ISSN 1346-5597

National Institute of Informatics

NII Technical Report

Cluster Gauss-Newton method for finding multiple approximate minimisers of nonlinear least squares problems with applications to parameter estimation of pharmacokinetic models

Yasunori Aoki, Ken Hayami, Kota Toshimoto, and Yuichi Sugiyama

NII-2020-002E
Apr. 2020

Cluster Gauss-Newton method for finding multiple approximate minimisers of nonlinear least squares problems with applications to parameter estimation of pharmacokinetic models

Yasunori Aoki[*] Ken Hayami[†] Kota Toshimoto[‡] Yuichi Sugiyama[§]

April 10, 2020

Abstract

Parameter estimation problems of mathematical models can often be formulated as nonlinear least squares problems. Typically these problems are solved numerically using iterative methods. The local minimiser obtained using these iterative methods usually depends on the choice of the initial iterate. Thus, the estimated parameter and subsequent analyses using it depend on the choice of the initial iterate. One way to reduce the analysis bias due to the choice of the initial iterate is to repeat the algorithm from multiple initial iterates (i.e. use a multi-start method). However, the procedure can be computationally intensive and is not always used in practice. To overcome this problem, we propose the Cluster Gauss-Newton (CGN) method, an efficient algorithm for finding multiple approximate minimisers of nonlinear-least squares problems. CGN simultaneously solves the nonlinear least squares problem from multiple initial iterates. Then, CGN iteratively improves the solutions from these initial iterates similarly to the Gauss-Newton method. However, it uses a global linear approximation instead of the Jacobian. The global linear approximations are computed collectively among all the iterates to minimise the computational cost. We use physiologically based pharmacokinetic (PBPK) models used in pharmaceutical drug development to demonstrate its use and show that CGN is computationally more efficient and more robust against local minima compared to the standard Levenberg-Marquardt method, as well as state-of-the art multi-start and derivative-free methods.

1 Introduction

The parameter estimation of mathematical models often boils down to solving nonlinear least squares problems. Hence, algorithms for solving nonlinear least squares problems are widely used in many scientific fields.

The most traditional least squares solver is the Gauss-Newton method [13, 4]. In practice, the Gauss-Newton method with regularisation (i.e., Levenberg-Marquardt (LM) method [25, 27]) or with the Trust-Region method [9] is often used. Recently, derivative-free methods, which do not explicitly use derivative information of the nonlinear function, have been developed. These methods are usually computationally more efficient as it avoids the costly computation of the derivatives. Also, they can be applied even to

[*]Visiting Associate Professor, National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda, Tokyo 101-8430, Japan, and Visiting Researcher, Sugiyama Laboratory, RIKEN Baton Zone Program, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan (yaoki@uwaterloo.ca).

[†]National Institute of Informatics, and The Graduate University for Advanced Studies (SOKENDAI), 2-1-2 Hitotsubashi, Chiyoda, Tokyo 101-8430, Japan (hayami@nii.ac.jp). The author was supported in part by JSPS KAKENHI Grant Number 15K04768 and 15H02968.

[‡]Sugiyama Laboratory, RIKEN Baton Zone Program, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan.

[§]Sugiyama Laboratory, RIKEN Baton Zone Program, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan.

problems where the mathematical models are 'black box'. The state of the art derivative-free algorithms are DFO-LS [8] and POUNDERS [36]. A comprehensive review of the derivative-free methods can be found in [24].

Another approach for obtaining a solution of nonlinear least squares problems is to directly minimise the sum of squared residuals (SSR) using generic optimisation algorithms for the scalar objective function. The most classical approach is to obtain a solution where the gradient of the SSR becomes zero using the Newton method. As it is usually too costly to compute the Hessian of the SSR, Quasi-Newton methods which approximate the Hessian are used. The commonly used Quasi-Newton method is the BFGS method [6, 10, 16, 31, 32]. Another approach which makes use of the Newton-type method for optimisation is Implicit Filtering [21] which combines grid search and the Newton method. In addition to these optimisation algorithms, we can use numerous global optimisation algorithms when bound constraints are given. For example, Surrogate Optimisation [17], Genetic Algorithm [15], Particle Swarm algorithm [22, 26], and DIRECT [20] are well known global optimisation algorithms.

Although there are a variety of algorithms to solve the nonlinear least squares problems as listed above, they mostly focus on finding one solution. To the best of our knowledge, there is very limited methodological development on algorithms for simultaneously finding multiple approximate minimisers of nonlinear least squares problems. For instance, when using the Levenberg-Marquardt method, the local algorithm often gives a local minimiser which depends on the choice of the initial iterate. To reduce the analysis bias due to the initial iterate used for a local algorithm, it is a good practice to repeatedly use the local algorithm with various initial iterates, as in multi-start methods [5]. Similarly, for problems where bound constraints are given, one can use global optimisation algorithms to find one of the global minimisers. On the other hand, if there are multiple global minisers, the global minimiser found can depend on the algorithm setting, for example, the random seed. Hence, it is beneficial to use global optimisation algorithms with various settings repeatedly if the uniqueness of the global minimiser is not guaranteed. The trivial bottleneck of repeatedly using these algorithms is the computation cost. In this paper, we propose a new method addressing this computational challenge of finding multiple approximate minimisers of nonlinear least squares problems.

Our algorithm development for finding multiple local minimisers of nonlinear least squares problems was motivated by a mathematical model of pharmaceutical drug concentration in a human body called the physiologically based pharmacokinetic (PBPK) model [34]. The PBPK model is typically a system of mildly nonlinear stiff ordinary differential equations (ODEs) with many parameters. This type of mathematical model is constructed based on the knowledge of the mechanism of how the drug is absorbed, distributed, metabolised and excreted. Given the complexity of this process and the limitation of the observations we can obtain from a live human subject, the model parameters cannot be uniquely identified from the observations, meaning that there are non-unique global minimisers to the nonlinear least squares problem. The estimated parameters of the PBPK model are used to simulate the drug concentration of the patient from whom we are often unable to test the drug on (e.g., children, pregnant person, a person with rare genetic anomaly) or to predict the experiment that is yet to be run (different amount of drug administration, multiple drug used at the same time). As the simulated drug concentration is used to predict the safety of the drug in these different scenarios, it is essential to consider multiple predictions based on multiple possible parameters that are estimated from the available observations. A motivating example is presented in Appendix A. Another reason why we want to obtain multiple sets of parameters is that we can understand which parameter cannot be estimated from the available data. This will motivate the pharmaceutical scientists to perform additional (e.g., in-vitro or in-animal) experiments to determine these parameters that were not estimable from the available data.

In [1, 2] we proposed the Cluster Newton (CN) method for obtaining multiple solutions of **a system of underdetermined nonlinear equations**. In recent years CN has been used in the field of pharmaceutical science [38, 11, 3, 33, 23, 28]. For example, [33] used the parameters estimated by CN to predict the adverse drug effect, [28] used the estimated parameters to predict the outcome of a clinical trial. However, based on these applications of CN, we observed the necessity to formulate the problem as a nonlinear least squares problem and make the method more robust against noise in the observed data. This is mainly because actual pharmaceutical data may contain measurement error and inconsistency coming from an inadequate model.

(See Appendix B.)

1.1 Nonlinear least squares problem of our interest

In this paper, we propose an algorithm for obtaining multiple approximate minimisers of nonlinear least squares problems

$$\min_{\boldsymbol{x}} ||\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{y}^*||_2^2 \tag{1}$$

which do not have a unique solution (global minimiser), that is to say, there exist $\boldsymbol{x}^{(1)} \neq \boldsymbol{x}^{(2)}$ such that

$$\min_{\boldsymbol{x}} ||\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{y}^*||_2^2 = ||\boldsymbol{f}(\boldsymbol{x}^{(1)}) - \boldsymbol{y}^*||_2^2 = ||\boldsymbol{f}(\boldsymbol{x}^{(2)}) - \boldsymbol{y}^*||_2^2 \,. \tag{2}$$

Here, $\boldsymbol{f}$ is a nonlinear function from $\mathbb{R}^n$ to $\mathbb{R}^m$, $\boldsymbol{x}^{(1)}$, $\boldsymbol{x}^{(2)} \in \mathbb{R}^n$ and $\boldsymbol{y}^* \in \mathbb{R}^m$. The nonlinear function $\boldsymbol{f}$ can be derived from a mathematical model, the vector $\boldsymbol{x}$ can be regarded as a set of model parameters which one wishes to estimate, and the vector $\boldsymbol{y}^*$ can be regarded as a set of observations one wishes to fit the model to. In the motivating problem of estimating parameters of PBPK models, since there are usually insufficient observations, the corresponding nonlinear least squares problem has multiple global minimisers. Therefore, we are interested in finding multiple minimisers instead of just one minimiser.

We shall call the following quantity the sum of squared residuals (SSR):

$$||\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{y}^*||_2^2 \,, \tag{3}$$

and use it for the quantification of the goodness of $\boldsymbol{x}$ as the approximation of the solution of the least squares problem (1).

1.2 Well known examples in pharmacokinetics

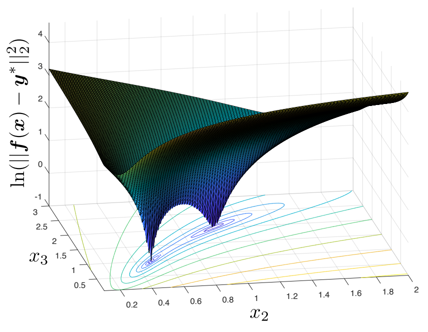
In this subsection, we present two simple pharmacokinetics parameter estimation problems where the corresponding nonlinear least squares problems have non-unique global minimisers. The first example is called 'flip-flop kinetics' and can be found in most standard textbooks in pharmacokinetics (e,g, [14]). Flip-flop kinetics occurs when estimating the pharmacokinetic parameters of the drug that is orally given (e.g., as a pill or a tablet) to patients based on the observation of the drug concentration in the blood plasma. The simplest mathematical model for this pharmacokinetics can be written as follows:

$$\begin{aligned}
\frac{\mathrm{d}u_1}{\mathrm{d}t} &= -Ka\,u_1 && (4)\\
\frac{\mathrm{d}u_2}{\mathrm{d}t} &= \frac{Ka\,u_1 - CL\,u_2}{V} && (5)\\
u_1(t=0) &= 100 && (6)\\
u_2(t=0) &= 0 && (7)
\end{aligned}$$

where

$$\begin{aligned}
CL &= 10^{x_1} && (8)\\
Ka &= 10^{x_2} && (9)\\
V &= 10^{x_3} \,. && (10)
\end{aligned}$$

For this problem u_2 corresponds to the drug concentration in the blood plasma, which is observable. It can be shown analytically that there are two distinct parameter sets that realise the same drug concentration time-course curve. Figure 1a shows the surface plot of the sum of squared residuals of the corresponding nonlinear least squares problem. As can be seen, there are two global minimisers for this nonlinear least squares problem.


(a) Flip-flop kinetics.

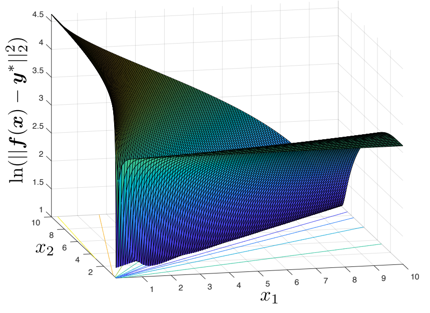

(b) Total amount of the drug in the body as observation.

Figure 1: Surface plots of the sum of squared residuals for pharmacokinetics model parameter estimation problems with non-unique global minimisers.

The second example is when an isotope labelled substance is given to patients via intervenes administration and pharmacokinetics parameters are estimated from the total amount of the substance in the body (i.e., not from the concentration). The simplest mathematical model for this pharmacokinetics can be written as follows:

$$\frac{\mathrm{d}u}{\mathrm{d}t} = -\frac{CL}{V}u \tag{11}$$
$$u(t=0) = 100 \tag{12}$$

where

$$CL = 10^{x_1} \tag{13}$$
$$V = 10^{x_2}. \tag{14}$$

For this problem, u corresponds to the total amount of the substance in the body, which is the observable quantity. In this case, there are infinite sets of parameters (x_1, x_2) which satisfy a linear relation, that realises the given time-course substance amount observation. Figure 1b is a surface plot of the sum of squared residuals of the corresponding nonlinear least squares problem. As can be seen, there is a line of global minimisers for this nonlinear least squares problem.

In this subsection, we have shown the simplest possible forms of this issue of non-unique global minimisers in pharmacokinetics. Thus, we would like to point out that one cannot assume the uniqueness of the global minimiser for more complex pharmacokinetic models, for example, the PBPK model of our interest.

2 Method: Algorithm

In this section we describe the proposed algorithm. We first introduce a rough concept using a toy example in Subsection 2.1 and then introduce the full algorithm in detail in Subsection 2.2.

2.1 Brief explanation of the algorithm

The aim of the proposed Cluster Gauss-Newton (CGN) algorithm is to efficiently find multiple approximate minimisers of the nonlinear least squares problem (1). We do so by first creating a collection of initial guesses which we call the 'cluster'. Then, we move the cluster iteratively using linear approximations of the nonlinear function $\boldsymbol{f}$, similarly to the Gauss-Newton method [4].

The unique idea in the CGN method is that the linear approximation is constructed collectively throughout the points in the cluster instead of using the Jacobian matrix which approximates the nonlinear function linearly at a point as in the Gauss-Newton or LM. By using points in the cluster to construct a linear approximation, instead of explicitly approximating the Jacobian, we minimise the computational cost for each iteration. In addition, by constructing linear approximation using non-local information, CGN is more likely to converge to approximate minimisers with smaller SSR than methods using the Jacobian.

In order to visualise the key differences between the proposed linear approximation (CGN) and the Jacobian (i.e., derivative) approaches, we consider the nonlinear function

$$f = \begin{cases} (x+1)^2 - 2\cos(10(x+1)) + 5 & \text{if } x < -1 \\ 3 & \text{if } -1 \le x \le 1 \\ (x-1)^2 - 2\cos(10(x-1)) + 5 & \text{if } x > 1 \end{cases} \tag{15}$$

(see Figure 2) and aim to find global minimisers. Any point $x \in [-1, 1]$ is a global minimiser of this problem. Hence, this problem has nonunique global minimisers. Let the points of the initial iterates be:

$$\begin{aligned} &x_1 = -6.3797853, \qquad x_2 = -4.1656025, \qquad x_3 = -3.6145728, \\ &x_4 = 2.0755468, \qquad x_5 = 4.1540421. \end{aligned} \tag{16}$$

We now compute the linear approximations used to move these points in the cluster to minimise the function f.

