Abstract—We consider the problem of Bayesian inference for parameters in non-linear regression models whereby the underlying unknown response functions are formed by a set of differential equations. Bayesian methods of inference for unknown parameters rely primarily on the posterior obtained by Bayes rule. For differential equation models, analytic and closed forms for the posterior are not available and one has to resort to approximations. We propose a two-stage Laplace expansion to approximate the marginal likelihood, and hence, the posterior to obtain an approximate closed form solution. The first Laplace expansion arises from integrating out the nuisance parameters consisting of the initial starting values of the differential equations. This is followed by another Laplace approximation for the parameters of interest to find the posterior mode. For large sample sizes, the method of inference borrows from non-linear regression theory for maximum likelihood estimates in which case the Bayesian method can be shown to be consistent. Our approach is exact in the limit and does not need the specification of an additional penalty parameter. Examples in this paper include the exponential model and SIR (Susceptible-Infected-Recovered) disease spread model.

Keywords—Ordinary differential equations, posterior computations, Laplace approximations, non-linear regression.

I. INTRODUCTION

Physical and biological processes described by ordinary differential equations are ubiquitous in the real world. These processes evolve over time by specification of certain initial conditions and parameter specifications that govern the temporal dynamics. These processes are frequently only partially observed, that is, observed over a sample of finite time points, and with underlying errors. An observed process with such characteristics is termed as arising from a differential equation model (DEM). DEMs are, thus, regression models whose mean functions are governed by ordinary differential equations. In the most general form, DEMs can be written in the regression set-up as

\[ y_i = x(t_i) + \epsilon_i, \]

where \( y_i \) is a \( p \)-dimensional vector of observations obtained at the \((n + 1)\) sampled time points \( T_0 \equiv t_0 < t_1 < \cdots < t_n \equiv T_1 \), \( \epsilon_i \) are independent and identically distributed errors, and the mean function of the regression model in (1), \( x(t) \in \mathbb{R}^p \) for each \( t \), is the solution of the set of ordinary differential equations

\[ \dot{x}(t) = f(x(t), \theta), \quad t \in [T_0, T_1], \]

where \( \dot{x}(t) \) denotes the component-wise first derivative of \( x(t) \) with respect to time \( t \), \( \theta \in \mathbb{R}^q \) are unknown parameters of the dynamical system, and \( f(x, \theta) \) is a \( p \)-dimensional smooth function of \( x \) and \( \theta \).

The complete function \( x(t) \) on \([T_0, T_1]\) is determined by an initial value, \( x(t_0) \equiv \alpha \) and specification of the parameter \( \theta \). Typically, \( \theta \) is the unknown parameter of interest and needs to be estimated; \( \alpha \) is necessary for the evolution of the dynamical system and for obtaining the observations \( y_i \), but is usually of secondary importance, i.e., nuisance parameters. Parameter inference for DEMs is complicated by the fact that analytic solutions to most differential equations are unknown. A further complication is the presence of \( \alpha \) as nuisance parameters in the inference of \( \theta \).

The differential equation (2) involves only the first order derivatives, but can be used to describe differential equations with higher orders of the derivative. For example, consider a DEM based on \( \dot{x}(t) = g(x, x, \theta) \). Letting \( z(t) = \dot{x}(t) \) and since \( \ddot{x}(t) = g(z, x, \theta) \), the second order derivative model can be expressed as vector dynamical system of order \( p = 2 \). We have

\[ \dot{X}(t) \equiv \begin{pmatrix} \dot{x}(t) \\ \dot{z}(t) \end{pmatrix} \equiv \begin{pmatrix} z(t) \\ g(z, x, \theta) \end{pmatrix} \equiv f_1(z, x, \theta), \]

where \( X(t) \equiv (x(t)^T z(t)^T)^T \), \( f_1(z, x, \theta) = z \) and \( f_2(z, x, \theta) = g(x, z, \theta) \). Thus, in this paper, we consider only first order differential equations and note that higher order derivative models can be accommodated in our framework with an appropriate reformulation. Further, we consider only autonomous systems: the function \( f(x, \theta) \) is a function of time \( t \) implicitly through \( x(t) \). Non-autonomous systems where \( f = f(x, \theta, t) \) will be covered in subsequent work.

Popular examples of DEMs as in (1) and (2) include the Lotka-Volterra predator-prey system (Alligood et al., 1997), the SIR (Susceptible, Infected, Recovered) model (Kermack and McKendrick, 1927) and the continuously stirred tank reactor (CSTR) model (Schmidt, 2005). The Lotka-Volterra equations are a set of differential equations describing the dynamics of predator-prey systems. The SIR model is a differential equation model that governs the spread of disease in a closed population by determining the numbers of susceptible, infected and recovered individuals at each time point \( t \). The CSTR model reflects that an object changes its surface temperature proportional to its relative temperature.

