Gradient Matching

Anran Wang

2022.10.12

Outline

Table of Contents

Model

The model for the state variables $x = (x_1, \ldots, x_d)^\top$ consists of an initial value problem

$$\begin{cases} \dot{x}(t) = F(t, x(t), \theta), \\ x(0) = x_0, \end{cases} \tag{1}$$

where F is a time-dependent vector field from $\mathbb{R}^d$ to $\mathbb{R}^d$, $d \in \mathbb{N}$, and $\theta \in \Theta$, Θ being a subset of a Euclidean space.
We want to estimate the parameter θ of the ordinary differential equation (1) from noisy observations at n points, $t_1 < \cdots < t_n$.

$$y_i = x(t_i) + \varepsilon_i, i = 1, \ldots, n,$$

where the ε_i are i.i.d centered random variables. The ODE is indexed by a parameter $\theta \in \Theta \subset \mathbb{R}^p$ and initial value x_0; the true parameter value is θ^* and the corresponding solution of (1) is x^*.

A simple example

- **Why we use Gradient Matching?**

First, we consider about a simple example

$$\dot{x}(t) = x(t)^{\top}\beta,$$

where β is the parameter we want to estimate.

A general idea is get n samples $(\tilde{\dot{x}}_i, \tilde{x}_i)$ by the observed value y_i and t_i, $i = 1, \ldots, n$, e.g. we can choose

$$(\tilde{\dot{x}}_i, \tilde{x}_i) = \left(\frac{y_{i+1} - y_{i-1}}{t_{i+1} - t_{i-1}},\ y_i \right).$$

Then we can solve for β using least squares.

But when the dimension d is very large, the method above doesn't work well.

Gradient Matching

So we can consider fitting x and $\dot{x}$ to a curve instead of n points. Then we minimize the error between $\hat{\dot{x}}$ and $F(t, \hat{x}, \theta)$.
Now, when the dimension d is very large, we can also work it well.

- **What is Gradient Matching?**

Gradient Matching is a method that

1. use suitable splines to get nonparametric curve fit to x, and get $\hat{x}, \hat{\dot{x}}$,
2. then, get the estimate $\hat{\theta}$ of θ by minimizing $\|\hat{\dot{x}} - F(t, \hat{x}, \theta)\|$, where $\|\cdot\|$ is a suitable norm.

For example, minimize the equation (8.1) of [1]:

$$\mathrm{ISSE}_1(\theta) = \int \left\| \hat{\dot{x}}(t) - F(t, \hat{x}, \theta) \right\|^2 dt.$$

Gradient Matching

And via equation (8.5) of [1], we define $\hat{\theta}$ to be the minimizer of the integrated squared error:

$$\mathrm{ISSE}(\theta) = \sum_{i=1}^{d} \int w_i(t) \left(\hat{\dot{x}}_i(t) - F_i(t, \hat{x}, \theta) \right)^2 dt. \tag{2}$$

This objective can be fairly readily extended to provide different weights to different components x_i of the model and even to provide more weight to some particular parts of the time domain.

Optimizing Gradient Matching

- **How Gradient Matching work?**

Since we expect that F will be nonlinear, (2) is usually not tractable. Consequently, we need to approximate it by a quadrature rule in which we evaluate the integrand at time points t_q for $q = 1, \ldots, Q$ and approximate the integral by a weighted sum

$$\widehat{\text{ISSE}}(\boldsymbol{\theta}) = \sum_{i=1}^{d} \sum_{q=1}^{Q} w_{iq} \left(\widehat{\dot{x}}_i \left(t_q\right) - F_i \left(t, \hat{x} \left(t_q\right), \theta\right) \right)^2 ,$$

where $w_{iq} := w_i(t_q)$.

We now observe that $\widehat{\text{ISSE}}(\boldsymbol{\theta})$ is a weighted least squares criterion and we can apply a **Gauss-Newton scheme** to minimize it just as in Chap 7 of [1].

Gauss-Newton algorithm

We start with some sensible guess $\hat{\boldsymbol{\theta}}_0$ for $\boldsymbol{\theta}$ and on iteration ℓ make the update

$$\hat{\boldsymbol{\theta}}^{\ell+1} = \hat{\boldsymbol{\theta}}^{\ell} - \left[\sum_{i=1}^{d} \partial_\theta F_i\left(\hat{\boldsymbol{\theta}}^{\ell}\right)^\top \mathbf{W}_i \partial_\theta F_i\left(\hat{\boldsymbol{\theta}}^{\ell}\right)\right]^{-1} \cdot \sum_{i=1}^{d} \partial_\theta F_i\left(\hat{\boldsymbol{\theta}}^{\ell}\right)^\top \mathbf{W}_i \left(\widehat{\dot{\mathbf{X}}}_i - F_i\left(\hat{\boldsymbol{\theta}}^{\ell}\right)\right).$$

where, for $i = 1, \ldots, d$ and $q = 1, \ldots, Q$,

- $F_i(\boldsymbol{\theta})$ is dim Q vector, with $[F_i(\boldsymbol{\theta})]_q = F_i\left(t_q, \hat{\mathbf{x}}\left(t_q\right), \boldsymbol{\theta}\right)$.
- $\partial_\theta F_i(\boldsymbol{\theta})$ is dim $Q \times p$ matrix, with $[\partial_\theta F_i(\boldsymbol{\theta})]_{qj} = \partial_{\theta_j} F_i\left(t_q, \hat{\mathbf{x}}\left(t_q\right), \boldsymbol{\theta}\right)$.
- $\hat{\dot{X}}_i$ is dim Q vector, with $[\hat{\dot{X}}_i]_q = \hat{\dot{x}}_i(t_q)$.
- $\mathbf{W}_i$ is $Q \times Q$ matrix, with the w_{iq} on the diagonal.

Example: Filling a container

In this subsection, we use two examples to show the diagnostic plots.
We do gradient matching with the refinery data Introduced in section 1.1.3 of [1], where we fitted a first-order forced linear ODE:

$$\dot{x} = \beta_0 + \beta_1 x + \alpha u(t).$$

Figure 8.3 presents the results of this estimate. We have shown both $\dot{\hat{x}}$ and our prediction of it from $\hat{x}$ as well as the error between the two. This example shows why gradient matching is appealing. Besides avoiding solving the differential equation, gradient matching also often involves a better conditioned optimization problem.

Example: Filling a container

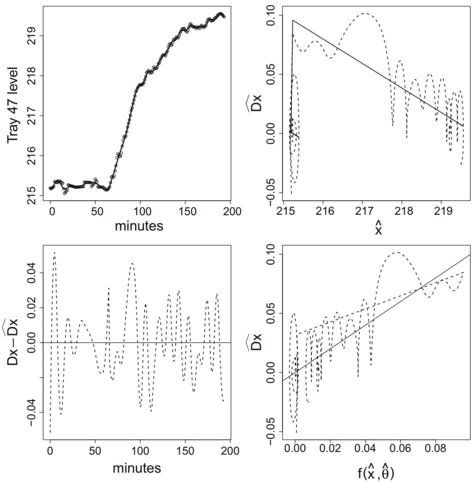


Fig. 8.3 Gradient matching on the refinery data. *Top left* level in Tray 47 and RS move of these data. *Top right* $\widehat{Dx}$ plotted against $\hat{x}$ (*dashed*) and the prediction of it (*solid*) using estimated parameters. *Bottom left* difference between $\widehat{Dx}$ and its prediction plotted over time. *Bottom right* predicted versus fitted $\widehat{Dx}$

Residual plot

The bottom left panel in Figure 8.3 presents the basic approach. We have plotted process residuals

$$r(t) = \hat{\dot{x}}(t) - F(t, \hat{x}, \theta).$$

We think of this as a lack-of-fit function which indicates how far $\hat{x}$ is from satisfying the ODE, even at the best possible parameter estimates. Figure 8.3 lets us examine r where we see what appears to be random oscillations. This also gives us the opportunity to examine whether F has been poorly chosen.

