficiently similar to individual dynamics/aggregate data situation and accordingly we do not consider aggregate dynamics models as a separate case but only offer some explicit comments as warranted.

In both generic estimation problems discussed, the underlying goal is the determination of the probability measure which describes the distribution of parameters across all members of the population. Thus, two main issues are of interest. First, a sensible framework must be established for each situation so that the estimation problem is meaningful. Thus we must decide what type of information and/or estimates are desired (e.g., mean, variance, complete distribution function, etc.) and determine how these decisions will depend on the type of data available. Second, we must examine what techniques are available for the computation of such estimates. Because the space of probability measures is an infinite dimensional space, we must make some type of finite dimensional approximations so that the estimation problem is amenable to computation. Moreover, here we are interested in frameworks for approximation and convergence which are not restricted to classic parametric estimates. We shall see that, while the individual dynamics/individual data and individual dynamics/aggregate data problems are fundamentally different, there are notable similarities in the approximation techniques (namely, use of discrete measures) common to both problems.

2 Generalized (Hierarchical) Individual Model

We begin by establishing a general individual mathematical model. In this presentation we use notation corresponding to ordinary differential equations, but it should be noted that the techniques described apply equally well to partial differential equations, delay differential equations, algebraic systems, etc. Assume the population under study consists of i individuals, $1 \leq i \leq N$. The dynamics for the i^{th} individual are assumed to be described by the mathematical model

$$\frac{dy_i}{dt} = g_i(t, y_i(t); \theta_i, \psi). \quad (3)$$

In order to simplify notation we assume y is scalar-valued. (The results presented are readily generalized if y is vector-valued.) The parameters θ_i are specific to each individual (e.g., rates of drug clearance, growth rates, etc.) and $\theta_i \in \Theta$ for all i where Θ is some admissible parameter set. The parameter $\psi \in \Psi$ is a parameter shared by all individuals in the population (e.g., environmental factors). In general, it is possible to have $\psi = \psi(t)$ but we will not consider that additional complexity here. The model solution will be represented by

$$y_i(t) = f_i(t; \theta_i, \psi). \quad (4)$$

We will assume $g_i = g$ for all i and hence $f_i = f$ for all i . This condition is almost always met in practice and thus the assumption is not particularly restrictive. Traditional inverse problems, focusing on a single model and data from a single subject, entail use of data to determine the parameters of the mathematical model (θ_i, ψ) which best fit the data in an appropriate sense. In this report, however, we are less interested in the individual parameters θ_i and more interested in the probability measure P which describes the distribution of these parameters across the total population. This probability distribution must be estimated simultaneously with the estimation of the parameter ψ which is common to all individuals in the population. The interplay between individual data and population level characteristics makes hierarchical modeling a natural framework in which to work.

Hierarchical modeling relates individual and population level parameters to individual characteristics of interest by establishing relationships between parameters at multiple stages. The first stage establishes a mathematical model for each individual in the population being studied, as well as a statistical model (to be discussed) to account for intraindividual variability in the data. The first stage mathematical model is given by Equation (4). In the second stage, fixed and random effects are identified which link individual and population level parameters, and a statistical model accounts for interindividual variability. Thus, for the individual parameters θ_i , we might have the mixed (i.e., both fixed and random) effects model

$$\theta_i = h_i(\phi, \varphi_i, \alpha_i), \quad (5)$$

where ϕ is a vector of fixed effects (i.e., effects common to all individuals in the population), φ_i is a vector of random effects (i.e., effects which vary from individual to individual in the population), and α_i is a (known) parameter in the mixed effects model for individual i . It is commonly assumed that $E[\varphi_i] = 0$ for all i . These assumptions are not absolutely necessary, but are not particularly restrictive since one can always redefine the functions h_i so that the condition holds. As an example, one frequently encounters linear mixed effects models

$$h(\phi, \varphi_i, \alpha_i) = A_i \phi + B_i \varphi_i \quad (6)$$

where $\alpha_i = (A_i, B_i)$, with ‘design matrices’ A_i and B_i (which are either fixed or known, but not estimated). It is common (and not overly restrictive [29, Sec. 5.3]) to assume B_i is the identity for all i . Finally, it is assumed $\text{Cov}(\varphi_i) = D$ for all i . A very general discussion of hierarchical modeling along with numerous examples can be found in [29, Chs. 3-4].

We remark that care should be taken to distinguish between the parameter ψ and the parameter ϕ . While both parameters are common to all members of the population, the parameter ψ appears directly in the individual model (4) and must be estimated regardless of which technique (see below) is chosen. The parameter ϕ is a part of the second stage model and is not considered in the event the parameters θ_i are estimated directly. As an example, consider the analog of (6) when $\theta_i \in \mathbb{R}$ (so that A_i and B_i are scalars). In a pharmacokinetics study, we might assume that individual absorption rates θ_i are directly proportional to body weight, but with some random (additive) variation between individuals. Hence we would have $\theta_i = A_i \phi + \varphi_i$ where A_i is the (known) weight of the i^{th} individual. The estimation problem is now to determine the rate constant ϕ as well as the distribution of the random effects φ_i . Since it is assumed that $E[\varphi_i] = 0$, it is useful (for linear mixed effects models) to think of ϕ as a kind of expected value of the parameter θ_i , with φ_i characterizing individual variability around this expectation. (Of course, in the simplest case, $\theta_i = \phi + \varphi_i$, and we have exactly $E[\theta_i] = \phi$, and $\text{Cov}(\theta_i) = \text{Cov}(\varphi_i) = D$.)

As has been mentioned several times, the ultimate goal of the inverse problems to be discussed in this report is to determine the probability distribution P describing the variability of the parameters θ_i in the population. To that end, the specification of the second stage (5) in the hierarchical model is not strictly necessary. In general, there is little difference between determining the distribution from which the θ_i are drawn or the distribution

from which the random effects φ_i are drawn as the knowledge of one can be used to draw conclusions regarding the other (particularly for the linear mixed effects model (6)). As will be seen below, some techniques in the mathematical and statistical literature (e.g., the *Nonparametric Maximum Likelihood* (NPML) of Section 3.1.4, the *Prohorov Metric Framework* (PMF) of Section 4.1) are designed for the nonparametric estimation of such distributions (for either θ_i or φ_i) without regard for which stage in the hierarchy is being estimated. On the other hand, some methods (e.g., the *global two stage* (GTS) method of Section 3.1.2 or the *first order* (FO) method of Section 3.1.3) are explicitly premised upon the full specification of the hierarchical model.

3 Individual Data

We now discuss estimation techniques when data has been collected for the state of an individual member of the population under study (and of course, when one has multiple such sets of data, generally one data set for each of a subset of individuals in the population). A common example is a pharmacokinetics study in which one has a mathematical model (e.g. (1)) describing renal absorption and clearance of a drug, and one or more (at distinct time points) concentration measurements for each individual in the study. Such situations have been commonly investigated in the statistical literature (see, e.g., [30] and the references therein) and many of the tools described in this section were originally developed by statisticians.

Data is collected from individual i at times $\{t_{ij}\}$, $1 \leq j \leq m_i$, where m_i is the number of data points available for individual i . Following standard inverse problem procedure [21, Ch. 3] the data can be considered as a random variable

$$Y_{ij} = f(t_{ij}; \theta_i, \psi) + \mathcal{E}_{ij} \quad (7)$$

where the $\mathcal{E}_{ij}$ are random variables representing error and/or deviation of the data from the model predictions. This is the statistical model corresponding to the mathematical model (3)-(4).

For the estimation problem, we must in some way establish (and/or assume) the statistical properties of the error random variables $\mathcal{E}_{ij}$. There are two distinct possibilities for doing so. One possibility is to specify a conditional probability density function which describes the likelihood of an observation (or set of observations) given a particular set of parameters for the individual mathematical model. This is typically done in the context of maximum likelihood estimation, but constitutes very strong assumptions on the form of the error random variables that are often difficult if not impossible to verify in practice. Alternatively, least squares estimation requires only that the first two statistical moments of these error random variables be specified. This is less restrictive than a complete likelihood specification, and is easier to at least partially verify by residual analysis [21, Ch. 3]. Moreover, least squares estimation is typically more robust to misspecification of the error properties [29, Ch. 2]. However, as we will see, there are some advantages in specifying the likelihood densities (when done correctly). In general, one has a trade off: less restrictive requirements lead to less powerful results, and conversely. For now, we specify only the first two moments of the error random variables. A likelihood formulation is revisited in Section 3.1.4 where it is necessary.

