Theorems and Calculations for Smoothing-based Profiled Estimation of Differential Equations

G. Hooker

Abstract

This report provides proofs of some theorems appearing in [Ramsay *et. al.* 2007] and also provides some of the calculations necessary to carry out the described procedure.

1 Introduction

[Ramsay *et. al.* 2007] details a method of estimating parameters for ordinary differential equations using smoothing spline technology. ODEs represent processes that transforms a set of m input functions $\mathbf{u}(t)$ into a set of d output functions $\mathbf{x}(t)$. Dynamic systems model output change directly by linking the output derivatives $\dot{\mathbf{x}}(t)$ to $\mathbf{x}(t)$ itself, as well as to inputs $\mathbf{u}$.

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}, \mathbf{u}, t|\boldsymbol{\theta}), \quad t \in [0, T]. \tag{1}$$

Vector $\boldsymbol{\theta}$ contains any parameters defining the system whose values are not known from experimental data, theoretical considerations or other sources of information. The task is then to estimate $\boldsymbol{\theta}$ from noisily observed data. In general, we assume that only some subset of the components of $\mathbf{x}$ have been measured and we denote the set of indices of these by $\mathcal{I}$ with associated measurements $\mathbf{y}_i$ taken at times $\mathbf{t}_i$.

Unfortunately, explicit solutions to (1) are rarely available and must be approximated numerically. Moreover, the fitting surfaces tend to be very rough and direct optimization methods tend to frequently find local minima. [Ramsay *et. al.* 2007] attempts to ameliorate both these problems.

The approach in [Ramsay *et. al.* 2007] belongs in the family of *collocation* methods that express the approximation $\hat{x}_i$ of x_i in terms a basis function expansion

$$\hat{x}_i(t) = \sum_k^{K_i} c_{ik}\phi_{ik}(t) = \mathbf{c}_i'\boldsymbol{\phi}_i(t), \tag{2}$$

where the number K_i of basis functions in vector $\boldsymbol{\phi}_i$ is chosen so as to ensure enough flexibility to capture the variation in the approximated function x_i and its derivatives.

The task is now the joint estimation of $\mathbf{c}$ and $\boldsymbol{\theta}$. This is done in a two-stage process, in the *inner optimization*, the $\mathbf{c}$ are chosen by minimizing the criterion

$$J(\mathbf{c}|\boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{\lambda}) = -\sum_{i \in \mathcal{I}} \ln g(\mathbf{y}_i|\mathbf{c}_i, \boldsymbol{\sigma}_i) + P(\hat{\mathbf{x}}|\boldsymbol{\theta}, \boldsymbol{\lambda}), \tag{3}$$

where g_i represents the likelihood for the observation $\mathbf{y}_i$ given $\hat{\mathbf{x}}_i(\mathbf{t}_i) = \mathbf{c}_i'\boldsymbol{\phi}_i(\mathbf{t}_i)$ and $\boldsymbol{\sigma}_i$ are (known) parameters defining g_i. We take a nonlinear penalty

$$P(\hat{\mathbf{x}}|\boldsymbol{\theta}, \boldsymbol{\lambda}) = \sum_{i=1}^{d} \lambda_i \int \left(\frac{d}{dt} x_i(t) - f_i(\mathbf{x}|\boldsymbol{\theta}, \mathbf{u}) \right)^2 dt \tag{4}$$

which explicitly controls the extent to which $\mathbf{x}$ may deviate from a solution to (1). This gives $\mathbf{c}(\boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$ which is then chosen to minimize:

$$H(\boldsymbol{\theta}, \boldsymbol{\sigma}|\boldsymbol{\lambda}) = -\sum_{i \in \mathcal{I}} \ln g(\mathbf{y}_i|\mathbf{c}_i(\boldsymbol{\theta}), \boldsymbol{\sigma}_i) \tag{5}$$

this is called the *outer optimization*. In practise, it is common to assume an uncorrelated error structure for the $\mathbf{y}_i$, leading to the error sum of squares criterion

$$g_i(\mathbf{y}_i|\hat{\mathbf{x}}_i, \boldsymbol{\sigma}_i) = -w_i \|\mathbf{y}_i - \hat{x}_i(\mathbf{t}_i)\|^2. \tag{6}$$

For the purposes of this report, we will take $\lambda_i = \lambda$ to be constant so that we can write

$$J(\mathbf{c}|\boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{\lambda}) = l(\mathbf{x}) + \lambda P(\mathbf{x}|\boldsymbol{\theta})$$

with

$$H(\boldsymbol{\theta}, \boldsymbol{\sigma}|\boldsymbol{\lambda}) = l(\mathbf{x})$$

for some l dependent on the data and $\boldsymbol{\sigma}$. This report examines the behavior of our estimate of $\boldsymbol{\theta}$ as λ is allowed to increase.

