Nonparametric estimation of a mixing distribution for a population of nonlinear stochastic dynamical systems

Alona Kryshchenko[1,2], David S. Bayard[3,2],
Michael N. Neely[4,2], Julian D. Otálvaro[2], Alan Schumitzky[5,2]
[1]Department of Mathematics,
California State University of Channel Islands,
1 University Dr., Camarillo, CA 93012, USA
alona.kryshchenko@csuci.edu
[2]Laboratory of Applied Pharmacokinetics and Bioinformatics,
Children's Hospital-Los Angeles,
4650 Sunset Blvd., Los Angeles, CA 90027, USA
[3]Jet Propulsion Laboratory
California Institute of Technology, Pasadena CA, 91109, USA;
dbay007@earthlink.net
[4]Pediatric Infectious Diseases,
Children's Hospital of Los Angeles, Keck School of Medicine,
University of Southern California, Los Angeles, CA 90027, USA
mneely@chla.usc.edu
juliandavid347@gmail.com
[5]Department of Mathematics,
University of Southern California, Los Angeles, CA 90089, USA
schumitzky@gmail.com

April 2, 2021

Abstract

In this paper, we develop a nonparametric maximum likelihood method for estimating the mixing distribution of a population of nonlinear stochastic dynamical systems. In our application to pharmacokinetics, this includes population models with process and measurement noise. Most research in mixing distributions only consider measurement noise. The advantages of the models with process noise are that, in addition to the measurement errors, the uncertainties in the model itself are taken into the account. For example, in pharmacokinetic models, errors in dose amounts, administration times, and timing of blood samples are typically not included. For linear stochastic models, we can use linear Kalman-Bucy filtering to calculate the likelihood of the observations. For nonlinear stochastic models, we use particle filtering to calculate the likelihood of the observations and then employ a convex optimization algorithm to find the nonparametric maximum likelihood estimate of the mixing distribution. We then use the directional derivatives of the estimated mixing distribution to show that the result found attains a global maximum. Several examples are given using simulated data from a one compartment pharmacokinetic nonlinear stochastic model with randomly changing parameters.

AMS Subject Classification: 65C35, 62P10, 49N45, 49N99, 68U20, 92C45

Key Words and Phrases: mixing distribution, stochastic models, nonparametric maximum likelihood, convex optimization, pharmacokinetic population models.

1 Introduction

In this paper, we develop a nonparametric maximum likelihood method for estimating the mixing distribution of a population of nonlinear stochastic dynamical systems. In our application to pharmacokinetics, this includes population models with process and measurement noise and randomly changing IOV parameters. Most research in mixing distributions only consider measurement noise. The advantages of the models with process noise are that, in addition to the measurement errors, the uncertainties in the model

itself are taken into the account. For example, in pharmacokinetic models, errors in dose amounts, administration times, and timing of blood samples are typically not included. For linear stochastic models, we can use linear Kalman-Bucy filtering to calculate the likelihood of the observations. For nonlinear stochastic models, we use particle filtering to calculate the likelihood of the observations and then employ a convex optimization algorithm to find the nonparametric maximum likelihood estimate of the mixing distribution. We then use the directional derivatives of the estimated mixing distribution to show that the result found attains a global maximum. Several examples are given using simulated data from a one compartment pharmacokinetic nonlinear stochastic model with process noise, measurement noise and randomly changing parameters.

The mixing distribution problem we consider can be stated precisely as follows. Let $Y_1, Y_2, ..., Y_N$ be a sequence of independent but not necessarily identically distributed random vectors. Each Y_i is a vector of one or more observations from each of N subjects in the population. Let $\theta_1, \theta_2, ..., \theta_N$ be sequence of independent and identically distributed random vectors belonging to a compact subset Θ of Euclidean space with common but unknown distribution F. The $\{\theta_i\}$ are not observed. It is assumed that the conditional densities $p(Y_i|\theta_i)$ are known, for $i = 1, ..., N$. The mixing distribution of Y_i with respect to F is given by $p(Y_i|F) = \int p(Y_i|\theta_i)dF(\theta_i)$. Because of independence of the $\{Y_i\}$, the likelihood of the mixing distribution of the all the Y_i with respect to F is given by

$$L(F) = p(Y_1, ..., Y_N|F) = \prod_{i=1}^{N} \int p(Y_i|\theta_i)dF(\theta_i) \qquad (1)$$

The mixing distribution problem is to maximize $L(F)$ with respect to all distributions F on Θ.

It is important later to note that $L(F)$ is a convex function of F. Further, it is shown in Lindsay, 1983 [L83] and Mallet, 1986 [M86],under simple hypotheses, that the global maximizer F^{ML} of $L(F)$ is a discrete distribution with at most N support points, where N is the number of subjects in the population and a support point is a vector of model parameter values with nonzero probabil-

ity.

It is common in the literature on mixing distributions to consider a deterministic model of the conditional density $p(Y_i|\theta_i)$, i.e. to consider Y_i to be a function of θ_i with additive measurement noise ν_i, which is assumed to be normally distributed with mean vector zero and known covariance matrix $V_i(\theta)$. In practice however the model for $p(Y_i|\theta_i)$ is not deterministic as it is affected by the random state-space process of generating Y_i. For example, in case of pharmacokinetic problems, errors in the dose amount and timing, so called process noise, are not included in the deterministic models. It is shown in Jelliffe et al. 1992 [JSG92], that the resulting drug concentrations are heavily influenced by these kinds of errors. The fundamental importance of our paper is that the method we describe is able to account for process and measurement noise in the models. In particular, we consider Y_i to be a vector of discrete measurements for a linear or nonlinear stochastic differential equation, where the state vector includes process noise and the measurement vector includes measurement noise.