The errors $\mathcal{E}_{ij}$ could be caused by a number of sources [29, 30], such as misspecifications and approximations made in obtaining the mathematical model, error and bias in the experimental measurement procedure itself, etc. Errors resulting from model misspecification, while important, are the subject of model selection criteria [11, 12, 24] and we do not discuss such topics here. Rather, we will assume (as is standard in much of the statistical literature) that the mathematical model has been carefully chosen and is an accurate description of the process under study. Thus, we make the assumption $E[\mathcal{E}_{ij}] = 0$ for all i, j and hence $E[Y_{ij}] = E[Y_{ij}|\theta_i, \psi] = f(t_{ij}; \theta_i, \psi)$. To further describe the error random variables, define

$$\vec{\mathcal{E}}_i = \begin{pmatrix} \mathcal{E}_{i1} \\ \vdots \\ \mathcal{E}_{im_i} \end{pmatrix}.$$

Then the covariance model for the individual data random variables is assumed to be described by

$$\text{Cov}(\vec{\mathcal{E}}_i|\theta_i, \psi) = R_i(\theta_i, \psi, \zeta_i), \quad (8)$$

where ζ_i is an additional set of parameters needed to describe the covariance and may be known or estimated. As above, it is typical to assume $R_i = R$ for all i . The data itself is considered to be a realization of the random variables Y_{ij} ,

$$y_{ij} = f(t_{ij}; \theta_i, \psi) + \epsilon_{ij}, \quad (9)$$

where ϵ_{ij} is a realization of the random variable $\mathcal{E}_{ij}$.

At this point, a number of assumptions have been made regarding the admissible forms of the mathematical and statistical models for the individual data. We pause briefly to indicate the wide applicability of the assumptions made so far. As mentioned above, the assumption that the error random variables have zero expectation (and hence, that the mathematical model is correctly specified) is standard in parameter estimation problems [7]. The form of the covariance model, while opaque in its general form in Equation (8) above, contains several common examples of interest. If the errors $\mathcal{E}_{ij}$ are independent, then the matrix R is diagonal. Additionally, in many parameter estimation problems, one assumes the individual errors $\mathcal{E}_{ij}$ are identically distributed with some constant variance, $\text{Var}(\mathcal{E}_{ij}) = \sigma_i^2$ for all j . Given the assumptions of independence and constant variance, the constant variance (CV) statistical model is

$$R(\theta_i, \psi, \zeta_i) = \sigma_i^2 I_{m_i} \quad (10)$$

where the variance parameter $\zeta_i = \sigma_i$ and I_{m_i} is the $m_i \times m_i$ identity matrix. Alternatively, there are many cases in which the magnitude of the error variance depends upon the magnitude of the observation. In such a situation, it is common to assume independent errors with a constant coefficient of variance, $\text{Var}(\mathcal{E}_{ij}) = \sigma_i^2 f^2(t_{ij}; \theta_i, \psi)$ for all j , and the constant coefficient of variance (CCV) statistical model is

$$R(\theta_i, \psi, \zeta_i) = \sigma_i^2 \left(\sum_j f^2(t_{ij}; \theta_i, \psi) \mathbf{e}_j \mathbf{e}_j^T \right) \quad (11)$$

where $\mathbf{e}_j$ is the j^{th} unit basis vector in $\mathbb{R}^{m_i}$ and again $\zeta_i = \sigma_i$. (That is, R is a diagonal matrix with entries $\sigma_i^2 f^2(t_{ij}; \theta_i, \psi)$, $1 \leq j \leq m_i$.) We remark that, just as was discussed above, for both statistical models (10) and (11), assumptions have been made regarding the first two moments of the (unknown) probability distribution from which the random variables $\mathcal{E}_{ij}$ are drawn, but no assumption has been made regarding the particular parametric family (e.g., normal) of the distribution. When maximum likelihood estimation is employed, one must make a further assumption to obtain a density function describing the probability of each observation (given the model solution). Significantly, the stronger likelihood assumptions allow for the possibility of proving the consistency of the maximum likelihood estimator [49, 50]. In several cases, least squares estimation and maximum likelihood estimation solve the same set of estimating equations, and thus can be considered equivalent [21, 26]. Given the robust properties of least squares estimators [29, Ch. 2] we prefer least squares estimation whenever possible.

3.1 Estimation Techniques

We now turn to consider various estimation techniques for individual data. Given the two stages (4) and (5) of the hierarchical model discussed above, it is obvious that there are two choices for characterizing the distribution of parameters in the total population. First, one could try to characterize the parameters θ_i (or their distribution) directly. Alternatively, one could attempt to determine the parameters ϕ and φ_i in the second stage of the hierarchical model and then use those parameters to infer information regarding the θ_i values. In general, this choice will have direct consequences for the computational technique used for the estimation problem, and will depend on the amount and type of data available.

In this section, we provide a brief overview of existing techniques for the estimation of population level characteristics of a parameter distribution. Many of these techniques are reviewed at length and with extensive references and historical context in [29, 30, 31, 46]. We do not go into great detail here except to provide an overview of more basic methods along with their assumptions, applicability, advantages, and shortfalls. We also do not consider fully parametric methods (e.g., $\theta_i \sim \mathcal{N}(\mu, \Sigma)$ for all i) as these are sufficiently similar to standard inverse problems. Many of the results presented below ignore the parameter ϕ (fixed for all individuals in the population) in the estimation problem. This is purely for notational convenience, given the goal of this

presentation to review techniques for the nonparametric estimation of measures. The estimation procedures discussed can be adapted to include the parameters ϕ and ψ and comments are included when warranted.

We begin with methods that do not actually estimate the underlying population probability distribution itself, but rather estimate its statistical moments. The naive methods of Section 3.1.1 focus on estimating only the population expectation or mean for the distribution from which the θ_i are drawn. Two-stage methods (Section 3.1.2) and methods based upon linearization (Section 3.1.3) estimate both the expectation and variance of the distribution for the parameter θ_i . While these methods do not provide estimates of the probability distribution itself, they can provide meaningful statistical details which may be sufficient in many applications of interest. Significantly, the methods discussed are not dependent upon a distributional form (e.g., normal) of the probability distribution and thus are more general than traditional parametric methods. Finally, we establish a stronger set of assumptions on the error random variables to form an estimate of the population distribution directly.

3.1.1 Naive Methods

One of the oldest techniques for treating data from multiple individuals is the *Naive Averaged Data* (NAD) technique [31]. In the event data for each individual was taken at the same set of time points (say, as part of a controlled experiment) so that $f(t_{ij}; \theta_i, \psi) = f(t_j; \theta_i, \psi)$, one can compute the average value

$$\bar{y}_j = \frac{1}{N} \sum_{i=1}^N y_{ij}$$

of the data collected at each time point. One may then use this averaged data to compute

$$\hat{\theta}_{NAD} = \arg \min_{\theta \in \Theta} \sum_j (f(t_j; \theta, \psi) - \bar{y}_j)^2. \quad (12)$$

A very similar technique, applicable even if the data was collected at different times for each individual, is the *Naive Pooled Data* (NPD) method [31, 44] in which all data is used as if it were collected from a single individual. The resulting estimate is

$$\hat{\theta}_{NPD} = \arg \min_{\theta \in \Theta} \sum_{i,j} (f(t_{ij}; \theta, \psi) - y_{ij})^2. \quad (13)$$

The NPD method is most advantageous when the collected data is sparse (few data points for each individual, and few individuals). For both the NAD and NPD methods, the parameter ψ could be estimated simultaneously with the estimation of $\hat{\theta}$. It is clear that both methods are reduced to minimization problems over finite-dimensional Euclidean space—a problem for which standard computational techniques exist.

While the ordinary least squares (OLS) framework for Equations (12) and (13) is simple, it is unclear how the resulting estimates ($\hat{\theta}_{NAD}$ or $\hat{\theta}_{NPD}$) should be interpreted. Implicit in the least squares formulation is the assumption that the averaged or pooled data is centered around the curve $f(t_j; \bar{\theta}, \psi)$ for some $\bar{\theta}$, and that the errors accounting for the deviations of the data from this curve are independent with constant variance (that is, a CV statistical model (10)). (Of course, the assumption of constant variance is flexible, as one could just as well use a CCV statistical model and a generalized least squares (GLS) framework [21, 30] in the place of equations (12) and (13)). At first glance, it would appear that the unknown parameter $\bar{\theta}$ represents the sample mean (e.g., $\bar{\theta} = \frac{1}{N} \sum_{i=1}^N \theta_i$) and hence can be interpreted as a statistic (based upon the given data) for the mean or expectation of the population distribution from which the θ_i are drawn. However, this is only true if the model f is linear in θ so that $E[f(t; \theta, \psi)] = f(t; E[\theta], \psi)$ and this is not generally the case for practical models of interest. Thus it is unlikely that the estimates $\hat{\theta}_{NAD}$ or $\hat{\theta}_{NPD}$ have any meaningful interpretation. Moreover, for the NPD method, data from an individual will be correlated and the statistical model for the data must be modified to account for this correlation. Even then, naive methods do not provide meaningful estimates of variance, and are generally outperformed by two-stage methods (see below) [23, 44, 46].