As noted earlier, DEMs involve unknown parameters in most applications which need to be estimated from the data. Bard (1974) proposed a nonlinear least squares approach using the approximated solution of DEMs by numerical integration. A gradient-based method was suggested for minimization of the least squares objective function, and it was found
that the solutions were sensitive to initial values, divergent or convergent to local minimums. Varah (1982) proposed a two step estimation approach. In the first step, the solution of DEM is approximated using cubic splines with fixed knots. In the second step, the parameters of the DEM are estimated by least squares with the distance between the differential equation model and the spline based solution in the first step as a measure of lack of fit. Ramsay and Silverman (2005) applied and developed this approach further to functional data analysis. They modified the first step by adding a roughness penalty measure and found the minimizer iteratively.

Ramsay et al. (2007) combined smoothing and estimation of parameters in DEMs by a parameter cascading method. They divided parameters into three different groups, i.e., levels. The regularization parameter (or penalty parameter) occupies the top most level followed by parameters in ordinary differential equation (called structural parameters in Ramsay et al. (2007)) at the intermediate level, and finally the spline regression coefficients at the lowest level. Ramsay’s approach is to first minimize the objective function for the regression coefficient parameters given at the higher levels, then minimize the objective function as a function of the structural parameters given the penalty parameter, and finally, minimize the objective function as a function of the penalty parameter. This method depends crucially on the penalty parameter as different values can given different smoothing results (and hence, different inference results for the structural parameters).

The development of Bayesian inference for DEMs has also been reported in the literature. Gelman et al. (1996) used a hierarchical Bayesian model to infer a pharmacokinetic model, which describes the individual flow system of the compound in the body. Huang et al. (2006) also used a hierarchical Bayesian method to estimate the parameters in a longitudinal HIV dynamic system. Campbell (2007) merged the collocation method (Ramsay, 2007) and combined parallel tempering (Geyer, 1991) as follows. They suggested to use collocation method to approximate the solution of the differential equation and employed parallel tempering to estimate the parameters in the differential equation. This method is called smooth functional tempering or collocation tempering. In the above papers, the Metropolis-Hastings algorithm is used for sampling parameter realizations from the posterior distribution which can be inefficient in time. Further, the above methods rely on the optimal choice of penalty parameters.

The approach advocated in this paper is fundamentally different from Ramsay et al. (2007) and previous Bayesian approaches in two ways. First, we propose an inferential framework for differential equation models by viewing it as a sub-problem of non-linear regression. Second, we propose a two-stage Laplace approximation to the marginal likelihood which result from integrating out parameters based on elicited prior distributions. The first Laplace approximation arises by integrating out the nuisance parameters consisting of the initial starting values of the differential equations. This is followed by another Laplace approximation for the parameters of interest to find the posterior mode. For large sample sizes, the method of inference borrows from non-linear regression theory for maximum likelihood estimates in which case the Bayesian and frequentist inferences become identical. Our approach is exact in the limit and does not need the specification of an additional penalty or regularization parameter. Thus, parameter inference theory is not dependent on values of the regularization parameters.

II. NON-LINEAR REGRESSION SET-UP

Consider the DEM of (1) and (2) with initial conditions \( a \) and unknown parameter specification \( \theta \). Based on (2), it follows that \( x(t) \) is a function of both \( a \) and \( \theta \) as well as time \( t \); thus, we write \( x(t) \equiv x(t,a,\theta) \) for initial condition \( a \) and parameter specification \( \theta \). The errors \( \epsilon_i, i = 0, 1, \cdots, n \) are assumed to be independent and identically distributed as \( p \)-variate normal random variables, with mean 0 and common unknown marginal variances \( \sigma^2 \), that is, \( \epsilon_i \sim N_p(0,\sigma^2 I_p) \) where \( I_p \) is the identity matrix of dimension \( p \times p \). The (log)likelihood of \((a,\theta,\sigma^2)\) is given by

\[
\ell(a,\theta,\sigma^2) = -\frac{1}{2\sigma^2} \sum_{i=0}^{n} (y_i - x_i(a,\theta))^T (y_i - x_i(a,\theta)) - \frac{(n+1)p}{2} \log \sigma^2 \quad (3)
\]

where \( x_i(a,\theta) \equiv x(t_i,a,\theta) \) is the vector \( x(t) \) evaluated at time point \( t_i \) for \( i = 0, 1, \cdots, n \). Clearly,

\[
x_0(a,\theta) \equiv a \quad (4)
\]

by definition. It follows that the set-up in (3) is that of non-linear regression since the covariate \( x_i(a,\theta) \) is generally not a linear function of \( a \) and \( \theta \).

