Generalized profiled estimation

Weixin Wang & Jianbin Tan

University of Science and Technology of China

2022.10.19

Content

Table of Contents

Introduction

Generalized profiled estimation

- In this two chapters, we will introduce a profiled estimation for parameter estimation in mechanistic models and discuss its statistical consistency.
- Related algorithms to improve computation efficiency will also be discussed.

Review of parameter estimation strategies

Review of current ordinary differential equation parameter estimation strategies

- Data fitting by numerical approximation of an initial value problem.
- Collocation methods or basis function expansions.

Review of parameter estimation strategies

Data fitting by numerical approximation of an initial value problem (NLS method)

- Approximate solutions of ODEs over a range $[t_0, t_1]$ use fixed initial values $x_0 = x(t_0)$.
- One numerical method: Runge–Kutta algorithm.
- Solve the sensitivity differential equations:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}\theta}\right) = \frac{\partial \mathbf{f}}{\partial \theta} + \frac{\partial \mathbf{f}}{\partial \mathbf{x}}\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}\theta}, \qquad \text{with} \quad \left.\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}\theta}\right|_{t=0} = 0.$$

- In the event that $x(0) = x_0$ must also be estimated, the corresponding sensitivity equations are:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}\mathbf{x}_0}\right) = \frac{\partial \mathbf{f}}{\partial \mathbf{x}}\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}\mathbf{x}_0}, \qquad \text{with} \quad \left.\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}\mathbf{x}_0}\right|_{t=0} = \mathbf{I}.$$

Review of parameter estimation strategies

NLS procedure has many problems:

- Computationally intensive
- Inaccuracy of the numerical approximation
- Size of the parameter set may be increased by the set of initial conditions that are needed

Review of parameter estimation strategies

Collocation methods or basis function expansions

- Express the approximation $\hat{x}_i$ of x_i in terms of a basis function expansion:

$$\hat{x}_i(t) = \sum_k^{K_i} c_{ik}\phi_{ik}(t) = \mathbf{c}_i'\phi_i(t)$$

- Procedure:
  - Each x_i is first estimated by data smoothing methods
  - Minimization of a least squares measure of the fit of $\mathrm{d}\hat{\mathbf{x}}/\mathrm{d}t$ to $\mathbf{f}(\hat{\mathbf{x}}, \mathbf{u}, t \mid \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$
  - Iteration's roughness penalty: $\|\mathrm{d}\hat{\mathbf{x}}/\mathrm{d}t - f(\hat{\mathbf{x}}, \mathbf{u}, t \mid \theta)\|$

Table of Contents

The Parameter Cascading Strategy for Estimating Parameters

The elements of parameter cascading

- Approximate the solution $x(t)$ to the differential equation by the basis function expansion

$$\hat{x}(t) = \sum_{k}^{K} c_k \phi_k(t) = \mathbf{c}^T \boldsymbol{\phi}(t).$$

- Constrain the coefficient vector $\mathbf{c}$ to be a smooth function $\mathbf{c}(\boldsymbol{\theta})$ of the parameters in vector θ that define the differential equation.
- Optimize the fit to observations $y_1, \ldots, y_n$ in vector $\mathbf{y}$ by $\hat{x}(\mathbf{t}) = \mathbf{c}(\boldsymbol{\theta})^T \boldsymbol{\phi}(\mathbf{t})$ with respect to $\boldsymbol{\theta}$ by optimizing a fitting criterion $H(\boldsymbol{\theta} \mid \mathbf{y})$.

Two Classes of Parameters

Main idea

- Replace the coefficients c_k in the basis expansion by values of a smooth function of the parameters in θ
- The status of the K-dimensional coefficient vector c as a set of parameters is eliminated
- The coefficient estimates are actually the values of the function $c(\hat{\theta})$at the optimal value of the θ

So why we do this?

Two Classes of Parameters

Why parameter cascading?

- Nuisance parameters? Profiled?
- Prevent over-fitting the data
- Reduce computation required to optimize the fit of the model to the data

We will see over and over again that forcing $c(\theta)$ to be a smooth function of the parameters will results in better parameter estimates with smaller confidence intervals without seriously harming the fit to the data that $c(\theta)$ defines.

Defining Coefficients as Functions of Parameters

Approach to defining $c(\theta)$

- Define an inner fitting criterion $J(\mathbf{c} \mid \boldsymbol{\theta})$ that optimizes $\mathbf{c}$ given any candidate value for the parameters in θ
- Then that is re-optimized each time $\boldsymbol{\theta}$ is changed in order to optimize an outer fitting criterion $H(\boldsymbol{\theta})$
- Higher priority on the estimation θ; Lower priority on the estimation of c and the $x(t)$

How can we find a way of blending together the information in the data and the information in the equation?

The Symmetric Relation Between the Data and the Differential Equation

Just capturing the prominent shape features

- Fix a value $\rho \in [0, 1)$, called a smoothing parameter or a bandwidth parameter, that **modulates the relative emphasis on fitting the data as opposed to satisfying the differential equation**
- By choosing ρ judiciously, we can see how much of the variation in the data can be accommodated by what is perhaps a too-simple dynamic model. If this is possible, then the residual variation is apt to give important cues about either how the model is to be elaborated, or about what kind and quantity of data would be more revealing.

Inner Optimization Criterion J

Define the inner optimization function $J(\mathbf{c} \mid \theta, \rho)$ (one dimension):

$$J(\mathbf{c} \mid \boldsymbol{\theta}, \rho) = (1-\rho)\sum_{j}^{n}\left[y_j - x(t_j)\right]^2/n + \rho\int_0^T \left\{\mathrm{D}^m x(t) - f[x(t)]\right\}^2 \,\mathrm{d}t/T$$

Remark:

- Squared residual fit measures
- The smoothing parameter $\rho \in [0, 1)$
- $\rho = 0 \Rightarrow$ ignore the differential equation entirely and concentrate solely on data-fitting
- $\rho \to 1 \Rightarrow$ place more and more emphasis on x being close to a solution to the differential equation
- $\rho < 1 \Rightarrow$ not want to ignore the data entirely

A solution to most differential equations is not uniquely defined without some additional information, and the data provides this information.

The Least Squares Cascade Coefficient Function

Some notations

- Differential equation residual function $\mathrm{D}^m x(t) - f[x(t) \mid \boldsymbol{\theta}]$ can be re-expressed as $\mathrm{L}x = 0$ where L is the differential operator form of the differential equation $\mathrm{D}^m x = f(x \mid \boldsymbol{\theta})$
- The fitting function is $\hat{x}(t) = \mathbf{c}^T \boldsymbol{\phi}(t)$, where the coefficient vector $\mathbf{c}$ and basis function vector $\boldsymbol{\phi}$ are of length K

Then inner criterion becomes:

$$J(\mathbf{c} \mid \boldsymbol{\theta}) = (1-\rho)(\mathbf{y} - \boldsymbol{\Phi}\mathbf{c})^T(\mathbf{y} - \boldsymbol{\Phi}\mathbf{c})/n + \rho \int_0^T \left[\mathbf{c}^T \mathrm{L}\boldsymbol{\phi}(t)\right] \left[\mathbf{c}^T \mathrm{L}\boldsymbol{\phi}(t)\right]^T \mathrm{d}t/$$

$$= (1-\rho)(\mathbf{y} - \boldsymbol{\Phi}\mathbf{c})^T(\mathbf{y} - \boldsymbol{\Phi}\mathbf{c})/n + \rho\mathbf{c}^T \left(\int_0^T [\mathrm{L}\boldsymbol{\phi}(t)] \left[\mathrm{L}\boldsymbol{\phi}(t)^T\right] \mathrm{d}t \right) \mathbf{c}/$$

$$= (1-\rho)(\mathbf{y} - \boldsymbol{\Phi}\mathbf{c})^T(\mathbf{y} - \boldsymbol{\Phi}\mathbf{c})/n + \rho\mathbf{c}^T\mathbf{R}(\boldsymbol{\theta})\mathbf{c}/T$$

where

$$\mathbf{R}(\theta) = \int_0^T [\mathrm{L}\boldsymbol{\phi}(t)] \left[\mathrm{L}\boldsymbol{\phi}(t)^T\right] \mathrm{d}t$$