3.1.2 Two-stage Methods

It is clear from the shortcomings of the naive methods (when viewed as attempts to estimate the first moment of the distribution) that extreme care must be taken so that the estimated parameter is meaningful. Given the hierarchical structure of the mathematical model (Equations (4) and (5)) and the fact that we are using data at the level of individuals to make inferences regarding population level structures, it is natural to think of the estimation problem itself in two stages. At the first stage, one can use regression with individual data to determine parameter estimates $\hat{\theta}_i$ for each individual. At the second stage, these individual estimates are combined in some way to estimate population level characteristics.

For simplicity, assume that intraindividual variability is characterized by the CV statistical model (10). Then for each individual, one can use ordinary least squares to obtain an estimate

$$\hat{\theta}_i = \arg \min_{\theta \in \Theta} \sum_j (f(t_{ij}; \theta, \psi) - y_{ij})^2 \quad (14)$$

of θ_i for each individual i . In the *standard two stage* (STS) method, one assumes that these estimates $\hat{\theta}_i$ are exact. Then, given the N ‘exact’ values $\{\hat{\theta}_i\}$, estimation of the population expectation and variance follows trivially from by computing the sample mean and variance for this sample of N individuals. Most obviously, the STS method is premised upon having sufficient data for each individual so that the parameter estimates $\hat{\theta}_i$ are accurate. Then, provided one has a sufficiently large sample of individuals (that is, N is sufficiently large), the STS method provides a reasonable estimate of the population expectation [23, 30, 31, 44]. Clearly, the STS method is heavily dependent upon the availability of a significant amount of data, and such a large number of data points is rarely attained in practice [46]. Regardless of the availability of data, the STS estimate of the population variance is often significantly biased upward [23]. The primary cause of this bias lies in the assumption that the estimates $\hat{\theta}_i$ are exact. In actual fact, the variability of these estimates is the result not only of interindividual variability (which is what we seek to estimate), but also of intraindividual variability and error in the estimation procedure itself. Thus the variance estimate produced by the STS method is the sum of these two sources of variability [29, pg. 138].

In order to correct for this bias, the *global two stage* (GTS) method uses the second stage of the hierarchical model to account for both sources of variance. In the first stage of the GTS algorithm, the estimates $\hat{\theta}_i$ are calculated exactly as in (14) above. Additionally, estimates of the individual covariance matrices $\hat{\Sigma}_i$ for the parameter estimates $\hat{\theta}_i$ are also computed. These covariance matrices can be calculated using asymptotic theory [21, 42] or via Monte Carlo techniques. Now assume that the second-stage model can be rewritten

$$h(\phi, \varphi_i, \alpha_i) = \tilde{h}(\phi, \alpha_i) + \varphi_i. \quad (15)$$

Though seemingly restrictive, this requirement is met in practice by a large number of hierarchical models [29, pgs. 142–143]. The key to the GTS method is to treat the N estimates $\hat{\theta}_i$ as observations (subject to error) of the second stage model. Thus

$$\hat{\theta}_i = \tilde{h}(\phi, \alpha_i) + \varphi_i + k_i$$

where k_i are realizations of some random variables κ_i satisfying $E[\kappa_i] = 0$ and $\text{Cov}(\kappa_i) = \hat{\Sigma}_i$ for all i . For the random effects, we have already (see Section 2) assumed $E[\varphi_i] = 0$ and $\text{Cov}(\varphi_i) = D$. Since D is unknown, it must be included in the parameters to be estimated. The goal then, is to determine the *extended least squares* (ELS) estimates

$$(\hat{\phi}, \hat{D}) = \arg \min \left\{ \sum_{i=1}^N \log |\hat{\Sigma}_i + D| + \sum_{i=1}^N \left(\hat{\theta}_i - \tilde{h}(\phi, \alpha_i) \right)^T \left(\hat{\Sigma}_i + D \right)^{-1} \left(\hat{\theta}_i - \tilde{h}(\phi, \alpha_i) \right) \right\}. \quad (16)$$

Assuming the asymptotic normality of the corresponding estimators, this criteria is equivalent to likelihood estimation. The estimated parameters $\hat{\phi}$ and $\hat{D}$ can be used with the additive random effects model (15) to estimate the statistical moments of the distributions for the parameters θ_i . The GTS method typically provides much better estimates of variance than does the STS method [46]. However, the reliability of the method depends upon the accuracy of the approximate covariance matrices $\hat{\Sigma}_i$. These approximations require a large number of

observations in order to obtain reasonable accuracy of the covariance matrix estimate. Hence, much like the STS method, the GTS method requires a significant availability of data, and this may be rarely attained in practice.

The STS and GTS methods can be easily generalized for non-CV errors by using a GLS estimate in the place of (14). Various other two stage methods have also been defined in the literature. For instance, one can use an EM-type algorithm to define an *iterative two stage* (ITS) method, or assume Bayesian priors and establish a *Bayesian two stage* (BTS) method. See [30, 31] for details. A major potential shortcoming of two-stage methods is that it can be hard to estimate a population parameter ψ , as this parameter is common to all individuals yet will appear in N distinct least squares problems. One solution is to make an initial guess for $\hat{\psi}$ (say, its NPD estimate) and fix its value while the individual estimates $\hat{\theta}_i$ are obtained. One may then fix the values of the $\{\hat{\theta}_i\}$ and update the estimate $\hat{\psi}$ via the NPD method. This process can be repeated iteratively, but it is not clear that the resulting estimate of $\hat{\psi}$ (or, indeed, of the set $\{\hat{\theta}_i\}$) will be accurate.

3.1.3 Linearization Methods

As we have seen, two-stage methods generally require a significant amount of data in order to accurately estimate the first two moments of the population distribution. However, in many cases (such as in observational studies) there are many individuals available but data is sparse for each individual. Ultimately, the primary shortcoming of two-stage methods is the requirement that individual models are identifiable. Meanwhile, the use of individual estimates in the first stage ignores any additional information which might be available through comparison to data from similar individuals. Conversely, the primary shortcoming of the naive methods was that individual differences were ignored as the data was combined.

A third set of methods attempts to strike a balance between these two shortcomings. Originally developed by Sheiner, et al., [44], the technique relies upon the linearization of the hierarchical model in terms of the random effects. First, if we define

$$Y_{i.} = \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{im_i} \end{pmatrix}$$

and define $t_{i.}$ and $\mathcal{E}_{i.}$ similarly, the hierarchical model can be combined with the statistical model to obtain

$$Y_{i.} = f(t_{i.}; h(\phi, \varphi_i, \alpha_i), \psi) + \mathcal{E}_{i.}.$$

It follows from previous assumptions (see (8) and discussion prior) that $E[\mathcal{E}_{i.}] = \vec{0}$ and $\text{Cov}(\mathcal{E}_{i.} | \theta_i, \phi) = R(h(\phi, \varphi_i, \alpha_i), \psi, \zeta_i)$. Alternatively, we may rewrite

$$Y_{i.} = f(t_{i.}; h(\phi, \varphi_i, \alpha_i), \psi) + R^{1/2}(h(\phi, \varphi_i, \alpha_i), \psi, \zeta_i) \tilde{\mathcal{E}}_{i.} \quad (17)$$

where $R^{1/2}$ is a Cholesky decomposition of the symmetric positive definite matrix R . The new error random variables $\tilde{\mathcal{E}}_{i.}$ satisfy $E[\tilde{\mathcal{E}}_{i.}] = \vec{0}$ and $\text{Cov}(\tilde{\mathcal{E}}_{i.} | \theta_i, \psi) = I_{m_i}$, the $m_i \times m_i$ identity matrix. This ‘canonical model’ [22] allows for the approximation not only of the nonlinear (in the random effects) model f but also of the nonlinear variance model. Using a Taylor expansion around the expectation (which we have assumed to be zero) of the random effects φ_i , we have

$$\begin{aligned} Y_{i.} &\approx f(t_{i.}; h(\phi, 0, \alpha_i), \psi) + \nabla_{\theta_i} f(t_{i.}; h(\phi, 0, \alpha_i), \psi) \nabla_{\varphi_i} h(\phi, 0, \alpha_i) \varphi_i \\ &\quad + R^{1/2}(h(\phi, 0, \alpha_i), \psi, \zeta_i) \tilde{\mathcal{E}}_{i.} \end{aligned} \quad (18)$$

We remark that the expansion of the mathematical model $f(t_{i.}; h(\phi, \varphi_i, \alpha_i), \psi)$ has been truncated at two terms while the expansion of the covariance matrix $R(h(\phi, \varphi_i, \alpha_i), \psi, \zeta_i)$ has been truncated after the first term. This is equivalent to the assumption that the terms $\varphi_i \tilde{\mathcal{E}}_{ij}$ are negligibly small [29].