Example: Rosenzweig–MacArthur model

We do gradient matching on the Rosenzweig–MacArthur model Introduced in section 7.4.6 of [1], which is:

$$\frac{dC}{dt} = \rho C(\kappa - C) - \frac{\gamma \beta CB}{\chi + C},$$
$$\frac{dB}{dt} = \frac{\beta CB}{\chi + C} - \delta B.$$

In Figure 8.4, we can see that gradient matching has resulted in trajectories with a longer period and smaller amplitude than the data exhibit and the estimate in Figure 7.5. This is partly the result of statistical inaccuracies in estimating $\hat{\dot{x}}$ and $\hat{x}$, but can also be due to error in the ODE as well as in our measurementss, something we discuss next.

Example: Rosenzweig–MacArthur model

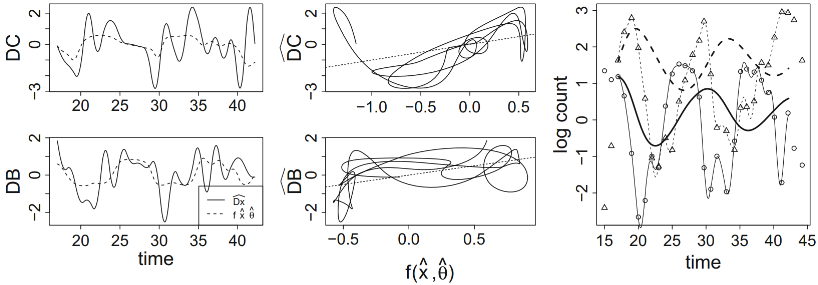


Fig. 8.4 Gradient matching on the chemostat data. The *left panel* plots $\widehat{\mathbf{Dx}}$ (*dashed*) and the fitted $\mathbf{f}(\hat{\mathbf{x}}, \hat{\boldsymbol{\theta}})$ (*solid*) as a function of time for C (*top*) and B (*bottom*). The *middle panel* presents $\widehat{\mathbf{Dx}}$ plotted against $\mathbf{f}(\hat{\mathbf{x}}, \hat{\boldsymbol{\theta}})$. The *right panel* plots log data and smooth (*thin lines*) along with the trajectories obtained by solving the ODE at parameters obtained from gradient matching (*thick lines*)

Example: Rosenzweig–MacArthur model

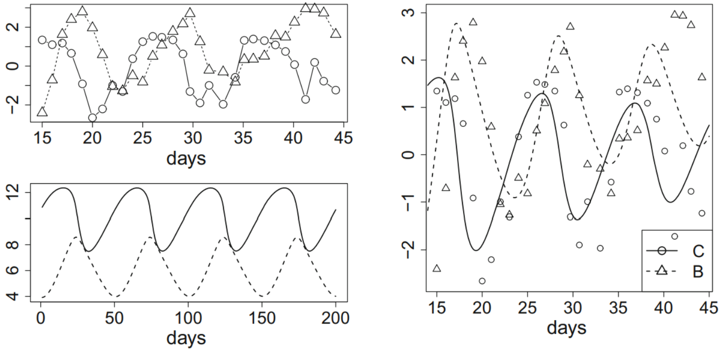


Fig. 7.5 Fits of chemostat data to the Rosenzweig–MacArthur model. *Left* a comparison of the logged data and solutions to (7.15) with initial parameters before some basic transformation. *Right* data and solutions at the finalized objective function

Example: Rosenzweig–MacArthur model

Figure 8.5 produces a sequence of plots. On the left, we have plotted $r(t)$ as a two-dimensional function of $\hat{C}(t)$ and $\hat{B}(t)$. The middle plot compares this to what we get when fitting a Lotka–Volterra model

$$\frac{dC}{dt} = \alpha C - \beta CB,$$
$$\frac{dB}{dt} = \gamma CB - \delta B.$$

The right panel shows Lotka–Volterra model using gradient matching.

We can see that comparing with the one of Lotka–Volterra model, $r(t)$ of Rosenzweig–MacArthur model is bigger when $|C|$ is big.

Example: Rosenzweig–MacArthur model

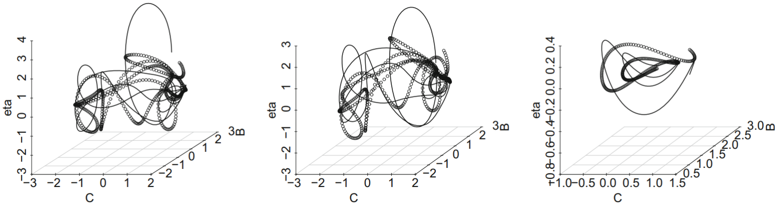


Fig. 8.5 Diagnostics for fitting the chemostat data. The *left panel* provides the lack of fit $\mathbf{r}(t)$ plotted against the estimated state variables using the Rosenzweig–MacArthur ODE. The lack of fit in the C equation is given by *lines*, in the B equation by *circles*. The middle panel uses a simpler Lotka–Volterra ODE. The right panel provides a prototype diagnostic resulting from fitting solutions to a Rosenzweig–MacArthur model with a Lotka–Volterra ODE

Conducting inference

To try and understand the variability in $\hat{\theta}$, we will consider taking one step of the Gauss-Newton algorithm that from $\hat{\theta}^{\ell}$ to $\hat{\theta}^{\ell+1}$ i.e. equation (6).
For $q = 1, \ldots, Q,\ i = 1, \ldots, d$, we will write out the difference in the last term as

$$\begin{aligned}\left[\widehat{\dot{X}}_i - F_i\left(\hat{\boldsymbol{\theta}}^{\ell}\right)\right]_q =& \widehat{\dot{X}}_{iq} - E\widehat{\dot{X}}_{iq} + E\widehat{\dot{X}}_{iq} \\ & - F_i\left(t_q, E\hat{X}_q, \hat{\boldsymbol{\theta}}^{\ell}\right) + F_i\left(t_q, E\hat{X}_q, \hat{\boldsymbol{\theta}}^{\ell}\right) - F_i\left(t_q, \hat{X}_q, \hat{\boldsymbol{\theta}}^{\ell}\right) \\ \approx& \widehat{\dot{X}}_{iq} - E\widehat{\dot{X}}_{iq} \\ & + \partial_{\mathbf{x}} F_i\left(t_q, \hat{X}_q, \hat{\boldsymbol{\theta}}^{\ell}\right)\left(\hat{X}_q - E\hat{X}_q\right) + E\widehat{\dot{X}}_{iq} - F_i\left(t_q, E\hat{X}_q, \hat{\boldsymbol{\theta}}^{\ell}\right),\end{aligned}$$

where the last of these terms only involves expectations, so has zero variance.