2 Theorems and Proofs

This theorem states and proves theorems appearing in [Ramsay *et. al.* 2007]. The essential import of these is that as λ increases, the parameter estimates we get tend to those that would have been gotten by optimizing exact solutions to (1).

We will assume that solutions to the inner optimization problem exist and are well defined, and therefore that there are objects $\mathbf{x}$ that satisfy $P(\mathbf{x}|\boldsymbol{\theta}) = 0$. This is guaranteed locally by the following theorem adapted from [Bellman, 1953]:

Theorem 2.1. *Let* $\mathbf{f}$ *be Lipschitz continuous and* $\mathbf{u}$ *differentiable almost everywhere, then the initial value problem:*

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}, \mathbf{u}, t|\theta), \ \mathbf{x}(t_0) = \mathbf{x}_0$$

has a unique solution.

2.1 Preliminaries

The following theorem is a well-known consequence of the method of Lagrange multipliers:

Theorem 2.2. *Suppose that* x_λ *minimizes* $F(x) + \lambda P(x)$*, then* x_λ *minimizes* $F(z)$ *for* $z \in \{x : P(x) < P(x_\lambda)\}$*. Moreover, for* $\lambda' > \lambda$*,* $P(x_{\lambda'}) \leq P(x_\lambda)$*.*

Two corollaries:

Corollary 2.1. *For* $\lambda' > \lambda$*,* $F(x_{\lambda'}) \geq F(x_\lambda)$*.*

Corollary 2.2. *If* $\exists x$ *such that* $P(x) = 0$*, then* $P(x_\lambda) \to 0$ *as* $\lambda \to \infty$*.*

follow immediately.

The proofs of Theorems 2.4 and 2.5 rely heavily on the following:

Theorem 2.3. *Let* $\mathcal{X}$ *and* $\mathcal{Y}$ *be metric spaces with* $\mathcal{X}$ *closed and bounded. Let* $g(x, \alpha) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ *be uniformly continuous in* x *and* α*, such that*

$$x(\alpha) = \operatorname*{argmin}_{x \in \mathcal{X}} g(x, \alpha)$$

is well defined for each α*. Then* $x(\alpha) : \mathcal{Y} \to \mathcal{X}$ *is continuous.*

We begin with two lemmas:

Lemma 2.1. *Let* $\mathcal{X}$ *be a closed and bounded metric space. Suppose that*

$$x^* = \operatorname*{argmin}_{x \in \mathcal{X}} g(x) \tag{7}$$

is well defined and $g(x)$ *is continuous. Then*

$$\forall \epsilon > 0, \ \exists \delta > 0 \text{ such that } \|x - x^*\| > \epsilon \Rightarrow f(x) - f(x^*) > \delta.$$

holds for all $x \in \mathcal{X}$*.*

Proof. Assume that the the statement is not true. That is, for some $\epsilon > 0$ we can find a sequence $x_n \in \mathcal{X}$ such that $\|x_n - x^*\| > \epsilon$ but $\|g(x_n) - g(x^*)\| < 1/n$. Since $\mathcal{X}$ is closed and bounded, it is compact and there exists a subsequence $x_{n'} \to x^{**} \neq x^*$ for some x^{**}. By the continuity of g, we have $g(x^{**}) = g(x^*)$ violating the assumption that (7) is well defined. □

Lemma 2.2. *Let $\mathcal{X}$ and $\mathcal{Y}$ be metric spaces and $g(x,\alpha) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be bounded below and uniformly continuous in α and x, then $j(\alpha) = min_{x\in\mathcal{X}} g(x,\alpha)$ is a continuous function.*

Proof. Assume $j(\alpha)$ is not continuous: that is, for some $\alpha \in \mathcal{Y}$, $\exists\epsilon > 0$ such that $\forall\delta > 0$, $\exists\alpha'$ with $|\alpha' - \alpha| < \delta$ and $|j(\alpha) - j(\alpha')| > \epsilon$.

By the uniformity of g in α across x, we can choose $\delta' > 0$ so that $|g(x,\alpha) - g(x,\alpha')| < \epsilon/3$ for all x when $|\alpha - \alpha'| < \delta'$. By assumption, we can find some such α' so that $|j(\alpha) - j(\alpha')| > \epsilon$. Without loss of generality, let $j(\alpha) < j(\alpha')$.