Thus we now have $E[Y_{i.}] \approx f(t_{i.}; h(\phi, 0, \alpha_i), \psi)$ and that

$$\text{Cov}(Y_{i.}) = V_i = R(h(\phi, 0, \alpha_i), \varphi, \zeta_i) + Z_i^T D Z_i$$

where

$$Z_i = \nabla_{\theta_i} f(t_{i.}; h(\phi, 0, \alpha_i), \psi) \nabla_{\varphi_i} h(\phi, 0, \alpha_i).$$

Estimation by first order (FO) methods is performed by assuming that the linearization (??) is exact. The method of Sheiner et al., [22, 45] is to find

$$(\hat{\phi}, \hat{D}) = \arg \min_i \left\{ \log |V_i| + (y_i - f(t_i; h(\phi, 0, \alpha_i), \psi))^T V_i^{-1} (y_i - f(t_i; h(\phi, 0, \alpha_i), \psi)) \right\}. \quad (19)$$

In some situations, the parameters ζ_i in the intraindividual covariance matrix R must also be estimated. As with two-stage methods, the estimates $\hat{\phi}$ and $\hat{D}$ can be used to estimate the statistical properties of the distributions for the θ_i . The ELS criteria (19) is equivalent to maximum likelihood estimation if the linearized model is exact and all errors are normal [29, 30]. Several variations of this FO method have been proposed: using GLS rather than ELS, using an iterative EM algorithm, or estimation based upon generalized estimating equations (GEE). See [30, 31] for an overview of the additional possibilities. Also, FO methods in a (normal) likelihood formulation are discussed in [29, pg. 173].

As discussed above, the FO methods strike a balance between pooled data methods and two-stage methods. Like the pooled data methods, estimation is based upon the central value of the model in the absence of any random effects (that is, $f(t; h(\phi, 0, \alpha_i), \psi)$). The higher order terms in the linearization (??) help to account for interindividual differences in a manner similar to the two-stage methods. Several computational comparisons have explored the relative merits of two-stage and FO methods. The primary advantage of the the FO methods are that they do not require the identifiability of the individual parameters θ_i , and thus provide reasonable estimates of the distributions for the parameters even with sparse data. Additionally, the parameter ψ can be estimated simultaneously according to Equation (19). Higher order moments have been derived [48] but do not typically provide much improvement for the additional mathematical complexity.

3.1.4 Nonparametric Maximum Likelihood

All of the methods discussed so far were designed to estimate only the first one or two moments of the population distribution from which the parameters θ_i (or, equivalently when h is a linear function, φ_i) are drawn. While this may often be sufficient when the true underlying distribution is approximately symmetric, non-symmetric distributions are not accurately characterized by their first two statistical moments. Meanwhile, parametric estimation of a non-symmetric measure is possible (e.g., by assuming a lognormal form) but may result in overly restrictive or incorrect assumptions on the underlying measure. Thus, there are problems of interest in which one needs a nonparametric estimate of the underlying probability distribution itself.

Of course, the methods discussed so far have been quite general in terms of their assumptions for the statistical model. To this point, we have only needed to know the first two statistical moments of the error random variables, and in return we have developed methods which estimate the first two statistical moments of a population distribution. An obvious question, then, is whether or not a stronger set of assumptions on the error process could lead to the more direct estimation of the underlying probability measure itself. This is the goal of the nonparametric maximum likelihood (NPML) method proposed by Mallet [38].

Likelihood Formulation

The NPML method is designed to work in a likelihood setting. Consider again the statistical model

$$Y_{ij} = f(t_{ij}; \theta_i, \psi) + \mathcal{E}_{ij}.$$

For simplicity, the discussion of the NPML method will focus on the estimation of the distribution of the parameters θ_i directly, rather than on the parameters ϕ and φ_i of the hierarchical model, though the NPML applies in either setting. While previous methods required only the specification of the first two statistical moments of the random variables $\mathcal{E}_{ij}$, we must now assume that the $\mathcal{E}_{ij}$ admit a density function. In this case, the observations y_{ij} (realizations of the random variables Y_{ij}) are characterized by a conditional density function

$$p(y_{ij}) = \lambda_{ij}(y_{ij}|t_{ij}; \theta_i, \psi). \quad (20)$$

More generally, the functions λ_{ij} may also depend on the entire set $\{y_{ij}\}$ of observations (if multiple observations or individuals are correlated). This case is not explicitly considered, but does not pose any theoretical problems

for the NPML. It will be assumed as above that $\lambda_{ij} = \lambda$ for all i and j . As an example, it is quite common to assume independent, normally distributed errors with constant variance, $\mathcal{E}_{ij} \sim \mathcal{N}(0, \sigma^2)$. Then the conditional density functions are

$$\lambda(y_{ij}|t_{ij}; \theta_i, \psi) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(f(t_{ij}; \theta_i, \psi) - y_{ij})^2}{2\sigma^2}\right).$$

Like the two-stage methods discussed above, the likelihood function is formed by considering the multiple (individual and population) stages of the sampling process. Let $\vec{y}_i = \{y_{ij}\}_{j=1}^{m_i}$, $\vec{t}_i = \{t_{ij}\}_{j=1}^{m_i}$. Then we can define the individual likelihood functions

$$l_i(\vec{y}_i|\vec{t}_i; \theta_i, \psi) = \prod_{j=1}^{m_i} \lambda_{ij}(y_{ij}|t_{ij}; \theta_i, \psi)$$

which describe the likelihood of the m_i data points taken from the i^{th} individual. If the individual data points are not independent, then the form of the functions l_i may be more complex. Now, assume that the individual parameters θ_i are all drawn from some (unknown) probability distribution $P = P(\theta)$. Then the mixing distribution for each individual is defined as

$$\begin{aligned} \tilde{l}_i(\vec{y}_i|\vec{t}_i; P, \psi) &= \int_{\Theta} l_i(\vec{y}_i|\vec{t}_i; \theta, \psi) dP(\theta) \\ &= E[l_i(\vec{y}_i|\vec{t}_i; \cdot, \psi) | P]. \end{aligned} \tag{21}$$

The mixing distributions $\tilde{l}_i$ describe the likelihood of a given set of observations (for individual i) for a given parameter θ , weighted by the likelihood of that particular value of θ . Finally, the population likelihood (which is to be maximized) is

$$\mathcal{L}(P, \psi | \{\vec{y}_i\}_{i=1}^N, \{\vec{t}_i\}_{i=1}^N) = \prod_{i=1}^N \tilde{l}_i(\vec{y}_i|\vec{t}_i; P, \psi) = \prod_{i=1}^N \int_{\Theta} l_i(\vec{y}_i|\vec{t}_i; \theta, \psi) dP(\theta).$$

The goal of maximum likelihood estimation is to determine the parameter (in this case, the probability measure P) which maximizes the likelihood of the given observations. The probability measures are taken in the set of probability measures on Θ , denoted $\mathcal{P}(\Theta)$. The problem to be solved is thus to find

$$(\hat{P}, \hat{\psi}) = \arg \max_{\mathcal{P}(\Theta) \times \Psi} \mathcal{L}(P, \psi | \{\vec{y}_i\}_{i=1}^N, \{\vec{t}_i\}_{i=1}^N) \tag{22}$$

Computational Techniques

Of course, this is an optimization problem over an infinite dimensional space, so finite dimensional approximation techniques must be considered. To that end, consider the N -dimensional linear model

$$\begin{aligned} Z &= L(\theta)x + \tilde{\mathcal{E}} \\ E[\tilde{\mathcal{E}}] &= 0 \\ \text{Cov}(\tilde{\mathcal{E}}) &= I_N, \end{aligned}$$

where the design matrix is

$$L(\theta) = \text{diag}\left(\sqrt{l_i(\vec{y}_i|\vec{t}_i; \theta, \psi)}\right).$$

Here, the ‘data’ Z depends linearly on ‘parameters’ x , and the goal is to choose the design (that is, find θ) to obtain the best possible (in an appropriate sense) estimate of x given observations of Z . Optimal design theory is well-studied (see [17] and the references therein) and involves maximizing some function of the Fisher Information Matrix. For the linear model above, the Fisher Information Matrix is

$$F(P) = \int_{\Theta} L(\theta)^T L(\theta) dP(\theta).$$

For a D-optimal design, one seeks to maximize the determinant $|F(P)|$ of the Fisher Information matrix over all possible probability measures P . Now, observe

$$|F(P)| = \prod_{i=1}^N \int_{\Theta} l_i(\vec{y}_i | \vec{t}_i; \theta, \psi) dP(\theta) = \mathcal{L}(P, \psi | \{\vec{y}_i\}_{i=1}^N, \{\vec{t}_i\}_{i=1}^N).$$

Thus the problem of maximizing the population likelihood is equivalent to finding a D-optimal design for a finite dimensional linear companion problem. As such, theorems and computational techniques can be borrowed from the theory of D-optimal design. The following theorem is stated by Mallet [38]:

Theorem 3.1. *If Θ is closed and L is continuous in θ , then the set of all information matrices,*

$$\{F(P) | P \in \mathcal{P}\},$$

is the convex closure of the set of information matrices corresponding to Dirac measures

$$\{F(P) | P = \Delta_{\theta^*}\} = \{L(\theta^*)^T L(\theta^*) | \theta^* \in \Theta\}.$$

Moreover, the D-optimal design matrix belongs to the boundary of this set. Hence for any D-optimal design $\hat{P}$,

$$F(\hat{P}) = \sum_{k=1}^M L(\theta_k)^T L(\theta_k) p_k,$$

with $M \leq N$.

