MARINO Simeone

PhD, Operations Research

Abstract PhD Thesis

Parameter Estimation of Ordinary Differential Equations Systems by Least Squares algorithms.

This dissertation will apply operations research and statistical modeling techniques in defining, implementing and solving parameter estimation and model validation phases of the model building process. The nonlinear programming formulation will be developed as a standard least squares (LS) problem where model fitting is entertained on the basis of experimental observations of the phenomenon under study. LS formulation will be modified in order to solve parameter estimation of differential equations systems. Yet, a tentative "statistically correct" formulation will be illustrated and implemented by a constrained LS problem, where asymptotic results of convergence and consistency of least squares estimators (LSE) will be given as statistical constraints of a general nonlinear programming problem.

Statistical experiments, generated by observations of the form

$$y_i = f(\mathbf{x}_i;\boldsymbol{\theta}) + \varepsilon_i, \quad i = 1,....,m, \tag{1.1}$$

are planned and studied. In (1.1) the observations are given by $(y_i, \mathbf{x}_i), (i = 1,....,m)$, where the response variable is represented by $y \in Y \subset R$ (it can be a vector-valued variable) and the explanatory variables or regressors are represented by $\mathbf{x}_i \in X \subset R^p\,(i = 1,...,m)$. $f(\mathbf{x}_i;\boldsymbol{\theta})$ is a non-random function defined on a parameter space Θ, where $\boldsymbol{\theta} \in \Theta \subseteq R^n$ is the unknown parameter to be estimated on the basis of the observations. $\varepsilon_i, (i = 1,....,m)$ are called the errors and they are supposed to be independent random variables with zero mean, constant variance and common distribution function not depending on $\boldsymbol{\theta}$. Y, X and Θ are open sets in R, R^p and R^n respectively. $f(\mathbf{x}_i;\boldsymbol{\theta})$ can represent both static and dynamic models, the latter being both continuous and discrete. Parameter estimation and curve fitting are usually solved by least squares (LS) formulations, where by choosing optimal parameter vectors, the "distance" between experimental data $\left(y_j,\ j = 1, 2,....,m\right)$

and model predictions $\left(\hat{y}_j,\ j=1,2,....,m\right)$ is minimized. Many nonlinear programming algorithms can numerically solve LS problems: among those, Gauss-Newton and Levenberg-Marquardt are frequently implemented.

Standard LS formulation, i.e.

$$\underset{\boldsymbol{\theta}}{\mathrm{Min}} \frac{\sum_{i=1}^{m}\left(y_i - f(\mathbf{x}_i;\boldsymbol{\theta})\right)^2}{m} = \underset{\boldsymbol{\theta}}{\mathrm{Min}} \frac{\sum_{i=1}^{m}\left(\varepsilon_i\right)^2}{m} = \underset{\boldsymbol{\theta}}{\mathrm{Min}}\,(\mathrm{Var}(\boldsymbol{\varepsilon})) \qquad (1.2)$$

where $Var(\boldsymbol{\varepsilon})$ is called the variance of the errors, requires the model to be specified in explicit form. Local and global convergence results for standard LS formulations exist, both for Gauss-Newton and for Levenberg-Marquardt algorithms.

The present work will extend those convergence results to a differential LS formulation, where the model of the phenomenon is specified in the following ordinary differential equation (ODE) form,

$$\begin{cases} \dot{y}(t) = g(t,y;\boldsymbol{\theta}) \\ y(t_0) = y_0 \end{cases} \qquad (1.3)$$

where $g(t,y;\boldsymbol{\theta}):(I\,x\,Y)\,x\,\Theta \to Y$, $I=[t_0,t_f]\subset R^{+}\cup\{0\}$, $Y\subset R$, $\Theta\subset R^{n}$.

Let suppose that

$$y(t,t_0,y_0,\boldsymbol{\theta}) \qquad (1.4)$$

is the explicit solution of (1.3), where $y(t,t_0,y_0,\boldsymbol{\theta}): I\,x\,(I\,x\,Y)\,x\,\Theta \to Y$, and $I=[t_0,t_f]\subset R^{+}\cup\{0\}$, $Y\subset R$, $\Theta\subset R^{n}$. If $y(t,t_0,y_0,\boldsymbol{\theta})$ is known and m measurements (t_j,y_j), $j=1,2,....,m$ of the phenomenon are collected, parameter estimation of (1.3) can be obtained by solving the following optimization problem

$$\underset{\boldsymbol{\theta}}{\mathrm{Min}}\ f(\boldsymbol{\theta}) = \|\, r(\boldsymbol{\theta})\,\|^2 \qquad (1.5)$$

where $r_j(\boldsymbol{\theta}) = y_j - y(t_j,t_0,y_0;\boldsymbol{\theta})$, $j=1,2,....,m$, and can be addressed as the standard least squares (**LS**) problem. If (1.3) cannot be solved explicitly, as it is often the case, a numerical approximation method must be used for estimating $y(t,t_0,y_0,\boldsymbol{\theta})$ in each experimental point t_i $(i=1,2,....,m)$.