Now, choose $x \in \mathcal{X}$ so that $g(x,\alpha) < j(\alpha) + \epsilon/3$. Then $g(x,\alpha') < j(\alpha) + 2\epsilon/3 < j(\alpha')$, contradicting $j(\alpha') = \min_{x\in\mathcal{X}} g(x,\alpha)$. □

Using these, we can now prove Theorem 2.3:

Proof. Let $\epsilon > 0$, by Lemma 2.1 there exists $\delta' > 0$ such that

$$g(x,\alpha) - g(x(\alpha),\alpha) < \delta' \Rightarrow \|x - x(\alpha)\| < \epsilon.$$

By Lemma 2.2, $j(\alpha)$ is continuous. Since $g(x,\alpha)$ is uniformly continuous , we can choose δ so that

$$|\alpha - \alpha'| < \delta \to |j(\alpha) - j(\alpha')| < \delta'/3 \text{ and } \forall x, |g(x,\alpha) - g(x,\alpha')| < \delta'/3$$

giving

$$\begin{aligned} |g(x(\alpha),\alpha) - g(x(\alpha'),\alpha)| &< |g(x(\alpha),\alpha) - g(x(\alpha'),\alpha')| + |g(x(\alpha'),\alpha') - g(x(\alpha'),\alpha)| \\ &= |j(\alpha) - j(\alpha')| + |g(x(\alpha'),\alpha') - g(x(\alpha'),\alpha)| \\ &< \delta/3 + \delta/3 \\ &< \delta \end{aligned}$$

from which we conclude $\|x(\alpha) - x(\alpha')\| < \epsilon$. □

2.2 The inner optimization

Theorem 2.4. *Let $\lambda_k \to \infty$ and assume that*

$$\mathbf{x}_k = \underset{\mathbf{x}\in(W^1)^n}{\operatorname{argmin}}\ l(\mathbf{x}) + \lambda_k P(\mathbf{x}|\boldsymbol{\theta})$$

is well defined and uniformly bounded over λ. Then $\mathbf{x}_k$ converges to $\mathbf{x}^*$ with $P(\mathbf{x}^*|\boldsymbol{\theta}) = 0$.

Proof. We first note that we can re-express $\mathbf{x}_k$ as

$$\mathbf{x}_k = \underset{\mathbf{x}\in(W^1)^n}{\operatorname{argmin}}\ (1-\alpha_k)l(\mathbf{x}) + \alpha_k P(\mathbf{x}_k|\boldsymbol{\theta}) \tag{8}$$

where $\alpha_k = \lambda_k/(1+\lambda_k) \to 1$.

By the continuity of point-wise evaluation in $(W^1)^n$, $l(\mathbf{x})$ is a continuous functional of $\mathbf{x}$ and $P(\mathbf{x}|\boldsymbol{\theta})$ is similarly continuous. Since the x_k lie in a bounded set $\mathcal{X}$, we have that

$$l(\mathbf{x}) < F^* \text{ and } P(\mathbf{x}|\boldsymbol{\theta}) < P^*$$

for all $\mathbf{x} \in \mathcal{X}$. Both $l(\mathbf{x})$ and $P(\mathbf{x}|\boldsymbol{\theta})$ are bounded below by 0 and we note that

$$g(\mathbf{x}, \alpha) = (1-\alpha)l(\mathbf{x}) + \alpha P(\mathbf{x}|\boldsymbol{\theta})$$

is uniformly bounded on $\mathcal{C}$ by 0 and $F^* + P^*$ and is therefore uniformly continuous in α and $\mathbf{x}$.

By Theorem 2.3,

$$\mathbf{x}(\alpha) = \underset{\mathbf{x}\in\mathcal{C}}{\operatorname{argmin}}\, g(\mathbf{x}, \alpha)$$

is a continuous function from $(0, 1)$ to $(W^1)^n$. Since $\|x(\alpha)\|$ is bounded by assumption, it is uniformly continuous. Since $\alpha_n \to 1$ is convergent, we must have that $\mathbf{x}_n = \mathbf{x}(\alpha_n) \to \mathbf{x}*$. By the continuity of $P(\mathbf{x}|\boldsymbol{\theta})$, $P(\mathbf{x}^*|\boldsymbol{\theta}) = 0$. □

Note that if it were possible to define $\mathbf{x}(\alpha)$ as a continuous function on $[0, 1]$, the need for a bound on $||\mathbf{x}(\alpha)||$ would be removed. However, since we do not expect $g(\mathbf{x}, 1) = P(\mathbf{x}|\boldsymbol{\theta})$ to have a well-defined minimum, boundedness is required to ensure that $\mathbf{x}(\alpha)$ has a limit as $\alpha \to 1$.

We can now go further when $P(\mathbf{x}|\boldsymbol{\theta})$ is given by (4) by specifying that $\mathbf{x}^*$ is the solution of the differential equations (1) that is obtained by minimizing squared error over the choice of initial conditions. To see this, we observe that Theorem 2.1 ensures that

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}, \mathbf{u}, t|\boldsymbol{\theta}).$$

with

$$\mathbf{x}(t_0) = \mathbf{x}_0$$

specifies a unique element of $(W^1)^n$. Let

$$\mathcal{F} = \{\mathbf{x}, P(\mathbf{x}|\boldsymbol{\theta}) = 0\},$$

then

$$\lim_{k\to\infty} l(\mathbf{x}_n) \leq \min_{\mathbf{x}\in\mathcal{F}} l(\mathbf{x}).$$