The minimizing value of c

Inner criterion

$$J(\mathbf{c} \mid \theta) = (1-\rho)(\mathbf{y}-\mathbf{\Phi c})^T(\mathbf{y}-\mathbf{\Phi c})/n + \rho \mathbf{c}^T \mathbf{R}(\theta)\mathbf{c}/T$$

In matrix calculus to work out that the minimizing value of c that satisfies the equation

$$\mathbf{c}(\theta) = \left[(1-\rho)\mathbf{\Phi}^T\mathbf{\Phi}/n + \rho \mathbf{R}(\theta)/T\right]^{-1}(1-\rho)\mathbf{\Phi}^T\mathbf{y}/n$$

More smooth as we increase ρ

This figure shows how the coefficients $c_3(\rho), c_6(\rho)$, and $c_9(\rho)$ vary as β ranges from 0.5 to 1.5 for the relatively low smoothing parameters $\rho = 0.27$, and for the high value $\rho = 0.98$

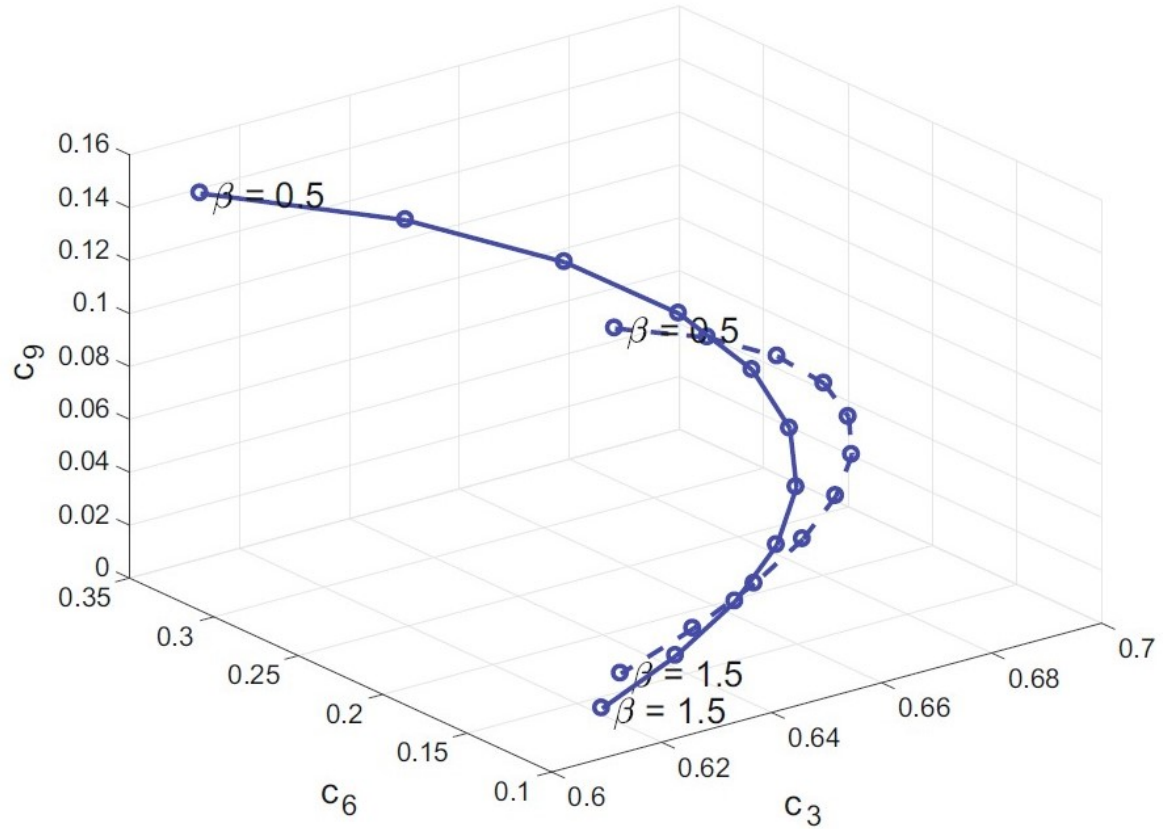


Fig. 9.2 The *curves* display three of the 13 coefficient functions $c_3(\beta)$, $c_6(\beta)$, and $c_9(\beta)$ defining $x(t)$ varying over values of β between 0.5 and 1.5 for the model and data in Fig. 9.1, and for values of ρ of 0.27 (*dashed*) and 0.98 (*solid*)

The Outer Fitting Criterion H

To complete the parameter cascading strategy, we specify an outer fitting criterion

$$H(\boldsymbol{\theta} \mid \rho) = G_0[\mathbf{y}, x(\mathbf{t}) \mid \boldsymbol{\theta}, \rho].$$

If we stick with least squares as our criterion, then

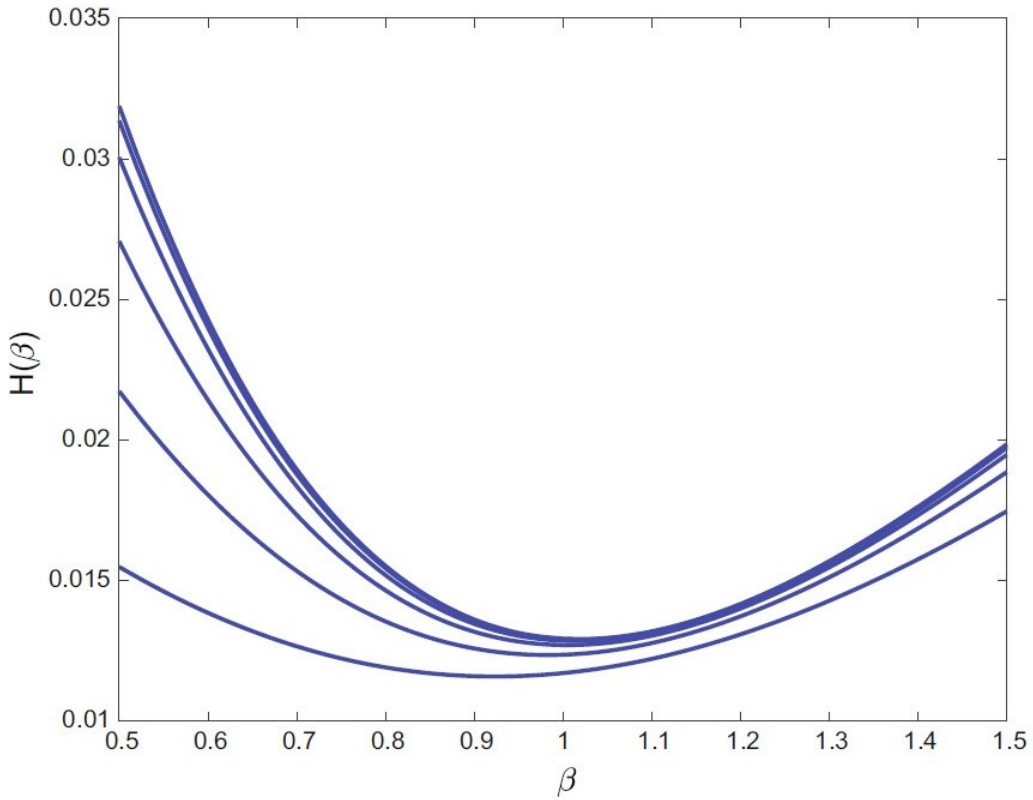
$$H(\boldsymbol{\theta} \mid \rho) = \sum_j [y_j - \hat{x}(t_j)]^2 = \sum_j \left[y_j - \mathbf{c}^T(\boldsymbol{\theta} \mid \rho)\boldsymbol{\phi}(t_j)\right]^2$$

Remark:

- Use the Gauss-Newton algorithm
- The shape of the criterion function also varies with the level of ρ

The shape of the criterion function also varies with the level of ρ

- As we increase ρ the height of H increases, its asymmetry increases, as does its curvature, but we also see that the minimizing values moves closer and closer to the true value $\beta = 1$.