This theorem relies on the identification (via the Riesz Representation Theorem) of the space $\mathcal{P}(\Theta)$ of probability measures on Θ as a subset of $C_B^*(\Theta)$, the dual of the space $C_B(\Theta)$ of bounded continuous functions on Θ . If one further assumes Θ is compact, then one obtains the equivalence (see [37, 40]) of the two sets above, i.e.,

$$\{F(P) | P \in \mathcal{P}\} = \text{conv}(\{L(\theta^*)^T L(\theta^*) | \theta^* \in \Theta\}),$$

and the closure is no longer needed. It is proven by Federov [32, Th. 2.2.3] that the information matrix corresponding to an optimal design lies on the boundary of the convex closure of the set of matrices corresponding to one-point designs. The index limitation ($M \leq N$) then follows from the identification of an $N \times N$ diagonal matrix with a subset of Euclidean N -space and Caratheodory's Theorem for the representation of points on the boundary of a convex set. Not only does this theorem guarantee the existence of an optimal solution for the NPML, it also specifies the finite dimensional form of an optimal solution. It can be shown [32, 37] that the optimal solution is unique in the likelihood geometry. (That is, if $\hat{P}_1$ and $\hat{P}_2$ are two optimal solutions, then $F(\hat{P}_1) = F(\hat{P}_2)$ and hence $\tilde{l}_i(\vec{y}_i | \vec{t}_i; \hat{P}_1, \psi) = \tilde{l}_i(\vec{y}_i | \vec{t}_i; \hat{P}_2, \psi)$ for all i .) Both Lindsay [37] and Mallet [38] offer conditions which can be used to assess whether the estimated measure $\hat{P}$ is itself unique.

In order to actually compute the estimate $\hat{P}$, Mallet provides an algorithm which is an extension of the work of Federov [32]. The algorithm attempts to find the $M \leq N$ nodes and corresponding weights which form the optimal measure. Interestingly, the algorithm stems from the Kiefer and Wolfowitz Equivalence Theorem [36] demonstrating the equivalence of D-optimal and G-optimal design. The algorithm itself can be found in [38, Sec. 4] and the references therein. See also [29, pgs. 195-196] for more details regarding the role of the Equivalence Theorem. We provide only an overview of the algorithm:

1. Fix an initial node $\theta \in \Theta$ and set $P = \Delta_{\theta}$. Compute $\mathcal{L}(P, \psi | \{\vec{y}_i\}_{i=1}^N, \{\vec{t}_i\}_{i=1}^N)$.
2. Compute

$$\theta_{new} = \arg \max_{\theta_{new}} d(\theta_{new} | P) = \arg \max_{\theta_{new}} \sum_{i=1}^N \frac{l_i(\theta_{new})}{\tilde{l}_i(P)},$$

where P is the current design, $l_i(\theta_{new}) = l_i(\vec{y}_i | \vec{t}_i; \theta_{new}, \psi)$, and $\tilde{l}_i(P) = \tilde{l}_i(\vec{y}_i | \vec{t}_i; P, \psi)$.

3. Update the set of nodes. At iteration K , $\{\theta_k\}_{k=1}^K = \{\theta_k\}_{k=1}^{K-1} \cup \{\theta_{new}\}$.
4. For the fixed set of nodes $\{\theta_k\}_{k=1}^K$, determine the new weights p_k for each node θ_k by maximizing the likelihood function over the set $\{p_k\}$.
5. If this new design/distribution estimate has more than N nodes, perform an index limitation step (described in [38, Sec. 4.2]) to reduce the design to N nodes, and store this as the updated design P .
6. Repeat steps 2.-5. until the algorithm converges, represented by obtaining $d(\theta_{new}|P) = N$, or when a maximum number of iterations has been reached.

The above algorithm creates a sequence of designs which are guaranteed to increase the resulting likelihood function at each step [32, 38]. Schumitzky [43] proposed an EM-type algorithm for computation of the index limited design. An another alternative algorithm proposed by Wynn [51] approximates the optimal measure by using a larger and larger number M of nodes, without regard for whether $M \leq N$. This is a similar technique to the Prohorov metric framework to be discussed in Section 4.1. In all cases, it is significant that the original infinite dimensional optimization problem of maximizing $\mathcal{L}(P)$ over all $P \in \mathcal{P}(\Theta)$ has been approximated by a finite dimensional problem which guarantees existence and convergence. The NPML can be extended to estimate a parameter ψ in an iterative leap-frogging method similar to the one discussed for two-stage methods: for a nominal value of $\hat{\psi}$, a measure $\hat{P}$ is constructed; this $\hat{P}$ is then fixed and an updated estimate of $\hat{\psi}$ is obtained, etc.

The NPML method has been shown to outperform two-stage methods in a computational study [39]. In fact, the NPML can accurately estimate the underlying distribution of the parameters θ_i even when only a single data point is available for each individual [38] so that the individual models are not identifiable (and hence, two-stage methods do not apply). Given the strong assumptions on the errors (by specifying a conditional density function), this is not entirely surprising. In fact, if the conditional density functions are correctly specified (which can be difficult to verify), then one actually obtains consistency [34, 35] of the NPML estimate. That is, as the number of individuals N goes to infinity, the sequence of estimates $\hat{P}_N$ converges pointwise to the true measure P at all points of continuity of the latter. (The notion of pointwise convergence is premised upon the condition that $\Theta \subset \mathbb{R}^p$. This is almost always the case in practice. For abstract spaces Θ , alternative characterizations are available via the Portmanteau Theorem [25, Th. 2.1].) The NPML does require the identification of sets of data taken from the same individual (so that the individual likelihoods $l_i(\vec{y}_i|\vec{t}_i; \theta_{new}, \psi)$ can be formed). Yet, it is interesting that the $M \leq N$ nodes $\{\theta_k\}$ do not necessarily correspond to the N underlying parameters θ_i [46, pg. 93].

Because of the assumed ability to track individuals, it is not clear how the NPML would perform on pooled data. While such a situation is similar to having only single data points (but from a larger number of individuals), the additional correlation between ‘individuals’ (because some individuals are the same) would need to be accounted for. It is also not clear if the restrictive assumptions on the error random variables could be relaxed with least squares estimation used in the place of maximum likelihood. It is common to establish the equivalence of least squares and likelihood estimators by demonstrating that they solve the same set of estimating equations (see, e.g., [21, Ch. 3] and the references therein). However, the typical estimating equations (first order optimality conditions) are obtained by taking the derivative of the likelihood function with respect to the unknown parameters, and it is not yet clear how this might be meaningfully done when the underlying parameter is a probability measure (see [10] and references therein for some efforts on this subject).

4 Aggregate Data and the Prohorov Metric Framework

We now turn our attention to the case where one has a mathematical model which describes the dynamics of one or more states of interest for an individual member of the population, but that data is only available at the aggregate (population) level. At first glance, this seems to be an unusual situation. In fact, such cases arise quite frequently when the physical or biological process of interest is modeled with some form of density function (e.g., size-structured density [8], electromagnetic polarization [16], viscoelastic relaxation [3], etc.). For instance, we have already seen the growth rate distribution model (2) (which is a type of size-structured density) which

features a probabilistically distributed intrinsic growth rate. These studies are familiar to applied mathematicians (see, e.g., [13, 14]) and the primary tools for treating such problems can be found in the mathematics literature.

The use of aggregate data requires a slightly different form for the statistical model compared to the model of the previous section. Recall the individual mathematical model (4) given by

$$y(t) = f(t; \theta, \psi).$$

Again, it is assumed that the parameter $\theta \in \Theta$ varies from individual to individual in the population and that some probability measure $P(\theta)$ characterizes the distribution of the parameter. The parameter $\psi \in \Psi$ is common to all individuals in the population. We remark that the index i previously used to track individuals within the population is no longer needed as individuals are not identifiable with the aggregate sampling of the population. Moreover, because individuals are not identifiable, we do not consider in any detail the hierarchical models discussed in the Section 2. Rather, we focus only on the determination of the probability distribution P over the parameters θ . For a fully specified hierarchical model with only aggregate data, this is no different from the estimation of a probability distribution over the random effects φ .