Gradient (LM)

For this nonlinear function, since the function is given in analytic form, we can obtain the gradient explicitly. In, practice, when f is given as a "black box", we can approximate the derivative by a finite difference scheme, for example, $f'(x_i) \approx \frac{f(x_i+\epsilon)-f(x_i)}{\epsilon}$. Then, the linear approximation at x_i can be written as

$f(x) \approx \frac{f(x_i+\epsilon)-f(x_i)}{\epsilon}(x - x_i) + f(x_i)$. Notice that it requires one extra evaluation of f at $x_i + \epsilon$ for each x_i. This number of extra function evaluation is, when evaluating a full gradient estimate, equal to the number of independent variables of f. (This is not the case if a directional derivative estimate is used.) If f is given by a system of ODEs, one may use the adjoint method to obtain the derivatives more efficiently. However, it requires solving an additional system of ODEs (the adjoint equation). More importantly, iterates of methods based on the gradient may converge to local minimisers, since they use local gradient information.

Cluster Gauss-Newton (CGN) method (proposed method)

In the proposed method, we construct a linear approximation for each point in the cluster while using the value of f at other points in the cluster to globally approximate the nonlinear function with a linear function. The influence of another point in the cluster to the linear approximation is weighted according to how close the point is to the point of approximation, i.e.,

$$\min_{a_{(i)}} \sum_{i \neq j} \left(d_{j(i)} \left((x_j - x_i) a_{(i)} - (f(x_j) - f(x_i)) \right) \right)^2 \tag{17}$$

where $a_{(i)}$ is the slope of the linear approximation at x_i and the linear approximation at x_i can be written as $f(x) \approx a_{(i)}(x - x_i) + f(x_i)$. There are many possibilities for the weight function $d_{j(i)}$. In this paper, we choose $d_{j(i)} = (x_j - x_i)^{-2\gamma}$ where $\gamma \ge 0$ ($\gamma = 0$ corresponds to uniform weight). Note that Equation(17) can also be regarded as a weighted least squares solution of a system of linear equations where the weight is chosen to be $d_{j(i)}$. The weight is motivated by the fact that we weight the information from the neighbouring points in the cluster more than the ones further away when constructing the linear approximation. Note that we do not require any extra evaluation of f for obtaining these linear approximations.

For the multi-dimensional nonlinear function $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$, when we have N points in the cluster, we solve the following linear least squares problem:

$$\min_{A_{(i)} \in \mathbb{R}^{m \times n}} \left\| D_{(i)} \left(\Delta X_{(i)} A_{(i)}^{\mathrm{T}} - \Delta Y_{(i)} \right) \right\|_{\mathrm{F}} \tag{18}$$

where $D_{(i)} = \mathrm{diag}(d_{1i}, ..., d_{Ni})$ is a diagonal matrix defining the weights, $\Delta X_{(i)} \in \mathbb{R}^{N \times n}$ is the difference between all the cluster points and $\boldsymbol{x}_i$, and $\Delta Y_{(i)} \in \mathbb{R}^{N \times m}$ is the difference between the nonlinear function

$\boldsymbol{f}$ evaluated at all the cluster points and at $\boldsymbol{x}_i$. The precise definition of these quantities and derivation of (18) is given in the next subsection.

For the one dimensional case (i.e., $m = n = 1$), the linear approximations at each point for the first ten iterations for both CGN and LM are shown in Figure 2. As can be seen, the gradient used in LM captures the local behaviour of the nonlinear function. The linear approximation used in CGN, on the other hand, captures the global behaviour of the nonlinear function. After nine iterations of the CGN, all the points reached the minimisers with smallest SSR. On the other hand, the LM converges to local minimisers whose SSR are not necessarily the minimum.

2.2 Detailed description of the algorithm

Next, we describe the proposed CGN algorithm in detail. In this subsection, we denote a scalar quantity by a lower case letter e.g., a, c, a matrix by a capital letter, e.g., A, or M, and a column vector by a bold symbol of a lower case letter, e.g., $\boldsymbol{v}$, $\boldsymbol{a}$, unless otherwise specifically stated. Super script T indicates the transpose. Hence, $\boldsymbol{v}^{\mathrm{T}}$ and $\boldsymbol{a}^{\mathrm{T}}$ are row vectors.

1) Pre-iteration process

1-1) Create initial cluster

The initial iterates of CGN, a set of vectors $\{\boldsymbol{x}_i^{(0)}\}_{i=1}^N$ are genrated using uniform random numbers in each component within the domain of initial estimate of the plausible location of global minimisers given by the user. The unique point of the CGN is that the user specifies the domain of the initial estimates instead of a point. In this paper, we assume that the domain of initial guess is given by the user by two sets of vectors $\boldsymbol{x}^{\mathrm{L}}$, $\boldsymbol{x}^{\mathrm{U}}$, and the value $x_{ji}^{(0)}$ is sampled from the uniform distribution between x_j^{L} and x_j^{U}, where $x_{ji}^{(0)}$ is the j th element of vector $\boldsymbol{x}_i^{(0)}$.

Note that this does **not** mean that all the following iterates $\boldsymbol{x}_i^{(k)}$ $(k \geq 1, i = 1, 2, \ldots, N)$ must satisfy $\boldsymbol{x}^L \leq \boldsymbol{x}_i^{(k)} \leq \boldsymbol{x}^U$.

Store the initial set of vectors in a matrix $X^{(0)}$, i.e.,

$$X^{(0)} = [\boldsymbol{x}_1^{(0)}, \boldsymbol{x}_2^{(0)}, \ldots, \boldsymbol{x}_N^{(0)}] \tag{19}$$

where the super script (0) indicates the initial iterate.

Evaluate the nonlinear function $\boldsymbol{f}$ at each $\boldsymbol{x}_i^{(0)}$ as $\boldsymbol{y}_i^{(0)} = \boldsymbol{f}(\boldsymbol{x}_i^{(0)})$ $(i = 1, 2, \ldots, N)$ and store in matrix $Y^{(0)}$, i.e.,

$$Y^{(0)} = \left[\boldsymbol{y}_1^{(0)}, \boldsymbol{y}_2^{(0)}, \ldots, \boldsymbol{y}_N^{(0)}\right]. \tag{20}$$

If the function $\boldsymbol{f}$ cannot be evaluated at $\boldsymbol{x}_i^{(0)}$, then re-sample $\boldsymbol{x}_i^{(0)}$ until $\boldsymbol{f}$ can be evaluated.

Compute the sum of squared residuals vector $\boldsymbol{r}^{(0)}$ i.e.,

$$r_i^0 = ||\boldsymbol{y}_i^0 - \boldsymbol{y}^*||_2^2 \quad (i = 1, 2, \ldots, N) \tag{21}$$

$$\boldsymbol{r}^{(0)} = \left[r_1^{(0)}, r_2^0, \ldots, r_N^0\right]^{\mathrm{T}}. \tag{22}$$

The concise pseudo-code for the creation of the initial cluster can be found in Algorithm 1.

1-2) Initialise regularisation parameter vector

Fill the regularisation parameter vector $\boldsymbol{\lambda}^{(0)} \in \mathbb{R}^N$, with the user-specified initial regularisation parameter λ_{init} .i.e.,

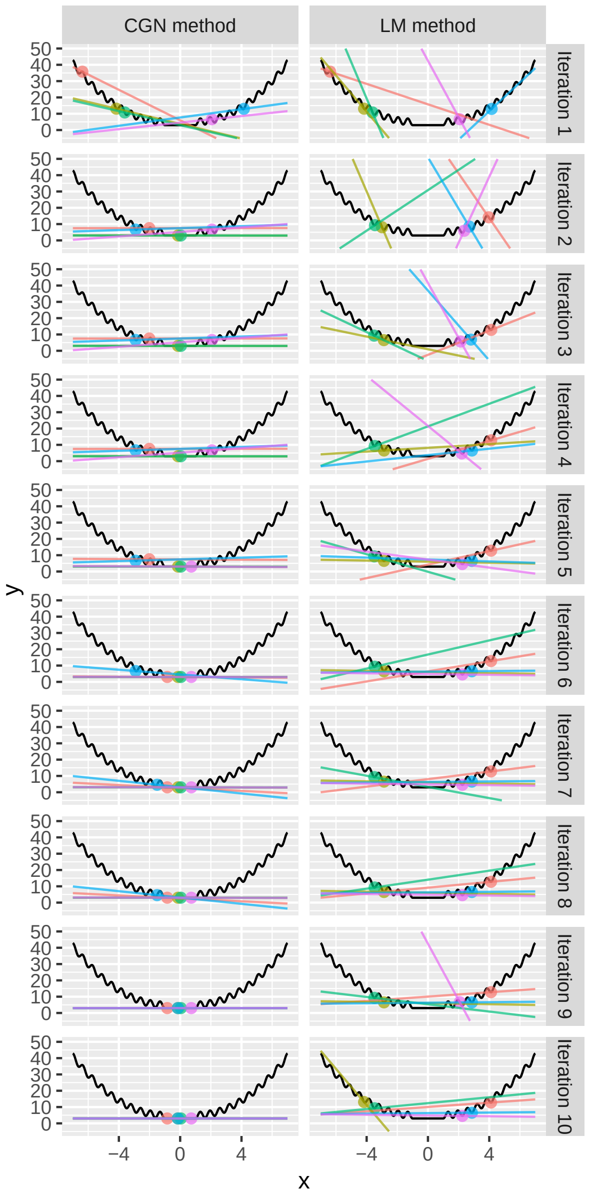


Figure 2: Schematic comparison of CGN and LM. Dots represent the iterates with their function values in each iteration and the lines represent linear approximations used to update the iterates. As can be seen in this figure, the linear approximation used in CGN follows the global trend of the function, while the slope used in LM captures the local feature of the function. As a result, when using CGN all the iterates converge to the minimisers with the smallest residual (in this case global minimisers), while all the iterates of LM converge to local minimisere whose residuals are not the minimum. In addition, for the 10 iterations presented in the figure, CGN required only 50 function evaluations while LM required 121 function evaluations since the slope is approximated using finite differences.

Algorithm 1 $(X^{(0)}, Y^{(0)}, \boldsymbol{r}^{(0)})$=**create_initialCluster**$(\boldsymbol{x}^{\mathrm{L}}, \boldsymbol{x}^{\mathrm{U}}, N, \boldsymbol{f}, \boldsymbol{y}^*)$

```
$X^{(0)} \leftarrow$ allocate memory for $n \times N$ matrix
$Y^{(0)} \leftarrow m \times N$ matrix of not a number (NaN)
for all $i = 1, ..., N$ do
  while $\boldsymbol{y}_i^{(0)}$ is not a NaN vector do
    for $j \leftarrow 1, n$ do
      Draw random number $x_{ij}^{(0)}$ from $\mathcal{U}(x_j^{\mathrm{L}}, x_j^{\mathrm{U}})$
    end for
    if $\boldsymbol{f}(\boldsymbol{x}_i)$ can be evaluated then
      $\boldsymbol{y}_i^{(0)} \leftarrow \boldsymbol{f}(\boldsymbol{x}_i)$
    else
      $\boldsymbol{y}_i^{(0)} \leftarrow$ NaN vector of length $m$
    end if
  end while
  $r_i^{(0)} \leftarrow ||\boldsymbol{y}_i^{(0)} - \boldsymbol{y}^*||_2^2$
end for
```

$$\boldsymbol{\lambda}^{(0)} = [\lambda_{\mathrm{init}}, \lambda_{\mathrm{init}}, ..., \lambda_{\mathrm{init}}]^{\mathrm{T}} \tag{23}$$

2) Main iteration

Repeat the following procedure until the user specified stopping criteria are met. We denote the iteration number as k, which starts from 0 and is incremented by 1 after each iteration.

2-1) Construct weighted linear approximations of the nonlinear function

We first construct a linear approximation around the point $\boldsymbol{x}_i^{(k)}$, s.t.,

$$\boldsymbol{f}(\boldsymbol{x}) \approx A_{(i)}^{(k)}(\boldsymbol{x} - \boldsymbol{x}_i^{(k)}) + \boldsymbol{f}(\boldsymbol{x}_i^{(k)}), \tag{24}$$

Here, $A_{(i)}^{(k)} \in \mathbb{R}^{m\times n}$ describes the slope of the linear approximation around $\boldsymbol{x}_i^{(k)}$.

The key difference of our algorithm compared to others is that we construct a Jacobian like matrix $A_{(i)}^{(k)}$ collectively using all the function evaluations of $\boldsymbol{f}$ in the previous iteration, i.e., we solve

$$\begin{aligned} &\min_{A\in\mathbb{R}^{m\times n}} \sum_{j=1}^{N} \left[d_{j(i)}^{(k)} \left|\left| \boldsymbol{f}(\boldsymbol{x}_j^{(k)}) - \left\{ A\left(\boldsymbol{x}_j^{(k)} - \boldsymbol{x}_i^{(k)}\right) + \boldsymbol{f}(\boldsymbol{x}_i^{(k)}) \right\} \right|\right|_2 \right]^2 \\ &= \min_{A\in\mathbb{R}^{m\times n}} \sum_{j=1}^{N} \left(d_{j(i)}^{(k)} \left|\left| \Delta\boldsymbol{y}_{j(i)}^{(k)} - A\Delta\boldsymbol{x}_{j(i)}^{(k)} \right|\right|_2 \right)^2 \end{aligned} \tag{25}$$

for $i = 1, ..., N$, where $d_{j(i)}^{(k)} \geq 0$, $j = 1, ..., N$ are weights. Here, $\Delta\boldsymbol{y}_{j(i)}^{(k)} = \boldsymbol{f}(\boldsymbol{x}_j^{(k)}) - \boldsymbol{f}(\boldsymbol{x}_i^{(k)}) \in \mathbb{R}^m$ and $\Delta\boldsymbol{x}_{j(i)}^{(k)} = \boldsymbol{x}_j^{(k)} - \boldsymbol{x}_i^{(k)} \in \mathbb{R}^n$. (Note that $\Delta\boldsymbol{y}_{i(i)}^{(k)} = \boldsymbol{0}$, $\Delta\boldsymbol{x}_{i(i)}^{(k)} = \boldsymbol{0}$.) Also, let

$$\Delta Y_{(i)}^{(k)} = \left[\Delta\boldsymbol{y}_{1(i)}^{(k)}, \Delta\boldsymbol{y}_{2(i)}^{(k)}, \ldots \Delta\boldsymbol{y}_{N(i)}^{(k)}\right] \in \mathbb{R}^{m\times N} \tag{26}$$

$$\Delta X_{(i)}^{(k)} = \left[\Delta\boldsymbol{x}_{1(i)}^{(k)}, \Delta\boldsymbol{x}_{2(i)}^{(k)}, \ldots \Delta\boldsymbol{x}_{N(i)}^{(k)}\right] \in \mathbb{R}^{n\times N}. \tag{27}$$

Note that $\boldsymbol{f}(\boldsymbol{x}_i^{(k)})$ are always computed at the previous iteration (e.g., as equation (20) when $k = 0$ and in Step 2-3 when $k > 0$). Hence, no new evaluation of $\boldsymbol{f}$ is required at this step.