Choosing the Smoothing Parameter ρ

A simple and effective answer is to estimate both θ and x for a wide range of ρ values so as to calculate the functional flow $\mathbf{x}(\rho)$ or parametric flow $\theta(\rho)$.

Degrees of freedom

$$df(\boldsymbol{\theta}) = \text{trace}\left[2\mathbf{M}(\boldsymbol{\theta}) - \mathbf{M}(\boldsymbol{\theta})\mathbf{M}(\boldsymbol{\theta})^T\right].$$

where

$$\mathbf{M}(\boldsymbol{\theta}) = (1-\rho)\boldsymbol{\Phi}\left[(1-\rho)\boldsymbol{\Phi}^T\boldsymbol{\Phi}/n + \rho\mathbf{R}(\boldsymbol{\theta})/T\right]^{-1}\boldsymbol{\Phi}^T/n$$

Remark:

- Refers to the effective dimensionality of the fitting function as a function of ρ.
- For $\rho \to 0$, it turns out that this measure converges to K, the number of basis functions.
- But as $\rho \to 1$, the measure converges to $K - \text{rank}(\mathbf{M}) = m$, which corresponds to the dimensionality of the solution space for the differential equation.

GCV criterion

GCV criterion

$$GCV(\rho) = \frac{n}{[n - df(\rho)]^2} \text{SSE}(\rho)$$

where $SSE(\rho)$ is the residual sum of squares

- the value of ρ that minimizes this GCV measure will provide a nearly optimal estimate of x most of the time if the residuals are reasonably close to being uncorrelated.

Multi-variable Systems

The general formulation:

$$\mathrm{D}^{M_i} x_i(t) = \sum_{k=1}^{d} \sum_{j=0}^{M_k - 1} b_{ijk} \beta_{ijk}(t \mid \boldsymbol{\theta}) \mathrm{D}^j x_k(t) + \sum_{\ell}^{*} a_{i\ell} \alpha_{i\ell}(t \mid \boldsymbol{\theta}) u_{i\ell}(t), \quad M_i \geq 0, \quad i = 1, \ldots, d.$$

The inner criterion J will now be

$$J(\mathbf{c} \mid \theta) = (\mathbf{y} - \mathbf{\Phi c})^T \mathbf{W}_1 (\mathbf{y} - \mathbf{\Phi c}) + \mathbf{c}^T \mathbf{W}_2 \mathbf{R}(\theta) \mathbf{c} + \rho \mathbf{c}^T \mathbf{W}_2 \mathbf{S}(\theta),$$

Then we have that

$$\mathbf{c}(\theta) = \left[\mathbf{W}_1 \mathbf{\Phi}^T \mathbf{\Phi} + \mathbf{W}_2 \mathbf{R}(\theta)\right]^{-1} \left[\mathbf{W}_1 \mathbf{\Phi}^T \mathbf{y} + \mathbf{W}_2 \mathbf{S}(\theta)\right]$$

Example

Analysis of the Head Impact Data (measure the effects of motorcycle accident on the driver's brain tissue)

- We analyzed the data using the three-parameter forced damped harmonic equation

$$\mathrm{D}^2 x(t) = -\beta_0 x(t) - \beta_1 \mathrm{D} x(t) + \alpha u(t)$$

Analysis of the Head Impact Data

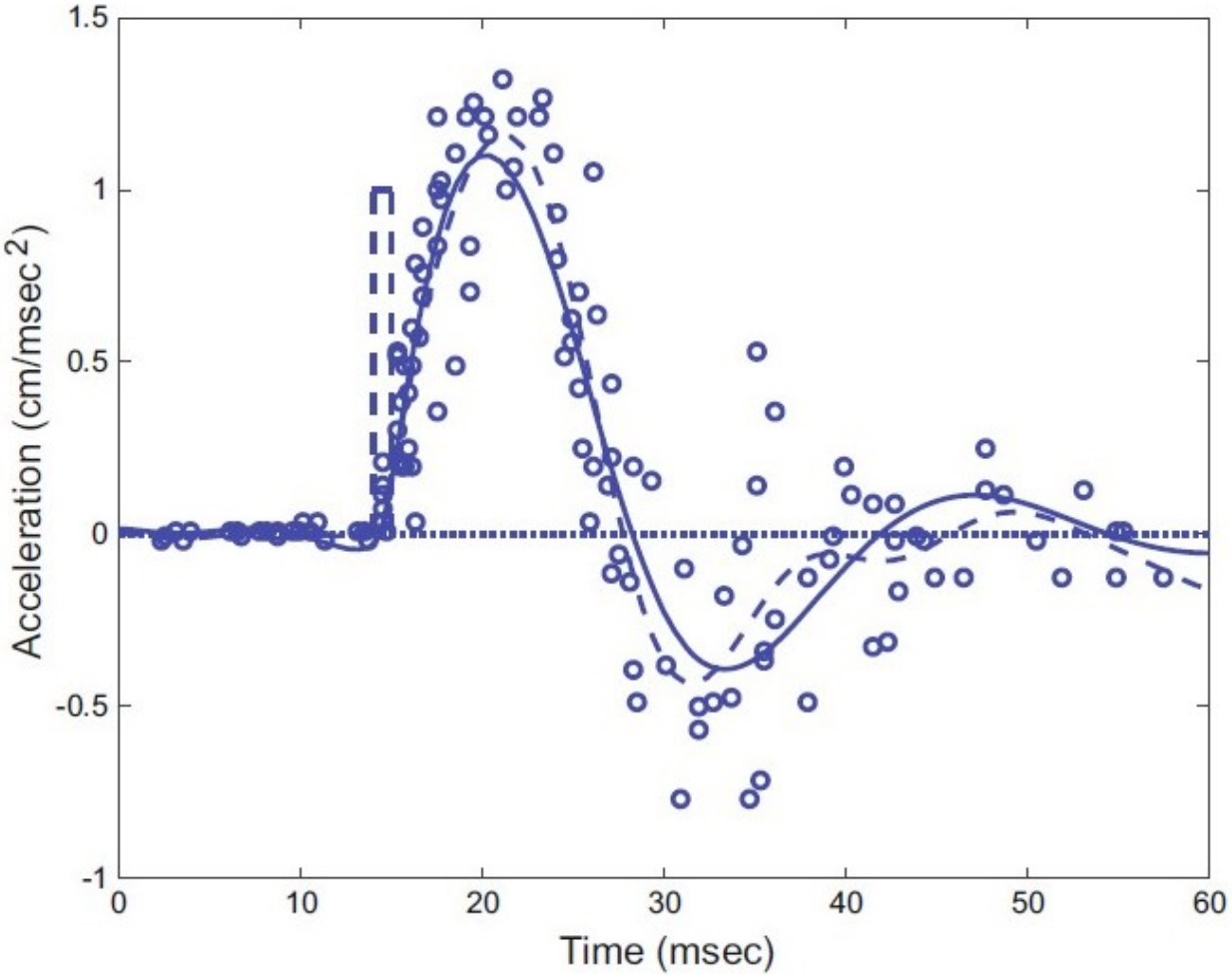


Fig. 9.5 The *circles* indicate observations of acceleration of brain tissue from five replications of an experiment involving striking the cranium of a corpse with a blunt object. The *box* was constructed to represent the impact itself spread over one time unit. The *dashed* and *solid lines* indicate the fits from the parameter cascading analysis of the data in Fig. 9.5 using the Eq. (9.22) and corresponding to $\rho = 0.5$ and $\rho = 0.998$, respectively