Conducting inference

We define

$$H(\boldsymbol{\theta}) := \sum_{i=1}^{d} \partial_\theta F_i(\boldsymbol{\theta})^\top W_i \partial_\theta F_i(\boldsymbol{\theta}),$$

$$J_i(\boldsymbol{\theta}) := H(\boldsymbol{\theta})^{-1} \left(\sum_{k=1}^{d} \partial_\theta F_k(\boldsymbol{\theta})^\top W_k \left[\mathrm{diag}\left(\partial_{\mathbf{x}} F_i \left(t_q, \hat{X}_q, \hat{\boldsymbol{\theta}}^\ell \right) \right)_{i=1}^{d} \mathbb{I}_{i=k} \right] \right), \tag{3}$$

where $\mathbb{I}_{i=k}$ is an identity matrix if $i = k$ and a matrix of zeros, otherwise. Then, putting things together we have

$$Var\left(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_0\right) \sim Var\left(\sum_{i=1}^{d} J_i(\tilde{\boldsymbol{\theta}}) \begin{bmatrix} \hat{X}_i - E\hat{X}_i \\ \widehat{\dot{X}}_i - E\widehat{\dot{X}}_i \end{bmatrix} \right),$$

where $\tilde{\theta}$ depend on $\hat{\boldsymbol{\theta}}_0$, but we can substitute $\hat{\boldsymbol{\theta}}$ in $J_i(\tilde{\boldsymbol{\theta}})$ and still produce an (asymptotically) correct answer.

Nonparametric smoothing variances

With the expression above, we can now calculate a variance as being

$$\operatorname{var}(\hat{\boldsymbol{\theta}}) \sim \sum_{i=1}^{d}\sum_{k=1}^{d} J_i(\hat{\boldsymbol{\theta}}) \begin{bmatrix} \operatorname{var}\left(\hat{X}_i, \hat{X}_k\right) & \operatorname{cov}\left(\hat{X}_i, \widehat{\dot{X}}_k\right) \\ \operatorname{cov}\left(\widehat{\dot{X}}_i, \hat{X}_k\right) & \operatorname{var}\left(\widehat{\dot{X}}_i, \widehat{\dot{X}}_k\right) \end{bmatrix} J_k(\hat{\boldsymbol{\theta}})^\top.$$

While x is generated by PENSSE (i.e. (19)) with spline basis $\phi(t)$, we have

$$\hat{x}_i(t) = \phi(t)^\top \left(\boldsymbol{\Phi}_i^\top \boldsymbol{\Phi}_i + \lambda_i \mathbf{P}\right)^{-1} \boldsymbol{\Phi}_i^\top Y_i,$$

$$\hat{\dot{x}}_i(t) = \dot{\phi}(t)^\top \left(\boldsymbol{\Phi}_i^\top \boldsymbol{\Phi}_i + \lambda_i \mathbf{P}\right)^{-1} \boldsymbol{\Phi}_i^\top Y_i.$$

where λ_i is a regularization parameter, and $\mathbf{P}$ based on $\phi(t)$, look section 8.2 in [1] for further reading.

Nonparametric smoothing variances

Y_i have the same variance as the residuals ε, our estimates just inherit this variance. In this case, we can write

$$\begin{aligned}
\operatorname{cov}\left(\hat{x}_i(t), \hat{x}_k(s)\right) &= \sigma_{ik} \phi(t)^\top \left(\mathbf{\Phi}_i^\top \mathbf{\Phi}_i + \lambda_i \mathbf{P}\right)^{-1} \mathbf{\Phi}_i^\top \mathbf{\Phi}_k \left(\mathbf{\Phi}_k^\top \mathbf{\Phi}_k + \lambda_k \mathbf{P}\right)^{-1} \phi(t)^\top, \\
\operatorname{cov}\left(\hat{x}_i(t), \widehat{\dot{x}}_k(s)\right) &= \sigma_{ik} \phi(t)^\top \left(\mathbf{\Phi}_i^\top \mathbf{\Phi}_i + \lambda_i \mathbf{P}\right)^{-1} \mathbf{\Phi}_i^\top \mathbf{\Phi}_k \left(\mathbf{\Phi}_k^\top \mathbf{\Phi}_k + \lambda_k \mathbf{P}\right)^{-1} \dot{\phi}(t)^\top, \\
\operatorname{cov}\left(\widehat{\dot{x}}_i(t), \widehat{\dot{x}}_k(s)\right) &= \sigma_{ik} \dot{\phi}(t)^\top \left(\mathbf{\Phi}_i^\top \mathbf{\Phi}_i + \lambda_i \mathbf{P}\right)^{-1} \mathbf{\Phi}_i^\top \mathbf{\Phi}_k \left(\mathbf{\Phi}_k^\top \mathbf{\Phi}_k + \lambda_k \mathbf{P}\right)^{-1} \dot{\phi}(t)^\top,
\end{aligned}$$

which allows us to construct the covariances above.

Then we use the refinery data to examine above.

Example: Filling a container

Table 8.1 Variances and confidence intervals after gradient matching on the refinery data

	Estimated variance	Confidence intervals	
		Lower	Upper
β_0	0.137091	7.6932	9.1743
β_1	0.000001	−0.0228	−0.0187
α	0.000059	−0.2101	−0.1795

Table of Contents

Model

In this section, we consider $\hat{x}_n, \hat{\dot{x}}_n, \hat{\theta}_n$ as estimates of $x, \dot{x}, \theta$ when the sample size is n.
The $L^q(w)$ norm on the space of integrable functions on $[0, 1]$ w.r.t. the measure w is

$$\|f\|_{q,w} = \left(\int_0^1 |f(t)|^q w(t) dt \right)^{1/q}, \ 0 < q \leq \infty. \tag{4}$$

We define the two-step estimator

$$\hat{\theta}_n = \arg\min_{\theta} R^q_{n,w}(\theta), \tag{5}$$

where $R^q_{n,w}(\theta) = \left\| \hat{\dot{x}}_n - F\left(t, \hat{x}_n, \theta\right) \right\|_{q,w}$.

Next, we consider the consistency and the asymptotic property of Gradient Matching under this model.

A Theorem about M-Estimators

First, we consider a theorem about M-Estimators.

Theorem 1. (Thm 5.7. of [2])

Let M_n be random functions and let M be a fixed function of θ such that for every $\varepsilon > 0$

$$\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \xrightarrow{\mathrm{P}} 0,$$

$$\sup_{\theta : d(\theta, \theta_0) \geq \varepsilon} M(\theta) < M(\theta_0).$$

Then any sequence of estimators $\hat{\theta}_n$ with $M_n(\hat{\theta}_n) \geq M_n(\theta_0) - o_P(1)$ converges in probability to θ_0.

Proof of Theorem 1

Proof of Theorem 1:

1. **First, we prove $M(\hat{\theta}_n) \xrightarrow{\text{P}} M(\theta_0)$.**
   For the uniform convergence of M_n to M, we have $M_n(\theta_0) \xrightarrow{\text{P}} M(\theta_0)$ i.e. $M_n(\theta_0) = M(\theta_0) + o_P(1)$. It follows that $M_n(\hat{\theta}_n) \geq M(\theta_0) - o_P(1)$. Then, we have

$$\begin{aligned} 0 \leq M(\theta_0) - M(\hat{\theta}_n) &\leq M_n(\hat{\theta}_n) - M(\hat{\theta}_n) + o_P(1) \\ &\leq \sup_{\theta} |M_n - M|(\theta) + o_P(1) \xrightarrow{\text{P}} 0. \end{aligned} \tag{6}$$

2. **Next, we prove $\hat{\theta}_n \xrightarrow{\text{P}} \theta_0$.**
   There exists for every $\varepsilon > 0$ a number $\eta > 0$ such that $M(\theta) < M(\theta_0) - \eta$ for every θ with $d(\theta, \theta_0) \geq \varepsilon$. Then, we have

$$P(d(\theta, \theta_0) \geq \varepsilon) \leq P\left(M(\hat{\theta}_n) < M(\theta_0) - \eta\right) \to 0,$$

   i.e. $\hat{\theta}_n \xrightarrow{\text{P}} \theta_0$.