The key idea here is that we weight the information of the function evaluation near $\boldsymbol{x}_i^{(k)}$ more than the function evaluation further away. That is to say, $d_{j(i)}^{(k)} > d_{j'(i)}^{(k)}$ if $||\boldsymbol{x}_j^{(k)} - \boldsymbol{x}_i^{(k)}|| < ||\boldsymbol{x}_{j'}^{(k)} - \boldsymbol{x}_i^{(k)}||$. The importance of this idea can be seen in the numerical experiment presented in Appendix C.

Noting that

$$\sum_{j=1}^{N} \left(d_{j(i)}^{(k)} \left|\left| A\Delta \boldsymbol{x}_{j(i)}^{(k)} - \Delta \boldsymbol{y}_{j(i)}^{(k)} \right|\right|_2 \right)^2 = \left|\left| \left(A\Delta X_{(i)}^{(k)} - \Delta Y_{(i)}^{(k)} \right) D_{(i)}^{(k)} \right|\right|_{\mathrm{F}}^2 \tag{28}$$

$$= \left|\left| D_{(i)}^{(k)} \left(\Delta X_{(i)}^{(k)\,\mathrm{T}} A^{\mathrm{T}} - \Delta Y_{(i)}^{(k)\,\mathrm{T}} \right) \right|\right|_{\mathrm{F}}^2 , \tag{29}$$

we can rewrite Equation (25) as

$$\min_{A_{(i)}^{(k)} \in \mathbb{R}^{m\times n}} \left|\left| D_{(i)}^{(k)} \left(\Delta X_{(i)}^{(k)\,\mathrm{T}} A_{(i)}^{(k)\mathrm{T}} - \Delta Y_{(i)}^{(k)\,\mathrm{T}} \right) \right|\right|_{\mathrm{F}}^2 \tag{30}$$

where

$$D_{(i)}^{(k)} = \mathrm{diag} \left(d_{1(i)}^{(k)}, d_{2(i)}^{(k)}, ..., d_{N(i)}^{(k)} \right) , \tag{31}$$

where $d_{l(i)}^{(k)} \geq 0$. In this paper, we choose the weights as

$$d_{j(i)}^{(k)} = \begin{cases} \left(\frac{1}{\sum_{l=1}^{n} ((x_{lj}^{(k)} - x_{li}^{(k)})/(x_l^{\mathrm{U}} - x_l^{\mathrm{L}}))^2} \right)^{\gamma} & \text{if } j \neq i \\ 0 & \text{if } j = i \end{cases} , \tag{32}$$

where $x_{lj}^{(k)}, x_l^{\mathrm{U}}, x_l^{\mathrm{L}}$ are the l th element of the vectors $\boldsymbol{x}_j^{(k)}, \boldsymbol{x}^{\mathrm{U}}, \boldsymbol{x}^{\mathrm{L}}$, respectively ($l = 1, ..., n$), and $\gamma \geq 0$ is a constant. We use this weighting scheme so that the "information" from the nonlinear function evaluation from the point closer to the point of approximation is more influential when constructing the linear approximation. The distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are normalised by the size of the domain of initial guess (i.e., $\boldsymbol{x}^{\mathrm{U}}$ and $\boldsymbol{x}^{\mathrm{L}}$). The effect of the weight $d_{j(i)}^{(k)}$ and the parameter γ and its necessity is analysed in Appendix C. The minimum norm solution of Equation (30) is given by

$$A_{(i)}^{(k)} = \Delta Y_{(i)}^{(k)} D_{(i)}^{(k)} \left(\Delta X_{(i)}^{(k)} D_{(i)}^{(k)} \right)^{\dagger} . \tag{33}$$

where † denotes the Moore-Penrose inverse.

If rank $\left(\Delta X_{(i)}^{(k)} D_{(i)}^{(k)} \right) = n$,

$$A_{(i)}^{(k)} = \Delta Y_{(i)}^{(k)} D_{(i)}^{(k)} \left(\Delta X_{(i)}^{(k)} D_{(i)}^{(k)} \right)^{\mathrm{T}} \left\{ \left(\Delta X_{(i)}^{(k)} D_{(i)}^{(k)} \right) \left(\Delta X_{(i)}^{(k)} D_{(i)}^{(k)} \right)^{\mathrm{T}} \right\}^{-1} . \tag{34}$$

Generically, $\mathrm{rank}\Delta X_{(i)}^{(k)} D_{(i)}^{(k)} = n$. $\mathrm{rank}\Delta X_{(i)}^{(k)} < n$ can happen when $x_{li}^{(k)} = c \quad (i = 1, 2, \ldots, N)$, i.e. when $\boldsymbol{x}_i^{(k)}$ lie in the same hyperplane $x_l = c$. This happens when the l-th component of $\boldsymbol{x}_i^{(k)} \quad (i = 1, 2, \ldots, N)$ are all equal.

The concise pseudo-code for the weighted linear approximation can be found in Algorithm 2.

2-2) Solve for $\boldsymbol{x}$ that minimises $||\boldsymbol{y}^* - (A_{(i)}^{(k)}(\boldsymbol{x} - \boldsymbol{x}_i^{(k)}) + \boldsymbol{f}(\boldsymbol{x}_i^{(k)}))||_2^2$

We now compute the next iterate $X^{(k+1)}$ using the matrices $\{A_{(i)}^{(k)}\}_{i=1}^{N}$ similarly to the Gauss-Newton method with Tikhonov regularisation (e.g., [4, 18]), i.e.,

$$\boldsymbol{x}_i^{(k+1)} = \boldsymbol{x}_i^{(k)} + \left(A_{(i)}^{(k)\mathrm{T}} A_{(i)}^{(k)} + \lambda_i^{(k)} I \right)^{-1} A_{(i)}^{(k)\mathrm{T}} (\boldsymbol{y}^* - \boldsymbol{y}_i^{(k)}) \qquad \text{for } i = 1, ..., N, \tag{35}$$

Algorithm 2 $A_{(i)}^{(k)}$=**construct_linearApproximation**$(i, X^{(k)}, Y^{(k)}, \boldsymbol{x}^{\mathrm{L}}, \boldsymbol{x}^{\mathrm{U}}, \gamma)$

```
$D_{(i)}^{(k)} \leftarrow$ allocate memory for diagonal matrix of $N \times N$
$\Delta X^{(0)} \leftarrow$ allocate memory for $n \times N$ matrix
$\Delta Y^{(0)} \leftarrow$ allocate memory for $m \times N$ matrix
for all $j = 1, ..., N$ do
    $d_{jj(i)}^{(k)} \leftarrow \begin{cases} \left(\frac{1}{\sum_{l=1}^{n}((x_{lj}^{(k)} - x_{li}^{(k)})/(x_l^{\mathrm{U}} - x_l^{\mathrm{L}}))^2}\right)^{\gamma} & \text{if } j \neq i \\ 0 & \text{if } j = i \end{cases}$
    $\Delta \boldsymbol{x}_j^{(k)} \leftarrow \boldsymbol{x}_j^{(k)} - \boldsymbol{x}_i^{(k)}$
    $\Delta \boldsymbol{y}_j^{(k)} \leftarrow \boldsymbol{y}_j^{(k)} - \boldsymbol{y}_i^{(k)}$
end for
$A_{(i)}^{(k)} \leftarrow \Delta Y_{(i)}^{(k)} D_{(i)}^{(k)} \left(\Delta X_{(i)}^{(k)} D_{(i)}^{(k)}\right)^{\dagger}$ {†: Moore-Penrose inverse.}
```

where $\boldsymbol{y}^*$ is the set of observations one wishes to fit the nonlinear function $\boldsymbol{f}$ to (cf. (1)), and $\boldsymbol{y}_i^{(k)} \equiv \boldsymbol{f}(\boldsymbol{x}_i^{(k)})$. For CGN, we require $\lambda_i^{(k)} > 0$. The necessity of the regularisation can be seen in the numerical experiment presented in Appendix F.

2-3) Update matrices X and Y and vectors $\boldsymbol{r}$ and $\boldsymbol{\lambda}$

Evaluate the nonlinear function $\boldsymbol{f}$ for each $\boldsymbol{x}_i^{(k+1)}$ as $\boldsymbol{y}_i^{(k+1)} = \boldsymbol{f}(\boldsymbol{x}_i^{(k+1)})$ $(\boldsymbol{i = 1, 2, \ldots, n})$, and store as $\boldsymbol{Y}^{(k+1)} = [\boldsymbol{y}_1^{(k+1)}, \boldsymbol{y}_2^{(k+1)}, \ldots, \boldsymbol{y}_N^{(k+1)}]$. Compute the sum of squared residuals vector as $\boldsymbol{r}^{(k+1)} = [r_1^{(k+1)}, r_2^{(k+1)}, \ldots, r_N^{(k+1)}]^{\mathrm{T}}$ where $r_i^{(k+1)} = \|y_i^{(k+1)} - y^*\|_2^2$ $(i = 1, 2, \ldots, N)$. Note that this process can be implemented embarrassingly parallelly.

If the residual r_i increases, we replace $\boldsymbol{x}_i^{(k+1)}$ by $\boldsymbol{x}_i^{(k)}$ and increase the regularisation parameter i.e.,

if $r_i^{(k)} < r_i^{(k+1)}$ or $\boldsymbol{f}(\boldsymbol{x}_i^{(k+1)})$ can not be evaluated, then let

$$\boldsymbol{x}_i^{(k+1)} = \boldsymbol{x}_i^{(k)} \tag{36}$$

$$\boldsymbol{y}_i^{(k+1)} = \boldsymbol{y}_i^{(k)} \tag{37}$$

$$\lambda_i^{(k+1)} = 10\,\lambda_i^{(k)} \tag{38}$$

else decrease the regularisation parameter, i.e.,

$$\lambda_i^{(k+1)} = \frac{1}{10}\lambda_i^{(k)}, \tag{39}$$

for each $i = 1, ..., N$.

There are various ways to update the regularisation parameter λ. In this paper we followed the strategy used in the Matlab's implementation of the Levenberg-Marquardt method.

In addition, in this step, we impose the stopping criteria for each point in the iteration. As can be seen in Equation (35), $\boldsymbol{x}_i^{(k+1)} \approx \boldsymbol{x}_i^{(k)}$ for large λ_i, so that we can expect very small update in $\boldsymbol{x}_i^{(k+1)}$. Hence, in order to mimic the minimum step size stopping criteria, we stop the update for i where $\lambda_i > \lambda_{\max}$.

The concise pseudo-code for updating matrices X and Y and vectors $\boldsymbol{r}$ and $\boldsymbol{\lambda}$ is given in Algorithm 3.

Algorithm 3
$(X^{(k+1)}, Y^{(k+1)}, \boldsymbol{r}^{(k+1)}, \boldsymbol{\lambda}^{(k+1)})$=**update_cluster**$(X^{(k)}, X^{(k+1)}, Y^{(k)}, \boldsymbol{r}^{(k)}, \boldsymbol{\lambda}^{(k)}, \lambda_{\max})$

```
$Y^{(k+1)} \leftarrow$ allocate memory for $m \times N$ matrix
$\boldsymbol{r}^{(k+1)} \leftarrow$ allocate memory for length $N$ vector
$\boldsymbol{\lambda}^{(k+1)} \leftarrow$ allocate memory for length $N$ vector
for all $i = 1, ..., N$ do
  if $\lambda_i^{(k)} \leq \lambda_{\max}$ and $\boldsymbol{f}(\boldsymbol{x}_i^{(k+1)})$ can be evaluated then
    $\boldsymbol{y}_i^{(k+1)} \leftarrow \boldsymbol{f}(\boldsymbol{x}_i^{(k+1)})$
    $r_i^{(k+1)} \leftarrow ||\boldsymbol{y}_i^{(k+1)} - \boldsymbol{y}^*||_2^2$
    if $r_i^{(k+1)} > r_i^{(k)}$ then
      $\boldsymbol{x}_i^{(k+1)} \leftarrow \boldsymbol{x}_i^{(k)}$
      $\boldsymbol{y}_i^{(k+1)} \leftarrow \boldsymbol{y}_i^{(k)}$
      $r_i^{(k+1)} \leftarrow r_i^{(k)}$
      $\lambda_i^{(k+1)} \leftarrow 10\,\lambda_i^{(k)}$
    else
      $\lambda_i^{(k+1)} \leftarrow \frac{1}{10}\lambda_i^{(k)}$
    end if
  else
    $\boldsymbol{x}_i^{(k+1)} \leftarrow \boldsymbol{x}_i^{(k)}$
    $\boldsymbol{y}_i^{(k+1)} \leftarrow \boldsymbol{y}_i^{(k)}$
    $r_i^{(k+1)} \leftarrow r_i^{(k)}$
    $\lambda_i^{(k+1)} \leftarrow 10\,\lambda_i^{(k)}$
  end if
end for
```

The influence of the choice of the initial value λ_{init} of the regularization parameter in equation (23) is studied in Appendix D.

We present a concise description of the CGN as a pseudo-code in Algorithm 4.