For Bayesian inference, all unknown parameters in the above model, \((a,\theta,\sigma^2)\), are given prior distributions. We take the following joint prior elicitation for \( a, \sigma^2 \) and \( \theta \):

\[
a \mid \sigma^2 \sim \pi_1 \equiv N_p(\mu_0, c\sigma^2 I_p), \quad \sigma^2 \sim \pi_2 \equiv IG(a_0, b_0), \quad \theta \sim \pi_3(\cdot),
\]

where \( IG(a_0, b_0) \) is the inverse gamma distribution with shape and scale parameters \( a_0, b_0 > 0 \), respectively (that is, with mean \( b_0/(a_0 - 1) \) if \( a_0 > 1 \), \( \pi_1 \) is the multivariate normal distribution with mean \( \mu_0 \) and covariance \( c\sigma^2 I_p \) where \( c > 0 \), and \( \pi_3 \) is an arbitrary prior for \( \theta \). The prior selection for \( \sigma^2 \) and \( a \) is guided by conjugacy considerations which enable some components of the posterior to be integrated in closed form. The choice of \( \pi_3 \) is left arbitrary. For large sample sizes, the effect of the joint prior specification in terms of \( \pi_1, \pi_2 \) and \( \pi_3 \) are minimal; most of the inference will be driven and guided by the likelihood component, \( \ell(a,\theta,\sigma^2) \), of the posterior. The conditional prior specification of \( a \) given \( \sigma^2 \) is justified by the fact that \( a \) represents the initial condition for \( x \), which is in the same domain as the set of observations, \( y = (y_0, y_1, \cdots, y_n) \). It is reasonable to expect that as the uncertainty in \( a \) should increase proportionately to the uncertainty in observing the \( y \) data.

The above prior elicitation gives an expression of the posterior for \((a,\theta,\sigma^2)\) as

\[
\pi(a,\theta,\sigma^2 \mid y) \propto e^{\ell(a,\theta,\sigma^2)} \pi_1(a \mid \sigma^2) \pi_2(\sigma^2) \pi_3(\theta) / m(y)
\]

(6)
where $m(y)$ is the marginal distribution after integrating out all unknown parameters $(a, \theta, \sigma^2)$. Bayesian inference is carried out based on this posterior distribution for $(a, \theta, \sigma^2)$.

### III. Two-Stage Laplace Approximation Approach

We note that several difficulties with the posterior in (6). As we have mentioned earlier, the likelihood $\ell(a, \theta, \sigma^2)$ is non-linear in $(a, \theta)$, and moreover, we cannot generally obtain a closed form solution for each $x_i(a, \theta)$. As a result, the posterior of $(a, \theta, \sigma^2)$ is not readily available in closed or simple analytical form for sampling. In this section, the cascading Laplace approximation approach is discussed as a way of working around these challenges.

The marginalization with respect to $(a, \theta, \sigma^2)$ can be viewed in three stages: (i) the marginalization with respect to $a$ for fixed $(\theta, \sigma^2)$, followed by (ii) the marginalization with respect to $\theta$ for each fixed $\sigma$, and finally, (iii) the marginalization with respect to $\sigma^2$. In the first step, the marginal posterior of $(\theta, \sigma^2)$ in (7), $\pi(\theta, \sigma^2 | y)$, is proportional to

$$
\int_{a \in \mathbb{R}} e^{\ell(a, \theta, \sigma^2)} \pi_1(a | \sigma^2) da \quad \pi_2(\sigma^2) \pi_3(\theta)
$$

where $g_n(a, \theta)$ is given by

$$
g_n(a, \theta) = \frac{1}{2n} \sum_{i=0}^{n} (y_i - x_i(a, \theta))^T (y_i - x_i(a, \theta)) + \frac{1}{2cn} (a - \mu_0)^T (a - \mu_0).
$$