Once the exact form of the conditional density $p(Y_i|\theta_i)$ has been determined, there are a number of algorithms that can be used for solving the mixing distribution problem, see Yamada et al. 2021 [YNB+21], and the references therein. In this paper we use the Primal Dual Interior Point Method of Convex Optimization, [YNB+21].

This paper is organized in the following way. In Section 2 we discuss the types of models considered based on the form of the conditional densities $\{p(Y_i|\theta_i)\}$. In Section 3 we show that the task of determining log likelihood can be reduced to a problem of calculating $p(y_{i(k+1)}|y_{i1}, ..., y_{ik})$, where $y_{i1}, ..., y_{ik}$ are measurements at times $t_1, t_2, ..., t_k$ for each individual subject i. We discuss briefly common simple regression models, which do not allow for the important process noise errors. Then in Sections 3.1 and 3.2 we introduce the main models of interest, where Y_i is the discrete measurement for a stochastic differential equation. These stochastic models accommodate process noise errors. For linear stochastic differential equations, the differential equations can be exactly represented by discrete time equations. The likelihood function can be calculated exactly in terms of a linear Kalman-Bucy filter [KSG15]. For nonlinear stochastic differential equations, when exact representations

are not possible, numerical discretization methods must be used. The likelihood function can be calculated approximately using a particle filter [ref.], among other methods, eg. Extended Kalman Filtering [MKD+07], and MCMC [DS13].

Equally important in this paper is the method for calculating the global optimum F^{ML}. This is discussed in Section 4. Our method is different from the popular methods in the literature, [DS13]. Our method is called Nonparametric Adaptive Grid (NPAG), see[YNB21]. It is based on modern convex analysis and adaptive discrete optimization. We note that there is a simple condition, which guarantees that a proposed solution F is indeed a global optimum. This is unique to convex optimization and is employed in this paper.

Finally in Section 5, we end with an important application of the paper to pharmacokinetic population models. We study a one-compartment model with process and measurement noise and give numerical examples. In particular, we simulate measurements with different amount of process and measurement noise and then compare the simulated distributions F to the estimated distributions F^{ML}, using a two-sample Kolmogorov-Smirnov test [P83]. The results show that sample from the simulated distribution F can not be distinguished from the sample from the estimated distributions F^{ML} at the 5 percent level of significance.

2 Models

The difficulty of the mixing distribution problem is determined by the form of the conditional densities $p(Y_i|\theta_i)$.

2.1 Nonlinear Regression Models

Most of the results in the literature for this mixing distribution problem assume a regression equation of the form

$$Y_i = h_i(\theta_i) + \nu_i, i = 1, ..., N \tag{2}$$

where h_i is a known vector function and ν_i is the normal measurement noise with mean vector zero and known covariance matrix $V_i(\theta)$. In this case $p(Y_i|\theta_i) = \eta(Y_i, h_i(\theta_i), V_i(\theta))$, where $\eta(Y, M, \Sigma)$

is the density of the multivariate normal distribution with mean vector M, covariance matrix Σ, evaluated at the vector Y.

2.2 Stochastic Differential Equation Models

In this paper we consider the mixing distribution problem in a much more general setting. It is assumed that the observation vector Y_i is the discrete output of a stochastic differential equation (SDE) as described below, where we have suppressed the subscript i for simplicity.

$$dx(t) = f(x, u, t, \theta)dt + g(x, u, t, \theta)dW(t), t \geq t_0, \tag{3a}$$

$$x(t_0) \sim N(\hat{x_0}(\theta), \Sigma_0(\theta)) \tag{3b}$$

$$y_k = h_k(x(t_k), u(t_k), \theta) + \nu_k, k = 1, ..., m \tag{3c}$$

In (3a) at time t, $x(t)$ is the state vector; $u(t)$ is a known piecewise continuous input; $\theta \in \Theta$ is a vector of subject-specific parameters; f and g are known continuous vector functions; $W(t)$ represents the standard Wiener process with the property that $W(0) = 0$ and for $t > 0$ and $d > 0$, $W(t+d) - W(t) \sim N(0, d)$, where $N(m, S)$ represents the multivariate normal distribution with mean vector m, covariance matrix S. In (3c) at time t_k, y_k is the noisy measurement vector; h_k is a known continuous vector function; and $\nu_k \sim N(0, V_k(\theta))$ is the vector measurement noise.

In the case when f and g are linear functions of their respective arguments and g does not depend on $x(t)$, the stochastic system of (3a),(3b) and (3c) is called linear. Otherwise the system is called nonlinear.

3 Likelihood Calculations

By the telescoping property of conditional densities we have:

$$p(Y_i|\theta) = p(y_{i1}, ..., y_{im}|\theta) = \prod_{k=0}^{m-1} p(y_{i(k+1)|y_{i1},...,y_{im},\theta}) \tag{4}$$

and therefore the log likelihood function can be written as

$$l(F) = log(L(F)) = log(p(Y_1, ..., Y_N|F))$$
$$= \sum_{i=1}^{N} log(\int \prod_{k=0}^{m-1} p(y_{i(k+1)}|y_{i1}, ..., y_{im}, \theta) dF(\theta)) \quad (5)$$

Let $I_{ik} = (y_{i1}, ..., y_{ik}; u_{i1}, ..., u_{ik})$ be a vector of all measurements and inputs. The crux of the likelihood calculation is in the calculation of $p(y_{i(k+1)}|I_{ik}, \theta)$ for an individual subject.