The aggregate state for the entire population can be defined in terms of an expectation

$$v(t; P, \psi) = \mathbb{E}[y(t; \cdot, \psi) | P] = \int_{\Theta} f(t; \theta, \psi) dP(\theta). \quad (23)$$

When aggregate data is collected, this represents a sampling from the aggregate population. Thus the data collection process is assumed to be represented by the random variables

$$V_j = v(t_j; P, \psi) + \mathcal{E}_j \quad (24)$$

where $\mathbb{E}[\mathcal{E}_j] = 0$ (again corresponding to the tacit assumption that the mathematical model is accurate). As in Section 3, we must also specify the covariance properties of the error random variables $\mathcal{E}_j$. For notational simplicity, we will make the assumption of independent, constant variance errors so that $\text{Var}(\mathcal{E}_j) = \sigma^2$ for all j . This is not absolutely necessary, as one could similarly assume a CCV statistical model, or even a more general form featuring a general covariance matrix R for the vector of errors as in (8). This will not pose any problem for the approximation results presented in Section 4.1, provided the resulting cost function J is a continuous function of P . In fact, one could even use likelihood in the place of least squares if the errors $\mathcal{E}_j$ admitted a known density.

4.1 Estimation Techniques

With the statistical models of the data established, we turn to a discussion of estimation techniques. As before, the goal is to determine the measure P describing the distribution of the parameters θ in the population. Methods for the estimation of the measure P have traditionally focused on a least squares framework. The CV statistical model (24) allows the use of the notation for ordinary least squares which is appealing in its simplicity. Thus we seek to determine

$$(P^*, \psi^*) = \arg \min_{\mathcal{P}(\Theta) \times \Psi} J(P, \psi) = \arg \min_{\mathcal{P}(\Theta) \times \Psi} \sum_{j=1}^m (v_j - v(t_j; P, \psi))^2 \quad (25)$$

where $J(P, \psi)$ is the OLS cost of the model given the parameters P and ψ and the data v_j are realizations of the random variables V_j . For more general aggregate sampling errors (e.g., CCV) the cost function J in (25) would need to be appropriately redefined.

The change of notation from the estimates $(\hat{P}, \hat{\psi})$ to the minimizers (P^*, ψ^*) deserves careful attention. In Section 3, computational methods for individual data were based upon either likelihood estimation (the NPML) or a least squares framework (two-stage methods, linearization methods) which is provably equivalent to likelihood estimation under an appropriate set of assumptions. In all cases (with the exception of naive methods, which pool or average the individual data) asymptotic consistency theories exist which establish the convergence of the estimators to the true population values as the number of sampled individuals increases. For two-stage and linearization methods (which are finite-dimensional optimization problems) consistency follows immediately from classical results (e.g., [49, 50]). These results have been extended for the NPML (an infinite-dimensional

optimization) [35]. However, it is not clear that the results of [35] apply to the optimization problem (25) because the unknown measure P does not play the role of a mixing distribution (see [35, eqn (2.1)] and compare with (21)). Moreover, this statement would remain true even if the estimate (25) were replaced by a likelihood estimate. As we are not aware of any proof of consistency for the present problem, we denote by (P^*, ψ^*) *minimizers* of the OLS cost function J , and (at least for now) distinguish between these minimizers and the estimates $(\hat{P}, \hat{\psi})$ of the actual population parameters. Assuming consistency, this distinction can be ignored. Encouragingly, simulation studies [8, 9] have found the minimizers (P^*, ψ^*) to be quite close to the true population distribution at least for the examples investigated.

The minimization (25) must be carried out over the infinite dimensional space $\mathcal{P}(\Theta)$ of probability measures on the space Θ . As such, one must establish a finite dimensional approximation and convergence theory. To that end, we introduce the *Prohorov Metric Framework* (PMF) developed over the past decade (a summary from a functional analytic perspective along with a discussion of other topologies for probability distributions can be found in [1]) for use in the estimation of a probability measure. This framework is more common in the applied mathematics literature compared to the methods for individual data discussed previously, and is named for the Russian probabilist Y.V. Prohorov, whose metric [41] features prominently in theoretical and computational efforts [2, 3, 5, 14, 20] establishing convergence results for the computational methods discussed.

We begin by defining the Prohorov metric on the space of probability measures. Consider the admissible parameter space Θ with its metric $\tilde{d}$ and the space of probability measures $\mathcal{P}(\Theta)$ on Θ . For any closed set $F \subset \Theta$, define

$$F^\epsilon = \{x \in \Theta : \inf_{y \in F} \tilde{d}(x, y) < \epsilon\}.$$

Then we have the following definition of the Prohorov metric [25, pg 237-238]:

Definition 4.1. *Let $F \subset \Theta$ be any closed set and define F^ϵ as above. For $P_1, P_2 \in \mathcal{P}(\Theta)$, the Prohorov metric is given by*

$$\rho(P_1, P_2) = \inf\{\epsilon > 0 : P[F] \leq P'[F^\epsilon] + \epsilon, F \text{ closed in } \Theta\}. \quad (26)$$

While several alternative (and equivalent) definitions of the Prohorov metric are possible [25, pgs. 236-237], it is clear from the definition above that its meaning and application are far from intuitive. Yet the following theorem provides a useful characterization:

Theorem 4.2. *Assume the parameter space Θ with its metric $\tilde{d}$ is separable. Then the Prohorov metric metrizes the weak* topology of the space $C_B^*(\Theta)$ (the topological dual of $C_B(\Theta)$) and the following are equivalent:*

- (i) $\rho(P_M, P) \rightarrow 0$;
- (ii) $\int_\Theta f(\theta) dP_M(\theta) \rightarrow \int_\Theta f(\theta) dP(\theta)$ for bounded, uniformly continuous functions f .

We remark that the separability of the metric space $(\Theta, \tilde{d})$ is not strictly necessary. The weak topology of probability measures (weak* topology) and the topology induced by the Prohorov metric are equivalent provided every P in $\mathcal{P}(\Theta)$ is a separable measure [25, Th 5, pg. 238]. We can be guaranteed of this property in the event Θ is separable [25, pg. 236]. From a functional analysis perspective, if we view $\mathcal{P}(\Theta) \subset C_B^*(\Theta, \mathbb{R})$, then convergence in the ρ topology is equivalent to weak* convergence in $\mathcal{P}(\Theta)$.

The following theorem can be used to establish the existence of a minimum to the least squares optimization problem.

Theorem 4.3. *Let Θ be separable when taken with its metric $\tilde{d}$. Then*

- (i) $(\mathcal{P}(\Theta), \rho)$ is a complete metric space iff $(\Theta, \tilde{d})$ is a complete metric space.
- (ii) If $(\Theta, \tilde{d})$ is a compact metric space, then $(\mathcal{P}(\Theta), \rho)$ is a compact metric space.

For the first property above the converse direction is true only in the event P has separable support for all $P \in \mathcal{P}(\Theta)$ [25, pg. 240] which is guaranteed by the separability of Θ . The forward direction holds in all cases. The second property is a consequence of Prohorov's Theorem [25, pg. 37]: if Θ is compact, then every collection of measures on Θ , specifically $\mathcal{P}(\Theta)$, is tight [25, pg. 9] and hence relatively sequentially compact by

Prohorov's Theorem. Moreover, if Θ is compact (and separable), then it is also complete and it follows from the first property above that the relative sequential compactness of $(\mathcal{P}(\Theta), \rho)$ is sequential compactness. While it is sequential compactness that will be of interest in establishing the hypotheses of the convergence theorem (Theorem 4.5), we remark that in a metric space (i.e., when $\mathcal{P}(\Theta)$ is metrized by ρ) there is no distinction between sequential compactness and compactness.

At this point, we have used the Prohorov metric to establish a measure over the space of probability measures. As an added benefit, when Θ is compact and separable (which is almost always the case in applications, particularly when Θ is a closed bounded subset of Euclidean space) the space of probability measures (taken with the Prohorov metric) is compact. Hence, provided the least squares cost functional J is continuous in P , we are guaranteed of the existence of a minimum.

We now turn our attention to strategies for the actual computation of the minimizing probability measure P^* . In order to make the estimation problem amenable to computational solution, the infinite dimensional space $\mathcal{P}(\Theta)$ must be approximated in a meaningful way. Traditional parametric methods, which assume the optimal measure has a particular distributional form, are not preferred as they are overly restrictive and will produce inaccurate results if the parametric form is misspecified. Rather, we have the following theorem:

Theorem 4.4. *Assume $(\Theta, \tilde{d})$ is a complete, separable metric space and $\mathcal{P}(\Theta)$ is the set of probability measures on Θ . Let $\Phi = \{\theta_k\}_1^\infty$ be a countable, dense subset of Θ . Define the collection of discrete measures*

$$\tilde{\mathcal{P}}(\Theta) = \left\{ P \in \mathcal{P}(\Theta) : P = \sum_{k=1}^M p_k \Delta_{\theta_k}, M \in \mathbb{N}, p_k \in \mathbb{R}^+, \sum_{k=1}^M p_k = 1 \right\}.$$

Then $\tilde{\mathcal{P}}(\Theta)$ is dense in $\mathcal{P}(\Theta)$.