Analysis of the Head Impact Data

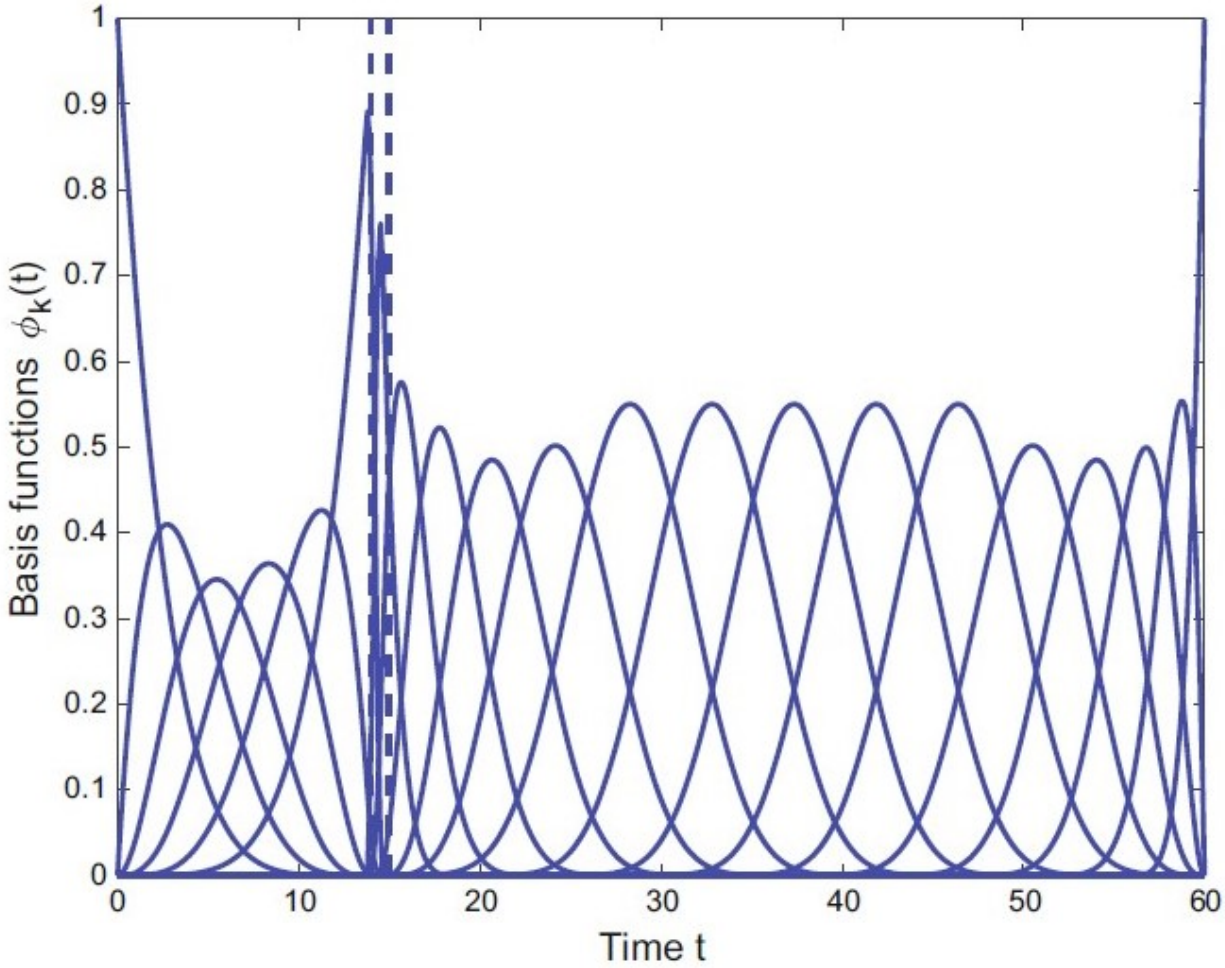


Fig. 9.6 The *solid lines* are the 21 order 6 B-spline basis functions active in a region around the box forcing function for the head impact data. The *vertical dashed lines* indicate the two locations of three coincident knots

Analysis of the Head Impact Data

Figure 9.7 shows how the parameter estimates vary over ρ values 0.50, 0.73, 0.88, 0.95, 0.98, 0.99 and 0.998. The values of parameters β_0, β_1, and α, along with two standard errors at $\rho = 0.998$ are 0.056 ± 0.011, 0.128 ± 065 and 0.383 ± 0.101, respectively. The GCV criterion was minimized at 0.556 for $\rho = 0.98$, corresponding to 10.3 degrees of freedom, and for $\rho = 0.998$, GCV = 0.574 corresponding to 6.4 degrees of freedom.

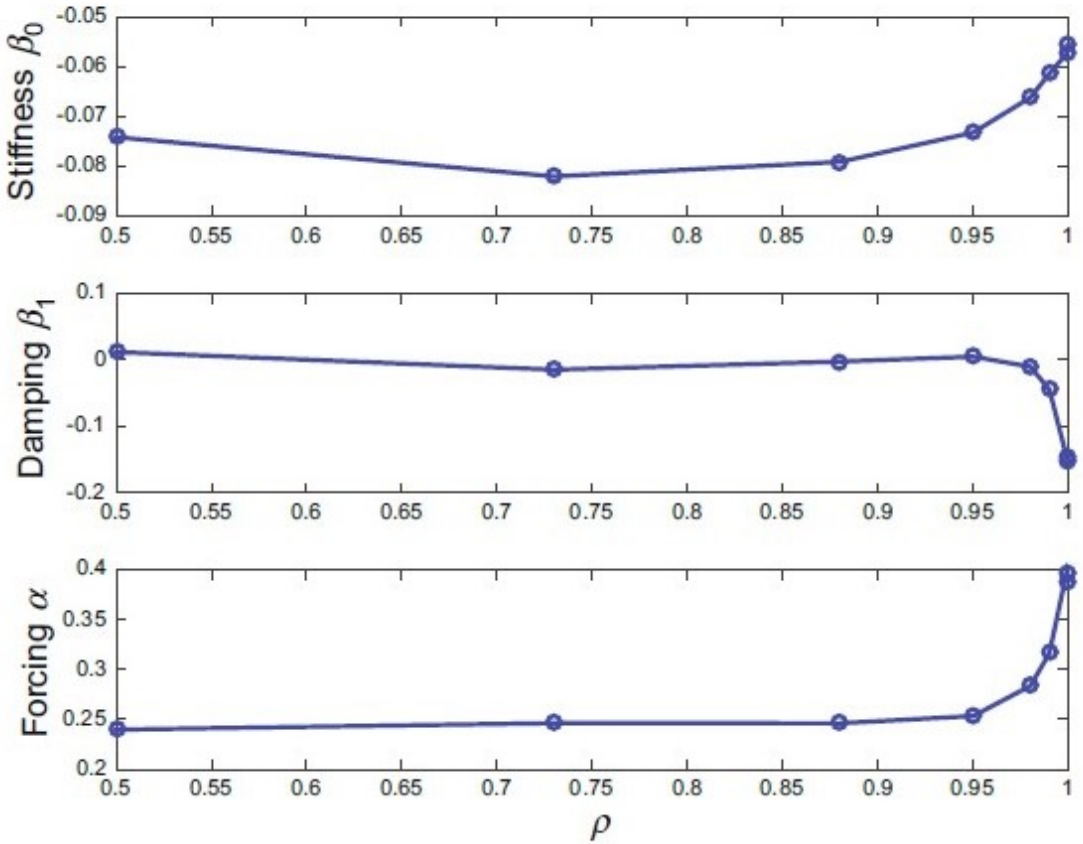


Fig. 9.7 The three parameter values as a function ρ, the largest value of which is 0.998