Consider

We expect that uniformly in $\theta \in \Theta$

$$R_{n,w}^{q}(\theta) \xrightarrow{P} R_{w}^{q}(\theta),$$

where

$$\begin{aligned} R_{w}^{q}(\theta) :&= \left\| \dot{x^*} - F\left(t, x^*, \theta\right) \right\|_{q,w} \\ &= \left(\int_0^1 \left| F\left(t, x^*, \theta^*\right) - F\left(t, x^*, \theta\right) \right|^q w(t) dt \right)^{1/q}. \end{aligned}$$

This discrepancy measure enables us to construct a consistent estimator $\hat{\theta}_n$.

Consistency of two-step estimators

Proposition 1. (Prop 2.1. of [3])

We suppose there exists a compact set $\mathcal{K} \subset \mathcal{X}$ *such that* $\forall \theta \in \Theta$, $\forall x_0 \in \mathcal{X}$, $\forall t \in [0, 1]$, $x_{\theta,x_0}(t)$ *is in* $\mathcal{K}$. *Then under the conditions*

1. w *is a positive continuous function on* $[0, 1]$,
2. *Uniformly in* $(t, \theta) \in [0, 1] \times \Theta$, $F(t, \cdot, \theta)$ *is* $K-$ *Lipschitz on* $\mathcal{K}$,
3. $\hat{x}_n$ *and* $\hat{\dot{x}}_n$ *are consistent, and* $\hat{x}_n(t) \in \mathcal{K}$ *almost surely,*

we have

$$\sup_{\theta \in \Theta} \left| R^q_{n,w}(\theta) - R^q_w(\theta) \right| = o_P(1). \tag{7}$$

Moreover, if

4. *(identifiability condition)* $\forall \epsilon > 0, \inf_{\|\theta - \theta^*\| \geq \epsilon} R^q_w(\theta) > R^q_w(\theta^*)$,

then, the two-step estimator is consistent, i.e. $\hat{\theta}_n - \theta^* = o_P(1)$.

Remarks of Proposition 1

Remarks of Proposition 1:

- Condition 2 $\iff$ For all $\theta \in \Theta$ and $x_1, x_2 \in \mathcal{K}$, we have

$$\|F(\cdot, x_1, \theta) - F(\cdot, x_2, \theta)\|_{q,w} \leq K \|x_1 - x_2\|_{q,w}.$$

- In condition 3, $\hat{x}_n$ and $\hat{\dot{x}}_n$ are consistent i.e. $\|\hat{x}_n - x^*\|_q \xrightarrow{P} 0$ and $\|\hat{\dot{x}}_n - \dot{x}^*\|_q \xrightarrow{P} 0$.
  This property depends on the property of the chosen spline.
- Via Theorem 1, we can find that with equation (5), (7) and condition 4, we have $\hat{\theta}_n \xrightarrow{P} \theta^*$.

Proof of Proposition 1

Proof of Proposition 1:

For all $\theta \in \Theta$, we have

$$\begin{aligned}\left|R^q_{n,w}(\theta) - R^q_w(\theta)\right| &= \left| \|\hat{\dot{x}}_n - F(\cdot, \hat{x}_n, \theta)\|_{w,q} - \| F(\cdot, x^*, \theta) - F(\cdot, x^*, \theta^*)\|_{q,w} \right| \\ &\leq \left\| \left(\hat{\dot{x}}_n - F(\cdot, \hat{x}_n, \theta)\right) + \left(F(\cdot, x^*, \theta) - F(\cdot, x^*, \theta^*)\right)\right\|_{q,w} \\ &\leq \left\| \hat{\dot{x}}_n - F(\cdot, x^*, \theta^*)\right\|_{q,w} + \|F(\cdot, \hat{x}_n, \theta) - F(\cdot, x^*, \theta)\|_{q,w}.\end{aligned}$$

- By condition 3, we have

$$\begin{aligned}\left\| \hat{\dot{x}}_n - F(\cdot, x^*, \theta^*)\right\|_{q,w} &= \|\hat{\dot{x}}_n - \dot{x}^*\|_{q,w} \\ &\leq M\|\hat{\dot{x}}_n - \dot{x}^*\|_q \xrightarrow{P} 0,\end{aligned}$$

where M is a upper bound for w.

Proof of Proposition 1

- By condition 2 and condition 3, for all $(t, \theta) \in [0, 1] \times \Theta$ we have

$$\begin{aligned}\|F(\cdot, \hat{x}_n, \theta) - F(\cdot, x^*, \theta)\|_{q,w} &\leq K\|\hat{x}_n - x^*\|_{q,w} \\ &\leq KM \|\hat{x}_n - x^*\|_q \xrightarrow{P} 0.\end{aligned}$$

Then, we have

$$\begin{aligned}\sup_{\theta\in\Theta} \left|R^q_{n,w}(\theta) - R^q_w(\theta)\right| \leq & \left\|\hat{\dot{x}}_n - F(\cdot, x^*, \theta^*)\right\|_{w,q} \\ & + \sup_{\theta\in\Theta} \|F(\cdot, \hat{x}_n, \theta) - F(\cdot, x^*, \theta)\|_{w,q} \\ \leq & M \left\|\hat{\dot{x}}_n - F(x^*, \theta^*)\right\|_q + KM \|\hat{x}_n - x^*\|_q = o_P(1).\end{aligned}$$

By Theorem 1, if condition 4 holds, we have $\hat{\theta}_n - \theta^* = o_P(1)$. □

Asymptotics of two-step estimators

In this subsection, we focus on the least squares criterion $R^2_{n,w}$, when the estimator of the derivative is $\hat{\dot{x}}_n := \dot{\hat{x}}_n$.

In that case, we show that the two-step estimator behaves as the sum of two linear functionals of $\hat{x}_n$ of different nature: a smooth and a non-smooth one.

We can set condition $w(0) = w(1) = 0$ to make the non-smooth part vanish, implying that the two-step estimator can have a parametric rate of convergence.