Algorithm 4
$X^{(k_{\max})}$ =**Cluster_Gauss-Newton_method**$(\boldsymbol{x}^{\text{L}}, \boldsymbol{x}^{\text{U}}, N, \boldsymbol{f}, \boldsymbol{y}^{*}, \lambda_{\text{init}}, \lambda_{\max}, \gamma, k_{\max})$

```
$\boldsymbol{r}^{(0)} \leftarrow$ allocate memory for length $N$ vector
for all $i \leftarrow 1, N$ do
    $\lambda_i^{(0)} = \lambda_{\text{init}}$
end for
$(X^{(0)}, Y^{(0)}, \boldsymbol{r}^{(0)})$=create_initialCluster$(\boldsymbol{x}^{\text{L}}, \boldsymbol{x}^{\text{U}}, N, \boldsymbol{f}, \boldsymbol{y}^{*})$
for $k \leftarrow 0, (k_{\max} - 1)$ do
    for all $i \leftarrow 1, N$ do
        if $\lambda_i^{(k)} \le \lambda_{\max}$ then
            $A_{(i)}^{(k)}$=construct_linearApproximation$(i, X^{(k)}, Y^{(k)}, \boldsymbol{x}^{\text{L}}, \boldsymbol{x}^{\text{U}}, \gamma)$
            $\boldsymbol{x}_i^{(k+1)} = \boldsymbol{x}_i^{(k)} + \left(A_{(i)}^{(k)\text{T}} A_{(i)}^{(k)} + \lambda_i^{(k)} I\right)^{-1} A_{(i)}^{(k)\text{T}} (\boldsymbol{y}^{*} - \boldsymbol{y}_i^{(k)})$
        end if
    end for
    $(X^{(k+1)}, Y^{(k+1)}, \boldsymbol{r}^{(k+1)}, \boldsymbol{\lambda}^{(k+1)})$
            =update_cluster$(X^{(k)}, X^{(k+1)}, Y^{(k)}, \boldsymbol{r}^{(k)}, \boldsymbol{\lambda}^{(k)}, \lambda_{\max})$
end for
```

3 Numerical Experiments

In this section we illustrate the advantages of the proposed CGN algorithm by numerical experiments on three PBPK models.

3.1 Numerical Experiment setup

In this subsection we specify the details of the numerical experiment set-up.

3.1.1 Mathematical models

For the numerical experiments, we used the following three published mathematical models of drug concentration in the blood of a human body (PBPK model). The time course drug concentration was simulated using the model, and random noise was added to mimic the observation uncertainties. The random noise was generated by a normal distribution with a standard deviation of 10% of the simulated concentration value. The simulated drug concentration was used as the test dataset, and multiple possible parameters were estimated from the test dataset using CGN and conventional methods. The implementations of all the examples are available as the supplementary material.

Example 1: Multi-dose problem
For this example, we consider the case where three different amounts of a drug is given (low-dose, medium-dose, and high-dose) orally as pills to a patient. This can be modelled using the same mathematical model as in the Motivating Example (cf. Appendix A and Appendix D). The initial

value problem of this mathematical model can be written as:

$$\begin{aligned}\frac{\mathrm{d}\boldsymbol{u}}{\mathrm{d}t} &= \boldsymbol{g}(\boldsymbol{u}, t; \boldsymbol{x})\,, && (40)\\ u_i(t=0) &= 0 \qquad \text{for } i = 1, ..., 17, 19, 20\,, && (41)\\ u_{18}(t=0) &= \begin{cases} 30000 & \text{for low-dose,}\\ 100000 & \text{for medium-dose,}\\ 300000 & \text{for high-dose.}\end{cases} && (42)\end{aligned}$$

In this mathematical model, $u_1(t)$ represents the drug concentration in blood at time t. For convenience, we denote the drug concentration at time t for the low-dose, medium-dose, and high-dose as $u_1^{\mathrm{l}}(t;\boldsymbol{x})$, $u_1^{\mathrm{m}}(t;\boldsymbol{x})$, and $u_1^{\mathrm{h}}(t;\boldsymbol{x})$, respectively. Note that the right hand side of the system of ODEs (40) depends on the parameter vector $\boldsymbol{x}$. Hence, the solution of the system of ODEs $\boldsymbol{u}$ depends not only on t but also on $\boldsymbol{x}$. For this example, we consider the case where the blood sample is taken at $t = 2, 3, 4, 6, 8, 12, 24, 36, 48, 72$, so that the nonlinear function $\boldsymbol{f}$ in (1) can be written as

$$\begin{aligned}\boldsymbol{f}(\boldsymbol{x}) \;=\; &[u_1^{\mathrm{l}}(2;\boldsymbol{x}), u_1^{\mathrm{l}}(3;\boldsymbol{x}), u_1^{\mathrm{l}}(4;\boldsymbol{x}), u_1^{\mathrm{l}}(6;\boldsymbol{x}), u_1^{\mathrm{l}}(8;\boldsymbol{x}), u_1^{\mathrm{l}}(12;\boldsymbol{x}), u_1^{\mathrm{l}}(24;\boldsymbol{x}), u_1^{\mathrm{l}}(36;\boldsymbol{x}),\\ &u_1^{\mathrm{l}}(48;\boldsymbol{x}), u_1^{\mathrm{l}}(72), u_1^{\mathrm{m}}(2;\boldsymbol{x}), u_1^{\mathrm{m}}(3;\boldsymbol{x}), u_1^{\mathrm{m}}(4;\boldsymbol{x}), u_1^{\mathrm{m}}(6;\boldsymbol{x}), u_1^{\mathrm{m}}(8;\boldsymbol{x}), u_1^{\mathrm{m}}(12;\boldsymbol{x}),\\ &u_1^{\mathrm{m}}(24;\boldsymbol{x}), u_1^{\mathrm{m}}(36;\boldsymbol{x}), u_1^{\mathrm{m}}(48;\boldsymbol{x}), u_1^{\mathrm{m}}(72), u_1^{\mathrm{h}}(2;\boldsymbol{x}), u_1^{\mathrm{h}}(3;\boldsymbol{x}), u_1^{\mathrm{h}}(4;\boldsymbol{x}), u_1^{\mathrm{h}}(6;\boldsymbol{x}),\\ &u_1^{\mathrm{h}}(8;\boldsymbol{x}), u_1^{\mathrm{h}}(12;\boldsymbol{x}), u_1^{\mathrm{h}}(24;\boldsymbol{x}), u_1^{\mathrm{h}}(36;\boldsymbol{x}), u_1^{\mathrm{h}}(48;\boldsymbol{x}), u_1^{\mathrm{h}}(72)]^{\mathrm{T}}.\end{aligned}$$

This example is based on [11].

Example 2: Intravenous injection and oral administration problem
For this example, we consider the case where the drug is administered via injection into the vein (i.v.) and the case where the drug is given orally as a pill (p.o.) to a patient. This was modelled using a mathematical model similar to the motivating example in Appendix A. The initial value problem of this mathematical model for i.v. administration can be written as:

$$\begin{aligned}\frac{\mathrm{d}\boldsymbol{u}}{\mathrm{d}t} &= \tilde{\boldsymbol{g}}(\boldsymbol{u}, t; \boldsymbol{x})\,, && (43)\\ u_1(t=0) &= \frac{30.488}{0.074 + 10^{x_4}}\,, && (44)\\ u_i(t=0) &= 0 \qquad \text{for } i = 2, 3, ..., 20\,, && (45)\end{aligned}$$

where x_4 is one of the parameters in the parameter vector $\boldsymbol{x}$. The initial value problem of this mathematical model for p.o. administration can be written as:

$$\begin{aligned}\frac{\mathrm{d}\boldsymbol{u}}{\mathrm{d}t} &= \tilde{\boldsymbol{g}}(\boldsymbol{u}, t; \boldsymbol{x})\,, && (46)\\ u_i(t=0) &= 0 \qquad \text{for } i = 1, ..., 20\,, && (47)\\ u_{18}\left(t = \frac{e^{x_{11}}}{2(1+e^{x_{11}})}\right) &= 30.488\,, && (48)\\ u_i\left(t = \frac{e^{x_{11}}}{2(1+e^{x_{11}})}\right) &= 0 \qquad \text{for } i = 1, ..., 17, 19, 20\,, && (49)\end{aligned}$$

where x_{11} is one of the parameters in the parameter vector $\boldsymbol{x}$. $\frac{e^{x_{11}}}{2(1+e^{x_{11}})}$ is the delay in the absorption of the drug (for example the time it takes from intake in the mouth to pill being dissolved in the stomach). Similarly to Example 1, the observable quantity, the drug concentration in the blood plasma is represented as $u_1(t;\boldsymbol{x})$. For convenience, we denote the drug concentration at time t when the drug is administrated by i.v. and p.o. administrations as $u_1^{\mathrm{i}}(t;\boldsymbol{x})$ and $u_1^{\mathrm{p}}(t;\boldsymbol{x})$, respectively. For this example, we consider the case where the blood sample is taken at $t = 0.0833, 0.1667, 0.25, 0.5, 0.75, 1, 1.5, 2, 3, 4, 6, 8$ when the drug is given as intravenous injection and the blood sample is taken at $t = 0.5, 1, 1.5, 2, 3, 4, 6, 8, 12, 14$ when the drug is given as a pill orally. The nonlinear function $\boldsymbol{f}$ can be written as

$$\begin{aligned}\boldsymbol{f}(\boldsymbol{x}) &= [u_1^i(0.0833;\boldsymbol{x}), u_1^i(0.1667;\boldsymbol{x}), u_1^i(0.25;\boldsymbol{x}), u_1^i(0.5;\boldsymbol{x}), u_1^i(0.75;\boldsymbol{x}), u_1^i(1;\boldsymbol{x}),\\ &\quad u_1^i(1.5;\boldsymbol{x}), u_1^i(2;\boldsymbol{x}), u_1^i(3;\boldsymbol{x}), u_1^i(4;\boldsymbol{x}), u_1^i(6;\boldsymbol{x}), u_1^i(8;\boldsymbol{x}), u_1^p(0.5;\boldsymbol{x}), u_1^p(1;\boldsymbol{x}),\\ &\quad u_1^p(1.5;\boldsymbol{x}), u_1^p(2;\boldsymbol{x}), u_1^p(3;\boldsymbol{x}), u_1^p(4;\boldsymbol{x}), u_1^p(6;\boldsymbol{x}), u_1^p(8;\boldsymbol{x}), u_1^p(12;\boldsymbol{x}), u_1^p(14;\boldsymbol{x})]^{\mathrm{T}}.\end{aligned}$$

This example is based on [37]. For this example, we used the upper and lower bounds of the initial cluster of CGN, which we know to include the known global minimiser (based on the parameter set we simulated the data from), as the bound constraints for the constrained optimisation algorithms. However, in practice, it is usually very hard to specify the upper and lower bounds of the parameters for the parameter estimation problem of our interest.

Example 3: Drug-drug interaction problem

For this example, we consider the case when a patient takes two different drugs. As is often stated in the instruction for drugs, if two drugs are taken together, they can potentially interact inside the body and cause undesirable effects. We model the cases where pitavastatin (for the ease of writing we shall refer to this drug as Drug A) is taken alone or with cyclosporin A (Drug B). The concentration of Drug A is modelled using a mathematical model similar to the Motivating Example (Appendix A) and Drug B is modelled using a simplified version of the model. The interaction of Drugs A and B in the liver compartment, when they are administered simultaneously is also modelled.

The initial value problem of the mathematical model for Drug A can be written as

$$\begin{aligned}\frac{\mathrm{d}\boldsymbol{u}}{\mathrm{d}t} &= \boldsymbol{g}(\boldsymbol{u}, t; \boldsymbol{x}) && (50)\\ u_i(t=0) &= 0 \qquad \text{for } i = 1, ..., 20\,, && (51)\\ u_{12}\left(t = \frac{e^{x_{18}}}{2(1+e^{x_{18}})}\right) &= 30.4971\,, && (52)\\ u_i\left(t = \frac{e^{x_{18}}}{2(1+e^{x_{18}})}\right) &= 0 \qquad \text{for } i = 1, ..., 11, 13, ..., 20\,. && (53)\end{aligned}$$

The initial value problem of the mathematical model for Drug A administered with Drug B can be written as:

$$\begin{aligned}\frac{\mathrm{d}\boldsymbol{u}}{\mathrm{d}t} &= \boldsymbol{h}(\boldsymbol{u}, t; \boldsymbol{x}) && (54)\\ u_i(t=0) &= 0 \qquad \text{for } i = 1, ..., 33\,, && (55)\\ u_{12}\left(t = \frac{e^{x_{18}}}{2(1+e^{x_{18}})}\right) &= 30.4971\,, && (56)\\ u_{33}\left(t = \frac{e^{x_{19}}}{2(1+e^{x_{19}})}\right) &= 2000\,, && (57)\\ u_i\left(t = \frac{e^{x_{18}}}{2(1+e^{x_{18}})}\right) &= 0 \qquad \text{for } i = 1, ..., 11, 13, .., 33\,, && (58)\\ u_i\left(t = \frac{e^{x_{19}}}{2(1+e^{x_{19}})}\right) &= 0 \qquad \text{for } i = 1, ..., 32\,. && (59)\end{aligned}$$

For convenience, we denote the concentration of Drug A at time t when only Drug A is administered as $u_1^{\mathrm{A}}(t;\boldsymbol{x})$. When both Drugs A and B are administered, we denote the drug concentration of Drug A as $u_1^{\mathrm{AB}}(t;\boldsymbol{x})$ and the concentration of Drug B as $u_{21}^{\mathrm{AB}}(t;\boldsymbol{x})$. For this example, we consider the case where the blood sample is taken at $t = 0.5, 1, 1.5, 2, 3, 5, 8, 12$, and for the case where both Drugs A and B are administered, we measure the drug concentration of both drugs. The nonlinear function $\boldsymbol{f}$

	Model structure	Parameters to be Estimated	Simulated Observations
Model 1	Three systems of nonlinear ODEs with 20 variables	11 parameters	30 observations
Model 2	Two systems of linear ODEs with 20 variables	11 parameters	22 observations
Model 3	Two systems of nonlinear ODEs with 20 and 33 variables	19 parameters	24 observations

Table 2: Summary of the mathematical description of the PBPK models used for the numerical experiments.

can be written as

$$\begin{aligned}\boldsymbol{f}(\boldsymbol{x}) \quad = \quad & [u_1^{\mathrm{A}}(0.5;\boldsymbol{x}), u_1^{\mathrm{A}}(1;\boldsymbol{x}), u_1^{\mathrm{A}}(1.5;\boldsymbol{x}), u_1^{\mathrm{A}}(2;\boldsymbol{x}), u_1^{\mathrm{A}}(3;\boldsymbol{x}), u_1^{\mathrm{A}}(5;\boldsymbol{x}), u_1^{\mathrm{A}}(8;\boldsymbol{x}),\\ & u_1^{\mathrm{A}}(12;\boldsymbol{x}), u_1^{\mathrm{AB}}(0.5;\boldsymbol{x}), u_1^{\mathrm{AB}}(1;\boldsymbol{x}), u_1^{\mathrm{AB}}(1.5;\boldsymbol{x}), u_1^{\mathrm{AB}}(2;\boldsymbol{x}), u_1^{\mathrm{AB}}(3;\boldsymbol{x}), u_1^{\mathrm{AB}}(5;\boldsymbol{x}),\\ & u_1^{\mathrm{AB}}(8;\boldsymbol{x}), u_1^{\mathrm{AB}}(12;\boldsymbol{x}), u_{21}^{\mathrm{AB}}(0.5;\boldsymbol{x}), u_{21}^{\mathrm{AB}}(1;\boldsymbol{x}), u_{21}^{\mathrm{AB}}(1.5;\boldsymbol{x}), u_{21}^{\mathrm{AB}}(2;\boldsymbol{x}), u_{21}^{\mathrm{AB}}(3;\boldsymbol{x}),\\ & u_{21}^{\mathrm{AB}}(5;\boldsymbol{x}), u_{21}^{\mathrm{AB}}(8;\boldsymbol{x}), u_{21}^{\mathrm{AB}}(12;\boldsymbol{x})]^{\mathrm{T}}\end{aligned}$$

This example is based on [39].