The first stage approximation is obtained by performing a Laplace expansion around the minimum point, $\hat{a}(\theta)$, of $g_n(a, \theta)$ with respect to $a$ for fixed $\theta$ and $\sigma^2$. For large $n$, the integral in (7) is approximated as

$$
\int_{a \in \mathbb{R}} e^{-g_n(a, \theta)/\sigma^2} da \approx \left(\frac{2\pi\sigma^2}{n}\right)^{p/2} e^{-g_n(\hat{a}(\theta), \theta)/\sigma^2} \times | \det(\partial^2 g_n/\partial a^2) |^{-1/2}
$$

where $\partial^2 g_n/\partial a^2$ denotes the second derivative matrix of $g_n$ with respect to $a = (a_1, a_2, \cdots, a_p)^T$ with the $(i,j)$-th element given by $\partial^2 g_n/\partial a_i \partial a_j$, for $i, j = 1, 2, \cdots, p$. Combining (7) and (9), the numerator for the posterior $\pi(\theta, \sigma^2 | y)$ can be approximated as

$$
\pi(\theta, \sigma^2 | y) \propto e^{-g_n(\hat{a}(\theta), \theta)/\sigma^2} \times | \det(\partial^2 g_n/\partial a^2) |^{-1/2} n^{p/2} (2\pi\sigma^2)^{-(n+1)p/2} \pi_2(\sigma^2) \pi_3(\theta).
$$

Subsequently, performing another integration with respect to $\sigma^2$ brings the posterior of $\theta$ equalling to

$$
\pi(\theta | y) \propto \frac{1}{(ng_n(\hat{a}(\theta), \theta) + b_0)^{n+1}} \times | \det(\partial^2 g_n/\partial a^2) |^{-1/2} \pi_3(\theta)
$$

as a function of $\theta$ up to constants of proportionality. Expression (11) is obtained from the standard conjugacy of the inverse gamma distribution to the normal likelihood for $\sigma^2$. For the purpose of the second Laplace expansion, we write

$$
\pi(\theta | y) \propto e^{-nh_n(\hat{a}(\theta), \theta)}
$$

where

$$
h_n(\hat{a}(\theta), \theta) = \frac{p}{2} \left[ 1 + \frac{1}{n} + 2\frac{a_0}{pn} \right] \log[g_n(\hat{a}(\theta), \theta)] + \frac{1}{2n} \log[| \det(\partial^2 g_n/\partial a^2) |] - \frac{1}{n} \log(\pi_3(\theta))
$$

The expression in (13) states that as $n \to \infty$, only the first term of $h_n$ remains non-negligible. Thus, the second Laplace approximation can be performed around the minimum, $\hat{\theta}$, of $g_n(\hat{a}(\theta), \theta)$ (since log is a monotonically increasing function), thus obtaining

$$
\pi(\theta | y) \propto \exp \left\{ -nh_n^*(\hat{a}(\theta), \hat{\theta}) \right\} \times \exp \left\{ -\frac{n}{2} (\theta - \hat{\theta})^T \left[ \frac{\partial^2 h_n^*}{\partial \theta^2} \right] (\theta - \hat{\theta}) \right\}
$$

where $h_n^*(\hat{a}(\theta), \hat{\theta}) = \log[g_n(\hat{a}(\theta), \hat{\theta})]$ and $\partial^2 h_n^*/\partial \theta^2$ is the matrix of second order partial derivatives of $h_n^*(\hat{a}(\theta), \hat{\theta})$ with respect to entries of $\theta = (\theta_1, \theta_2, \cdots, \theta_q)^T$. The first term on the right hand side of (14) is a constant function of $\theta$ whereas the second term is the pdf of a multivariate normal density with mean $\hat{\theta}$ and covariance matrix given by $\left( n \frac{\partial^2 h_n^*}{\partial \theta^2} \right)^{-1}$.

### IV. Numerical Procedure

The methodology outlined in the previous sections rests on the availability of $\hat{a}(\theta)$, the minimizer of $g_n(a, \theta)$ for every fixed $\theta$, and $\hat{\theta}$, the minimizer of $g_n(\hat{a}(\theta), \theta)$ as a function of $\theta$. This section describes how to numerically obtain estimates of $\hat{\theta}$ and $\hat{a}(\theta)$ using the non-linear regression framework given in Section II. The objective function for non-linear regression is minimized using the Gauss-Newton method and variants thereof. Consider the objective function

$$
\frac{1}{n} \sum_{i=0}^{n} (y_i - x_i(\tau))^T (y_i - x_i(\tau))
$$

where each observation $y_i = (y_{i1}, y_{i2}, \cdots, y_{ip})^T \in R^p$ is to be regressed onto the covariate $x_i(\tau) = (x_{i1}(\tau), x_{i2}(\tau), \cdots, x_{ip}(\tau))^T \in R^p$ depending on unknown parameters $\tau = (\tau_1, \tau_2, \cdots, \tau_q)^T \in R^q$. In the Gauss-Newton method, the objective function is minimized by first considering a first order Taylor expansion of each $x_i(\tau)$,