Notations

Here are some notations:

- We define $D_1F(x,\theta)$ and $D_2F(x,\theta)$ as the differentials of F at (x,θ) w.r.t. x and θ.
- We adopt the notation $D_iF^*, i=1,2$ for the functions $t \mapsto D_iF\left(x^*(t),\theta^*\right), i=1,2$ and $\hat{D_iF}, i=1,2$ for the functions $t \mapsto D_iF\left(\hat{x}(t),\hat{\theta}\right)$.
- We adopt the notation $D_{12}F^*$ for $t \mapsto D_1D_2F\left(x^*(t),\theta^*\right)$.
- We consider the approximate Hessian matrix J^* of $R_w^2(\theta)$ at $\theta=\theta^*$:

$$J^* := \int_0^1 \left(D_2F\left(t,x^*(t),\theta^*\right)\right)^\top D_2F\left(t,x^*(t),\theta^*\right) w(t)dt. \tag{8}$$

Asymptotic representation of two-step estimators

Proposition 2. (Prop3.1. of [3])

We suppose that D_{12} exists and w is differentiable. We introduce the two linear operators $\Gamma_{s,w}$ and $\Gamma_{b,w}$ defined by

$$\Gamma_{s,w}(x) = \int_0^1 \left(D_2 F^{*\top}(t) D_1 F^*(t) w(t) + \frac{d}{dt}\left(D_2 F^*(t) w(t)\right) \right) x(t) dt,$$

and

$$\Gamma_{b,w}(x) = w(0) D_2 F^{*\top}(0) x(0) - w(1) D_2 F^{*\top}(1) x(1).$$

If $D_1 F, D_2 F$ are Lipschitz in (x, θ), J^* is invertible, and $\hat{x}_n, \dot{\hat{x}}_n$ are (resp.) consistent estimators of x^* and $\dot{x}^*$, then

$$\hat{\theta}_n - \theta^* = J^{*-1}\left(\Gamma_{s,w}\left(x^*\right) - \Gamma_{s,w}\left(\hat{x}_n\right) + \Gamma_{b,w}\left(x^*\right) - \Gamma_{b,w}\left(\hat{x}_n\right)\right) + o_P(1). \quad (9)$$

Sketch proof of Proposition 2

Sketch proof of Proposition 2:

In this proof, we use $\hat{\theta}, \hat{x}$ to abbreviate $\hat{\theta}_n, \hat{x}_n$.

1. **Basic form of** $(\hat{\theta} - \theta^*)$ **i.e.** (10)**.**

   For the definition of $\hat{\theta}$ i.e. (5), we have

$$
\begin{aligned}
0 &= \nabla_\theta R^2_{n,w}(\hat{\theta}) \\
&= \int_0^1 \left(D_2 F(\hat{x}, \hat{\theta})\right)^\top (\dot{\hat{x}} - F(\hat{x}(t), \hat{\theta}))w dt \\
&= \int_0^1 \left(D_2 F(\hat{x}, \hat{\theta})\right)^\top \left(\dot{\hat{x}} - \dot{x}^* + F^* - F\left(\hat{x}, \theta^*\right) + F\left(\hat{x}, \theta^*\right) - F(\hat{x}, \hat{\theta})\right) w dt.
\end{aligned}
$$

   By the Lagrange formula, we have

$$
\begin{aligned}
0 = \int_0^1 \left(D_2 F(\hat{x}, \hat{\theta})\right)^\top \Bigg(& \left(\dot{\hat{x}} - \dot{x}^*\right) \\
& + D_1 F\left(\tilde{x^*}, \theta^*\right)\left(x^* - \hat{x}\right) + D_2 F\left(\hat{x}, \tilde{\theta^*}\right)\left(\theta^* - \hat{\theta}\right)\Bigg) w dt,
\end{aligned}
$$

Sketch proof of Proposition 2

with $\tilde{x^*}$ and $\tilde{\theta^*}$ being random points between x^* and $\hat{x}$, and θ^* and $\hat{\theta}$. Then, an asymptotic expression for $(\theta^* - \hat{\theta})$ is

$$
\begin{aligned}
(\theta^* - \hat{\theta}) \int_0^1 \hat{D}_2 F^\top D_2 F(\hat{x}, \tilde{\theta^*}) w dt = & - \left(\int_0^1 \hat{D}_2 F^\top \left(\dot{\hat{x}} - \dot{x^*} \right) w dt \right. \\
& \left. + \int_0^1 \hat{D}_2 F^\top D_1 F \left(\tilde{x^*}, \theta^* \right) \left(x^* - \hat{x} \right) w dt \right) \\
:= & - G_n,
\end{aligned} \tag{10}
$$

where we define G_n as the negative value of RHS.

Sketch proof of Proposition 2

2. **Asymptotic form of** $(\hat{\theta} - \theta^*)$ **i.e.** (13).

We define

$$H_n := \int_0^1 (D_2F^*)^\top \left(\left(\dot{\hat{x}} - \dot{x}^*\right) + D_1F^* \left(x^* - \hat{x}\right)\right) w dt. \tag{11}$$

Then we have

$$\begin{aligned} &\|G_n - H_n\|_2 \xrightarrow{P} 0, \\ &\| \int_0^1 \hat{D_2F}^\top D_2F\left(\hat{x}, \theta^*\right) dt - J^*\|_2 \xrightarrow{P} 0, \end{aligned} \tag{12}$$

where the L_2 norm $\|A\|_2 = \int_0^1 Tr(A^\top(t)A(t))dt$.
We can prove (12) by using the continuous mapping theorem, and a further proof of (12) is in [3].

Sketch proof of Proposition 2

The asymptotic behavior of $(\theta^* - \hat{\theta})$ is then given by

$$\begin{aligned}\theta^* - \hat{\theta} =& J^{*-1} H_n + o_P(1) \\ =& J^{*-1} \int_0^1 (D_2 F^*)^\top \left(\left(\dot{\hat{x}} - \dot{x}^* \right) + D_1 F^* \left(x^* - \hat{x} \right) \right) w dt + o_P(1).\end{aligned} \tag{13}$$

3. **Simplify H_n.**
   Then, we define

$$\Gamma(x) := \int_0^1 D_2 F^{*\top} D_1 F^* x(t) w dt - \int_0^1 (D_2 F^*)^\top \dot{x}(t) w dt,$$

so that, we have

$$H_n = \Gamma(x^*) - \Gamma(\hat{x}). \tag{14}$$

Sketch proof of Proposition 2

Since $D_{12}F$ and $\dot{w}$ exist, we have

$$\int_0^1 (D_2F^*(t))^\top \dot{x}(t)w(t)dt = \left[D_2F^{*\top}xw\right]_0^1 - \int_0^1 \frac{d}{dt}\left(D_2F^*(t)w(t)\right)x(t)dt.$$

Γ is the sum of the linear functionals $\Gamma_{s,w}, \Gamma_{b,w}$:

$$\begin{aligned}\Gamma(x) = &\int_0^1 \left(D_2F^{*\top}D_1F^*w(t) + \frac{d}{dt}\left(D_2F^*(t)w(t)\right)\right)x(t)dt \\ &+ D_2F^{*\top}(0)x(0)w(0) - D_2F^{*\top}(1)x(1)w(1) \\ =&\Gamma_{s,w}(x) + \Gamma_{b,w}(x).\end{aligned} \tag{15}$$

Bring (14), (15) back to (13), we get

$$\hat{\theta}_n - \theta^* = J^{*-1}\left(\Gamma_{s,w}\left(x^*\right) - \Gamma_{s,w}\left(\hat{x}_n\right) + \Gamma_{b,w}\left(x^*\right) - \Gamma_{b,w}\left(\hat{x}_n\right)\right) + o_P(1).$$

B-spline

Next, we consider $\hat{x}_n$ as B-spline based on $y_i,\ i = 1, ..., n$.

For a fixed integer $k \geq 2$, we denote $\mathbb{S}(\boldsymbol{\xi}_n, k)$ the space of spline functions of order k with knots $\boldsymbol{\xi} = (0 = \xi_0 < \xi_1 < \cdots < \xi_{L+1} = 1)$, where $\boldsymbol{\xi}_n = \max_{1 \leq i \leq L_n+1}(\xi_i - \xi_{i-1})$.