In the regression case of Eq. (2), $p(y_{i(k+1)}|I_{ik}, \theta) = p(y_{i(k+1)}|\theta)$, and the problem is much simpler. In the general case of Eqs. (3a) - (3c) , the calculation of $p(y_{i(k+1)}|I_{ik}, \theta)$ is a problem of nonlinear filtering. For the application to population parmacokinetics, Klim et al. 2009 [KMK09], approximate this calculation with the extended Kalman filter. Donnet and Samson, 2014 [DS14] use MCMC methods. Approximations by particle filtering has been determined to be the most accurate, see Crisan and Doucet, 2002 [CD02] and Gordon, Salmond and Smith [GSS93].

3.1 Continuous state-discrete observations linear stochastic model

In this subsection, we consider the linear stochastic case. Assume we focus on an individual subject. The subscript i will be suppressed. Now consider Eqs. (3a) - (3c), and assume f, g and h are linear matrix/vector functions. Eqs. (3a) - (3c) then become

$$dx(t) = A(t, \theta)x(t)dt + B(t, \theta)u(t)dt + \sigma dW(t), t \geq 0; \quad (6a)$$
$$x(t_0) \sim \ N(x_0(\theta), \Sigma_0(\theta)), \quad (6b)$$
$$y_k = C_k(\theta)x(t_k) + v_k, k = 1, ..., m, \quad (6c)$$

where $A(t, \theta)$, $B(t, \theta)$ and $C_k(\theta)$ are known continuous matrices. Now assume $u(t)$ is piece-wise constant with $u(t) = u_{k+1}$ on the interval $[t_k, t_{k+1}]$. Then, using the Ito formula, Eq.(6a) can be integrated over the interval $[t_k, t_{k+1}]$ to give an exact discrete time system [J80, p. 199)]:

$$x_{k+1} = A_k(\theta)x_k + B_k(\theta)u_k + w_{k+1}(\theta); x(t_0) \sim N(x_0(\theta), \Sigma_0(\theta)) \quad (7)$$

where $x_k = x(t_k)$; $A_k(\theta) = \Phi(t_{k+1}, t_k, \theta)$ is the fundamental matrix of the homogeneous part of Eq.(6a),

$$\begin{aligned} B_k(\theta) &= \int_{t_k}^{t_{k+1}} \Phi(t_k, s, \theta) B(s, \theta) ds; \\ w_{k+1}(\theta) &= \int_{t_k}^{t_{k+1}} \Phi(t_{k+1}, s, \theta) \sigma dW(s); \end{aligned} \tag{8}$$

and $w_{k+1}(\theta)$ is a zero mean Gaussian sequence with the covariance matrix

$$E[w_{k+1}(\theta) w_{k+1}(\theta)^T] = \int_{t_k}^{t_{k+1}} \sigma^2 \Phi(t_{k+1}, s, \theta) \Phi^T(t_{k+1}, s, \theta) ds, \tag{9}$$

The main theoretical result for calculating $p(y_{k+1}|I^k, \theta)$ in the linear stochastic system of Eqs.(6)-(9) is given by the following Proposition, which is proved in Kumar and Varaiya (1986) [KV86].

Proposition: *Define*

$$\begin{aligned} x_{k+1|k}(\theta) &= E[x_{k+1}|Y^k] \\ \Sigma_{k+1|k}(\theta) &= E[(x_{k+1} - x_{k+1|k})(x_{k+1} - x_{k+1|k})^T|Y^k] \end{aligned} \tag{10}$$

where $Y^k = (y_1, ..., y_k)$. *Then the conditional density* $p(y_{k+1}|Y^k, \theta)$ *is normal with mean vector* $C_{k+1}(\theta) x_{k+1|k}$ *and covariance matrix* $C_{k+1}(\theta) \Sigma_{k+1|k} C_{k+1}(\theta)^T + V_{k+1}$.

Note: In[KV86] the above Proposition is proved in the case that the control u_k is a nonlinear feedback function of Y^k .

3.2 Linear SDE vs Nonlinear SDE

The linear SDE given by Eqs.(6a)-(6c) benefit from two properties:

1. The continuous time equation Eq.(6a) has an exact discrete time solution Eq. (7).

2. The likelihood function given by Eq.(5) has an exact solution.

The likelihood function can be calculated by linear Kalman-Bucy filtering [KSG15].

In the nonlinear case of Eqs.(3a)-(3c), these benefits do not hold except for a few special examples. So instead of 1. we have

1'. The continuous time equation Eq.(6a) has to be calculated by a numerical approximation.

To implement this we use the Euler-Maruyama method (E-M) [KP93]

And instead of 2. we have

2'. The likelihood function calculation is a problem in nonlinear filtering.

To implement this we use the Particle Filtering (PF) method described in Schoen et al. 2005 [SGN05].

Nonlinear pharmacokinetic example: First consider the simple deterministic one-compartment model with bolus input B_0

$$\begin{aligned} dX/dt = -KX(t), t > 0, \\ X(0) = B_0 \end{aligned} \tag{11}$$

where K is the 'elimination rate constant'. Adding process noise and writing in standard SDE form we have:

$$\begin{aligned} dX(t) = -KX(t)dt + \sigma_1 dW_1, t > 0, \\ X(0) = X_0 \end{aligned} \tag{12}$$

In reality, the so-called the "Elimination Rate Constant' K is not constant but varies randomly about its mean. We model K by the SDE

$$\begin{aligned} dK = -(K–K_0)dt + \sigma_2 dW_2, \\ K(0) = K_0 \end{aligned} \tag{13}$$

Eq.(13) is "mean-reverting", which means that the random solution K tends to its mean value, which is appropriate for an Elimination Rate Constant . Now Eq.(13) is linear, but combined with Eq.(12) we have a nonlinear system. Note for later purposes, if $\sigma_2 = 0$, then $K(t) \equiv K_0$

The measurement equation we consider is

$$\begin{aligned} Y(t_k) &= X(t_k)/V + \sigma_3 v_k, \\ v_k &\sim N(0,1), k = 1, 2, \ldots, N \end{aligned} \tag{14}$$

where $Vol = V$ is the "volume of distribution".