The mathematical description of each PBPK model used for the numerical experiments is summarised in Table 1.

3.1.2 Computation environment

All computational experiments were performed using Matlab 2019a on 3.1 GHz Intel Core i5 processors with MacOS version 10.13.6. When using the algorithms implemented in Python we used Python version 3.7. All results of the numerical experiments were summarised and visualised using ggplot2 version 2.2.1 [35] in R version 3.3.2.

3.1.3 The initial set of vectors $\{\boldsymbol{x}_i^{(0)}\}_{i=1}^N$

We generated the initial set of vectors $\{\boldsymbol{x}_i^{(0)}\}_{i=1}^N$ randomly based on Algorithm 1. The range of the initial cluster was set by the pharmacologically likely parameter range based on a priori knowledge defined by domain specialists (e.g., from the values obtained from the animal experiments or lab experiments).

We stored and used this initial set of vectors as the initial cluster for the CGN and also as initial iterates for the other nonlinear least squares solvers and optimisation algorithms which were compared.

Notice that this range of the initial cluster does not necessary contain the global minimiser as many pharmacokinetics parameter can be few order of magnitudes different between species. For the sake of comparison with the constrained optimisation algorithms, we constructed Example 2 so that the initial range contains the set of parameters that we used to create the dataset. This ensures that a global minimiser is in the domain where the initial cluster is made. When comparing with the constrained optimisation algorithms, we used this domain as the bound constraint for the parameter search for these algorithms.

3.1.4 ODE solver

For Examples 1 and 3 the nonlinear function evaluations require the numerical solution of stiff systems of ODEs. We used the ODE15s solver [30] with the default settings to solve these ODEs (relative tolerance 10^{-3}, absolute tolerance 10^{-6}). We observed that for some set of parameters, the ODE solver can get stuck in an infinite loop. Here, we set the timeout, where if the ODE evaluation takes longer than 5 seconds, it terminates the solution process and returns a not-a-number vector.

3.1.5 Setting for the Cluster Gauss-Newton (CGN) method:

We used the following parameters unless stated otherwise:d

$$
\begin{aligned}
N &= 250 & (60)\\
\lambda_{\text{init}} &= 0.01 & (61)\\
\lambda_{\max} &= 10^{10} & (62)\\
k_{\max} &= 100. & (63)
\end{aligned}
$$

Here $k_{\max}$ is the maximum number of iterations of the method. We used the initial set of vectors $\{\boldsymbol{x}_i^{(0)}\}_{i=1}^N$ described in Section 3.1.3 as the initial cluster. Note that we chose λ_{init} to be consistent with the default setting of the LM implementation (*lsqnonlin* in Matlab.

3.2 Algorithms Compared

Here, we list the conventional and recently developed algorithms that we compared our proposed algorithm with. We used the default setting of the algorithms unless stated otherwise.

3.2.1 Nonlinear least squares solvers

We compare the proposed algorithm against gradient-based and gradient-free nonlinear least squares algorithms. We used these algorithms to find multiple approximate minimisers of the nonlinear least square problem of our interest by repeatedly applying these algorithms with various initial iterates. Namely; we used each parameter vector in $\{\boldsymbol{x}_i^{(0)}\}_{i=1}^N$, which were randomly generated in the CGN method as the initial iterate, and excecuted each algorithm repeatedly to obtain N sets of estimated parameter vectors.

Levenberg-Marquardt (LM) method: The conventional gradient-based nonlinear least squares solver, Levenberg-Marquardt method [4] implemented in the *lsqcurvefit* function in Matlab.

We use LM with the default setting as well as setting 'FiniteDifferenceStepSize' to be the square root of the default accuracy of the ODE solve (e.g.,
FiniteDifferenceStepSize = $\sqrt{\text{AbsTol}} = \sqrt{10^{-6}}$). Note that the
FiniteDifferenceStepSize of the default setting is 10^{-6}.

Trust-Region (TR) method: The conventional gradient-based nonlinear least squares solver, Trust-Region method [9] implemented in the *lsqcurvefit* function in Matlab.

DFO-LS method: A gradient-free nonlinear least squares solver DFO-LS
method[7]. We obtained the Python code of DFO-LS Version 1.0.2 from `http://people.maths.ox.ac.uk/robertsl/dfols/` (last accessed on June 6th 2019). To make the numerical method for solving the model ODEs exactly the same, we called the Matlab implementation of the nonlinear function through MATLAB Engine API for Python.

libensemble with POUNDERS: We used the libensemble algorithm [19] that was available at the developer branch of `https://github.com/Libensemble/libensemble` (last accessed April 25th 2019) together with petsc version 3.10.4 available as `https://www.mcs.anl.gov/petsc/index.html`. We used POUNDERS as the nonlinear least squares solver. As libensemble was implemented in Python, we used the MATLAB Engine API to call the nonlinear function. POUNDERS algorithm utilises bound constraints for the search domain, and we tested this algorithm only for Example 2.

3.2.2 Optimisation algorithms

We compared the proposed method against the optimisation algorithms that are designed to minimise a function which takes a vector quantity as an input and scalar quantity as output. We applied these algorithms to our nonlinear least squares problem by minimising SSR defined in Equation (3)

Quasi-Newton method: We used the Quasi-Newton method implemented in Matlab's Optimisation toolbox as the *fminunc* function. This implementation uses the BFGS formula to update the approximate Hessian.

The following algorithms require bound constraints, and they were tested only for Example 2.

Implicit Filter method: This is another gradient-free optimisation algorithm. The code was downloaded from `https://archive.siam.org/books/se23/` (last accessed June 7th 2019). We used the 'least-squares' option with a budget of 200.

Surrogate optimisation method: We used the Surrogate optimisation algorithm implemented in Matlab's Global Optimisation Toolbox as the *surrogateopt* function. We repeated this algorithm 250 times with distinct random-seed to obtain 250 global minimisers of the nonlinear-least squares problem of our interest.

Particle Swarm We used the Particle Swarm optimisation algorithm implemented in Matlab's Global Optimisation Toolbox as the *particleswarm* function. We repeated this algorithm 250 times with distinct random seeds to obtain 250 global minimisers of the nonlinear least squares problem of our interest.

Genetic Algorithm: We used the Genetic algorithm implemented in Matlab's Global Optimisation Toolbox as the *ga* function. As one run of this algorithm required over 100,000 function evaluations for our examples, we did not repeat this algorithm to obtain multiple global minimisers.

DIRECT: A sampling algorithm DIRECT [20]. The Matlab implementation was obtained from `https://searchcode.com/codesearch/view/12449743/` (last accessed July 30th 2019). As one run of this algorithm required over 100,000 function evaluations for our examples, we did not repeat this algorithm to obtain multiple global minimisers.

3.3 Results

We compared the proposed CGN method with various existing algorithms. We compared the 'speed' of the algorithm by the number of function evaluations required and we compared the 'quality' of the minimisers found by the SSR (smaller the SSR the better minimiser). As we can see in Table 3, the dominant part of the computation time was spent by the nonlinear function evaluations in CGN. Hence, we claim that the number of function evaluations is a fair way to compare the computation cost.

	Example 1	Example 2	Example 3
Nonlinear function evaluation	1,890.598	852.116	1,779.362
Computation of linear approximation	3.610	6.929	11.021
Output results as csv files	3.647	3.663	4.129
Total computation	1,898.393	863.172	1,795.192

Table 3: Ingredients of the computation time for CGN (seconds).

We compared the speed for finding acceptable minimisers. We define the acceptable minimisers as the minimisers with SSR less than the SSR of the parameter that was used to generate the test dataset. In Figure 3, we show the number of acceptable minimisers found given the total number of function evaluations. As can be seen, CGN is significantly faster than any other method for finding acceptable minimisers.

As this analysis depends on how we define acceptable minimisers, we plotted the number of minimisers found for a given SSR threshold in Figure 4. For this analysis each algorithm was run until its default stopping criterion (given in the reference of each method) was met. We summarised the number of function evaluations that each algorithm required to meet its default stopping criterion in Table 4.

As can be seen in Figure 4, for Examples 1 and 3, CGN finds more approximate minimsers with small SSR than all the conventional algorithms. For Example 2, as we know that at least one global minimiser is enclosed in the domain defined by $\boldsymbol{x}^{\mathrm{U}}$ and $\boldsymbol{x}^{\mathrm{L}}$, we were able to apply optimisation algorithms with bound constraints (libsensemble with POUNDERS, implicit filter, surrogate optimisation, and particle swarm). For this problem, particle swarm obtained slightly more approximate minimisers with small SSR than CGN. However, as can be seen in Table 4, particle swarm required significantly more function evaluations than the CGN (particle swarm: 3,646,900 v.s. CGN: 6,768). Aside from the particle swarm, CGN outperformed all the other methods in the number of acceptable minimisers found.

In addition to above rigorous comparison, we applied two global optimisation algorithms, Genetic algorithm (GA) and DIRECT to Example 2. GA and DIRECT required 120,600 and 100,063 nonlinear function evaluations, respectively, to obtain one solution each. The SSR of the solutions found by GA and DIRECTwere 0.0197 and 0.244, respectively, while the minimum SSR found by the CGN is 0.0188. Hence, even to find just one solution, CGN is faster and more accurate than these two methods.

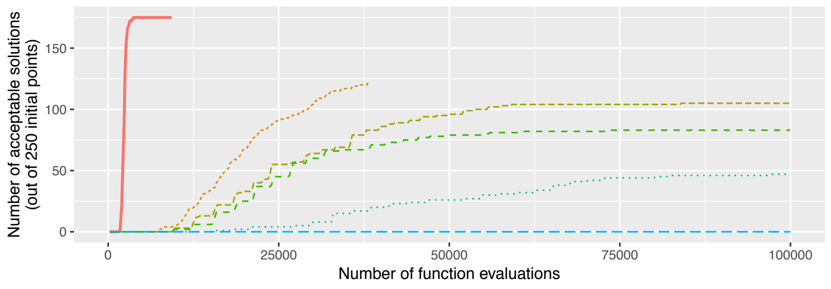
Algorithm	Example 1	Example 2	Example 3
CGN method	7,782	6,768	8,452
DFO LS method	57,577	36,804	70,854
LM method	72,400	43,359	153,184
libensemble POUNDERS	–	19,068	–
LM Def method	49,822	43,359	81,710
Imp Filter method	–	51,559	–
Trues-Region method	94,308	62,136	103,260
Quasi-Newton method	124,762	63,316	151,774
Surrogate Opt method	–	137,500	–
Particle Swarm method	–	3,646,900	–

Table 4: Number of total function evaluations.

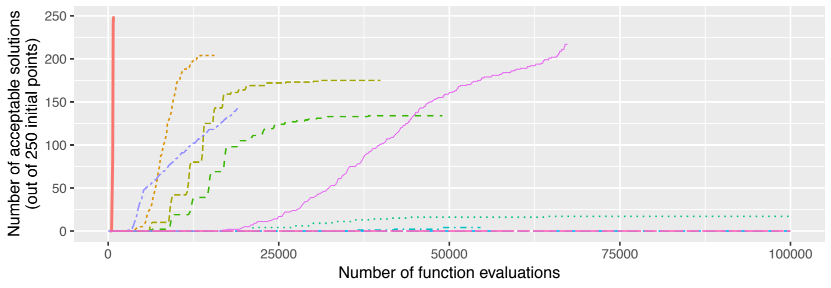
4 Concluding Remarks

We proposed the Cluster Gauss-Newton (CGN)
method, a new derivative free method specifically designed for finding multiple approximate minimisers of a nonlinear least squares problem. The development of this algorithm was motivated by the parameter estimation of physiologically based pharmacokinetic (PBPK) models that appears in pharmaceutical drug development. The particular nature of the model, where the model is over parameterised, and consideration of multiple possible parameters is necessary, motivated us to develop the new method. The fact that our algorithm obtains multiple approximate minimisers collectively makes it significantly more computationally efficient compared to existing nonlinear least squares solvers or optimisation algorithms. In addition, we observed that in general, CGN obtains minimisers with smaller sum of squared residuals (SSR) than existing algorithms. We demonstrated these advantages using three examples that come from real world drug development projects.