Table of Contents

The Setup for Parameter Cascading

- In the least squares data-fitting case

$$J(\mathbf{c} \mid \boldsymbol{\theta}, \lambda) = \sum_{j}^{n} [y_j - x(t_j)]^2 + \lambda \int_0^T \{\mathrm{D}^m x(t) - f[x(t), \mathbf{u}(t) \mid \boldsymbol{\theta}]\}^2 \, \mathrm{D}t$$

where the bandwidth or smoothing parameter $\lambda > 0$ controls the emphasis on fitting the equation but the data-fitting first term remains fixed

The outer criterion

- The outer criterion $H(\boldsymbol{\theta} \mid \lambda)$ defines fit in terms of only the parameter vector θ. The total derivative of H with respect to θ must allow for the fact that H may depend both directly on $\boldsymbol{\theta}$ and indirectly via its dependency on $\mathbf{c}(\boldsymbol{\theta})$:

$$\frac{dH}{d\boldsymbol{\theta}} = \frac{\partial H}{\partial \boldsymbol{\theta}} + \left(\frac{\partial H}{\partial \mathbf{c}}\right)\left(\frac{d\mathbf{c}}{d\boldsymbol{\theta}}\right)$$

- Although for most fitting criteria, including least squares, H has no direct dependency at all on $\boldsymbol{\theta}$, so that $\partial H/\partial \boldsymbol{\theta} = 0$ and can be dropped from the total derivative equation. It is the derivative $d\mathbf{c}/d\boldsymbol{\theta}$ that is the missing link, since we no longer have an explicit expression for $\mathbf{c}$.

Parameter Cascading Computations

Numerical methods must be used to achieve the optimum of J (implicit differentiation or the Implicit Function Theorem)

- Suppose that we have optimized J using a high quality numerical optimization strategy to obtain an optimum value c
- $\partial J/\partial \mathbf{c} \approx 0$ to a high level of accuracy
- We now compute the total derivative $d/d\boldsymbol{\theta}$ of the gradient of J with respect to θ, which, since $\partial J/\partial \mathbf{c} = 0$, will also be 0 :

$$\frac{d}{d\boldsymbol{\theta}}\left(\frac{\partial J}{\partial \mathbf{c}}\right) = \frac{\partial^2 J}{\partial \mathbf{c}\partial \boldsymbol{\theta}} + \left(\frac{\partial^2 J}{\partial \mathbf{c}^2}\right)\left(\frac{d\mathbf{c}}{d\boldsymbol{\theta}}\right) = 0$$

Implicit Function Theorem

$$\frac{d}{d\boldsymbol{\theta}}\left(\frac{\partial J}{\partial \mathbf{c}}\right) = \frac{\partial^2 J}{\partial \mathbf{c}\partial \boldsymbol{\theta}} + \left(\frac{\partial^2 J}{\partial \mathbf{c}^2}\right)\left(\frac{d\mathbf{c}}{d\boldsymbol{\theta}}\right) = 0$$

so that

$$\frac{d\mathbf{c}}{d\boldsymbol{\theta}} = -\left(\frac{\partial^2 J}{\partial \mathbf{c}^2}\right)^{-1}\left(\frac{\partial^2 J}{\partial \mathbf{c}\partial \boldsymbol{\theta}}\right)$$

and, substituting this into (10.3) we see that

$$\frac{dH}{d\boldsymbol{\theta}} = \frac{\partial H}{\partial \boldsymbol{\theta}} - \left(\frac{\partial H}{\partial \mathbf{c}}\right)\left(\frac{\partial^2 J}{\partial \mathbf{c}^2}\right)^{-1}\left(\frac{\partial^2 J}{\partial \mathbf{c}\partial \boldsymbol{\theta}}\right).$$

Then we can use gradient descent method or Gauss-Newton method to optimize.

Nonlinear Systems and Other Fitting Criteria

- We use the following expression of the inner fitting criterion involving a general data-fitting function G_J, in the special case where all x_i trajectories are expanded using the same set ϕ of K basis functions:

$$J(\mathbf{c} \mid \boldsymbol{\theta}, \lambda) = G_J[\mathbf{Y}, \mathbf{C}(\boldsymbol{\theta} \mid \lambda)\boldsymbol{\phi}] + \sum_{i}^{d} \lambda_i \int_0^T \{\mathrm{D}x_i(t) - f[\mathbf{x}(t), \mathbf{u}_i(t) \mid \boldsymbol{\theta}]\}^2 \mathrm{D}$$

- The coefficient matrix $\mathbf{C}$ will, in this simple case, have K rows and d columns. In this formulation, also for simplicity, we retain the integrated squared residuals in the penalty term, but we certainly envisage other penalty structures that might be more appropriate.

Nonlinear Systems and Other Fitting Criteria

- The corresponding expression for the outer criterion H is

$$H(\boldsymbol{\theta} \mid \lambda) = G_H[\mathbf{Y}, \mathbf{C}(\boldsymbol{\theta} \mid \boldsymbol{\lambda})\boldsymbol{\phi}].$$

- However, the two data fitting functions G_J and G_H may in many applications **be the same.**

Approximating the sampling variation of $\hat{\theta}$ and $\hat{c}$

Let $\mathbf{\Sigma}$ be the variance-covariance matrix for $\mathbf{y}$. Making explicit the dependence of H on the data $\mathbf{y}$ by using the notation $H(\theta \mid \mathbf{y})$, the estimate $\hat{\theta}(\mathbf{y})$ of θ is the solution of the stationary equation $\partial H(\theta \mid \mathbf{y})/\partial\theta = 0$.

The usual δ-method that is employed in non-linear least squares produces a variance estimate of the form

$$\mathrm{var}_{\mathrm{GN}}\{\hat{\boldsymbol{\theta}}(\mathbf{y})\} \approx \sigma^2 \left\{ \left(\frac{\mathrm{d}\hat{\mathbf{x}}}{\mathrm{d}\boldsymbol{\theta}}\right)' \left(\frac{\mathrm{d}\hat{\mathbf{x}}}{\mathrm{d}\boldsymbol{\theta}}\right) \right\}^{-1}$$

by making use of the approximation

$$\frac{\mathrm{d}^2 H}{\mathrm{d}\theta^2} \approx \left(\frac{\mathrm{d}\hat{\mathbf{x}}}{\mathrm{d}\boldsymbol{\theta}}\right)' \left(\frac{\mathrm{d}\hat{\mathbf{x}}}{\mathrm{d}\boldsymbol{\theta}}\right)$$