3.3 Numerical Experiments

To illustrate the novel aspects of our stochastic population analysis, we sometime assume $V \equiv 1$. The population parameter is now K_0 and we are interested how K_0 varies over the population. Since our population parameter K_0 is 1-dimensional, the graph of the gradient of the log-likelihood is more informative.

Otherwise we assume that for each subject V is constant but varies over the populations with a normal distribution

$$V \sim 1.0 + 0.2N(0,1) \tag{15}$$

We will do a number of experiments to show that our numerical methods are consistent with the exact solutions. For our approximate examples, we will treat equations (12) - (13) as a nonlinear SDE and use the Euler-Maruyama (E-M) method for solution (not to be confused with the Dempster-Laird EM algorithm of statistics)

Now let us consider the general nonlinear vector valued SDE:

$$\begin{aligned} dX &= F(X)dt + G(X)dW, \\ X(0) &= X_0 \end{aligned} \tag{16}$$

For Eq.(16), the Euler-Maruyama method then gives the discrete time system:

$$\begin{aligned} X(t_{k+1}) = X(t_k) + F(X(t_k))\Delta_k + G(X(t_k))(W(t_{k+1}) – W(t_k)), \\ \Delta_k = t_{k+1} – t_k \end{aligned} \tag{17}$$

We consider our original problem (12) - (14) on the time interval [0, 1] with measurements at the 5 points $t = \{0.2, 0.4, 0.6, 0.8, 1\}$. For the discretized system we choose a small number Δ so that

the SDE evaluation times are at $s = \{0, \Delta, 2\Delta, \ldots, m\Delta\}$, where $m = 1/\Delta$. Assume $0.2 = L\Delta$, then the measurements occur at the times $M = \{L\Delta, 2L\Delta, 3L\Delta, 4L\Delta, 5L\Delta\}$.

One of the problems with comparing exact solutions vs approximate solutions is that the dynamics are evaluated at every Δ time units, while the measurement are evaluated only at every $L\Delta$ time units. To approximate the solution of (12) - (13) at the discrete values of $[0, 1]$, the increment Δ must be small. On the other hand, the additional times the system has to be evaluated gives rise to additional numerical calculations and errors. So a compromise must be chosen, see [DS14].

Just like the problem of making Δ too small, there is a similar problem of making the number of particles in PF too large, see [DS14]

Reflecting Boundary. Also note that the solution $K(t)$ to Eq.(13) may become negative, even if $K_0 > 0$. To mitigate against this possibility, we set $[-\infty, 0]$ as a "Reflecting Boundary" as follows: For a particular subject, let $K(j), j = 1, 2..., j_1$ be the discrete values for E-M method in Eq.(17) until $K(j_1 + 1) < 0$, in which case set $K(j_1 + 1) = K(j_1)$. Then continue the E-M method until $K(j_2 + 1) < 0$, in which case set $K(j_2 + 1) = K(j_2)$. Etc. See [LS84] for justification.

4 Optimization of the Likelihood

Of equal importance in calculating the maximum likelihood estimate is the optimization of the likelihood function in Eq.(5) with respect to F.

4.1 Nonparametric Maximum Likelihood by Convex Optimization

In this paper, the optimization of $L(F)$ in Eq. (1) will be done by convex programming. Consider a set of points $\Theta' = (\theta_1, ..., \theta_n)$ in Θ. Then the objective function $L(F_n)$ restricted to Θ' is given by

$$L(F_n) = \prod_{i=1}^{N} \sum_{k=1}^{n} w_k p(Y_i|\theta_k) \tag{18}$$

where F_n is a discrete distribution with support points on Θ' and corresponding weights (w_k).

The problem of maximzing $L(F_n)$ with respect to (w_k) is convex. We solve this problem by the Primal-Dual Interior-Point method, for details see Yamada et al. 2021 [YNB+21]. This method is very fast and allows for high dimensional n.

Let F_n^* be the maximizer of $L(F_n)$ and F^{ML} be the maximizer of $L(F)$. If Θ is compact and the probabilties $p(Y_i|\theta_k)$ are piecewise continuous, then it can be shown that $L(F_n^*)$ tends to $L(F^{ML})$ as n goes to infinity, see [BW12]. Further the distribution F_n^* converges to F^{ML} in the weak topology [BW12]. This implies, for example, that the moments of F_n^* (like means and variances) converge to the corresponding moments of F^{ML}.

For the low dimensional numerical experiments in Section 5, we can choose our subset Θ' such that the difference between the log-likelihoods $log(L(F_n^*))$ and $log(L(F^{ML}))$ is less than 1%, see Section 4.2 for details. For higher dimensional problems, our Nonparametric Maximum Likelihood method will require our Adapted Grid method. For complete details see Yamada et al. 2021 [YNB+15].

In all of our numerical experiments, the discrete subset Θ' will consist of quasi-random Halton points [H64]. The term *Nhalton* will be the number of Halton points.

Note: Halton points are deterministic low-discrepancy sequences used to generate points in space for numerical methods such as Monte Carlo simulations. Halton points have better uniform covering properties than uniform random points.