Since l is a continuous functional on $(W^1)^n$, and $P(\mathbf{x}^*|\boldsymbol{\theta})0$, we must have

$$l(\mathbf{x}^*) = \min_{\mathbf{x}\in\mathcal{F}} l(\mathbf{x}).$$

By the assumption that the solutions to (8) are well defined and bounded, this specifies a unique set of initial conditions $\mathbf{x}_0^*$ such that

$$\dot{\mathbf{x}}^*(t) = \mathbf{f}(\mathbf{x}^*, \mathbf{u}, t|\boldsymbol{\theta}).$$

with

$$\mathbf{x}^*(t_0) = \mathbf{x}_0^*.$$

2.3 The outer optimization

Theorem 2.5. *Let $\mathcal{X} \subset (W^1)^n$ and $\Theta \subset \mathbb{R}^p$ be bounded. Let*

$$\mathbf{x}_{\boldsymbol{\theta},\lambda} = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{X}} l(\mathbf{x}) + \lambda P(\mathbf{x}|\boldsymbol{\theta})$$

be well defined for each $\boldsymbol{\theta}$ and λ, define $\mathbf{x}^*_{\boldsymbol{\theta}}$ to be such that

$$l(\mathbf{x}^*_{\boldsymbol{\theta}}) = \min_{\mathbf{x}:P(\mathbf{x}|\boldsymbol{\theta})=0} l(\mathbf{x})$$

and let

$$\boldsymbol{\theta}(\lambda) = \operatorname*{argmin}_{\boldsymbol{\theta}\in\Theta} l(\mathbf{x}_{\boldsymbol{\theta},\lambda}) \text{ and } \boldsymbol{\theta}^* = \operatorname*{argmin}_{\boldsymbol{\theta}\in\Theta} l(\mathbf{x}^*_{\boldsymbol{\theta}})$$

also be well defined for each λ. Then

$$\lim_{\lambda\to\infty} \boldsymbol{\theta}(\lambda) = \boldsymbol{\theta}^*$$

Proof. The proof is very similar to that of Theorem 2.4. Setting $\alpha = \lambda/(1+\lambda)$

$$g(\mathbf{x}, \alpha, \boldsymbol{\theta}) = (1-\alpha)l(\mathbf{x}) + \alpha P(\mathbf{x}|\boldsymbol{\theta})$$

is uniformly continuous in α, $\boldsymbol{\theta}$ and $\mathbf{x}$. As observed in Theorem 2.4, $\mathbf{x}_{\boldsymbol{\theta},\lambda}$ can be equivalently written as

$$\mathbf{x}_{\boldsymbol{\theta},\alpha} = \operatorname*{argmin}_{\mathbf{x}\in(W^1)^k} g(\mathbf{x}, \alpha, \boldsymbol{\theta}).$$

with $\alpha\lambda/(1+\lambda)$. By Theorem 2.3, $\mathbf{x}_{\boldsymbol{\theta},\alpha}$ is continuous in $\boldsymbol{\theta}$ and α. On the set $\mathcal{X}$, therefore, $l(\mathbf{x})$ is uniformly continuous in $\mathbf{x}$ and $\mathbf{x}_{\boldsymbol{\theta},\alpha}$ is uniformly continuous in $\boldsymbol{\theta}$ and α. $l(\mathbf{x}_{\boldsymbol{\theta},\alpha})$ is therefore uniformly continuous in $\boldsymbol{\theta}$ and α. Under the assumption that $\boldsymbol{\theta}(\alpha)$ is well defined for each α, we can now employ Theorem 2.3 again to give us that $\theta(\alpha)$ is continuous in α and the boundedness of Θ provides uniform continuity.
Assume that

$$\tilde{\boldsymbol{\theta}} = \lim_{\alpha\to 1} \boldsymbol{\theta}(\alpha) \neq \boldsymbol{\theta}^*$$

and in particular $\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| > \epsilon$. From Lemma 2.1 there must exist a $\delta > 0$ such that

$$l(\mathbf{x}^*_{\boldsymbol{\theta}^*}) < l(\mathbf{x}^*_{\boldsymbol{\theta}}) - \delta.$$

for all $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| > \epsilon/2$. Since $\boldsymbol{\theta}(\alpha)$ is uniformly continuous in α, there is some a such that $\|\boldsymbol{\theta}(\alpha) - \boldsymbol{\theta}^*\| > \epsilon/2$ for all $\alpha > a$. Now by the uniform continuity of $l(\mathbf{x}_{\boldsymbol{\theta},\alpha})$ in α and $\boldsymbol{\theta}$, we can choose $a_1 > a$ so that

$$\left| l(\mathbf{x}_{\boldsymbol{\theta}(\alpha),\alpha}) - l(\mathbf{x}^*_{\boldsymbol{\theta}}) \right| < \delta/3$$

for all $\alpha > a_1$. By the same uniform continuity, we can choose $\alpha > a_1$ so that

$$|l(\mathbf{x}_{\boldsymbol{\theta}^*,\alpha}) - l(\mathbf{x}^*_{\boldsymbol{\theta}^*})| < \delta/2$$

giving

$$l(\mathbf{x}_{\boldsymbol{\theta}^*,\alpha}) < l(\mathbf{x}_{\boldsymbol{\theta}(\alpha),\alpha})$$

contradicting the definition of $\boldsymbol{\theta}(\alpha)$. Finally, note that α is also uniformly continuous in λ and $\lim_{\lambda\to\infty} \alpha(\lambda) = 1$. □