$$
x_i(\tau) \approx x_i(\tau_0) + \left[ \frac{\partial x_i}{\partial \tau} \right] (\tau - \tau_0)
$$

around a pre-specific value $\tau = \tau_0$, say. In (16), $[\partial x_i/\partial \tau]$ is the matrix of first order partial derivatives of $x_i(\tau)$ with respect to entries of $\tau$, that is, the $(u,v)$-th entry of $[\partial x_i/\partial \tau]$ is $\partial x_{iu}(\tau)/\partial \tau_v$ for $u = 1, 2, \cdots, p$ and $v = 1, 2, \cdots, s$. Substituting (16) in (15) and minimizing with respect to $\tau$ given $\tau_0$, a closed form solution for the unique global minimizer is obtained as

$$
\tau_1 = \tau_0 + J^{-1} \ell_0
$$
where \( J = \sum_{n=0}^{t_n} (\frac{\partial r}{\partial \tau})^T (\frac{\partial x}{\partial \tau}) \) and \( t_n = \sum_{n=0}^{t_n} (\frac{\partial r}{\partial \tau})^T (y_i - x_i(\tau_n)) \). An iterative procedure is started at \( \tau = \tau_0 \), and at the \((k+1)\)-th step, \( k \geq 0 \), \( \tau_{k+1} \) is obtained from \( \tau_k \) using the updating formula in (17). A variant of the basic Gauss-Newton method is the Levenberg-Marquardt algorithm where the matrix \( J \) is replaced by \( J + \lambda \text{diag}(J^T J) \). The damping factor \( \lambda > 0 \) is adjusted at each iteration to optimally converge to a (local) minimum. In MATLAB, the Gauss-Newton and Levenberg-Marquardt algorithms are implemented in the \texttt{lsqnonlin} function.

The Gauss-Newton (and Levenberg-Marquardt) optimization procedure is utilized twice in a nested fashion for our present situation. The outer call to \texttt{lsqnonlin} is made for finding \( \hat{\theta} \) based on the covariates \( \{x_i(\hat{a}, \theta), i = 0, 1, 2, \ldots, n\} \). Within each outer call, an inner call to \texttt{lsqnonlin} is made for finding \( \hat{a}(\theta) \) based on the covariates \( \{x_i(\theta), i = 0, 1, 2, \ldots, n\} \) which are now viewed as functions of \( a \) only for fixed \( \theta \). The Gauss-Newton and Levenberg-Marquardt algorithms require numerical evaluation of the partial derivatives \( [\partial x_i/\partial \tau] \) in (17). For the inner call, \( \tau = a, x_i(\tau) = x_i(\theta) \) and the partial derivative matrix \([\partial x_i/\partial \tau] = [\partial x_i/\partial a] \) which consists of the entries \([\partial x_{i*}/\partial a_j] \) as the \((v,j)\)-th element, for \( v, j = 1, 2, \ldots, p \).

Now fix \( j \) and consider numerical evaluation of each column vector \( \partial x_i/\partial a = (\partial x_{i1}/\partial a, \ldots, \partial x_{ip}/\partial a)^T \) of \([\partial x_{i*}/\partial a] \). Since \( x \equiv x(t, a, \theta) \) satisfies the ordinary differential equation (2) with initial condition (4), it follows that \( \partial x/\partial a_j \) satisfies

\[
\left( \frac{\partial x_i}{\partial a_j} \right) = \sum_{k=1}^{p} \frac{\partial f(x(t, a, \theta), \theta)}{\partial x_k} \frac{\partial x_k}{\partial a_j} = \frac{\partial f}{\partial x} \left( \frac{\partial x_i}{\partial a_j} \right),
\]

(18)
a non-autonomous linear system of ordinary differential equations with initial condition \( \partial x_0/\partial a_j = e_j \) where \( e_j \) is the vector with 1 in the \( j \)-th position and zero otherwise; also, \([\partial f]/[\partial x] = (p \times p) \) matrix for which the \((v,v')\)-th entry is given by \( \partial f_v/\partial x_{v'} \) for \( v, v' = 1, 2, \ldots, p \). The system of equations can be solved numerically using Euler’s method of finite difference for given \( \theta \), i.e.,

\[ x_{i+1}(a, \theta) = x_i(a, \theta) + h_t f(x_i(a, \theta), \theta) \]  

where \( h_t = t_{i+1} - t_i \). The approximation can be accurate enough by dividing the interval \( h_t \) into \( m \) segments. Subsequently, we take \( \partial x_i/\partial a_j = \partial x(t_i, a, \theta)/\partial a_j \) evaluated at \( a = \hat{a}(\theta) \) for the entries of matrix \([\partial x_{i*}/\partial a] \).