4.2 Directional derivative conditions to check optimality

One method to check that F_n^* has converged to a global maximum is described in Lindsay 1983 [L83]. It uses the so-called directional derivative to check if a current distribution F is in fact optimal. (Convexity theory is unique in this sense that a proposed maximizer can be checked for optimality). For this purpose we need the directional derivative $D(F, \theta)$ of F in the direction of the atomic distribution $\delta(\theta)$. It follows $D(F, \theta)$ is defined by

$$D(F, \theta) = \sum_{i=1}^{N} \frac{p(Y_i|\theta)}{p(Y_i|F)} - N, \theta \in \Theta \tag{19}$$

Theorem (Lindsay 1983 [L83],): F^{ML} is the distribution that maximizes $L(F)$ in Eq. (1) with respect to all distributions on Θ if and only if

$$max[D(F^{ML}, \theta) : \theta \in \Theta] = 0 \tag{20}$$

Moreover to be optimal, the support of F^{ML} must be contained in the set of θ for which the function $D(F^{ML}, \theta)$ attains a global maximum.

In addition, Lindsay 1983 [L83] has a very practical result which can be used to test how close any distribution F is to F^{ML}. Namely,

$$|log(L(F^{ML}) - log(L(F))| \leq N \cdot log(1 + d/N) \tag{21}$$

where $d = max[D(F, \theta) : \theta \in \Theta]$.

In our applications, we apply this criterion to the distribution F_n^*, i.e, the optimal F on the discrete set Θ'. To check that $log(L(F_n^*))$ is within 1% of $log(L(F^{ML})$ we devide the r.h.s. of Eq.(21), by $|log(L(F_n^*)|$. This implies Eq.(22) since

$$|log(L(F_n^*))| \leq |log(L(F^{ML})| \tag{22}$$

Note: In Experiments 1 and 2 below, we run our NPML algorithms to convergence by using the Adaptive Grid part of our NPAG program [YNB+21]. In this way we can check empirically that the theoretical results based on the discrete subset Θ' are in fact valid on the whole space Θ.

5 Applications to Pharmacokinetic Population Analysis

Consider again our one compartment PK model for drug concentration $X(t)$ with additive process noise and random "Elimination Rate Constant" $K(t)$. Our SDE is then given by Eqs. (12) and (13), with measurement Eq. (14).

In all experiments the number of subjects will equal 100. Further, the initial condition K_0 will vary randomly over the population with a bimodal distribution
$F(K_0) = 0.5 * N(m_1, s_1^2) + 0.5 * N(m_2, s_2^2)$
with $m_1 = 0.5, m_2 = 1.5, s_1 = 0.05, s_2 = 0.15$.
We note that the estimated distributions also show bimodality.

In Experiments 2 and 3, Vol will be a known constant equal to 1. This allows for the graph of $D(F, \theta)$ to be very instructive.

In Experiments 1 and 4, Vol will be constant but unknown for each subject, and will vary randomly over the population with distribution:

$F(Vol) = 1 + 0.2 * N(0, 1)$

Again, the population analysis problem is to estimate the distribution of population parameter vector (K_0, Vol) given the data of Eq. (14) for each subject.

Note: For population pharmacokinetic models, there are two types of parameters which are relevant for our numerical Experiments. A parameter is said to have property IIV (Inter-Individual Variability), if the parameter is constant for each subject but varies randomly over the population. Examples are the above parameters $Vol = V$ and $Kel(0) = K_0$. The estimation of the distribution of IIV parameters is the subject of pharmacokinetic population analysis.

A parameter is said to have property IOV (Inter-Occasion Variability), if for each subject, the parameter varies randomly over the course of therapy. An example of an IOV parameter is the above $Kel = K(t)$

Initially, an IOV parameter was defined as a stepwise constant function with random jumps at arbitrary occasions [KS93]. Lavielle and Delattre questioned the propriety of this definition and proposed instead that an IOV parameter should be modeled as continuous stochastic process defined by a Stochastic Differential Equation (SDE) [LD12]. Deng et al. 2016 [DPK16]. proposed the same model and we adopt this definition.

5.1 Testing accuracy of the estimated distributions

In all of our experiments, we check the accuracy of our algorithms by testing the hypothesis that the original distribution F is well approximated by the estimated distribution F_n^* . To accomplish this we use the Two Sample Kolmogorov-Smirnov (KS) test. KS tests the Null Hypothesis that independent samples from F and F_n^* are drawn from the same underlying distribution (at a 5% level of significance).

Further we use Eq. (22) to check that

$$|log(L(F^{ML}) - log(L(F_n^*))|/|log(L(F_n^*))| \leq 1\% \tag{23}$$

Experiments 2 and 3 have only one random variable, i.e. K_0. In this case the Matlab program "kstest2.m" accomplished the KS test. One the other hand Experiments 1 and 4 have two random variables, i.e. (Kel, Vol). In this case Matlab has no 2-dimensional KS test. For this purpose we use the 2-dimensional KS test of Peacock [P83].

Both KS tests require independent samples from F and F_n^*. For independent samples from F we use what was generated. On the other hand, our method defines F_n^* as a discrete distribution. To generate independent samples from F_n^*, we use the Matlab program "gendist.m" which generates M independent samples from a given discrete distribution. In all cases we take M equal to 100 (the number of subjects).

We will consider 4 numerical experiments:

Experiment1 :$\sigma\mathbf{1} = \mathbf{1}, \sigma\mathbf{2} = \mathbf{0}, \mathbf{Vol} \sim \mathbf{random}$
Experiment2 :$\sigma\mathbf{1} = \mathbf{1}, \sigma\mathbf{2} = \mathbf{0}, \mathbf{Vol} = \mathbf{1}$
Experiment3 :$\sigma\mathbf{1} = \mathbf{1}, \sigma\mathbf{2} = \mathbf{1/2}, \mathbf{Vol} = \mathbf{1}$
Experiment4 :$\sigma\mathbf{1} = \mathbf{1}, \sigma\mathbf{2} = \mathbf{1/2}, \mathbf{Vol} \sim \mathbf{random}$

There are three "constants" which determine the accuracy of the program:

1) The step size (*delt*) in the E-M method

2) The number of particles (*Nparticle*) in PF

3) The number of support points (*Nhalton*) in the feasible region Θ

In Experiments 1 and 4, the feasible region is all (Kel, Vol) in the set Θ =[0.35, 2]$\times$[0.35, 2.0] .