A function s in $\mathbb{S}(\boldsymbol{\xi}_n, k)$ is a polynomial of degree $k-1$, on each interval $[\xi_i, \xi_{i+1}]$, $i = 0, \ldots, L$ and s is in $C^{k-2} \cdot \mathbb{S}(\boldsymbol{\xi}_n, k)$ is a space of dimension $L + k$.

B-splines can be defined recursively from the augmented knot sequence $\boldsymbol{\tau} = (\tau_j, j = 1, \ldots, L + 2k)$ with $\tau_1 = \cdots = \tau_k = 0, \tau_{j+k} = \xi_j, j = 1, \ldots, L$ and $\tau_{L+k+1} = \cdots = \tau_{L+2k} = 1$.

B-spline

We note $B_{i,k'}$, the i^{th} B-spline basis function of order k' $(1 \le k' \le k)$ with the corresponding knot sequence $\boldsymbol{\tau}$. The B-spline basis of order $k' = 1, \ldots, k$ are linked then by the recurrence equation:

$$\forall i = 1, \ldots, L + 2k - 1, \forall t \in [0, 1], B_{i,1}(t) = 1_{[\tau_i, \tau_{i+1}]}(t),$$

and $\forall i = 1, \ldots, L + 2k - k', \forall k' = 2, \ldots, k, \forall t \in [0, 1]$,

$$B_{i,k'}(t) = \frac{t - \tau_i}{\tau_{i+k'-1} - \tau_i} B_{i,k'-1}(t) + \frac{\tau_{i+k'} - t}{\tau_{i+k'} - \tau_{i+1}} B_{i+1,k'-1}(t).$$

B-spline

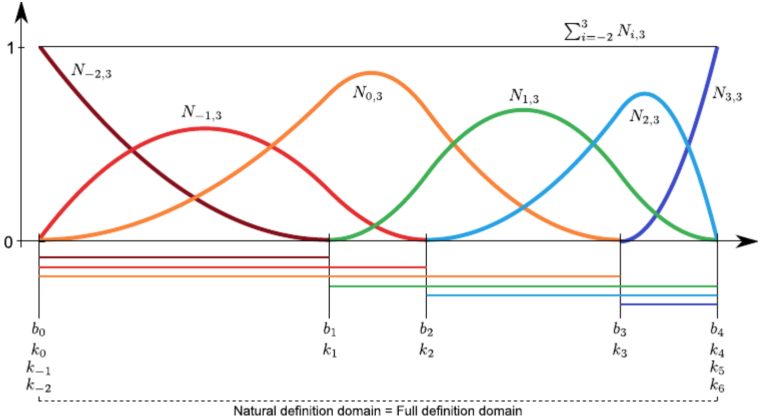


Figure: B-splines with order k=3.

B-spline

We prescribe some notations:

- $\mathbf{B} := (B_{1,k}, B_{2,k}, \ldots, B_{L+k,k})^\top$ is B-spline basis function of order k,
- $\mathbf{Y}_n := (\mathbf{Y}_1 \ldots \mathbf{Y}_d)$ is the $n \times d$ matrix of observations,
- $\mathbf{B}_n := \left(B_{j,k}(t_i)\right)_{1\le i\le n, 1\le j\le L+k}$ is the design matrix.

Then, the estimator $\hat{x}_n$ we consider is written componentwise in the basis of B-splines:

$$\forall i \in \{1, \ldots, d\}, \forall t \in [0,1], \hat{x}_{i,n}(t) = \mathbf{B}^\top(t)\hat{\mathbf{c}}_{i,n}. \tag{16}$$

We estimate the coefficient matrix C_n by least-squares

$$\hat{\mathbf{c}}_{i,n} = \arg\min_{\mathbf{c}\in\mathbb{R}^{L+k}} \sum_{j=1}^{n} \left(y_{ij} - \mathbf{B}(t_j)^\top \mathbf{c}\right)^2, i = 1, \ldots, d, \tag{17}$$

then we get $\hat{\mathbf{c}}_{i,n} = \left(\mathbf{B}_n^\top \mathbf{B}_n\right)^+ \mathbf{B}_n^\top \mathbf{Y}_i$.

General results given by reference 21 in [3] ensure that $\|\hat{x}_n - x^*\|_2 \xrightarrow{P} 0$ **for sequences of suitably chosen** $\mathbb{S}(k, \boldsymbol{\xi}_n)$.

Asymptotic of regression splines

Proposition 3. (Prop 3.2. of [3])

Let $(\boldsymbol{\xi}_{n,i})_{i=0}^{L_n+1}$ *be a sequence of knot sequences of length* $L_n + 2$*, and* $K_n = \dim\left(\mathbb{S}\left(\boldsymbol{\xi}_n, k\right)\right) = L_n + k$*, with* $k \geq 2$*. We suppose that* $L_n \to \infty$ *is such that* $n^{1/2}\left|\boldsymbol{\xi}_n\right| \to 0$ *and* $n\left|\boldsymbol{\xi}_n\right| \to \infty$*, where* $\boldsymbol{\xi}_n = \max_{1\leq i\leq L_n+1}\left(\xi_{n,i} - \xi_{n,i-1}\right)$*.*
If $a : [0, 1] \to \mathbb{R}$ *is in* C^1*, and* x^* *is in* $C^\alpha, 2 \leq \alpha \leq k$*,* $\Gamma(x^*) = \int_0^1 a(s)x^*(s)ds$ *then we have:*

- $\Gamma\left(\hat{x}_n\right) - \Gamma(x^*) = O_P\left(n^{-1/2}\right)$ *and* $\sqrt{n}\left(\Gamma\left(\hat{x}_n\right) - \Gamma(x^*)\right)$ *is asymptotically normal,*
- $\forall t \in [0, 1], \hat{x}_n(t) - x^*(t) = O_P\left(n^{-1/2}\left|\boldsymbol{\xi}_n\right|^{-1/2}\right)$*, and* $Var\left(\hat{x}_n(t)\right)^{-1/2}\left(\hat{x}_n(t) - x^*(t)\right)$ *is asymptotically normal,* $t \in [0, 1]$*.*

The proof of the proposition can be seen in [3]. We can **take the proposition back to** (9), then we get the asymptotic normality of the two-step estimator.

Asymptotic normality of the two-step estimator

Theorem 2. (Thm 3.1. of [3])

With the conditions

1. *The conditions of proposition 1 are satisfied,*
2. *F is a C^m vector field w.r.t (θ, x) $(m \geq 1)$, such that $D_1 F, D_2 F$ are Lipschitz w.r.t (θ, x), and $D_{12}F$ exists,*
3. *J^* of the asymptotic criterion $R_w^2(\theta)$ evaluated at θ^* is nonsingular.*

Let $\hat{x}_n \in \mathbb{S}(\boldsymbol{\xi}_n, k)$ a regression spline with $k \geq 2$, such that $n^{1/2} |\boldsymbol{\xi}_n| \to 0$ and $n |\boldsymbol{\xi}_n| \to \infty$, then the two-step estimator $\hat{\theta}_n = \arg\min_\theta R_{n,w}^2(\theta)$ is asymptotically normal and

- *if $w(0) = w(1) = 0$, then $\left(\hat{\theta}_n - \theta^*\right) = O_P\left(n^{-1/2}\right)$,*
- *if $w(0) \neq 0$ or $w(1) \neq 0$, then $\left(\hat{\theta}_n - \theta^*\right) = O_P\left(n^{-1/2} |\boldsymbol{\xi}_n|^{-1/2}\right)$. The optimal rate of convergence for the Mean Square Error is obtained for $K_n = O\left(n^{1/(2m+3)}\right)$, then $\left(\hat{\theta}_n - \theta^*\right) = O_P\left(n^{-(m+1)/(2m+3)}\right)$.*