3 Matrix calculations for profiling

The calculations used throughout [Ramsay *et. al.* 2007] have been based on matrices defined in terms of derivatives of J and H with respect to $\boldsymbol{\theta}$ and $\mathbf{c}$. In many cases, these matrices are non-trivial to calculate and expressions for their entries are derived here. For these calculations, we have assumed that the outer criterion, H is a straight-forward weighted sum of squared errors and only depends on $\boldsymbol{\theta}$ through $\mathbf{x}$.

3.1 Inner optimization

Using a Gauss-Newton method, we require the derivative of the fit at each observation point:

$$\frac{dx_i(t_{i,k})}{d\mathbf{c}_i} = \Phi_i(t_{i,k})$$

where $\Phi_i(t_{i,k})$ is the vector corresponding to the evaluation of all the basis functions used to represent x_i evaluated at $t_{i,k}$. This gradient of x_i with respect to $\mathbf{c}_j$ is zero.

A numerical quadrature rule allows the set of errors to be augmented with the evaluation of the penalty at the quadrature points and weighted by the quadrature rule:

$$(\lambda_i v_q)^{1/2} \left(\dot{x}_i(t_q) - f_i(\mathbf{x}(t_q), \mathbf{u}(t_q), t_q|\boldsymbol{\theta})\right)$$

Each of these then has derivative with respect to $\mathbf{c}_j$:

$$\begin{aligned}(\lambda_i v_q)^{1/2} \left(\dot{x}_i(t_q) - f_i(\mathbf{x}(t_q), \mathbf{u}(t_q), t_q|\boldsymbol{\theta})\right) I(i=j) D\Phi_i(t_q) \\ - \left(\sum_{k=1}^{n} (\lambda_i v_q)^{1/2} \frac{df_k}{dx_j} \left(\dot{x}_i(t_q) - f_i(\mathbf{x}(t_q), \mathbf{u}(t_q), t_q|\boldsymbol{\theta})\right)\right) \Phi_j(t_q)\end{aligned}$$

and the augmented errors and gradients can be used in a Gauss-Newton scheme. $I()$ is used as the indicator function of its argument.

3.2 Outer optimization

As in the inner optimization, in employing a Gauss-Newton scheme, we merely need to write a gradient for the point-wise fit with respect to the parameters:

$$\frac{d\mathbf{x}(t_{i,k})}{d\boldsymbol{\theta}} = \frac{d\mathbf{x}(t_{i,k})}{d\mathbf{c}}\frac{d\mathbf{c}}{d\boldsymbol{\theta}}$$

where $d\mathbf{x}(t_i)/d\mathbf{c}$ has already be calculated and

$$\frac{d\mathbf{c}}{d\boldsymbol{\theta}} = -\left[\frac{d^2J}{d\mathbf{c}^2}\right]^{-1}\frac{d^2J}{d\mathbf{c}d\boldsymbol{\theta}}$$

by the implicit function theorem.

Hessian matrix $d^2J/d\mathbf{c}^2$ may be expressed as a block form, the (i,j)th block corresponding to the cross-derivatives of the coefficients in the ith and jth components of $\mathbf{x}$. This block's (p,q)th entry is given by:

$$\begin{aligned}\left(\sum_{k=1}^{n_i}\phi_{ip}(t_{i,k})\phi_{jq}(t_{i,k}) + \lambda\int\phi_{ip}(t)\phi_{jq}(t)dt\right)I(i=j)\\ -\lambda_i\int\dot{\phi}_{ip}(t)\frac{df_i}{dx_j}\phi_{jq}(t)dt - \lambda_j\int\phi_{ip}(t)\frac{df_i}{dx_j}\dot{\phi}_{jq}(t)dt\\ +\int\phi_{ip}(t)\left[\sum_{k=1}^{n}\lambda_k\left(\frac{d^2f_k}{dx_idx_j}(f_k-\dot{x}_k(t)) + \frac{df_k}{dx_i}\frac{df_k}{dx_j}\right)\right]\phi_{jq}(t)dt\end{aligned}$$

with the integrals evaluated by numeric integration. The arguments to $f_k(\mathbf{x},\mathbf{u},t|\boldsymbol{\theta})$ have been dropped in the interests of notational legibility.