Approximating the sampling variation of $\hat{\theta}$ and $\hat{c}$

By applying the implicit function theorem to $\partial H/\partial\theta$ as a function of $\mathbf{y}$, we may say that for any $\mathbf{y}$ in $\mathcal{N}$ there is a value $\hat{\theta}(\mathbf{y})$ satisfying $\partial H/\partial\theta = 0$. By taking the $\mathbf{y}$-derivative of this relation, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{y}}\left(\left.\frac{\mathrm{d}H}{\mathrm{d}\boldsymbol{\theta}}\right|_{\hat{\theta}(\mathbf{y})}\right) = \left.\frac{\mathrm{d}^2 H}{\mathrm{d}\boldsymbol{\theta}\mathrm{d}\mathbf{y}}\right|_{\hat{\theta}(\mathrm{y})} + \left.\frac{\mathrm{d}^2 H}{\mathrm{d}\boldsymbol{\theta}^2}\right|_{\hat{\theta}(\mathrm{y})} \frac{\mathrm{d}\hat{\theta}}{\mathrm{d}\mathbf{y}} = 0,$$

where

$$\frac{\mathrm{d}^2 H}{\mathrm{d}\theta^2} = \frac{\partial^2 H}{\partial\theta^2} + \frac{\partial^2 H}{\partial\theta\partial\hat{\mathbf{c}}}\frac{\partial\hat{\mathbf{c}}}{\partial\theta} + \left(\frac{\partial\hat{\mathbf{c}}}{\partial\theta}\right)'\frac{\partial^2 H}{\partial\hat{\mathbf{c}}\partial\theta} + \left(\frac{\partial\hat{\mathbf{c}}}{\partial\theta}\right)'\frac{\partial^2 H}{\partial\hat{\mathbf{c}}^2}\frac{\partial\hat{\mathbf{c}}}{\partial\theta} + \frac{\partial H}{\partial\hat{\mathbf{c}}}\frac{\partial^2\hat{\mathbf{c}}}{\partial\theta^2}$$

and

$$\frac{\mathrm{d}^2 H}{\mathrm{d}\boldsymbol{\theta}\mathrm{d}\mathbf{y}} = \frac{\partial^2 H}{\partial\theta\partial\mathbf{y}} + \frac{\partial^2 H}{\partial\hat{\mathbf{c}}\partial\mathbf{y}}\frac{\partial\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}} + \frac{\partial^2 H}{\partial\boldsymbol{\theta}\partial\hat{\mathbf{c}}}\frac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}} + \left(\frac{\partial\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}}\right)'\frac{\partial^2 H}{\partial\hat{\mathbf{c}}^2}\frac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}} + \frac{\partial H}{\partial\hat{\mathbf{c}}}\frac{\partial^2\hat{\mathbf{c}}}{\partial\theta\partial\mathbf{y}}$$

Approximating the sampling variation of $\hat{\theta}$ and $\hat{c}$

Solve this equation

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{y}}\left(\left.\frac{\mathrm{d}H}{\mathrm{d}\boldsymbol{\theta}}\right|_{\hat{\theta}(\mathbf{y})}\right) = \left.\frac{\mathrm{d}^2H}{\mathrm{d}\boldsymbol{\theta}\mathrm{d}\mathbf{y}}\right|_{\hat{\theta}(\mathrm{y})} + \left.\frac{\mathrm{d}^2H}{\mathrm{d}\boldsymbol{\theta}^2}\right|_{\hat{\theta}(\mathrm{y})}\frac{\mathrm{d}\hat{\theta}}{\mathrm{d}\mathbf{y}} = 0,$$

We obtain the first derivative of $\hat{\boldsymbol{\theta}}$ with respect to $\mathbf{y}$:

$$\frac{\mathrm{d}\hat{\boldsymbol{\theta}}}{\mathrm{d}\mathbf{y}} = -\left(\left.\frac{\partial^2 H}{\partial\theta^2}\right|_{\hat{\boldsymbol{\theta}}(\mathrm{y})}\right)^{-1}\left(\left.\frac{\partial^2 H}{\partial\boldsymbol{\theta}\partial\mathbf{y}}\right|_{\hat{\boldsymbol{\theta}}(\mathbf{y})}\right)$$

Approximating the sampling variation of $\hat{\theta}$ and $\hat{c}$

Let $\boldsymbol{\mu} = E(\mathbf{y})$; the first-order Taylor series expansion for $\mathrm{d}\hat{\boldsymbol{\theta}}/\mathrm{d}\mathbf{y}$ is

$$\frac{\mathrm{d}\hat{\boldsymbol{\theta}}}{\mathrm{d}\mathbf{y}} \approx \frac{\mathrm{d}\hat{\boldsymbol{\theta}}}{\mathrm{d}\boldsymbol{\mu}} + \frac{\mathrm{d}^2\hat{\boldsymbol{\theta}}}{\mathrm{d}^2\boldsymbol{\mu}}(\mathbf{y} - \boldsymbol{\mu}).$$

When $\mathrm{d}^2\hat{\boldsymbol{\theta}}/\mathrm{d}^2\mu$ is uniformly bounded, we can take the expectation on both sides of approximation and derive $E(\ \mathrm{d}\hat{\theta}/\mathrm{d}\mu) \approx E(\ \mathrm{d}\hat{\theta}/\mathrm{d}\mathbf{y})$. We can also approximate $\hat{\theta}(\mathbf{y})$ by using the first-order Taylor series expansion:

$$\hat{\theta}(\mathbf{y}) \approx \hat{\theta}(\boldsymbol{\mu}) + \frac{\mathrm{d}\hat{\theta}}{\mathrm{d}\boldsymbol{\mu}}(\mathbf{y} - \boldsymbol{\mu})$$

Approximating the sampling variation of $\hat{\theta}$ and $\hat{c}$

We derive

$$\text{var}\{\hat{\theta}(\mathbf{y})\} \approx \left(\frac{\mathrm{d}\hat{\theta}}{\mathrm{d}\mu}\right) \Sigma \left(\frac{\mathrm{d}\hat{\theta}}{\mathrm{d}\mu}\right)' \approx \left(\frac{\mathrm{d}\hat{\theta}}{\mathrm{d}\mathbf{y}}\right) \Sigma \left(\frac{\mathrm{d}\hat{\theta}}{\mathrm{d}\mathbf{y}}\right)',$$

since

$$E\left(\frac{\mathrm{d}\hat{\theta}}{\mathrm{d}\mu}\right) \approx E\left(\frac{\mathrm{d}\hat{\theta}}{\mathrm{d}\mathbf{y}}\right)$$

Approximating the sampling variation of $\hat{\theta}$ and $\hat{c}$

Similarly, the sampling variance of $\hat{\mathbf{c}}\{\hat{\boldsymbol{\theta}}(\mathbf{y})\}$ is estimated by

$$\text{var}[\hat{\mathbf{c}}\{\hat{\boldsymbol{\theta}}(\mathbf{y})\}] = \left(\frac{\mathrm{d}\hat{\mathbf{c}}}{\mathrm{d}\mathbf{y}}\right) \boldsymbol{\Sigma} \left(\frac{\mathrm{d}\hat{\mathbf{c}}}{\mathrm{d}\mathbf{y}}\right)',$$

where

$$\frac{\mathrm{d}\hat{\mathbf{c}}}{\mathrm{d}\mathbf{y}} = \frac{\mathrm{d}\hat{\mathbf{c}}}{\mathrm{d}\hat{\boldsymbol{\theta}}} \frac{\mathrm{d}\hat{\boldsymbol{\theta}}}{\mathrm{d}\mathbf{y}} + \frac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}}$$

Table of Contents

A new efficient method for ODEs linear in parameters

Description of the algorithm

- We iteratively generate a sequence of estimates $\left(b^{(m)}, \alpha^{(m)}\right)$ with $\hat{u}^{(m)}$ given by

$$\hat{\boldsymbol{u}}^{(m)}(t) = \sum_{j=-1}^{p+1} \boldsymbol{b}_j^{(m)} B_j(t).$$