Sketch proof of Theorem 2

Sketch proof of Theorem 2:

- From proposition 2 and proposition 3, we can claim the asymptotic normality of $\sqrt{n}\left(\Gamma_{s,w}\left(\hat{x}_n\right)-\Gamma_{s,w}\left(x^*\right)\right)$ and of $\sqrt{n\left|\xi_n\right|}\left(\Gamma_{b,w}\left(\hat{x}_n\right)-\Gamma_{b,w}\left(x^*\right)\right)$.
  When $w(0)=w(1)=0$, there is only the parametric part, but when $\Gamma_{b,w}$ does not vanish (i.e. $w(0)\neq 0$ or $w(1)\neq 0$) the nonparametric part with rate $\sqrt{n\left|\xi_n\right|}$ remains.
- Theorem 2.1 in reference 44 of [3] gives $E\left(\left(\hat{x}_n(t)-x^*(t)\right)^2\right)=O\left(\left|\xi_n\right|^{m+1}\right)$ (because x^* is C^{m+1}) and $\operatorname{Var}\left(\hat{x}_n(t)\right)=O_P\left(n^{-1}\left|\xi_n\right|^{-1}\right)$ so the optimal rate is reached for $\left|\xi_n\right|=O\left(n^{-1/(2m+3)}\right)$ and is $O\left(n^{-(2m+2)/(2m+3)}\right)$.

$\square$

Table of Contents

Principal Differential Analysis (PDA)

In this section, we focus on the iPDA algorithm in [4].
We consider the model

$$\frac{dx}{dt}(t) + w_x x(t) + w_u u(t) = 0.$$

Then we consider $\hat{x}_n$ as spline based on $y_i,\ i = 1, ..., n$ (e.g. B-spline).
The regular PDA algorithm gets $\hat{x}_n$ that

$$\hat{x} = arg\min \sum_{i=1}^{n} (y(t_i) - \hat{x}(t))^2. \tag{18}$$

While $\hat{x}$ is the B-spline, $\hat{x}$ we get from (18) is the same as the one get from (16) with (17). Or we can gets $\hat{x}$ that (PENSSE)

$$\hat{x} = arg\min \left[\sum_{i=1}^{n} (y(t_i) - \hat{x}(t))^2 + \lambda_{HOD} \int \left(\frac{d^2\hat{x}}{dt^2}(t) \right)^2 dt \right]. \tag{19}$$

A model-based roughness penalty

The iterative part of the iPDA algorithm is based on a model-based roughness penalty that adds to (18).

With the known estimates $\hat{w}_x, \hat{w}_u, \hat{u}$ of w_x, w_u, u, We get $\hat{x}$ that

$$\hat{x} = arg\min\left[\sum_{i=1}^{n}(y(t_i) - \hat{x}(t))^2 + \lambda_{ODE}\int\left(\frac{d\hat{x}}{dt}(t) + \hat{w}_x\hat{x}(t) + \hat{w}_u\hat{u}(t)\right)dt\right]. \tag{20}$$

The penalty term with the weighting coefficient λ_{ODE} uses the residuals of the *ODE* model.

Iteratively refined principal differential analysis (iPDA)

Following are the steps of the iPDA algorithm:

1. Estimate the model parameters using the fitted splines and their derivatives as in standard PDA, i.e. solute the equation (18) or (19).
2. Solute $\hat{\theta}$ by Gradient Matching (e.g. solute the equation (5)), then get $\hat{w}_x, \hat{w}_u, \hat{u}$ by $\hat{\theta}$.
3. Obtain an improved spline fit using a model-based roughness penalty to ensure that the fitted splines are smooth and physically reasonable, i.e. solute the equation (20) and update $\hat{x}$.
4. Solute and update $\hat{\theta}$ by Gradient Matching.
5. Iterate between steps 3 and 4 until parameter estimates converge.

Example: CSTR

We use continuous stirred-tank reactor model (CSTR) to explore the merits of PDA and iPDA relative to traditional nonlinear least-squares regression. CSTR model has the form

$$\frac{\mathrm{d}C_{\mathrm{A}}}{\mathrm{d}t} = \frac{F}{V}\left(C_{\mathrm{A}_0} - C_{\mathrm{A}}\right) - k_{\mathrm{ref}} \exp\left(-\frac{E}{R}\left(\frac{1}{T} - \frac{1}{T_{\mathrm{ref}}}\right)\right) C_{\mathrm{A}}. \tag{21}$$

With variables:

1. C_A is the dynamic response of the concentration of reactant A, with steady state operating point $C_{\mathrm{As}} = 0.576\mathrm{kmol\ m}^{-3}$, and the inlet reactant concentration $C_{A_0} = 2.0\mathrm{kmol\ m}^{-3}$.
2. T is the temperature, with steady state operating point $T_{\mathrm{s}} = 332$ K.

Remark: steady state operating point s means $dC_A/dt|_{at\ s} = 0$.

Example: CSTR

The kinetic parameters to be estimated:

3. $k_{\text{ref}} = k_0 \exp\left(-E/R\, T_{\text{ref}}\right)$ is the value of the kinetic rate constant evaluated at reference temperature T_{ref} .
4. E/R, where E is the activation energy and R is the Boltzmann ideal gas constant.

And the other parameters:

5. T_{ref} is the reference temperature. We can estimate it by $\hat{k}_{\text{ref}}$ and $\widehat{E/R}$.
6. F is the reactant feed rate, which steady at $F_s = 0.05\text{m}^3\ \text{min}^{-1}$.
7. V is the constant volume, $V = 1.0\text{m}^3$.

Example: CSTR

The 1-order Taylor expansion of (21) at the steady state operating point has the form

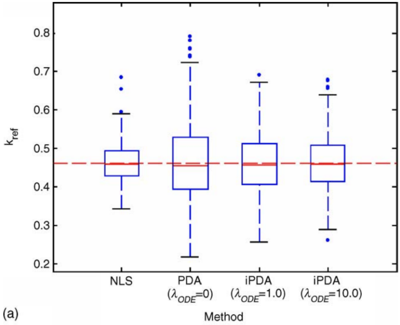
$$\frac{\mathrm{d}C_{\mathrm{A}}'}{\mathrm{d}t} + w_{\mathrm{C}} C_{\mathrm{A}}' + w_{\mathrm{T}} T' = 0, \tag{22}$$

where $C_{\mathrm{A}}' = C_{\mathrm{A}} - C_{\mathrm{As}}$, $T' = T - T_{\mathrm{s}}$. The constant coefficients w_{C} and w_{T} are related to the original model parameters by

$$\begin{aligned} w_{\mathrm{C}} &= \frac{F_{\mathrm{s}}}{V} + k_{\mathrm{ref}} \exp\left(-\frac{E}{R}\left(\frac{1}{T_{\mathrm{s}}} - \frac{1}{T_{\mathrm{ref}}}\right)\right), \\ w_{\mathrm{T}} &= k_{\mathrm{ref}} \frac{E}{R} \frac{C_{\mathrm{As}}}{T_{\mathrm{s}}^2} \exp\left(-\frac{E}{R}\left(\frac{1}{T_{\mathrm{s}}} - \frac{1}{T_{\mathrm{ref}}}\right)\right). \end{aligned}$$

Next, we consider algorithms based on the ODE (22).