We can similarly express the cross-derivatives $d^2J/d\mathbf{c}d\boldsymbol{\theta}$ as a block vector, the ith block corresponding to the coefficients in the basis expansion for the ith component of $\mathbf{x}$. The pth entry of this block can now be expressed as:

$$\lambda_i\int\frac{df_i}{d\boldsymbol{\theta}}\phi_{ip}(t)dt - \int\left(\sum_{k=1}^{n}\lambda_k\left[\frac{d^2f_k}{dx_id\boldsymbol{\theta}}(f_k-\dot{x}_k(t)) + \frac{df_k}{dx_i}\frac{df_k}{d\boldsymbol{\theta}}\right]\right)\phi_{i,p}(t)dt$$

3.3 Estimating the variance of $\hat{\boldsymbol{\theta}}$

The variance of the parameter estimates is calculated using

$$\frac{d\hat{\boldsymbol{\theta}}}{d\mathbf{y}} = -\left[\frac{d^2H}{d\boldsymbol{\theta}^2}\right]^{-1}\frac{d^2H}{d\boldsymbol{\theta}d\mathbf{y}},$$

where

$$\frac{d^2H}{d\boldsymbol{\theta}^2} = \frac{\partial^2H}{\partial\boldsymbol{\theta}^2} + 2\frac{\partial^2H}{\partial\hat{\mathbf{c}}\partial\boldsymbol{\theta}}\frac{\partial\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}} + \left(\frac{\partial\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}}\right)'\frac{\partial^2H}{\partial\hat{\mathbf{c}}^2}\frac{\partial\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}} + \frac{\partial H}{\partial\hat{\mathbf{c}}}\frac{\partial^2\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}^2}, \tag{9}$$

and

$$\frac{d^2H}{d\boldsymbol{\theta}d\mathbf{y}} = \frac{\partial^2H}{\partial\boldsymbol{\theta}\partial\mathbf{y}} + \frac{\partial^2H}{\partial\hat{\mathbf{c}}\partial\mathbf{y}}\frac{\partial\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}} + \frac{\partial^2H}{\partial\boldsymbol{\theta}\partial\hat{\mathbf{c}}}\frac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}} + \frac{\partial^2H}{\partial\hat{\mathbf{c}}^2}\frac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}}\frac{\partial\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}} + \frac{\partial H}{\partial\hat{\mathbf{c}}}\frac{\partial^2\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}\partial\mathbf{y}}. \tag{10}$$

The formulas (9) and (10) for $d^2H/d\boldsymbol{\theta}^2$ and $d^2H/d\boldsymbol{\theta}d\mathbf{y}$ involve the terms $\partial\hat{\mathbf{c}}/\partial\mathbf{y}$, $\partial^2\hat{\mathbf{c}}/\partial\boldsymbol{\theta}^2$ and $\partial^2\hat{\mathbf{c}}/\partial\boldsymbol{\theta}\partial\mathbf{y}$. In the following, we derive their analytical formulas by the Implicit Function Theorem. We introduce the following convention, which is caller *Einstein Summation Notation*. If a Latin index is repeated in a term, then it is understood as a summation with respect to that index. For instance, instead of the expression $\sum_i a_i x_i$, we merely write $a_i x_i$.

- $\frac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}}$
  Similar as the deduction for $d\hat{\mathbf{c}}/d\boldsymbol{\theta}$, we obtain the formula for $\partial\hat{\mathbf{c}}/\partial\mathbf{y}$ by applying the Implicit Function Theorem:

$$\frac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}} - \left[\frac{\partial^2 J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial\mathbf{c}^2}\bigg|_{\hat{\mathbf{c}}}\right]^{-1}\left[\frac{\partial^2 J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial\mathbf{c}\partial\mathbf{y}}\bigg|_{\hat{\mathbf{c}}}\right]. \tag{11}$$

- $\frac{\partial\mathbf{c}^2}{\partial\boldsymbol{\theta}\partial\mathbf{y}}$
  By taking the second derivative on both sides of the identity $\partial J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})/\partial\mathbf{c}|_{\hat{\mathbf{c}}} = 0$ with respect to $\boldsymbol{\theta}$ and y_k, we derive:

$$\begin{aligned}
&\frac{d^2}{d\boldsymbol{\theta}dy_k}\left(\frac{\partial J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial\mathbf{c}}\bigg|_{\hat{\mathbf{c}}}\right)\frac{\partial^3 J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial\mathbf{c}\partial\boldsymbol{\theta}\partial y_k}\bigg|_{\hat{\mathbf{c}}} + \frac{\partial^3 J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial\mathbf{c}\partial\boldsymbol{\theta}\partial c_i}\bigg|_{\hat{\mathbf{c}}}\frac{\partial\hat{c}_i}{\partial y_k} \\
&+\frac{\partial^3 J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial\mathbf{c}^2\partial y_k}\bigg|_{\hat{\mathbf{c}}}\frac{\partial\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}} + \frac{\partial^3 J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial\mathbf{c}^2\partial c_i}\bigg|_{\hat{\mathbf{c}}}\frac{\partial\hat{c}_i}{\partial y_k}\frac{\partial\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}} + \frac{\partial^2 J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial\mathbf{c}^2}\bigg|_{\hat{\mathbf{c}}}\frac{\partial^2\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}\partial y_k} \\
&= 0
\end{aligned} \tag{12}$$