- (i) The iteration is initialized with a spline $\hat{u}^{(0)}$ obtained by smoothing the data. More precisely, $\hat{u}^{(0)}$ is given by equation (3.3) with $b^{(0)}$ the minimum of $E_{\mathrm{D}}(b)$.
- (ii) For each $m = 1, 2, \ldots$, we obtain $\left(\boldsymbol{b}^{(m)}, \boldsymbol{\alpha}^{(m)}\right)$ from $\hat{\boldsymbol{u}}^{(m-1)}$ by minimizing $E\left(b, \alpha, \hat{u}^{(m-1)}\right)$ with respect to $(b, \boldsymbol{\alpha})$. This is a linear least squares problem which is carried out in one step using QR decomposition.
- (iii) We define $\hat{\boldsymbol{u}}^{(m)}$ from $\boldsymbol{b}^{(m)}$ using equation (3.3). (iv) If for some preset tolerance δ we have $\left\| b^{(m)} - b^{(m-1)} \right\| < \delta$ then terminate, otherwise return to step (ii).

A new efficient method for ODEs linear in parameters

Properties of the solution

- When the iteration terminates, we have $b^{(m)} = b^{(m-1)}$ to some pre-specified numerical tolerance. Thus

$$\hat{E}_{\mathrm{M}}\left(\boldsymbol{b}^{(m)}, \boldsymbol{\alpha}, \hat{\boldsymbol{u}}^{(m-1)}\right) = \hat{E}_{\mathrm{M}}\left(\boldsymbol{b}^{(m)}, \boldsymbol{\alpha}, \hat{\boldsymbol{u}}^{(m)}\right) = E_{\mathrm{M}}\left(\boldsymbol{b}^{(m)}, \boldsymbol{\alpha}\right)$$

  and hence

$$\hat{E}\left(\boldsymbol{b}^{(m)}, \boldsymbol{\alpha}, \hat{\boldsymbol{u}}^{(m-1)}\right) = \tilde{E}\left(\boldsymbol{b}^{(m)}, \boldsymbol{\alpha}\right).$$

- Since $\left(\boldsymbol{b}^{(m)}, \boldsymbol{\alpha}^{(m)}\right)$ minimizes$\hat{E}\left(\boldsymbol{b}, \boldsymbol{\alpha}, \hat{u}^{(m-1)}\right)$ with respect to $(\boldsymbol{b}, \boldsymbol{\alpha})$, we see that $\boldsymbol{\alpha}^{(m)}$ is a minimum of $\tilde{E}\left(\boldsymbol{b}^{(m)}, \boldsymbol{\alpha}\right)$ with respect to $\boldsymbol{\alpha}$. In general, it does not appear to be possible to ensure that $\tilde{E}(b, \alpha)$ is also minimized with respect to b but in practice the distinction between $\hat{E}_{\mathrm{M}}$ and E_{M} appears to have negligible effect.

Table of Contents

Notion

We model the unknown concentrations of the s th component of the dynamical system at time t with a linear combination of kernels $k(.,.)$ from some function family $\mathcal{F}$:

$$g_s(t; \boldsymbol{b}_s) = \sum_{j=1}^{n} b_{sj} k(t, t_j)$$

We denote by b_s the vector of kernel regression coefficients b_{sk} and define $\boldsymbol{B} = \left(\boldsymbol{b}_1^\top, \ldots, b_r^\top\right)$, where r denotes the total number of components in the system.

Objective function

We estimate B along with θ by minimisation of the following objective function

$$E(\boldsymbol{\theta}, \boldsymbol{B}) = \sum_{s=1}^{r} \left(\sum_{i=1}^{n} \left[g_s\left(t_i; \boldsymbol{b}_s\right) - y_{si} \right]^2 \right) + \rho \sum_{s=1}^{r} \left(\sum_{i=1}^{n} \left[\dot{g}_s\left(t_i; \boldsymbol{b}_s\right) - f_s\left(\boldsymbol{g}\left(t_i, \boldsymbol{B}\right), \boldsymbol{\theta}\right) \right]^2 \right)$$

where $\boldsymbol{g}\left(t_i, \boldsymbol{B}\right) = \left(g_1\left(t_i; \boldsymbol{b}_1\right), \ldots, g_r\left(t_i; \boldsymbol{b}_r\right)\right)^\top$, and $\rho \geq 0$ is a regularization parameter.

The gradient matching term

$$\dot{g}_s(t; b_s) = \sum_{i=1}^{n} b_{si} \frac{dk(t, t_i)}{dt} = \sum_{i=1}^{n} b_{si} \dot{k}(t, t_i)$$

We need to solve this inference problem

$$\{\hat{\boldsymbol{\theta}}, \hat{B}\} = \text{argmin}_{\boldsymbol{\theta}, \boldsymbol{B}} E(\boldsymbol{\theta}, \boldsymbol{B})$$

Step 1

Initialisation of regression parameters and optimisation of kernel hyperparameters

- Regularised loss function

$$\mathcal{L}\left(\boldsymbol{b}_s, \boldsymbol{\varphi}_s; \lambda_s\right) = \sum_{i=1}^{n} \left(g_s\left(t_i; \boldsymbol{b}_s\right) - y_{si}\right)^2 + \left\|\boldsymbol{g}_s\right\|^2$$

  where the dependence on φ_s is via k_s (which has not been made explicit in the notation), and the regularisation term $\left\|\boldsymbol{g}_s\right\|^2$ is the squared norm in $\mathcal{H}_s$, $\left\|\boldsymbol{g}_s\right\|^2 = \lambda_s b_s^T \boldsymbol{K}_s b_s$, which contains regularisation parameter $\lambda_s \geq 0$, and K_s is the Gram matrix with entries $k_s\left(t_k, t_i\right)$.

- The minimisation of $\mathcal{L}\left(b_s, \varphi_s; \lambda_s\right)$ with respect to b_s for given φ_s and λ_s is a convex optimisation problem with solution

$$b_s = \left(\boldsymbol{K}_s + \lambda_s \boldsymbol{I}\right)^{-1} \boldsymbol{y}_s$$

Step 2

Initialisation of ODE parameters using gradient matching

Setting B fixed at the values obtained from Step 1 , the ODE parameters θ are obtimised by minimising the objective function E using a standard optimisation routine (e.g. trust region or quasi Newton).

$$E(\boldsymbol{\theta}, \boldsymbol{B}) = \sum_{s=1}^{r} \left(\sum_{i=1}^{n} \left[g_s\left(t_i; \boldsymbol{b}_s\right) - y_{si} \right]^2 \right) + \rho \sum_{s=1}^{r} \left(\sum_{i=1}^{n} \left[\dot{g}_s\left(t_i; \boldsymbol{b}_s\right) - f_s\left(\boldsymbol{g}\left(t_i, \boldsymbol{B}\right), \boldsymbol{\theta}\right) \right]^2 \right)$$

Step 3

Minimisation of the combined objective function with convergence acceleration

- First, we define the following modified objective function

$$\tilde{E}(\boldsymbol{\theta}, \boldsymbol{B}, \tilde{\boldsymbol{B}}) = \sum_{s=1}^{r} \left(\sum_{i=1}^{n} \left[g_s\left(t_i; \boldsymbol{b}_s\right) - y_{si} \right]^2 \right)$$
$$+ \rho \sum_{s=1}^{r} \left(\sum_{i=1}^{n} \left[\dot{g}_s\left(t_i; \boldsymbol{b}_s\right) - f_s\left(\boldsymbol{g}\left(t_i, \tilde{\boldsymbol{B}}\right), \boldsymbol{\theta} \right) \right]^2 \right)$$

- Note that $\tilde{E}(\boldsymbol{\theta}, \boldsymbol{B}, \boldsymbol{B}) = E(\boldsymbol{\theta}, \boldsymbol{B})$.