For the outer loop, we have \( \tau = \theta \) and \( x_i(\tau) = x_i(\hat{a}(\theta), \theta) \). The element-wise expressions for the first order partial derivative matrix \([\partial x_i(\hat{a}(\theta), \theta)]/\partial \theta \) are as follows: For \( u = 1, 2, \ldots, q \), the \( u \)-th partial derivative of \( \partial x_{iv}/\partial \theta_u \) is given by

\[
\frac{\partial x_{iv}}{\partial \theta_u} = \sum_{k=1}^{p} \frac{\partial x_{iv}(\hat{a}(\theta), \theta)}{\partial a_k} \frac{\partial a_k}{\partial \theta_u}
\]

(20)
for \( k = 1, 2, \ldots, p \). In matrix notation, we write

\[
\frac{\partial x_i}{\partial \theta_u} = \left[ \frac{\partial x_i}{\partial \theta} \right] \left( \frac{\partial \hat{a}(\theta)}{\partial \theta_u} \right)
\]

(21)
where \( \partial x_i/\partial \theta_u = (\partial x_{i1}/\partial \theta_u, \partial x_{i2}/\partial \theta_u, \ldots, \partial x_{ip}/\partial \theta_u)^T \) is the \( p \times 1 \) vector of partial derivatives of \( x_i \) with respect to \( \theta_u \), \( \partial \hat{a}(\theta)/\partial \theta_u = (\partial \hat{a}_1(\theta)/\partial \theta_u, \partial \hat{a}_2(\theta)/\partial \theta_u, \ldots, \partial \hat{a}_p(\theta)/\partial \theta_u)^T \) is the \( p \times 1 \) vector of partial derivatives of \( \hat{a}(\theta) \) with respect to \( \theta_u \), and \( \{\partial x_i/\partial a\} \) is the \( p \times p \) matrix with \((v,j)\)-th element \( \partial x_{iv}/\partial a_j \) for \( v, j = 1, 2, \ldots, p \) described earlier. Note that each term in the vector \( \partial \hat{a}(\theta)/\partial \theta_u \) cannot be obtained in closed form either. However, from the fact that \( \hat{a}(\theta) \) satisfies

\[ \theta_n \partial \hat{a}(\theta)/\partial \theta_n = \theta_n \partial \hat{a}(\theta)/\partial \theta_n = 0 \]

for each \( j = 1, 2, \ldots, p \). Using matrix notation, we can write the above \( p \) equations as

\[
\sum_{k=1}^{p} \frac{\partial^2 g_n(\hat{a}(\theta), \theta)}{\partial a_k \partial \theta_u} \frac{\partial \hat{a}_k(\theta)}{\partial \theta_u} + \frac{\partial^2 g_n(\hat{a}(\theta), \theta)}{\partial \theta_u^2} \theta_n = 0.
\]

(22)
where \( \frac{\partial^2 g_n}{\partial a^2} \) is the matrix of second order partial derivatives of \( g_n \) with respect to \( a \) in which the \((v,v')\)-th entry is given by \( \frac{\partial^2 g_n}{\partial a_v \partial a_{v'}} \) and \( \frac{\partial^2 g_n}{\partial \theta_u^2} \theta_n \) is the \( p \times 1 \) vector given by \( \left( \frac{\partial^2 g_n}{\partial a_1 \partial a_1}, \frac{\partial^2 g_n}{\partial a_1 \partial a_2}, \ldots, \frac{\partial^2 g_n}{\partial a_p \partial a_{p-1}}, \frac{\partial^2 g_n}{\partial a_p \partial \theta_n} \right)^T \).

The computation of these quantities can be done by using the relation (8). Substituting (22) in (21) together with a numerical solution for \( \{\partial x_{i*}/\partial a\} \) using (18) allows us to evaluate \( \partial x_i/\partial \theta_u \) for each \( i = 1, 2, \ldots, p \) and \( u = 1, 2, \ldots, q \).