In Experiments 2 and 3 there is only one parameter to be updated, namely K_0. Therefore the feasible region is all Kel in the set Θ =[0.35, 2].

Accuracy Requirements

In all of our experiments, we solve for F_n^* on a sufficiently dense discrete subset Θ' contained in Θ. In this case we can use the convex program described in Section 4.1 without adding the more complicated Adaptive Grid method. We have shown in these Experiments that the maximum likelihoods on Θ' and Θ differ by less that 1 percent.

In Experiments 1 and 2, using the appropriate KS test, we show that KS accepts the Null Hypothesis that independent samples from the approximate F_n^* and the exact F^{ML} are drawn from the same underlying distribution (at a 5 percent level of significance).

In Experiments 3 and 4, we show that KS accepts the Null Hypothesis that independent samples from the approximate F_n^* and the initial distribution F are drawn from the same underlying distribution (at a 5 percent level of significance).

Finally, in Experiments 1 and 2, using the theory of Section 4.2, we show that the difference between the Log-Likelihoods $L(F_n^*)$ and $L(F^{ML})$ differ by less than 1 percent. We will refer to the above conditions as the *Accuracy Requirements*

For 100 subjects in the population, we show that we can achieve the above *Accuracy Requirements* with the following initial conditions:

Initial Conditions:
i) $delt = 0.002$
ii) $Nparticle = 1000$
iii) $Nhalton = 4000$ for the feasible set $\Theta = [0.35, 2] \times [0.35, 2.0]$
iv) $Nhalton = 1000$ for the feasible set $\Theta = [0.35, 2]$

We first consider

Experiment1 :$\sigma\mathbf{1} = \mathbf{1}, \sigma\mathbf{2} = \mathbf{0}, \mathbf{Vol} \sim \mathbf{random}$
This experiment is used only to check the accuracy of the our Euler-Maruyama (E-M) method for solving the SDE of Eqs. (12)-(13) and our Particle Filter (PF) method of calculating likelihoods. As $\sigma_2 = 0$, Eqs. (12)-(13) are now linear and we can find the exact solution to the SDE and the exact calculation of the log-likelihood. Using the program in [KSG+15] for the exact solution and our E-M and PF programs for the approximate solution, we verify that our *Accuracy Requirements* are satisfied. The initial distribution is given in Figure 1 and the estimated distribution is given in Figure 2. Note that the estimated distribution captures the bi-modality of the original distribution. We compare the program in [KSG+15] which uses the linearity of Eqs (12)-(14) to give exact solutions to the SDE and Kalman filtering to calculate the exact likelihoods with our current program which uses E-M and PF for the same purposes. Both programs use the same optimizer, the same observations and the same initial grid on the feasible set $[0.35, 2] \times [0.35, 2]$.

Relative to the *Accuracy Requirements*, the *Initial Conditions* specifically are sufficient so that the resulting Log-Likelihoods $L(F_n^*)$ and $L(F^{ML})$ differ by less than 1 percent. Further the 2-dimensional KS test of Peacock accepts the Null Hypothesis that independent samples from the approximate F_n^* and the exact F^{ML} are drawn from the same underlying distribution (at a 5 percent level of significance).

As this is sufficient accuracy, we will use the above initial conditions

for our Experiments 1 and 4.

In Experiments 2 and 3, $Vol = 1$ and there is only one IIV parameter K_0 which we assume is in the feasible set [0.3:2]. In this case *Initial Condition* iii) is replaced by

iv) $Nhalton = 1000$ on the feasible set $[0.35, 2]$

Experiment2 : $\sigma\mathbf{1} = \mathbf{1}, \sigma\mathbf{2} = \mathbf{0}, \mathbf{Vol} = \mathbf{1}$

The conditions for this population analysis experiment are the same as Experiment 1, except that now we assume $Vol = 1$. Again we verify hat the *Accuracy Requirements* are satisfied. Since there is only one updated parameter, K_0, the graph of the gradient $D(F^{ML}, \theta)$ as a function of θ in Figure 3 is more instrctive. Recall that F_n^{ML} is infact the nonparametric maximum likelihood method estimate of the the population distribution if and only if

D_max=max$[D(F_n^{ML}, \theta) : \theta \in \Theta] = 0$

Further, from Eq. (21) the difference between the estimated log likelihood and the maximum log likelihood is less than $Nlog(1 + D_max/N)$.
In this example D_max was calculated to be D_max=3.4*e-07.

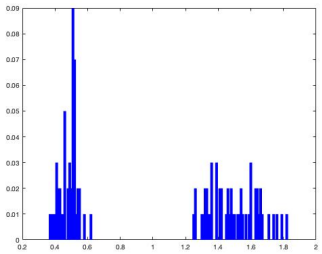

Figure 1: 100 Subjects - Initial Kel Distribution

Our (NPML) estimate is the discrete distribution F^{ML} given graphically in Fig. 2.

Note: To test the Hypothesis that $F = F^{ML}$ we use the Two Sample KS Matlab program "kstest2.m".