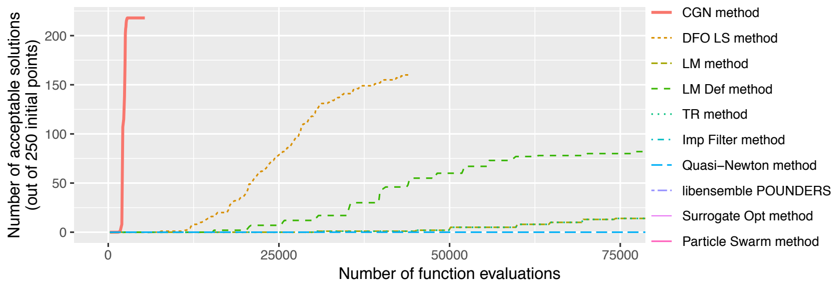
By minimising the assumption on the nonlinear function, where it can be a “black box”, we have ensured the ease of use and implementation for those who may not have a substantial background in mathematics or scientific computing. We believe this advantage of the proposed method will be appreciated by potential users of the algorithm in industry. In this paper, we used the pharmacokinetics



(a) Example 1

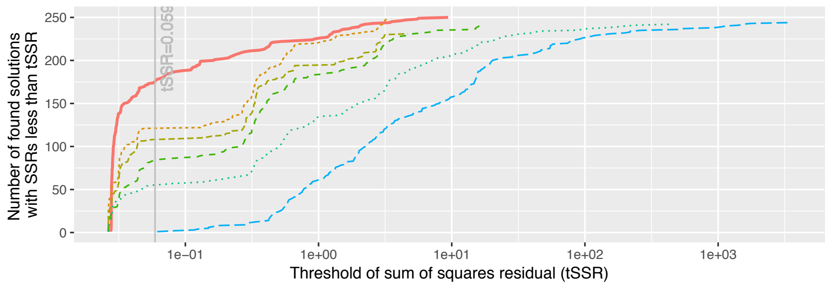


(b) Example 2

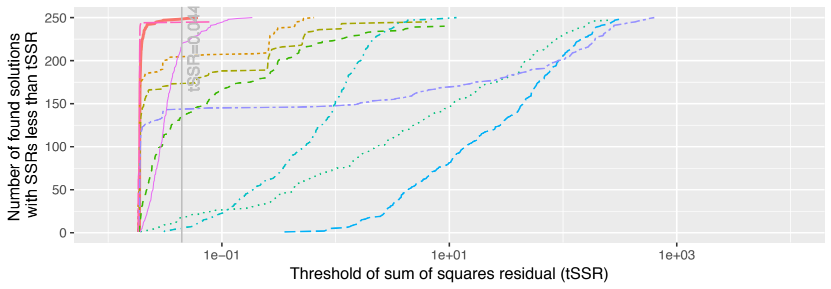


(c) Example 3

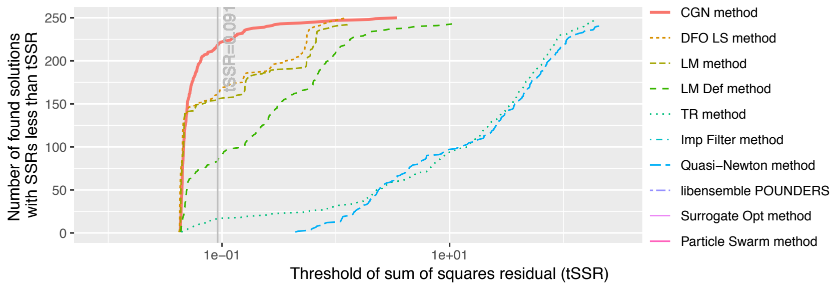
Figure 3: Number of acceptable minimisers (out of 250) found by each method for given total number of function evaluations. Acceptable minimisers are defined by the minimisers with SSR less than the SSR of the parameter that was used to generate the test dataset. (Example1: SSR < 0.0592, Example2: SSR < 0.0444, Example3: SSR < 0.0915)



(a) Example 1



(b) Example 2



(c) Example 3

Figure 4: Number of solutions (out of 250) found by each method for given accuracy threshold (tSSR). Smaller SSR indicates more accurate solution to the nonlinear-least squares problem.

models as examples. However, as we do not assume any particular form of the nonlinear function, we believe the proposed method can be used for many other mathematical models in various scientific fields.

Software for the CGN method proposed in this paper is available at `https://sourceforge.net/p/cluster-gauss-newton-method/code/`. GUI software is available at `http://www.bluetree.me/CGNmethod_for_PBPKmodels`. Matlab code is available at `https://www.mathworks.com/matlabcentral/fileexchange/68798`.

5 Acknowledgement

We would like to thank Professor Akihiko Konagaya for giving us the opportunity to initiate this research. We would like to thank Dr. Jeffrey Larson for his help on numerical experiment using libensemble. We would like to thank Professor Hiroshi Yabe and Dr. Ryota Kobayashi for useful discussions. We would also like to thank the referees for valuable advice which helped to greatly improve the paper. Ken Hayami was supported in part by JSPS KAKENHI Grant Number 15K04768. Yasunori Aoki is currently employed by AstraZeneca.

A Motivating Example

In this appendix, we introduce a motivating example to illustrate how the proposed method could be used in practice. We consider a scenario where a newly developed drug is tested for the first time in a human. Before the drug is given to a human, the biochemical properties of the drug are studied in-test-tube (in-vitro) and in-animal experiments. However, how the drug behaves in the human body is still uncertain. Based on the results of in-vitro and in-animal experiments, the team decides that 100mg is a safe amount of the drug to be given to a human and the experiment is conducted with a healthy normal volunteer, and the drug concentration in the blood plasma is measured at points in time. Using these measurements, we estimate the multiple possible model parameters which can be used to simulate various scenarios. The following workflow can be envisioned:

1: Construct a mathematical model based on the understanding of the physiology and biochemical properties of the drug.
In this example, we use the model presented in [34]. The mathematical model is depicted in Figure 5, and it can be written as a system of nonlinear ordinary differential equations with 20 variables. There are two types of model parameters in this model: physiological parameters and kinetic parameters. Examples of the physiological parameters are the sizes of the organs or the blood flow rates between the organs. As the human physiology is well studied and these parameters usually do not depend on the drugs, we can assume these parameters to be known. The kinetic parameters, such as, how fast the drug gets excreted from the body or how easily it binds to tissues are the parameters that depend on the drug and usually are not very well known. Before the first in-human experiment, the drug development team characterises these parameters using an organ in test-tubes or by administering the drug to an animal. However, these parameters can differ from animal to human, so we do not have a very accurate estimate of these parameters. The differences in these parameters between a human and an animal can be several orders of magnitude. Figure 6 depicts plots of the drug concentration simulation where the kinetic parameters are sampled within a reasonable range of the parameters. As can be seen in Figure 6, we cannot obtain any useful information just by randomly sampling the kinetic parameters from the feasible range.

2: Sample multiple possible model parameter sets that fit the model prediction of the drug concentration to the observed data from the 100mg experiment. We now use the observed data from the experiment where 100mg of the drug was given to a human. The red dots in Figure 7 depict the observed data. The left panel of Figure 7 shows some of the simulation results using the parameter sets of the initial iterate of CGN. The right panel of Figure 7 shows some of the

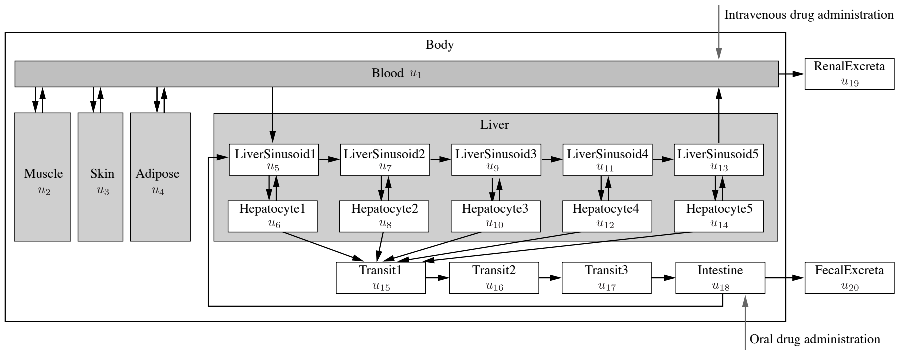


Figure 5: A schematic diagram of a physiologically based pharmacokinetic model. (Arrows represent the movement of the drug to a different part of the body. Variables u_i are the drug concentration or the amount of the drug in each compartment. The body is divided into Blood, Muscle, Skin, Adipose, Liver and Intestine. The liver is further divided into ten compartments to model the complex drug behaviour in the liver. The intestine is divided into four compartments, three transit compartments and one intestine compartment, to model the time it takes for the drug to reach the intestine.)

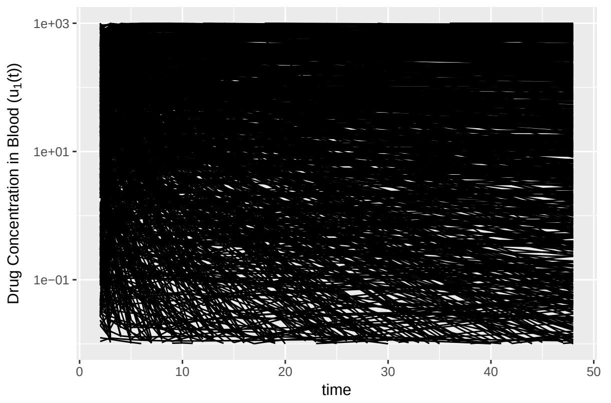


Figure 6: Simulation of the drug concentration in the blood plasma using the parameters that are naively sampled from the range of possible kinetic parameters. Note that these simulations do not give any useful information.

simulation results after 20 iterations of CGN. As can be seen in Figure 7, CGN can find multiple sets of parameters that fit the observed data. The parameter values are depicted in the box plots in Figure 8. As can be seen in Figure 8, after 20 iterations of CGN, the distribution of some of the parameters shrinks significantly suggesting these parameters can be identified from the observations while the distribution of some of the parameters are unchanged indicating that these parameters cannot be identified from the observation. In Figure 9, we show scatter plots of the parameters found by CGN. As can be seen in Figure 9, even if the parameter cannot be identified from the observation, some nonlinear relationships can occasionally be identified between the parameters.

B Relation with previous work

We initially proposed the Cluster Newton (CN) method for sampling multiple solutions of a system of nonlinear equations where the number of unknowns is more than the number of observations [1, 2, 12].

The parameter fitting problems presented in [38, 33, 23] uses CN to fit complex PBPK models to

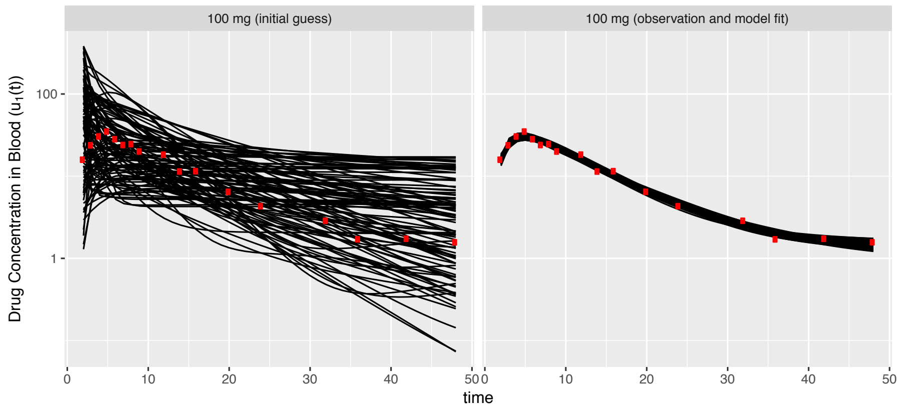


Figure 7: Plot of the simulation of drug concentration (black solid line) with observations (red dot). Simulation results are based on the parameters for the initial iterate for CGN and the parameters found after 20 iterations of CGN. In the left panel, the simulation results based on the top 100 sets of the parameters (out of 1000 parameter sets in the cluster) from the initial cluster are shown. In the right panel, the simulations results based on the top 100 sets of the parameters (out of 1000 parameter sets in the cluster) after 20 iterations of CGN are shown. Hence, top 100 means the 100 smallest sum of squared residuals (SSR) between the simulation and observed data.

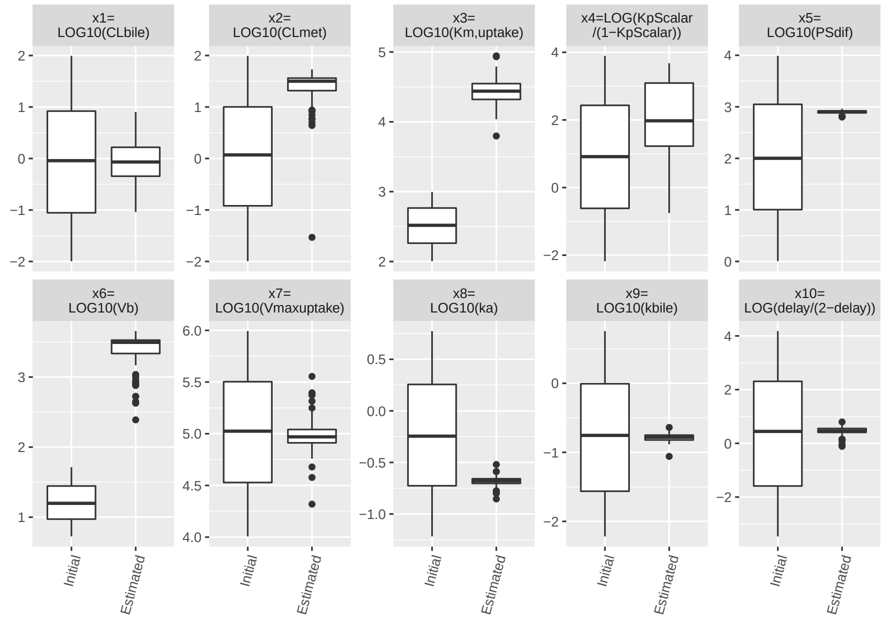


Figure 8: Box plots of top 100 parameters (the same parameter sets as the one used to plot Figure 7) from the initial cluster (left) and the cluster after 20 iterations of CGN (right). Note the distributions of $x5, x8, x9, x10$ clearly shrunk after 20 iterations while the distribution of $x4$ did not change noticeably. (Box plot: The edges of the boxes are the 75th and 25th percentiles. The line in the box is the median, and the whiskers extend to the largest and the smallest value within the 1.5 times the inter-quartile range. Dots are the outliers outside the whiskers.)