A proof of the theorem can be found in [2]. For each $M \in \mathbb{N}$, define

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

That is, for each M , $\mathcal{P}_M(\Theta)$ is the set of all atomic probability measures with nodes placed at the first M elements in the enumeration of the countable dense subset Φ of Θ . Note that $\mathcal{P}_M(\Theta) \subset \tilde{\mathcal{P}}(\Theta) \subset \mathcal{P}(\Theta)$ for all M .

Finally, the following theorem provides the desired convergence results.

Theorem 4.5. *Assume Θ is compact and separable and consider $\mathcal{P}(\Theta)$ taken with the Prohorov metric. If the mapping $P \rightarrow J(P)$ is continuous, then there exists a sequence of minimizers $\{P_M\}$ of $J(P)$ over $\mathcal{P}_M$. Moreover, this sequence has at least one convergent subsequence. The limit (as $M \rightarrow \infty$) P^* of such a subsequence minimizes J over $\mathcal{P}(\Theta)$.*

This theorem follows immediately from the theoretical framework and convergence theorems of [19] and the results above. The power of this theorem is that it transforms an infinite dimensional minimization problem (over $\mathcal{P}(\Theta)$) into an M dimensional problem. In practice, M nodes from the dense, countable subset $\{\theta_k\}_1^\infty$ are fixed in advance and then the M weights p_k^M are estimated in a minimization problem. Thus the least squares problem (25), i.e.,

$$(P^*, \psi^*) = \arg \min_{\mathcal{P}(\Theta) \times \Psi} J(P, \psi) = \arg \min_{\mathcal{P}(\Theta) \times \Psi} \sum_{j=1}^m (v_j - v(t_j; P, \psi))^2,$$

is approximated by the family of finite-dimensional problems

$$\begin{aligned} (P_M^*, \psi_M^*) &= \arg \min_{\mathcal{P}_M(\Theta) \times \Psi} \sum_{j=1}^m (v_j - v(t_j; P, \psi))^2 \\ &= \arg \min_{\mathbb{R}^M \times \Psi} \sum_{j=1}^m \left(v_j - \left(\sum_{k=1}^M f(t_j; \theta_k^M, \psi) p_k^M \right) \right)^2. \end{aligned} \tag{27}$$

In (27) we are seeking optimal weights $\bar{p}^M = (p_1^M, \dots, p_M^M)^T \in \widetilde{\mathbb{R}^M} = \{\bar{p}^M | p_k^M \in \mathbb{R}^+, \sum_{k=1}^M p_k^M = 1\}$ which will then characterize the optimal discrete measure (since the nodes are assumed to be fixed in advance). This leads to a constrained quadratic optimization problem which is eminently amenable to parallel computational methods [14]. We have replaced ψ^* with ψ_M^* to emphasize the possible dependence of the structural parameter on the approximations made in obtaining the finite-dimensional estimation problem, though this difference is expected to be small for reasonable approximations. The methods based on these ideas have been successfully used in a wide range of applied problems [2, 3, 5, 8, 9, 14, 20].

It is clear from the statement of Theorem 4.5 that uniqueness of the minimizer $\hat{P}$ is not guaranteed (as there could be multiple convergent subsequences). For a fixed set of nodes $\{\theta_k\}_{k=1}^M$, it is possible to establish first-order necessary conditions for the uniqueness of the computational problem [2, pg. 102]. But these should not be confused with conditions for global uniqueness of the minimizing measure P^* . Unlike the NPML method for individual data, the parameter ψ does not require a separate minimization process, and can be estimated simultaneously with the M weights. Of course, some care must be taken in choosing the M nodes (as well as in choosing the value of M) to ensure the resulting estimate is reasonable and computationally stable. Values of M under 50 have been studied computationally in [2, 8].

5 Final Remarks

In this presentation a variety of computational techniques have been reviewed and theoretical results stated (when available) for the nonparametric estimation of a probability measure in an inverse problem. Some methods (two-stage, linearization) focus only on the estimation of some specific statistical moments of the underlying distribution. These can certainly be useful in practice and are widely applicable in that no form is explicitly assumed for the underlying population probability distribution. The more interesting methods, the NPML estimator and the PMF-based method, attempt to estimate the the probability distribution directly, thus requiring a rich body of theoretical results. Significantly, these two methods apply to fundamentally different problems—the NPML when data is available for individuals, and the PMF when only aggregate data is available. A key distinction between the two situations is the role played by the probability measure to be estimated. For individual data, the unknown measure plays the role of a mixing distribution among multiple individuals each with a fixed mathematical model. For aggregate data, the unknown measure becomes a part of the mathematical model when the expectation is taken (see equation (23)). Though seemingly simple, this is an inherent difference between the formulations and this difference is key when examining similarities (or lack thereof) in the theory between the different formulations based on data type.

The theoretical results for the NPML and PMF share the remarkable feature that the underlying population distribution (regardless of its smoothness) can be approximated by a discrete probability measure. This is all the more fascinating since these results are derived from different theoretical bases and developed to solve different problems. Of course, if one has *a priori* knowledge of smoothness of the underlying measure, additional techniques have been developed. For individual data, approximation by a convex combination of atomic measures in the NPML can be replaced by a smooth density function which still maintains many of the desired nonparametric properties [27, 28, 33]. For aggregate data, the Prohorov metric framework can be restricted to the subset of distribution functions which are absolutely continuous. The underlying density function can then be estimated nonparametrically via splines [20].

The issues in an aggregate model, aggregate data situation deserve a few more comments. As previously mentioned, the measure estimation procedure is the same as for the individual model, aggregate data case. In both problems, the probability measure appears as part of the model (and thus is present inside the squared residuals, i.e., inside the inner portion of $\sum_{j=1}^m (v_j - v(t; P, \psi))^2$ —again, this is in contrast to the mixing distributions (21)), while the *form* of the model $v(t; P, \psi)$ will certainly differ depending upon whether one has an individual or aggregate model. In the first case the solution is given by (23) while in the latter case one could have the more complicated form

$$v(t; P, \psi) = \mathbb{E}[y(t; \cdot, P, \psi) | P] = \int_{\Theta} f(t; \theta, P, \psi) dP(\theta). \quad (28)$$

Of course the computation of forward solutions to yield $f(t; \theta, P, \psi)$ could be extremely challenging (special cases

and examples are given in [3, 5, 15, 16, 20]). However, we emphasize that the approximation and estimation framework discussed in Section 4.1 remains unchanged.

Several open questions remain. For the NPML, the likelihood framework can be especially restrictive given the difficulty of verifying a particular density function assumption (as in (20)) for the error random variables. Ideally, an analog of the NPML could be defined in a framework where only the first two statistical moments of the errors are specified. Meanwhile, for the PMF, while one only needs to specify the first two moments of the error random variables, it is not clear that the resulting least squares estimator is consistent (this would follow immediately if one found an equivalent likelihood formulation). There are also currently no node selection algorithms for PMF methods, nor do we yet have any sense of whether we can impose an index limitation step akin to that of Mallet [38].

For more traditional parametric and finite dimensional parameter estimation, methods for determining confidence intervals on parameters include asymptotic methods [21, 42] and bootstrapping [18]. The issue of finding confidence intervals becomes much more complicated once the object of interest is a probability distribution rather than a single constant parameter. For instance, asymptotic theory requires the computation of sensitivity matrices, which in the case of distribution estimation requires one to take a derivative with respect to a probability measure. This is not an easy task since the space of probability measures is not a linear space, nor is it finite dimensional. Previous work [10] examined the use of directional derivatives, and other work [9] has used standard asymptotic theory to put confidence intervals on each node in the finite dimensional approximation. The Portmanteau Theorem [25, Th. 2.1], which establishes several equivalent notions of the convergence of two probability measures, may provide useful techniques for the visualization of confidence intervals around a probability measure. This is an important avenue for future study in order to provide robust and trustworthy measure estimates in inverse problems. Of course, a meaningful confidence interval is premised upon a consistent estimator, and some work is still required to prove the consistency of the PMF estimator.

Overall, we have given a general outline of model types, data types, and the methods one might use to tackle measure estimation for a particular situation. It should be clear at this point that there is no universal method; the particular situation and research goals will dictate the choice of estimation procedure. We hope to have bridged some of the gap between measure estimation procedure development in statistics and that in applied mathematics. The different methods and approaches help inform and refine understanding of each other, and bring key questions (like consistency) into focus for all solution approaches.

Acknowledgements

This research was supported in part by the National Institute of Allergy and Infectious Diseases under Grant Number R01AI071915-07, in part by the U.S. Air Force Office of Scientific Research under grant number FA9550-12-1-0188, and in part by the Department of Education with a GAANN Fellowship (to ZRK) under grant number P200A070386.