Example: CSTR



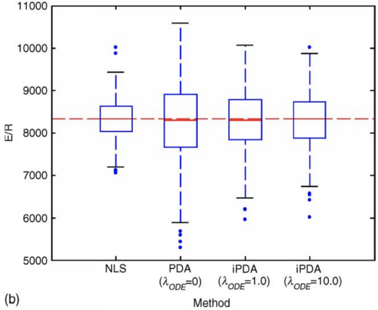
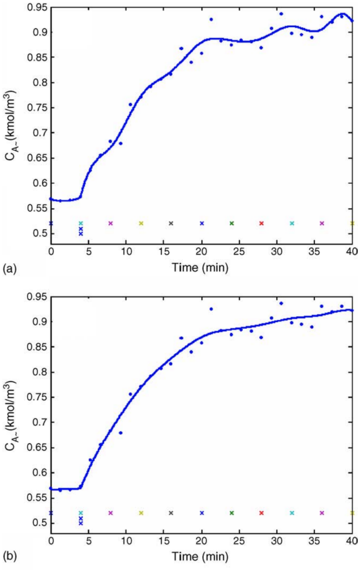


Fig. 6. Effect of iterative PDA penalty weight on estimates of (a) k_{ref} and (b) E/R obtained using the nonlinear CSTR model. Concentration was measured every 80 s with a standard deviation of 0.016 kmol m^{-3} ($I = 4.0$ min, three coincident knots at $t = 4.0$ min for PDA).

Example: CSTR



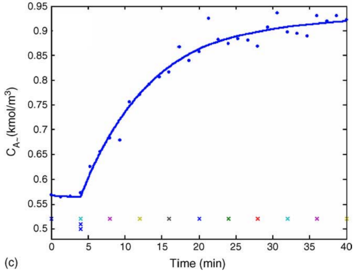


Fig. 7. Spline approximations using (a) regular PDA, (b) iPDA with $\lambda_{ODE} = 1.0$, and (c) iPDA with $\lambda_{ODE} = 10.0$. Concentration was measured every 80 s with a standard deviation of 0.016 kmol m^{-3} ($I = 4.0$ min, three coincident knots at $t = 4.0$ min).

Table of Contents

Appendix: Symbol table

Notation	Introduction
x	The function we consider about. (dim d)
$\dot{x}$	Derivative of x i.e. dx/dt.
θ	The parameter we force on. (dim p)
$\hat{x}, \hat{\dot{x}}, \hat{\theta}, \hat{\theta}^{\ell}$	Estimates, $\hat{\theta}^{\ell}$ is estimate of θ on iteration ℓ.
$\hat{x}_n, \hat{\dot{x}}_n, \hat{\theta}_n$	Estimates when the sample size is n
$x^*, \dot{x}^*, \theta^*$	Actual values.
$F(t, x(t), \theta)$	$\dot{x}(t) = F(t, x(t), \theta)$ is the ODE we consider.
y	The observed value, $y_i = x(t_i) + \varepsilon_i, i = 1, \ldots, n$.
w, w_i, w_{iq}, w_x, w_u	Suitable weight, $w_{iq} := w_i(t_q)$.
$\lambda_i, \lambda_{HOD}, \lambda_{ODE}$	Parameters of penalty terms.
$\mathrm{ISSE}(\theta)$	The integrated squared error i.e. (2).

Appendix: Symbol table

Notation	**Introduction**
$\widehat{\mathrm{ISSE}}(\theta)$	An approximate of $\mathrm{ISSE}(\theta)$, by a weighted sum at $t_q, q = 1, \ldots, Q$.
$F_i(\theta)$	Dim Q vector, with $[F_i(\theta)]_q = F_i\left(t_q, \hat{x}\left(t_q\right), \theta\right)$.
$\partial_\theta F_i(\theta)$	$Q \times p$ matrix, $[\partial_\theta F_i(\theta)]_{qj} = \partial_{\theta_j} F_i\left(t_q, \hat{x}\left(t_q\right), \theta\right)$.
$\hat{\dot{X}}_i$	Dim Q vector, with $[\hat{\dot{X}}_i]_q = \hat{\dot{x}}_i(t_q)$.
$\mathbf{W}_i$	$Q \times Q$ matrix, with the w_{iq} on the diagonal.
$r(t)$	The process residuals,$r(t) = \hat{\dot{x}}(t) - F(t, \hat{x}, \theta)$.
$H(\theta), J_i(\theta)$	Functions built to conducting inference, i.e. (3).
$\\| \cdot \\|_{q,w}$	The $L^q(w)$ norm, i.e. (4).
$R^q_{n,w}(\theta)$	An error, $R^q_{n,w}(\theta) := \\|\hat{\dot{x}}_n - F\left(t, \hat{x}_n, \theta\right)\\|_{q,w}$.
$R^q_w(\theta)$	$R^q_w(\theta) := \\|\dot{x}^* - F\left(t, x^*, \theta\right)\\|_{q,w}$.
$D_i F(x, \theta), i = 1, 2$	the differentials of F w.r.t. x and θ.

Appendix: Symbol table

Notation	Introduction
$D_iF^*, i=1,2$	The functions $t \mapsto D_iF\left(x^*(t), \theta^*\right), i=1,2$.
$\hat{D_i}F, i=1,2$	The functions $t \mapsto D_iF(\hat{x}(t), \hat{\theta})$.
$D_{12}F^*$	The function $t \mapsto D_1D_2F\left(x^*(t), \theta^*\right)$.
J^*	Approximate Hessian matrix of $R_w^2(\theta)$ at θ^* i.e. (8).
$\Gamma_{s,w}, \Gamma_{b,w}, \Gamma$	Linear operators defined in prop 2.
G_n, H_n	Equations in prop 2, i.e. (10) and (14).
$\mathbb{S}(\xi_n, k)$	The space of spline functions of order k and ξ_n is the maximum interval between knots.
$\phi(t)$, $\boldsymbol{\Phi}_i$	Spline basis, and spline basis at $t_j, j=1,\ldots,n$.
$B_{i,k'}(t)$	The i-th basis of order k' B-spline.
$\mathbf{B}$	Order k B-spline basis $\mathbf{B} := \left(B_{1,k}, B_{2,k}, \ldots, B_{L+k,k}\right)^\top$.
$\mathbf{B}_n$	$\mathbf{B}_n := \left(B_{j,k}\left(t_i\right)\right)_{1\le i\le n, 1\le j\le L+k}$ is the design matrix.

References

[1] Giles Hooker James Ramsay. "Dynamic Data Analysis". In: (2017). DOI: 10.1007/978-1-4939-7190-9.

[2] Manfred Denker Rabi Bhattacharya. "Asymptotic Statistics". In: (1990). DOI: 10.1007/978-3-0348-9254-4.

[3] Nicolas J-B. Brunel. "Parameter estimation of ODE's via nonparametric estimators". In: *Electronic Journal of Statistics* 2.none (2008), pp. 1242–1267. DOI: 10.1214/07-EJS132.

[4] A.A. Poyton et al. "Parameter Estimation in Continuous-Time Dynamic Models Using Principal Differential Analysis". In: *Computers Chemical Engineering* 30 (Feb. 2006), pp. 698–708. DOI: 10.1016/j.compchemeng.2005.11.008.