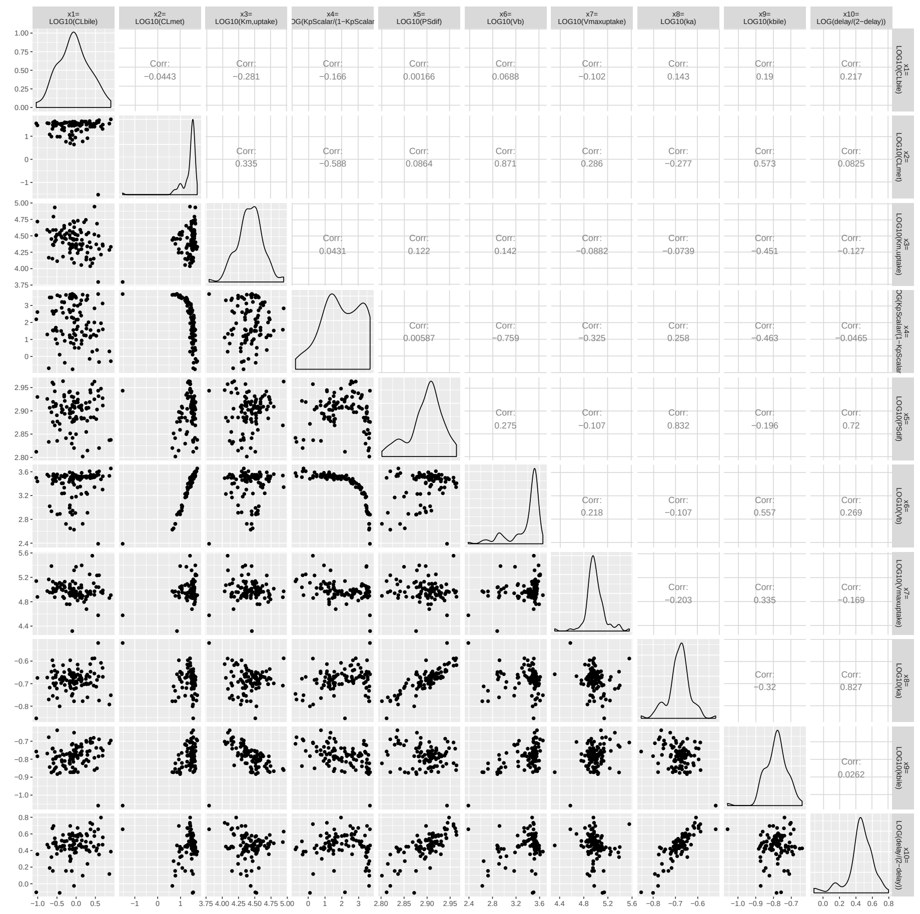


Figure 9: Scatter plots of parameters. As can be seen in this example, we can find parameter-parameter correlations of some of the parameters found by CGN. (parameters whose corresponding SSR is the least 100 within the cluster of 1,000, i.e., the same parameter sets as the one used to plot the right hand side of Figure 7.)

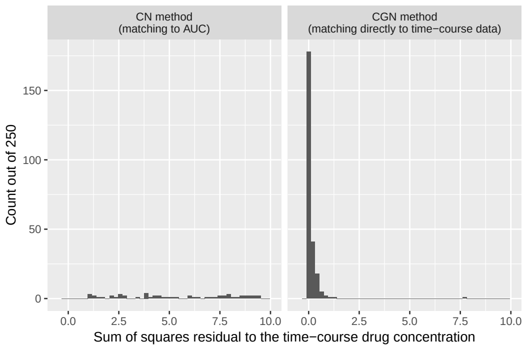


Figure 10: Comparison of the accuracy of the model fit of the Motivating Example (cf. Appendix A) using the approach employed in [38, 33] using the Cluster Newton (CN) method and the proposed Cluster Gauss-Newton (CGN) method.

the drug (and its metabolite) concentration measurements over time. As the drug concentration was measured repeatedly from the patient, there is a larger number of observations than the number of unknown parameters. CN assumes an underdetermined problem, where the number of observations is less than the number of parameters. To use CN, in [38, 33, 23], the authors constructed a summary value called Area Under the time-concentration Curve (AUC) to reduce the number of observations and fit the model to the summary value using CN. The AUC is essentially the time integral of the drug concentration from the time drug is administered to infinity. After finding multiple possible parameters that match the AUC of the model and the observations, Yoshida et al. and Toshimoto et al. [38, 33] selected the parameter sets that fit the time-course drug concentrations reasonably well from these parameter sets found using CN. Kim et al. [23] used a parameter set obtained using CN as the initial iterate for the Levenberg-Marquardt method and fitted the model to the time-course drug concentration data.

Based on the current use of CN, we identified two bottlenecks of the workflow employed in [38, 33, 23]. Firstly, CN solves a system of underdetermined system of nonlinear **equations**. Hence, the model needs to be constructed in a way that the observation and the model prediction match **exactly**. Hence, if there is a model misspecification or significant measurement error that influence the summary values (e.g. AUC) sufficiently so that there is no model prediction that exactly matches the observation derived summary value, then CN breaks down.

Secondly, in order to obtain the parameter sets that reasonably fit the original data (e.g., time-course concentration data), we need to obtain many parameter sets that fit the summary values (e.g., AUC). This is simply because we need to randomly sample from (number of parameters)-(number of summary values) dimension space to obtain the desired parameter sets. As a result, [33] had to find 500 000 parameter sets that fit the summary value (AUC) using CN and then was able to obtain 30 parameter sets that reasonably fit the original data (time-course drug concentration).

To overcome these bottlenecks, we suggested these authors to formulate these parameter estimation problems as nonlinear least squares problem and we created the Cluster Gauss-Newton method (CGN). The new method efficiently obtains multiple possible parameters by solving a nonlinear least squares problem so that the method does not break down even if the model does not exactly match the observation. Also, the new algorithm does not require the number of observations to be less than the number of parameters so that there is no need to summarise the observation and can directly use the original observations (e.g., time-course concentration measurements). As can be seen in Figure 10, the new CGN algorithm enables us to sample multiple possible parameters significantly more accurately than CN.

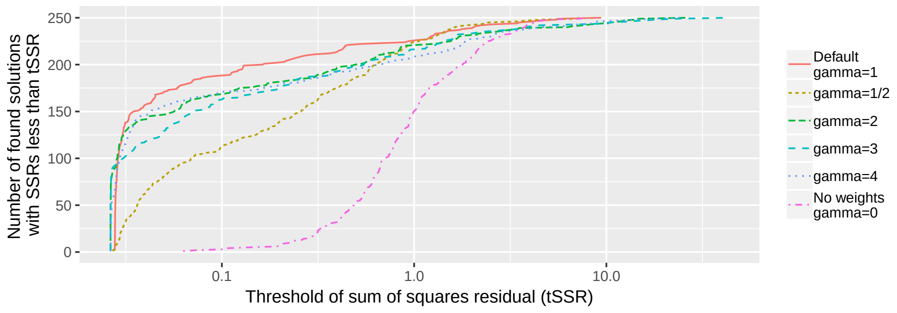


Figure 11: Number of solutions found for given accuracy threshold (tSSR) using various weights for the linear approximation (i.e., various γ). Smaller SSR indicates more accurate solution to the nonlinear-least squares problem.

C Numerical experiments on the influence of the weight of the linear approximation

To illustrate the influence of the weight $d^{(k)}_{j(i)}$ of the weighted linear approximation in 2-1) of the algorithm (equations (25) and (32)), we conducted numerical experiments using Example 1 of section 3. For this numerical experiment, we varied the parameter $\gamma \geq 0$ in Equation (32):

$$d^{(k)}_{j(i)} = \begin{cases} \left(\frac{1}{\sum_{l=1}^{n}((x^{(k)}_{lj}-x^{(k)}_{li})/(x^{\mathrm{U}}_{l}-x^{\mathrm{L}}_{l}))^2}\right)^{\gamma} & \text{if } j \neq i \\ 0 & \text{if } j = i \end{cases}. \tag{64}$$

In Figure 11, we show the number of solutions found by CGN for given accuracy threshold. As can be seen from the figure, the weight for the linear approximation improves the accuracy and the speed of CGN. Note that $\gamma = 0$, which corresponds to giving equal weights to all the cluster points is not optimal. In this example, $\gamma \geq 1$ gave good results.

D ODE expressions of the PBPK model used as the motivating example

In this subsection we explicitly write out the mathematical model used in the motivating example in Appendix A as a system of ODEs. The other PBPK models used in this paper can be similarly written as a system of ODEs. For the model parameters appearing in these ODEs, we follow the usual notations used in pharmacokinetics and we will provide brief descriptions of these parameters. For more detailed description of these parameters, we refer the reader to the standard textbooks in pharmacokinetics (e.g., [29]).

ODE for the blood compartment

$$\begin{aligned}\frac{\mathrm{d}u_1}{\mathrm{d}t} &= \frac{Q_h(u_{13}-u_1) - CL_r u_1 - Q_m(u_1 - \frac{u_2}{Kp_m Kp_{scalar}}) - Q_s(u_1 - \frac{u_3}{Kp_s Kp_{scalar}})}{V_b} \\ &\quad - \frac{Q_a(u_1 - \frac{u_4}{Kp_a Kp_{scalar}})}{V_b}\end{aligned}$$

Q_h is the blood flow rate through the liver. The unit is typically L/hr.
CL_r is the rate of the drug being excreted as urine. The unit is typically L/hr.

V_b is the volume of the blood compartment. The unit is typically L.

Q_m, Q_s, Q_a are the blood flow rates to muscle, skin, and adipose, respectively. The unit is typically L/hr.

V_m, V_s, V_a are the volumes of muscle, skin, and adipose compartments, respectively. The unit is typically L.

Kp_m, Kp_s, Kp_a are the partition coefficients of the drug for the muscle, skin and adipose, respectively. The partition coefficient is the ratio between the drug concentration in the blood plasma and the tissue at equilibrium when the tissue is in contact with the blood. We usually assume that these values can be measured by *in-vitro* experiments. Hence, they are usually fixed parameters in the model.

Kp_{scalar} is a scaling factor for the partitioning.

ODE for the muscle compartment

$$\frac{\mathrm{d}u_2}{\mathrm{d}t} = \frac{Q_m}{V_m}\left(u_1 - \frac{1}{Kp_m Kp_{\text{scalar}}}u_2\right) \tag{65}$$

ODE for the skin compartment

$$\frac{\mathrm{d}u_3}{\mathrm{d}t} = \frac{Q_s}{V_s}\left(u_1 - \frac{1}{Kp_s Kp_{\text{scalar}}}u_3\right) \tag{66}$$

ODE for the adipose (fat) compartment

$$\frac{\mathrm{d}u_4}{\mathrm{d}t} = \frac{Q_a}{V_a}\left(u_1 - \frac{1}{Kp_a Kp_{\text{scalar}}}u_4\right) \tag{67}$$

ODEs for the liver compartments

For a compartment representing the blood vessel in the liver (Liver Sinusoid 1)

$$\frac{\mathrm{d}u_5}{\mathrm{d}t} = -\frac{\frac{\text{Vmax}_{\text{uptake}}}{Km_{\text{uptake}}+u_5} + f_b\,PS_{\text{dif}}}{V_{hc}}u_5 + \frac{f_h\,PS_{\text{dif}}}{V_{hc}}u_6 + \frac{Q_h(u_1 - u_5) + ka\,u_{18}}{\frac{V_{hc}}{5}} \tag{68}$$

ka is the rate of the drug being absorbed from the intestines and transported to the liver through the portal vein. The unit is typically L/hr.

V_{hc} is the volume of the blood in the blood vessel.

PS_{diff} is the diffusion constant between the liver sinusoid compartment and the hepatocyte compartment.

f_b is the ratio of the drug that is not bound to the protein in blood in the blood vessel in the liver.

f_h is the ratio of the drug that is not bound to the protein in the liver cells. Only the portion of the drug that is not bound to the protein can permeate in and out of the liver cells.

For compartments representing the blood vessel in the liver (Liver Sinusoid 2, 3, 4, 5)

$$\frac{\mathrm{d}u_{7+2i}}{\mathrm{d}t} = -\frac{\frac{\text{Vmax}_{\text{uptake}}}{Km_{\text{uptake}}+u_{7+2i}} + f_b\,PS_{\text{dif}}}{V_{hc}}u_{7+2i} + \frac{f_h\,PS_{\text{dif}}}{V_{hc}}u_{8+2i} + \frac{Q_h(u_{5+2i} - u_{7+2i})}{\frac{V_{hc}}{5}}$$

for $i = 0, 1, 2, 3$.

$\text{Vmax}_{\text{uptake}}$ and Km_{uptake} are constants for the Michaelis-Menten kinetics where the active uptake of the drug from the blood to the liver cells by membrane transport protein (also known as "transporter") is modelled.

Compartments representing liver cells (Hepatocyte 1, 2, 3, 4, 5)

$$\frac{\mathrm{d}u_{6+2i}}{\mathrm{d}t} = \frac{\frac{\mathrm{Vmax}_{\mathrm{uptake}}}{Km_{\mathrm{uptake}}+u_{5+2i}} + f_b\, PS_{\mathrm{dif}}}{V_{he}} u_{5+2i} - \frac{f_h(PS_{\mathrm{dif}} + (\mathrm{CL}_{\mathrm{met}} + \mathrm{CL}_{\mathrm{bile}}))}{V_{he}} u_{6+2i}$$

for $i = 0, 1, 2, 3, 4$.

V_{he} is the volume of liver cells.

ODE for the transit compartment 1

$$\frac{\mathrm{d}u_{15}}{\mathrm{d}t} = f_h\, \mathrm{CL}_{\mathrm{bile}} \frac{u_6 + u_8 + u_{10} + u_{12} + u_{14}}{5} - k_{\mathrm{bile}}\, u_{15} \tag{69}$$

u_{15} is the amount of the drug in the transit compartment 1. The transit compartment can be thought of as a fictitious compartment that introduces the delay of the drug delivery to to the intestines.

$\mathrm{CL}_{\mathrm{met}}$ is a clearance (speed that drug is cleared out) due to the metabolisation of the drug of the subject. The unit is typically L/hr.

$\mathrm{CL}_{\mathrm{bile}}$ is a clearance due to the biliary excretion of the drug of the subject. The unit is typically L/hr.

k_{bile} is a diffusion coefficient between the transit compartments. The unit is typically 1/hr.

f_h is the fraction of the drug cleared by liver, it is typically unit-less.

ODE for the transit compartment 2

$$\frac{\mathrm{d}u_{16}}{\mathrm{d}t} = k_{\mathrm{bile}}\, (u_{15} - u_{16}) \tag{70}$$

u_{16} is the amount of the drug in the transit compartment 2.

k_{bile} is a diffusion constant where the drug is diffused into bile, the unit is typically 1/hr.