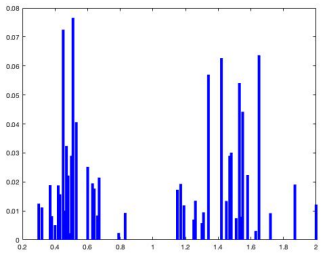

Figure 2: 100 Subjects - Estimated Kel Distribution - $\sigma_2 = 0$

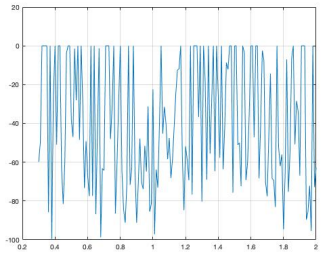

Figure 3: Gradient of Log-Likelihood

Experiment3 : $\sigma\mathbf{1} = \mathbf{1}, \sigma\mathbf{2} = \mathbf{1/2}, \mathbf{Vol} = \mathbf{1}$

In this population analysis experiment we assume $\sigma_2 = 1/2$. All other details are the same as in Experiment 2. The trajectories of $X(t)$ and $K(t)$ are given in Fig. 4 and Fig. 5. There are now two IIV paramters $K0$ and V.

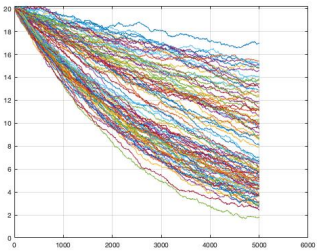

Figure 4: 100 Concentrations

In Experiment 3,our SDE system of Eqs (12)-(14) is nonlinear and we can not compare our likelihood results with exact values as before. To check the correctness of our algorithm, we still can

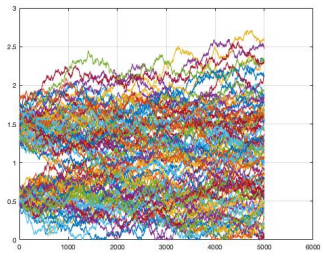

Figure 5: 100 Elimination Rate Constants

compare the initial distribution of F in Fig. 1 with the estimated NPML distribution F^{ML} in Fig. 6.

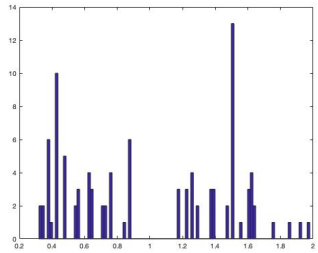

Figure 6: 100 Subjects - Estimated Kel Distribution - $\sigma_2 = 0.5$

Using the Kolmogov-Smirnov test, we again accept the hypothesis: $F = F^{ML}$ at the 5% level of significance.

The Gradient of the Log-Likelihood is now two dimensional and is shown to have the required properties for convergence.

Experiment4 : $\sigma\mathbf{1} = \mathbf{1}, \sigma\mathbf{2} = \mathbf{1/2}$, **Vol** $\sim$ **random**

In this experiment we assume $Vol \sim$ random. All other details are the same as in Experiment 3.

In this experiment our SDE system in Eqs. (12)-(14) is nonlinear and we can not compare our likelihood results with exact values as before. Further, for some realizations $K(t)$ can become negative. To mitigate this problem, we set $[-\infty, 0]$ to be a Reflecting Boundary as described above.

To check the correctness of our algorithm, we still can again compare the initial distribution of F in Fig. 1 with the estimated

NPML distribution F^{ML} . Using the 2-dimensional KS Peacock test, we again accept the hypothesis: $F = F^{ML}$ at the 5% level of significance.

Further, as discussed in Section 4.2, the two-dimension Gradient of the Log-Likelihood has the required properties for convergence of the algorithm.

6 Conclusions

In this paper, we developed a nonparametric maximum likelihood (NPML) method for estimating the mixing distribution of a population of nonlinear stochastic dynamical systems. Our main application is to population pharmacokinetics, which include pharmacokinetic models with process and measurement noise and randomly changing parameters. Our models are defined by a system of stochastic differential equations (SDE). We solve the SDEs numerically using the Euler-Maruyama (E-M) method. We calculate the corresponding likelihoods using a Particle Filter (PF). The accuracy of our NPML algorithm depends on 3 quantites: 1) the number and size of the discretetization steps in solving the SDE, 2) the number and placement of the initial grid of support points and 3) the number of particles in the PF. We test the accuracy of our NPML method in a number of ways: final likelihood values and final distributions.

All our Experiments have 100 subjects in the population. Our PK model is one-compartment with IV bolus, additive process noise, Volume of Distribution Vol changing from subject to subject (inter-subject variabilty IIV) and the Elimination Rate Constant Kel changes from subject to subject (IIV) and randomly changes for each subject during therapy (inter-occasion variability IOV). In Experiment 1, we only consider IIV changes in Vol and Kel. The resulting model is linear and we can compare our NPML results in likelihood and distribution with the exact answers. In the rest of our numerical Experiments, our model is nonlinear and we check likelihood accuracy using the theoretical properties of the likelihood gradient $D(F, \theta)$ and compare distributions usin KS. In Experiments 2 and 3, we assume Vol is known but Kel is both IIV

and IOV. In Experiment 4 we allow IIV in Vol and both IIV and IOV in Kel. In this case Kel can become negative. To mitigate this problem we use a Reflecting Boundary for the Kel SDE.

7 Future Work

The Euler-Maruyama method for solving SDEs is slow and inaccurate without taking a very fine discretization. In the sequal to this paper, we will use a much better SDE solver. Then using the Adaptive Grid part of NPAG, we can run our numerical Experiments 2-4 to convergence and verify the convergence criterion used for the initial set of support points.

For practical purposes, one of the best SDE solvers is the program SRIW1 in the Julia software package [BEK+17]. SRIW1 is an adaptive stochastic Runge-Kutta method [RN17]. Fortunately, Julia programs can be called from R, which is the primary language used in our existing Pmetrics modeling and simulation package [NGY+12] and will be the primary language of our new SDE NPML programs.

8 Acknowledgments

The authors wish to thank Walter Yamada for many helpful comments and for his contributions to the NPAG algorithm.

This work was supported in part by grants from NIH: RR11526, GM65619, GM068968, EB005803, EB001978, 320 HD070886.