V. POSTERIOR INFERENCE

The posterior \( \pi(a, \theta, \sigma^2 | y) \) is approximated by the two-stage Laplace expansion as described in Section III. More specifically, the sampling densities comprise

\[
\theta \sim N_q \left( \hat{\theta}, \left( n \frac{\partial^2 h_n^a}{\partial \theta^2} \right)^{-1} \right)
\]

(23)
\[
\sigma^2 \mid \theta \sim IG(a_1, b_1)
\]

(24)
\[
a \mid \theta, \sigma^2 \sim N_p \left( \hat{\theta}, \left( \sigma^2 \frac{\partial^2 g_n}{\partial a^2} \right)^{-1} \right)
\]

(25)
where \( IG(a_1, b_1) \) is the inverse Gamma distribution with shape and scale parameters \( a_1 = (n + 1)p/2 + a_0 \) and \( b_1 = n g_n + b_0 \), respectively, and the mean and covariance matrices for the multivariate normal distributions are as those defined in (9) and (12) in Section III. We denote the joint density of the sampling distribution in (23-25) by \( \pi_s(\theta, \sigma^2) = \pi_{s1}(\theta) \pi_{s2}(\sigma^2 \mid \theta) \pi_{s3}(a \mid \theta, \sigma^2) \) where \( \pi_{sj}, j = 1, 2, 3, \) are the densities given in (23-25).

Note that \( \pi_{sj}, j = 1, 2, 3 \) are not the actual posterior densities but are an approximation of the true ones. The justification of this approximation is motivated from large sample theory, that is, as \( n \to \infty \). The Laplace approximation becomes exact in the limit as \( n \to \infty \), and therefore, the joint approximate posterior \( \pi_{s3} \times \pi_{s2} \times \pi_{s1} \) becomes a more effective approximation of \( \pi(a, \theta, \sigma^2 \mid y) \) for large \( n \). The main advantage of the approximate posterior is that the
component densities are available in closed form which is not the case for the actual posterior. Thus, inference is easily made based on the approximate posterior compared to the actual one.

VI. EXAMPLES

Two examples are presented to demonstrate the methodology outlined in this paper. These are based on different specifications of the dynamical system for $x(t)$ illustrating the extent of non-linearities present in each case.

A. Exponential Model

We start with one of the simplest DEM. The first example is based on the exponential model which satisfies the ordinary differential equation model $x(t) = \theta x(t)$ with initial condition $x(t_0) = a$. Here $p = q = 1$ and we have a closed form solution for $x$, namely, $x(t) = ae^{\theta t}$ for $t \in [T_0, T_1]$. Data $y = (y_0, y_1, \ldots, y_n)$ is simulated from (1) and (2) with true parameter values $a = 1$, $\sigma^2 = 1$ and $\theta = -0.5$. We set the initial time point $t_0 = 0$ and step size $t_{i+1} - t_i = 0.01$ for $i = 0, 1, \ldots, n$ where $n = 199$. Thus, 200 observations are given, and Fig. 1 represents the observations and true mean function. The prior distributions are chosen as (6) with $a_0 = 0.1, b_0 = 100, c = 100, \mu_0 = 0$ and $\theta \sim \text{Unif}(-100, 100)$, which are non-informative priors.

Based on data, we ran the approximate posterior methodology on our simulated data. Approximate posterior samples from the joint posterior density, $(\theta_i, \sigma^2_i, a_i), i = 1, 2, \ldots, M$ can be obtained as follows. For $i = 1, 2, \ldots, M$, generate $\theta_i$ from $\pi_{\theta_i}$ in (23). Given $\theta_i$, generate $\sigma^2_i$ from (24), and finally given $\theta_i$ and $\sigma^2_i$, generate $a_i$ from (25). Histograms for the marginal posterior distributions for each of the parameters $\theta$, $a$ and $\sigma^2$ are shown in Fig. 2 for $M = 500$. Posterior inference for the means, standard deviations (standard errors) and credible sets are obtained from these histograms for $\sigma^2$ and $a$; for $\theta$, inference components can be obtained from standard normal theory.

The Bayesian methodology of inference gives posterior mean, standard deviation and credible intervals for the parameters $(\theta, \sigma^2, a)$ in Table I. Note that the posterior means are close to the true values of the parameters and the 95% credible interval covers the true value of the parameters.

B. SIR model

Kermack and McKendrick (1927) suggested a SIR model as known as the Kermack-McKendrick model which describes spread of infectious diseases. They considered the number of susceptible, infected and recovered people in the fixed population. The differential equations are of the form:

$$
\dot{x}_1(t) = -\theta_1 x_1(t) x_2(t),
\dot{x}_2(t) = \theta_1 x_1(t) x_2(t) - \theta_2 x_2(t),
\dot{x}_3(t) = \theta_2 x_2(t).
$$

Fig. 1: The solid line is the true mean function $ae^{\theta t}$ as a function of time from the exponential model with $a = 1$, $\theta = -0.5$. The points are the generated data $y$ when $\sigma^2 = 1$.