  Solving for $\frac{\partial^2\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}\partial y_k}$, we obtain the second derivative of $\hat{\mathbf{c}}$ with respect to $\boldsymbol{\theta}$ and y_k:

$$\begin{aligned}
\frac{\partial^2\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}\partial y_k} &= -\left[\frac{\partial^2 J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial\mathbf{c}^2}\bigg|_{\hat{\mathbf{c}}}\right]^{-1}\left[\frac{\partial^3 J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial\mathbf{c}\partial\boldsymbol{\theta}\partial y_k}\bigg|_{\hat{\mathbf{c}}} + \frac{\partial^3 J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial\mathbf{c}\partial\boldsymbol{\theta}\partial c_i}\bigg|_{\hat{\mathbf{c}}}\frac{\partial\hat{c}_i}{\partial y_k}\right. \\
&+ \left.\frac{\partial^3 J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial\mathbf{c}^2\partial y_k}\bigg|_{\hat{\mathbf{c}}}\frac{\partial\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}} + \frac{\partial^3 J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial\mathbf{c}^2\partial c_i}\bigg|_{\hat{\mathbf{c}}}\frac{\partial\hat{c}_i}{\partial y_k}\frac{\partial\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}}\right]
\end{aligned} \tag{13}$$

- $\frac{\partial^2\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}^2}$
  Similar to the deduction of $\partial^2\hat{\mathbf{c}}/\partial\boldsymbol{\theta}\partial y_k$, the second partial derivative of $\mathbf{c}$ with respect to $\boldsymbol{\theta}$ and θ_j is:

$$\begin{aligned}
\frac{\partial^2\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}\partial\theta_j} &= -\left[\frac{\partial^2 J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial\mathbf{c}^2}\bigg|_{\hat{\mathbf{c}}}\right]^{-1}\left[\frac{\partial^3 J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial\mathbf{c}\partial\boldsymbol{\theta}\partial\theta_j}\bigg|_{\hat{\mathbf{c}}} + \frac{\partial^3 J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial\mathbf{c}\partial\boldsymbol{\theta}\partial c_i}\bigg|_{\hat{\mathbf{c}}}\frac{\partial\hat{c}_i}{\partial\theta_j}\right. \\
&+ \left.\frac{\partial^3 J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial\mathbf{c}^2\partial\theta_j}\bigg|_{\hat{\mathbf{c}}}\frac{\partial\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}} + \frac{\partial^3 J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial\mathbf{c}^2\partial c_i}\bigg|_{\hat{\mathbf{c}}}\frac{\partial\hat{c}_i}{\partial\theta_j}\frac{\partial\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}}\right]
\end{aligned} \tag{14}$$

When estimating ODE's, we define $J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})$ as (3) and $H(\boldsymbol{\theta},\hat{\mathbf{c}}(\boldsymbol{\theta})|\mathbf{y})$ as (5), and further write the above formulas in terms of the basis functions in $\boldsymbol{\Phi}$ and the

functions $\mathbf{f}$ on the right side of the differential equation. For instance, $d^2J/d\mathbf{c}^2$ is a block-diagonal matrix with the ith block being $w_i\Phi_i(\mathbf{t}_i)^T\Phi_i(\mathbf{t}_i)$ and $dJ/d\mathbf{c}$ is a block vector containing blocs $-w_i\Phi_i(\mathbf{t}_i)^T(\mathbf{y}_i - x_i(\mathbf{t}_i))$.

The three-dimensional array $\partial^3 J/\partial\mathbf{c}\partial c_p\partial c_q$ can be written in the same block vector form as $\partial^2 J/\partial\mathbf{c}\partial\boldsymbol{\theta}$ with the uth entry of the kth block given by

$$\begin{aligned}
&\int\left(\sum_{l=1}^{n}\lambda_l\left[\frac{d^2f_l}{dx_idx_j}\frac{df_l}{dx_k}+\frac{d^2f_l}{dx_idx_k}\frac{df_l}{dx_j}+\frac{d^2f_l}{dx_jdx_k}\frac{df_l}{dx_i}\right]\right)\phi_{ip}(t)\phi_{jq}(t)\phi_{ku}(t)dt\\
&\quad+\int\sum_{l=1}^{n}\lambda_l\left(\frac{d^3f_k}{dx_idx_jdx_k}(f_l-\dot{x}_l(t))\right)\phi_{ip}(t)\phi_{jq}(t)\phi_{ku}(t)dt\\
&\quad-\lambda_i\int\frac{d^2f_i}{dx_jdx_k}\dot{\phi}_{ip}(t)\phi_{jq}(t)\phi_{ku}(t)dt-\lambda_j\int\frac{d^2f_j}{dx_idx_k}\phi_{ip}(t)\dot{\phi}_{jq}(t)\phi_{ku}(t)dt\\
&\qquad\qquad-\lambda_k\int\frac{d^2f_k}{dx_idx_j}\phi_{ip}(t)\phi_{jq}(t)\dot{\phi}_{ku}(t)dt
\end{aligned}$$