9 References

[BW12] Banks HT, Thompson WC. Least Squares Estimation of Probability Measures in the Prohorov Metric Framework. Center for Research in Scientific Computation Tech Rep, CRSC-TR12-21, North Carolina State University, Raleigh, NC, 2012.

[BEK+17] Bezanson J, Edelman A, Karpinski S, Shah VB. Julia: A fresh approach to numerical computing. SIAM review, 59(1), 65-98, 2017.

[CD02] Crisan CD, Doucet A. A survey of convergence results on particle filtering methods for practitioners. IEEE Trans. Signal Process 2002, 50, 736-746.

[DPK16] Deng C, Plan EL, Karlsson MO, Approaches for modeling within subject variabilityin pharmacometric count data analysis: dynamic inter-occasion variability and stochastic differential equations. J Pharmacokinet Pharmacodyn (2016) 43:305–314 DOI 10.1007/s10928-016-9473-1

[DS13] Donnet S and Samson A. A review on estimation of stochastic differential equations for pharmacokinetic/pharmacodynamic models. Advanced Drug Delivery Reviews, Elsevier, 2013, pp.1. 10.1016/j. addr.2013.03.005, hal-00777774

[DS14] Donnet S, Samson A. Using PMCMC in EM algorithm for stochastic mixed models: theoretical and practical issues. 2014, hal-00950760

[H64] Halton, J. Algorithm 247: Radical-inverse quasi-random point sequence. Communications of the ACM, 7: 701-701, doi:10.1145/355588.365104.

[GSS93] Gordon NJ, Salmond DJ, Smith AFM. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. IEE PROCEEDINGS-F,Vol.140,No.2, APRIL 1993

[Ja70] Jazwinski AH. Stochastic processes and filtering theory. (1970), Academic Press, New York

[Je92] Jelliffe, RW, Schumitzky, A, Van Guilder, M. Nonpharmacokinetic Clinical Factors Affecting Aminoglycoside Therapeutic Precision, Drug Investigation",1992",4, https://doi.org/10.1007/BF03258374

[KS93] Karlsson MO, Sheiner LB. The importance of modeling

interoccasion variability in population pharmacokinetic analyses. J Pharmacokinet Biopharm 1993;21:735–50.

[KMK+09] Klim S, Mortensen SB, Kristensen DN, Overgaard R, Madsen H. (2009) Population stochastic modeling (PSM) – An R package for mixed-effects models based on stochastic differential equations. Computer Methods and Programs in Biomedicine 94: 279-289.

[KP92] Kloeden E, Platen E. Numerical Solution of Stochastic Differential Equations Applications of Mathematics book series (SMAP, volume 23) 1992.

[KSG+15] Kryshchenko A, Schumitzky A, van Guilder M, Neely MN. Nonparametric estimation of a mixing distribution for a family of linear stochastic dynamical systems. arXiv:1509.04350, 2015.

[KV86] Kumar PR, Varaiya P (1986) Stochastic Systems: Estimation, Identification, and Adaptive Control, Chapter 7, Section 3)].. Prentice-Hall, Englewood Cliffs, New Jersey.

[LD12] Lavielle M, Delattre M. On the Use of Stochastic Differential Mixed Effects Models for Modeling Inter Occasion Variability. Presented at the Population Approach Group Europe, Venice, Italy, 2012.

[L83] Lindsay B. The Geometry of Mixture Likelihoods: A General Theory. Annals of Statistics 1983, 11, 1, 86-94.

[LS84] Lions PL, Sznitman AS. Stochastic differential equations with reflecting boundary conditions. Communications Pure and Applied Mathematics. 1984 https://doi.org/10.1002/cpa.3160370408

[M86]Mallet, A. A Maximum Likelihood Estimation Method for Random Coefficient Regression Models, Biometrika 1986, 73,3, 645-656.

[NGY+12] Neely MN, van Guilder MG, Yamada WM, Schumitzky A, Jelliffe RW. Accurate detection of outliers and subpop-

ulations with Pmetrics, a nonparametric and parametric pharmacometric modeling and simulation package for R. Therapeutic Drug Monitoring 2012;34:467–76. https://doi.org/10.1097/FTD.0b013e31825c4ba6.

[O03] Oksendal B. Stochastic Differential Equations: An Introduction with Applications. Springer, (Universitext) July 15, 2003

[P83] Peacock JA. Two-dimensional goodness-of-fit testing in astronomy. Monthly Notices of the Royal Astronomical Society, Volume 202, Issue 3, March 1983, Pages 615–627, https: //doi.org/10.1093/mnras/202.3.615

[RN 17] Rackauckas C, Nie Q. DifferentialEquations.jl – A performant and feature-rich ecosystem for solving differential equations in Julia J. Open Res. Softw. 2017; 5: 15

[SGN05] Schoen T, Gustafsson F, Nordlund PJ. Marginalized particle filters for mixed linear/nonlinear state-space models. IEEE Transactions on Signal Processing, 53(7):2279–2289, 2005.

[TNB+13] Tatarinova T, Neely MN, Bartroff J, van Guilder M, . Yamada W, Bayard D, Jelliffe J, Leary R, Chubatiuk A, Schumitzky A. Two general methods for population pharmacokinetic modeling: non-parametric adaptive grid and non-parametric Bayesian", journal, J. Pharmacokinetics and Pharmacodynamics, 2013,40, 89-99

[YNB+21] Yamada WM, Neely MN, Bartroff J, Bayard DS, Burke JV, Guilder M van, et al. An Algorithm for Non-parametric Estimation of a Multivariate Mixing Distribution with Applications to Population Pharmacokinetics. Pharmaceutics 2021;13:42. https://doi.org/10.3390/pharmaceutics13010042.