So, given $y(t_0)=y_{t_0}$

$$y(t_0+t) = y_{t_0} + \int_{t_0}^{t} \dot{y}\,dt = y_{t_0} + \int_{t_0}^{t} g(t,y;\boldsymbol{\theta})\,dt \ , \ t = t_1, t_2, \ldots, t_m$$

or equivalently, if it is assumed that $t_0 = 0$ ($y(t_0) = y(0) = y_0$) a numerical approximation of

$$y(t) = y_0 + \int_{0}^{t} \dot{y}\,dt = y_0 + \int_{0}^{t} g(t,y;\boldsymbol{\theta})\,dt \ , \ t = t_1, t_2, \ldots, t_m$$

must be calculated.

Thus, for each step of the optimization algorithm, current values of $\boldsymbol{\theta}$, i.e. $\boldsymbol{\theta}^{(k)}$ are used to simulate the trajectory of (1.3) by

$$\hat{y}(t_i) = y_0 + \int_{0}^{t_i} \dot{y}\,dt = y_0 + \int_{0}^{t_i} g(t,y;\boldsymbol{\theta}^{(k)})\,dt \ , \ i = 1, 2, \ldots, m \qquad (1.6)$$

Let define (1.6) by

$$NM(t_j;\boldsymbol{\theta}) = \hat{y}(t_j;\boldsymbol{\theta}) \ , \ j = 1, 2, \ldots..m \ , \ y(t_0) = y_0 \qquad (1.7)$$

where $NM(t_j;\boldsymbol{\theta})$ represents the general numerical method applied for calculating y in each of the *m* points of the experiment and

$$NM(t_j;\boldsymbol{\theta}^{(k)}) = \hat{y}(t_j) \ , \ j = 1, 2, \ldots..m \ , \ y(t_0) = y_0 \qquad (1.8)$$

which indicates the specific numerical-solution of (1.3) for $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$.

So, for each step of the optimization algorithm a call to (1.7) is performed and a vector of solutions $\{\hat{y}(t_j)\}_{\ j=1,2,\ldots..m}$ is calculated by (1.8). The objective function resulting from this new formulation is then modified as follows

$$\underset{\boldsymbol{\theta}}{\text{Min}} \ F(\boldsymbol{\theta}) = \left\| r(\boldsymbol{\theta}) \right\|^2 = \sum_{j=1}^{m} [y_j - NM(t_j;\boldsymbol{\theta})]^2 \qquad (1.9)$$

and it will be called the differential least squares problem.

In the present work, classical explicit fixed-step Runge Kutta method is used for implementing the numerical integration method (1.7). It is shown how the formulation (1.9) satisfies the hypothesis required for local convergence of both Gauss-Newton and Levenberg-Marquardt algorithm if the function $g(t, y;\boldsymbol{\theta})$ of (1.3) possesses certain properties. Moreover, stability properties for the differential equation are indeed required for assessing that the problem (1.9) is well posed and bounded. Changing the objective function in order to fulfill the coercivity condition attains also global convergence.

Moreover, the above results can be extended to explicit variable step Runge Kutta methods.

Another contribution of the dissertation is to establish a link between statistical model building and operations research techniques in order to formulate a tentative “statistically correct” procedure for parameter estimation and model validation. From an operational research point of view, the one and only goal is to find the global minimum of (1.5) or (1.9). From a statistical point of view, when a minimum is found, a validation phase by means of residuals analysis should be performed. This validation phase can be implemented in the nonlinear programming formulation. This is done by first defining statistical conditions over the residuals and, after that, by translating these conditions in equally powerful constraints for the optimization problem, so that the estimation and validation phases are performed simultaneously. The idea is that each estimate satisfies automatically, through the constraints, the required statistical properties. While there is a general agreement on the statistical properties that the residuals from an estimated model should satisfy, there has been hardly any work done on methods to constrain the parameter space, so that correct estimates can be obtained. Standard approach develops LSE without constraints, addressing the validation phase separately and they will be indicated as UNLLS, i.e. Unconstrained Non Linear Least Squares. An existing version of constrained nonlinear least squares (CNLLS)[1] is illustrated and improved by introducing two new statistical constraints which takes in account respectively the skewness and the kurtosis of the distribution of the residuals.

A general CNLLS for parameter estimation of models given both in the explicit form (1.1) and of ODEs (1.3) is described and 400 simulations will be run on different explicit and differential models, taken mostly from molecular biology and population dynamics.

The last contribution of the work has been to formulate two new mathematical models on two real-life applications in order to test both the differential LS implementation and the CNLLS on real data. One application will strengthen the latter conclusion (lower variance of UNLLS), comparing UNLLS and CNLLS formulation performed on data relating to the uptake of an artificial substrate in rat liver. Another

[1] Bartolozzi F., A. De Gaetano, E. Di Lena, S. Marino, L. Nieddu, G. Patrizi, *Operational research techniques in medical treatment and diagnosis: A review*, **European Journal of Operational Research**, Volume 121, Issue 3, 16 March 2000, Pages 435-466 *Ibidem*, 2000

application, namely the glucose-insulin dynamical system, will perform only UNLLS in the differential implementation. It underlines how the dynamic properties of any suggested model ought to be formally investigated in order to subject to closer scrutiny and meaningful numerical estimation those models which are known structurally to possess desirable overall characteristics, like stability, positivity or boundedness of solutions and the like.