assuming c_p is a coefficient in the basis representation of x_i and c_q a corresponds to x_j. The array $\partial^3 J/\partial\mathbf{c}\partial\theta_i\partial\theta_j$ is also expressed in the same block form with entry p in the kth block being:

$$\begin{aligned}
&\int\left(\sum_{l=1}^{n}\lambda_l\left[\frac{d^2f_l}{d\theta_id\theta_j}\frac{df_l}{dx_k}+\frac{d^2f_l}{d\theta_idx_k}\frac{df_l}{d\theta_j}+\frac{d^2f_l}{d\theta_jdx_k}\frac{df_l}{d\theta_i}\right]\right)\phi_{kp}(t)dt\\
&\quad+\int\sum_{l=1}^{n}\lambda_l\left(\frac{d^3f_k}{dx_kd\theta_id\theta_j}(f_l-\dot{x}_l(t))\right)\phi_{kp}(t)dt-\lambda_k\int\frac{d^2f_k}{d\theta_id\theta_k}\phi_{kp}(t)dt.
\end{aligned}$$

$\partial^3 J/\partial\mathbf{c}\partial c_p\partial\theta_i$ is in the same block from, with the qth entry of the jth block being:

$$\begin{aligned}
&\int\left(\sum_{l=1}^{n}\lambda_l\left[\frac{d^2f_l}{d\theta_idx_j}\frac{df_l}{dx_k}+\frac{d^2f_l}{d\theta_idx_k}\frac{df_l}{dx_j}+\frac{d^2f_l}{dx_jdx_k}\frac{df_l}{d\theta_i}\right]\right)\phi_{kp}(t)\phi_{jq}(t)dt\\
&\quad+\int\sum_{l=1}^{n}\lambda_l\left(\frac{d^3f_k}{dx_jdx_kd\theta_i}(f_l-\dot{x}_l(t))\right)\phi_{kp}(t)\phi_{jq}(t)dt\\
&\quad-\lambda_j\int\frac{d^2f_j}{d\theta_idx_k}\dot{\phi}_{jq}(t)\phi_{kp}(t)dt-\lambda_k\int\frac{d^2f_k}{d\theta_idx_j}\phi_{jq}(t)\dot{\phi}_{kp}(t)dt
\end{aligned}$$

where c_p corresponds to the basis representation of x_k.

Similar calculations give matrix $d^2H/d\boldsymbol{\theta}d\mathbf{y}$ explicitly as:

$$\begin{aligned}
&\frac{d\hat{\mathbf{c}}}{d\boldsymbol{\theta}}^T\left[\frac{\partial^2H}{\partial\hat{\mathbf{c}}\partial\mathbf{y}}+\frac{\partial^2H}{\partial\mathbf{c}^2}\frac{d\hat{\mathbf{c}}}{d\mathbf{y}}\right]\\
&\qquad-\frac{\partial H}{\partial\mathbf{c}}\left[\frac{\partial^2H}{\partial\mathbf{c}^2}\right]^{-1}\left\{\sum_{p,q=1}^{N}\frac{d\hat{c}_p}{d\boldsymbol{\theta}}^T\frac{\partial^3J}{\partial\mathbf{c}\partial c_p\partial c_q}\frac{d\hat{c}_q}{d\mathbf{y}}+\sum_{p=1}^{N}\frac{\partial^3J}{\partial\mathbf{c}\partial c_p\partial\boldsymbol{\theta}}\frac{d\hat{c}_p}{d\mathbf{y}}\right\}
\end{aligned}$$

with $d\hat{\mathbf{c}}/d\mathbf{y}$ given by

$$-\left[\frac{\partial^2 J}{\partial \mathbf{c}^2}\right]^{-1} \frac{\partial^2 J}{\partial \mathbf{c} \partial \mathbf{y}}$$

and $\partial^2 J/\partial \mathbf{c} d\mathbf{y}$ being block diagonal with the ith block containing $w_i \Phi_i(\mathbf{t}_i)$.

References

[Ramsay *et. al.* 2007] Ramsay, J. O., G. Hooker, D. Campbell and J. Cao (2007). Parameter Estimation for Differential Equations: A Generalized Smoothing Approach *Journal of the Royal Statistical Society, Series B* to appear.

[Bellman, 1953] R. Bellman, 1953, *Stability Theory of Differential Equations*, Dover, New York.