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 Kathleen 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, D. M. Bortz and S. E. Holte, Incorporation of variability into the mathematical modeling of viral delays in HIV infection dynamics, *Math. Biosciences*, **183** (2003), 63–91.
- [5] H.T. Banks, David Bortz, Gabriella Pinter, and Laura Potter, Modeling and imaging techniques with potential for application in bioterrorism, in *Bioterrorism: Mathematical Modeling Applications in Homeland*

- Security*, H.T. Banks and C. Castillo-Chavez, eds, *Frontiers in Applied Mathematics* **28**, SIAM, Philadelphia, PA, 2003.
- [6] H. T. Banks, L. W. Botsford, F. Kappel, and C. Wang, Modeling and estimation in size structured population models, in *Mathematical Ecology*, T. G. Hallam, L. J. Gross, and S. A. Levin, eds, World Scientific Publ, Singapore, 1988, pp. 521–541.
 - [7] H.T. Banks, M. Davidian, J. Samuels, and K.L. Sutton, An inverse problem statistical methodology summary, CRSC-TR08-01, North Carolina State University, January, 2008; Chapter 11 in *Mathematical and Statistical Estimation Approaches in Epidemiology*, G. Chowell, et al., eds., Berlin Heidelberg New York, 2009, pp. 249–302.
 - [8] 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.
 - [9] H.T. Banks and J.L. Davis, Quantifying uncertainty in the estimation of probability distributions, *Mathematical Biosciences and Engineering*, **5** (2008), 647–667.
 - [10] H.T. Banks, S. Dediu, and H.K. Nguyen, Sensitivity of dynamical systems to parameters in a convex subset of a topological vector space, *Mathematical Biosciences and Engineering* **4** (2007), 403–430.
 - [11] H.T. Banks and B.G. Fitzpatrick, Inverse problems for distributed systems: statistical tests and ANOVA, LCDS/CSS Report 88-16, Brown University, July, 1988; *Proc. International Symposium on Math. Approaches to Envir. and Ecol. Problems*, Springer Lecture Notes in Biomath., **81** (1989), 262–273.
 - [12] H.T. Banks and B.G. Fitzpatrick, Statistical methods for model comparison in parameter estimation problems for distributed systems, CAMS Tech. Rep. 89-4, Univ. of Southern California, September, 1989; *J. Math. Biol.*, **28** (1990), 501–527.
 - [13] 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.
 - [14] H.T. Banks, B.G. Fitzpatrick, Laura K. Potter, and Yue Zhang, Estimation of probability distributions for individual parameters using aggregate population observations, CRSC-TR98-06, North Carolina State University, January, 1998; *Stochastic Analysis, Control, Optimization and Applications* (1998), W.McEneaney, G. Yin, and Q. Zhang, eds., Birkhäuser, (1998), pp. 353–371.
 - [15] H.T. Banks and N. L. Gibson, Well-posedness in Maxwell systems with distributions of polarization relaxation parameters, *Applied Math. Letters*, **18** (2005), 423–430.
 - [16] 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.
 - [17] H.T. Banks, K. Holm, and F. Kappel, Comparison of optimal design methods in inverse problems, CRSC-TR10-11, North Carolina State University, July, 2010; *Inverse Problems*, **27** (2011), 075002.
 - [18] H.T. Banks, K. Holm, and D. Robbins, Standard error computations for uncertainty quantification in inverse problems: Asymptotic theory vs. bootstrapping, CRSC-TR09-13, North Carolina State University, June, 2009; *Mathematical and Computer Modelling*, **52** (2010), 1610–1625.
 - [19] H.T. Banks and K. Kunisch, *Estimation Techniques for Distributed Parameter Systems*, Birkhauser, Boston, 1989.
 - [20] 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.
 - [21] H.T. Banks and H.T. Tran, *Mathematical and Experimental Modeling of Physical and Biological Processes*, CRC Press, Boca Raton, 2009.

- [22] S.L. Beal and L.B. Sheiner, Estimation population pharmacokinetics, *CRC Critical Reviews in Biomedical Engineering*, **8** (1982), 195–222.
- [23] S.L. Beal and L.B. Sheiner, Methodology of population pharmacokinetics, Ch. 4 in *Drug Fate and Metabolism: Methods and Techniques*, Volume 5, Edward R. Garrett and Jean L. Hirtz, eds., Marcel Dekker, University of Michigan, 1985.
- [24] K.P. Burnham and D.R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (2nd Edition), Springer, New York, 2002.
- [25] P. Billingsley, *Convergence of Probability Measures*, Wiley & Sons, New York, 1968.
- [26] R.J. Carroll and D. Ruppert, *Transformation and Weighting in Regression*, Chapman & Hall, New York, 1988.
- [27] M. Davidian and A.R. Gallant, Smooth nonparametric maximum likelihood estimation for population pharmacokinetics, with application to quinidine, *J. Pharmacokinetics and Biopharmaceutics*, **20** (1992), 529–556.
- [28] M. Davidian and A.R. Gallant, The nonlinear mixed effects model with a smooth random effects density, *Biometrika*, **80** (1993), 475–488.
- [29] M. Davidian and D. Giltinan, *Nonlinear Models for Repeated Measurement Data*, Chapman and Hall, London, 1995.
- [30] 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.
- [31] E.I. Ette and P.J. Williams, Population pharmacokinetics II: Estimation methods, *Annals of Pharmacotherapy*, **38** (2004), 1907–1915.
- [32] V. Fedorov, *Theory of Optimal Experiments*, New York, Academic Press, 1972.
- [33] A.R. Gallant and D. Nychka, Semi-nonparametric maximum likelihood estimation, *Econometrica*, **55** (1987), 363–390.
- [34] B. Hoadley, Asymptotic properties of maximum likelihood estimators for the independent not identically distributed case, *Ann. Math. Stat.*, **42** (1971), 1977–1991.
- [35] J. Kiefer and J. Wolfowitz, Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters, *Ann. Math. Stat.*, **27** (1956), 887–906.
- [36] J. Kiefer and J. Wolfowitz, The equivalence of two extremum problems, *Canadian J. Mathematics*, **12** (1960), 363–366.
- [37] B. Lindsay, The geometry of mixture likelihoods: a general theory, *Ann. Stat.*, **11** (1983), 86–94.
- [38] A. Mallet, A maximum likelihood estimation method for random coefficient regression models, *Biometrika*, **73** (1986), 645–656.
- [39] A. Mallet, France Mentré, Jean-Louis Steimer, and François Lokiec, Nonparametric maximum likelihood estimation for population pharmacokinetics with application to cyclosporine, *J. Pharmacokinetics and Biopharmaceutics*, **16** (1988), 311–327.
- [40] R. Phelps, *Lectures on Choquet's Theorem*, Van Nostram, Princeton, 1955.
- [41] Y.V. Prohorov, Convergence of random processes and limit theorems in probability theory, *Theor. Probability Appl.*, **1** (1956), 157–214.
- [42] G.A. Seber and C.J. Wild, *Nonlinear Regression*, Wiley, Hoboken, 2003.

- [43] A. Schumitzky, Nonparametric EM algorithms for estimating prior distributions, *Applied Mathematics and Computation*, **45** (1991), 143–157.
- [44] L.B. Sheiner and S.L. Beal, Evaluation of methods for estimating population pharmacokinetic parameters. I. Michaelis-Menten model: Routine clinical pharmacokinetic data, *Journal of Pharmacokinetics and Biopharmaceutics*, **8** (1980), 553–571.
- [45] L.B. Sheiner, B. Rosenberg, and K.L. Melmon, Modelling of individual pharmacokinetics for computer-aided drug dosage, *Computers and Biomedical Research*, **5** (1972), 441–459.
- [46] J.L. Steimer, A. Mallet, and F. Mentre, Estimating interindividual pharmacokinetic variability, in *Variability in Drug Therapy: Description, Estimation, and Control*, M. Rowland et al., eds., Raven Press, New York, 1985.
- [47] J. Sinko and W. Streifer, A new model for age-size structure of a population, *Ecology*, **48** (1967), 910–918.
- [48] E.G. Vonesh, H. Wang, L. Nie, and D. Majumdar, Conditional second-order generalized estimating equations for generalized linear and nonlinear fixed-effects models, *J. American Statistical Association*, **97** (2002), 271–283.
- [49] A. Wald, Note of the consistency of the maximum likelihood estimate, *Ann. Math. Stat.*, **20** (1949), 595–601.
- [50] J. Wolfowitz, On Wald’s proof of the consistency of the maximum likelihood estimate, *Ann. Math. Stat.*, **20** (1949), 601–602.
- [51] H.P. Wynn, The sequential generation of D-optimal experimental designs, *Ann. of Math. Stat.*, **41** (1970), 1655–1664.