ODE for the transit compartment 3

$$\frac{\mathrm{d}u_{17}}{\mathrm{d}t} = k_{\mathrm{bile}}\, (u_{16} - u_{17}) \tag{71}$$

u_{17} is the amount of the drug in the transit compartment 3.

ODE for the intestine compartment

$$\frac{\mathrm{d}u_{18}}{\mathrm{d}t} = k_{\mathrm{bile}}\, u_{17} - \frac{ka}{FaFg} u_{18} \tag{72}$$

u_{18} is the amount of the drug in the intestine.

$FaFg$ is the fraction of the drug that gets absorbed by the intestines.

See also Appendix E

E Re-parameterisation

For the mathematical model we introduced in Appendix D, some of the parameters may be known for the drug of interest. For the motivating example presented in Appendix A, we assumed the following parameters to be known:

$$CL_r = 0, \quad FaFg = 0.55, \quad Kp_a = 0.086, \quad Kp_m = 0.113 \tag{73}$$

$$Kp_s = 0.478, \quad Q_a = 15.61, \quad Q_h = 86.94, \quad Q_m = 44.94 \tag{74}$$

$$Q_s = 17.99, \quad V_a = 10.01, \quad V_{hc} = 1.218, \quad V_{he} = 0.469 \tag{75}$$

$$V_m = 30.03, \quad V_s = 7.77, \quad f_b = 0.00617, \quad f_h = 0.012 \tag{76}$$

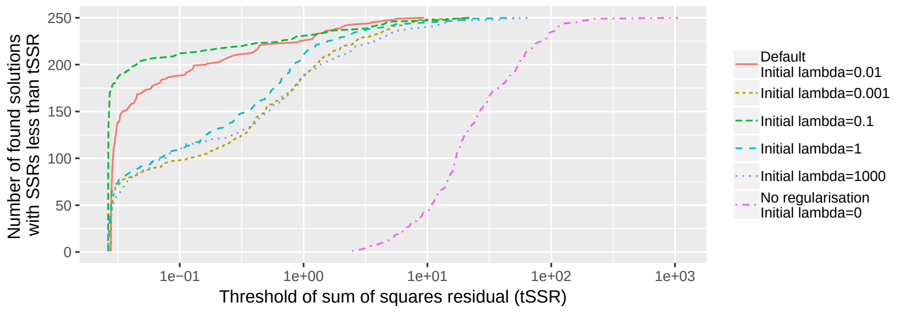


Figure 12: Number of solutions found for given accuracy threshold (tSSR) using various initial lambda (λ_{init}). Smaller SSR indicates more accurate solution to the nonlinear-least squares problem.

Also, most of the parameters have known physiologically or chemically possible range (e.g., clearances cannot be negative). Thus, for those parameters, we re-parameterise to impose known constraints. For the motivating example presented in Appendix A, we re-parameterised as in the following:

$$\text{CL}_{\text{bile}} = 10^{x_1}, \qquad \text{CL}_{\text{met}} = 10^{x_2}, \qquad Km_{\text{uptake}} = 10^{x_3}, \tag{77}$$

$$Kp_{\text{Scalar}} = \frac{e^{x_4}}{1+e^{x4}} \qquad PSdif = 10^{x_5}, \qquad Vb = 10^{x_6}, \tag{78}$$

$$Vmax_{\text{uptake}} = 10^{x_7}, \qquad ka = 10^{x_8} \qquad k_{\text{bile}} = 10^{x_9}. \tag{79}$$

F Numerical experiments on the influence of the regularisation

To illustrate the necessity and the influence of the regularisation in 2-2) of the algorithm (equation (35)), we conducted numerical experiments using Example 1 of section 3. We varied the initial value of the regularisation coefficient λ_{init} and tested CGN. Figure 12 shows the number of solutions found by CGN for given accuracy threshold. As can be seen from the figure, regularisation is necessary for CGN to perform well. For this example, $\lambda_{\text{init}} = 0.1$ gave a good result.

G Application to discontinuous nonlinear functions

The Jacobian based local optimisation methods cannot solve nonlinear least squares problems if the nonlinear function is discontinuous. On the other hand, the Cluster Gauss-Newton (CGN) method uses the linear approximation of the nonlinear function to capture the global behaviour of the function. Hence, it does not require the nonlinear function to be continuous.

For the following numerical experiment, we consider the case where the nonlinear function is not continuous. This example is constructed to show that CGN can solve nonlinear least squares problems that the conventional Jacobian based method cannot solve. We create such a nonlinear function by rounding the nonlinear function of Example 1 ((40)-(42)) to the first decimal place.

As can be seen in Figure 13, the CGN method was able to find many accurate solutions. On the other hand, the LM failed to find any reasonable solution. Unlike derivative based LM, a derivative-free method (DFO LS method) occasionally finds reasonable solutions.

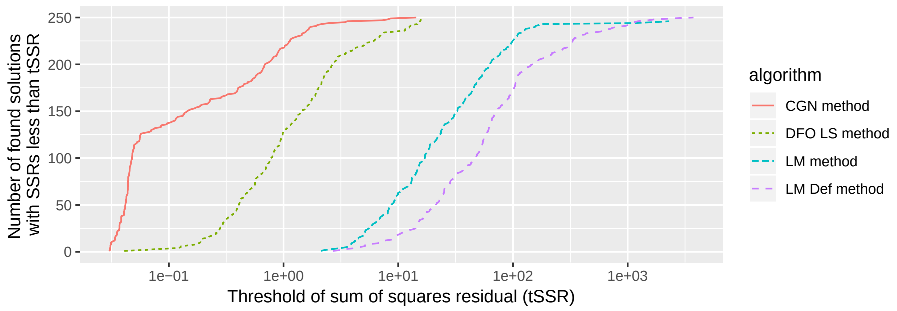


Figure 13: Example 1 when the nonlinear function is rounded to one decimal place: Number of solutions (out of 250) found by the various methods for given accuracy threshold (tSSR). Smaller SSR indicates more accurate solution to the nonlinear-least squares problem.

References

[1] Y. Aoki, K. Hayami, H. De Sterck, and A. Konagaya. Cluster Newton method for sampling multiple solutions of an underdetermined inverse problem: Parameter identification for pharmacokinetics. *NII Technical Reports*, (2):1–38, 2011.

[2] Y. Aoki, K. Hayami, H. De Sterck, and A. Konagaya. Cluster Newton method for sampling multiple solutions of underdetermined inverse problems: Application to a parameter identification problem in pharmacokinetics. *SIAM J. Sci. Comput.*, 36(1):B14–B44, 2014.

[3] S. Asami, D. Kiga, and A. Konagaya. Constraint-based perturbation analysis with Cluster Newton method: A case study of personalized parameter estimations with irinotecan whole-body physiologically based pharmacokinetic model. *BMCSyst. Biol.*, 2017.

[4] Å. Björck. *Numerical Methods for Least Squares Problems.* SIAM, Philadelphia, 1996.

[5] C. G. E. Boender, A. R. Kan, G. Timmer, and L. Stougie. A stochastic method for global optimization. *Math. Program.*, 22(1):125–140, 1982.

[6] C. G. Broyden. The convergence of a class of double-rank minimization algorithms 1. General considerations. *IMA J. Appl. Math.*, 6(1):76–90, 1970.

[7] C. Cartis, J. Fiala, B. Marteau, and L. Roberts. Improving the flexibility and robustness of model-based derivative-free optimization solvers. *ACM Trans. Math. Software*, (3):32:1–41, 2019.

[8] C. Cartis and L. Roberts. A derivative-free Gauss–Newton method. *Math. Program. Comput.*, 11(4):631–674, 2019.

[9] A. R. Conn, N. I. Gould, and P. L. Toint. *Trust Region Methods.* SIAM, Philadelphia, 2000.

[10] R. Fletcher. A new approach to variable metric algorithms. *Comput. J.*, 13(3):317–322, 1970.

[11] Y. Fukuchi, K. Toshimoto, T. Mori, K. Kakimoto, Y. Tobe, T. Sawada, R. Asaumi, T. Iwata, Y. Hashimoto, K. ichi Nunoya, H. Imawaka, S. Miyauchi, and Y. Sugiyam. Analysis of nonlinear pharmacokinetics of a highly Albumin-bound compound: Contribution of Albumin-mediated hepatic uptake mechanism. *J. Pharm. Sci.*, 2017.

[12] P. Gaudreau, K. Hayami, Y. Aoki, H. Safouhi, and A. Konagaya. Improvements to the cluster Newton method for underdetermined inverse problems. *J. Comput. Appl. Math.*, 283:122–141, 2015.

[13] C. F. Gauss. *Theory of the Motion of the Heavenly Bodies Moving About the Sun in Conic Sections: A Translation of Gauss's" Theoria Motus." With an Appendix.* Little, Brown and Company, Boston, 1857.

[14] M. Gibaldi and D. Perrier. Drugs and the pharmaceutical sciences. In *Pharmacokinetics*, volume 15, pages 445–449. Marcel Dekker New York, 1982.

[15] D. E. Goldberg and J. H. Holland. Genetic algorithms and machine learning. *Mach. Learn.*, 3(2):95–99, 1988.

[16] D. Goldfarb. A family of variable-metric methods derived by variational means. *Math. Comp.*, 24(109):23–26, 1970.

[17] H.-M. Gutmann. A radial basis function method for global optimization. *J. Global Optim.*, 19(3):201–227, 2001.

[18] P. C. Hansen. *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*, volume 4. SIAM, Philadelphia, 2005.

[19] S. Hudson, J. Larson, S. M. Wild, D. Bindel, and J.-L. Navarro. libEnsemble user manual. Technical Report Revision 0.5.1, Argonne National Laboratory, 2019.

[20] D. R. Jones, C. D. Perttunen, and B. E. Stuckman. Lipschitzian optimization without the Lipschitz constant. *J. Optim. Theory Appl.*, 79(1):157–181, 1993.

[21] C. T. Kelley. *Implicit Filtering*, volume 23 in Software Environments and Tools. SIAM, Philadelphia, 2011.

[22] J. Kennedy and R. Eberhart. Particle swarm optimization (PSO). In *Proc. IEEE International Conference on Neural Networks, Perth, Australia*, pages 1942–1948, 1995.

[23] S.-J. Kim, K. Toshimoto, Y. Yao, T. Yoshikado, and Y. Sugiyama. Quantitative analysis of complex drugdrug interactions between Repaglinide and Cyclosporin A /Gemfibrozil using physiologically based pharmacokinetic models with in vitro transporter/enzyme inhibition data. *J. Pharm. Sci.*, may 2017.

[24] J. Larson, M. Menickelly, and S. M. Wild. Derivative-free optimization methods. *Acta Numer.*, 28:287–404, 2019.

[25] D. W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(2):431–441, 1963.

[26] E. Mezura-Montes and C. A. C. Coello. Constraint-handling in nature-inspired numerical optimization: past, present and future. *Swarm and Evolutionary Computation*, 1(4):173–194, 2011.

[27] J. J. Moré. The Levenberg-Marquardt algorithm: Implementation and theory. In *Watson G.A. (eds) Numerical Analysis, Lecture Notes in Mathematics, vol. 630, Springer, Berlin, Heidelberg*, pages 105–116. 1978.

[28] T. Nakamura, K. Toshimoto, W. Lee, C. K. Imamura, Y. Tanigawara, and Y. Sugiyama. Application of PBPK modeling and virtual clinical study approaches to predict the outcomes of CYP2D6 genotype-guided dosing of tamoxifen. *CPT Pharmacometrics Syst. Pharmacol.*, jun 2018.

[29] M. Rowland, T. N. Tozer, H. Derendorf, and G. Hochhaus. *Clinical pharmacokinetics and pharmacodynamics: concepts and applications.* Wolters Kluwer Health/Lippincott William & Wilkins Philadelphia, PA, 2011.

[30] L. F. Shampine and M. W. Reichelt. The MATLAB ODE Suite. *SIAM J. Sci. Comput.*, 41(3):1–22, 1997.

[31] D. F. Shanno. Conditioning of quasi-Newton methods for function minimization. *Math. Comp.*, 24(111):647–656, 1970.

[32] D. F. Shanno and P. C. Kettler. Optimal conditioning of quasi-Newton methods. *Math. Comp.*, 24(111):657–664, 1970.

[33] K. Toshimoto, A. Tomaru, M. Hosokawa, and Y. Sugiyama. Virtual clinical studies to examine the probability distribution of the AUC at target tissues using physiologically-based pharmacokinetic modeling: Application to analyses of the effect of genetic polymorphism of enzymes and transporters on Irinotecan Ind. *Pharm. Res.*, 34(8):1584–1600, aug 2017.

[34] T. Watanabe, H. Kusuhara, K. Maeda, Y. Shitara, and Y. Sugiyama. Physiologically based pharmacokinetic modeling to predict transporter-mediated clearance and distribution of Pravastatin in humans. *J. Pharmacol. Exp. Ther.*, 2009.

[35] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York, 2016.

[36] S. M. Wild. Chapter 40, POUNDERS in TAO: Solving Derivative-Free Nonlinear Least-Squares Problems with POUNDERS. In *Advances and Trends in Optimization with Engineering Applications*, pages 529–539. SIAM, Philadelphia, 2017.

[37] Y. Yao, K. Toshimoto, S.-J. Kim, T. Yoshikado, and Y. Sugiyama. Quantitative analysis of complex drug-drug interactions between Cerivastatin and metabolism/transport inhibitors using physiologically based pharmacokinetic modeling. *Drug Metab. Dispos.*, 46(7):924–933, 2018.

[38] K. Yoshida, K. Maeda, H. Kusuhara, and A. Konagaya. Estimation of feasible solution space using Cluster Newton Method: application to pharmacokinetic analysis of irinotecan with physiologically-based pharmacokinetic models. *BMC systems biology*, 7 Suppl 3(Suppl 3):S3, 2013.

[39] T. Yoshikado, K. Yoshida, N. Kotani, T. Nakada, R. Asaumi, K. Toshimoto, K. Maeda, H. Kusuhara, and Y. Sugiyama. Quantitative analyses of hepatic OATP-mediated interactions between statins and inhibitors using PBPK modeling with a parameter optimization method. *Clin. Pharmacol. Ther.*, 100(5):513–523, 2016.