Fig. 2: Histograms of sample realizations from marginal posterior distributions for (a) $\theta$, (b) $\sigma^2$ and (c) $a$.

TABLE I: Posterior mean, standard deviation and credible intervals for the parameters $(\theta, \sigma^2, a)$ in exponential model.

<table>
<thead>
<tr>
<th></th>
<th>Posterior mean</th>
<th>Median</th>
<th>Standard deviation</th>
<th>95% credible set</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta$</td>
<td>-0.3164</td>
<td>-0.3164</td>
<td>0.1155</td>
<td>(-0.5428, -0.0900)</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>0.9682</td>
<td>0.9604</td>
<td>0.0963</td>
<td>(0.7945, 1.1813)</td>
</tr>
<tr>
<td>$a$</td>
<td>1.0195</td>
<td>1.0235</td>
<td>0.0630</td>
<td>(0.8929, 1.1408)</td>
</tr>
</tbody>
</table>
The approximate posterior is that it is available in closed posterior distribution for DEMs. The main advantage of stage Laplace approximation approach to get an approximate posterior of unknown parameters. We propose a two-differential equations is difficult due to the intractability of the parameters of interest $\theta$. Here, $x_1(t)$ is the number of susceptible people, $x_2(t)$ is the number of people infected and $x_3(t)$ is the number of people who have recovered from the disease. Note, $\dot{x}_1(t) + \dot{x}_2(t) + \dot{x}_3(t) = 0$ because the population size is fixed. The parameter $\theta_1$ indicates the infection rate, and $\theta_2$ indicates the recovery rate. In this case $p = 3$ and $q = 2$.

We choose $\alpha \equiv (a_1, a_2, a_3) = (499, 1, 1)^T$, $\sigma^2 = 1$ and $\theta \equiv (\theta_1, \theta_2) = (0.001, 0.1)^T$. The step size $t_{i+1} - t_i$ is chosen as 1, and 100 observations $(y_0, y_1, \ldots, y_{100})$ are obtained. Since the observation $y_i$ cannot be negative, we use the value $y_i = \max(x_i + \epsilon, 0)$ as a simulation data. The simulated data and approximated mean function are drawn at Fig. 3. The prior distributions are chosen as (6) with $a_0 = 0.1, b_0 = 100, c = 1000, \mu_0 = y_0$ and $\theta_k \sim \text{Unif}(0, 100)$ for using non-informative priors.

Histograms for the marginal posterior distributions for each of the parameter of interest $\theta_j$, $j = 1, 2$ are shown in Fig. 4 for $M = 500$; histograms for $a$ components and $\sigma^2$ are not given here but can be obtained similarly. For $\theta$ components, inference can be obtained from standard bivariate (and multivariate for $q > 2$) normal theory. For $\sigma^2$ and $a$, although not presented here, means, standard deviations (standard errors) and credible sets are obtained from samples from the posterior. The posterior mean, standard deviation and credible intervals for the parameters of interest $(\theta_1, \theta_2)$ is given in Table II.

VII. CONCLUSION AND DISCUSSION

Bayesian inference for DEMs governed by ordinary differential equations is difficult due to the intractability of the posterior of unknown parameters. We propose a two-stage Laplace approximation approach to get an approximate posterior distribution for DEMs. The main advantage of the approximate posterior is that it is available in closed form, and thus, easy to obtain inference from. We illustrate our inference methodology for two examples of DEMs. The proposed approximate posterior is close to the true posterior especially for large sample sizes as is the case in our examples.

REFERENCES

TABLE II: Posterior mean, standard deviation and credible intervals for the parameters of interest $\theta$ in SIR model. The true parameters are $(\theta_1, \theta_2) = (0.001, 0.1)$.

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>Posterior mean</th>
<th>Median</th>
<th>Standard deviation</th>
<th>95% credible set</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_1$</td>
<td>$0.99899 \times 10^{-3}$</td>
<td>$0.99906 \times 10^{-3}$</td>
<td>$2.3179 \times 10^{-6}$</td>
<td>$(0.99445, 1.00354) \times 10^{-3}$</td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>0.1001</td>
<td>0.1001</td>
<td>$1.5638 \times 10^{-4}$</td>
<td>$(0.0997, 0.1003)$</td>
</tr>
</tbody>
</table>