Step 3

We now carry out the following iteration until reaching a zero-gradient point:

- Given B and θ, minimize $\tilde{E}(\theta, B^*, B)$ with respect to B^*, i.e. find $B_{\text{new}} = \text{argmin}_{B^*} \tilde{E}(\theta, B^*, B)$
- Set $\boldsymbol{B} = \boldsymbol{B}_{\text{new}}$ and minimise $\tilde{E}(\boldsymbol{\theta}, \boldsymbol{B}_{\text{new}}, \boldsymbol{B}_{\text{new}})$ wrt $\boldsymbol{\theta}$, i.e. find $\boldsymbol{\theta}_{new} = \text{argmin}_{\theta} \tilde{E}(\boldsymbol{\theta}, \boldsymbol{B}_{\text{new}}, \boldsymbol{B}_{\text{new}})$

Theorem

Theorem 1

Let $\mathcal{G}$ denote the set of functions defined by equation (12). Assume $\mathcal{G}$ is contained in the solution space of the ODEs in the sense that $\forall g \in \mathcal{G} \exists \theta \in \mathbb{R} : \dot{g} = f(g, \theta)$. Then each parameter adaptation step of the algorithm described above, $(\boldsymbol{B}, \boldsymbol{\theta}) \rightarrow (\boldsymbol{B}_{new}, \theta_{new})$, implies that $E(\boldsymbol{\theta}_{new}, \boldsymbol{B}_{new}) \leq E(\boldsymbol{\theta}, \boldsymbol{B})$, and the iteration converges to a zero gradient point of $E(\boldsymbol{\theta}, \boldsymbol{B})$.

Proof

Proof. The first step of the algorithm implies that

$$\tilde{E}(\boldsymbol{\theta}, \boldsymbol{B}_{\text{new}}, \boldsymbol{B}) \leq \tilde{E}(\boldsymbol{\theta}, \boldsymbol{B}, \boldsymbol{B}) = E(\boldsymbol{\theta}, \boldsymbol{B}) \quad (18)$$

For the second step, note that $\exists \theta^* \in \mathbb{R}$ such that $\dot{g}(t, \boldsymbol{B}_{\text{new}}) = f(g(t, \boldsymbol{B}_{\text{new}}), \boldsymbol{\theta}^*)$, by assumption of the theorem. This implies that

$$\|\dot{\boldsymbol{g}}(t, \boldsymbol{B}_{\text{new}}) - f(\boldsymbol{g}(t, \boldsymbol{B}_{\text{new}}), \boldsymbol{\theta}^*)\|^2 = 0 \quad \forall t$$

Proof

Hence $\tilde{E}(\boldsymbol{\theta}, \boldsymbol{B}_{\text{new}}, \boldsymbol{B}) =$

$$= \sum_{i=1}^{N} \left\{ \|\boldsymbol{y}(t_i) - \boldsymbol{g}(t_i, \boldsymbol{B}_{new})\|^2 + \right.$$

$$\left. \rho \|\dot{\mathrm{g}}(t_i, \boldsymbol{B}_{\text{new}}) - f[\boldsymbol{g}(t_i, \boldsymbol{B}), \boldsymbol{\theta}]\|^2 \right\}$$

$$\geq \sum_{i=1}^{N} \left\{ \|\boldsymbol{y}(t_i) - \boldsymbol{g}(t_i, \boldsymbol{B}_{\text{new}})\|^2 \right\} + 0$$

$$= \sum_{i=1}^{N} \left\{ \|\boldsymbol{y}(t_i) - \boldsymbol{g}(t_i, \boldsymbol{B}_{\text{new}})\|^2 + \right.$$

$$\left. \rho \|\dot{\mathrm{g}}(t_i, \boldsymbol{B}_{\text{new}}) - f[\boldsymbol{g}(t_i, \boldsymbol{B}_{new}), \boldsymbol{\theta}^*]\|^2 \right\}$$

$$= \tilde{E}(\boldsymbol{\theta}^*, \boldsymbol{B}_{\text{new}}, \boldsymbol{B}_{\text{new}})$$

This implies that on completion of the second step of the algorithm, $\boldsymbol{\theta}_{\text{new}} = \text{argmin}_{\theta} \tilde{E}(\boldsymbol{\theta}, \boldsymbol{B}_{\text{new}}, \boldsymbol{B}_{\text{new}})$

Proof

We have $\tilde{E}(\boldsymbol{\theta}_{\text{new}}, \boldsymbol{B}_{\text{new}}, \boldsymbol{B}_{\text{new}}) = \tilde{E}(\boldsymbol{\theta}^*, \boldsymbol{B}_{\text{new}}, \boldsymbol{B}_{\text{new}})$, and hence

$$\begin{aligned} E(\boldsymbol{\theta}_{\text{new}}, \boldsymbol{B}_{\text{new}}) &= \tilde{E}(\boldsymbol{\theta}_{\text{new}}, \boldsymbol{B}_{\text{new}}, \boldsymbol{B}_{\text{new}}) \\ &= \tilde{E}(\boldsymbol{\theta}^*, \boldsymbol{B}_{\text{new}}, \boldsymbol{B}_{\text{new}}) \leq \tilde{E}(\boldsymbol{\theta}, \boldsymbol{B}_{\text{new}}, \boldsymbol{B}) \quad (19) \end{aligned}$$

Combining equations (18) and (19), we get:

$$E(\theta_{\text{new}}, \boldsymbol{B}_{\text{new}}) \leq \tilde{E}(\boldsymbol{\theta}, \boldsymbol{B}_{\text{new}}, \boldsymbol{B}) \leq E(\boldsymbol{\theta}, \boldsymbol{B})$$

which completes the first part of the proof. The iteration is continued until $\nabla_{\boldsymbol{\theta}} \tilde{E}(\boldsymbol{\theta}, \boldsymbol{B}, \boldsymbol{B}) = \nabla_{\boldsymbol{B}} \tilde{E}(\boldsymbol{\theta}, \boldsymbol{B}, \boldsymbol{B}) = 0$. Since $\tilde{E}(\boldsymbol{\theta}, \boldsymbol{B}, \boldsymbol{B}) = E(\boldsymbol{\theta}, \boldsymbol{B})$, this is a zero-gradient point of the original objective function $E(\theta, B)$.

Computational costs

Table 1. **Comparison of computational costs.** The table shows the computational costs for a single iteration of two optimisation algorithms compared in our study, and an alternative method discussed in Section 2, using the data generated from eq. (20).

Method	CPU time
Direct minimisation of $E(\boldsymbol{\theta}, \boldsymbol{B})$	$599.8s$
Proposed acceleration algorithm (RKG3)	$6.7s$
RKHS method by Gonzalez et al. (GON)	$4.2s$

Reference

- **James Ramsay, Giles Hooker**: "Dynamic Data Analysis - Modeling Data with Differential Equations"
- **Mu Niu, Simon Rogers, Maurizio Filippone, Dirk Husmeier**: "Fast Inference in Nonlinear Dynamical Systems using Gradient Matching"
- **J. O. Ramsay, G. Hooker, D. Campbell and J. Cao** : "Parameter estimation for differential equations: a generalized smoothing approach"
- **Daniel Brewer, Martino Barenco**: "Fitting ordinary differential equations to short time course data"

END

Thank you very much for your valuable time!