Functional Data Analysis in Matlab and R

James Ramsay, Professor, McGill U., Montreal
Hadley Wickham, Grad student, Iowa State, Ames, IA
Spencer Graves, Statistician, PDF Solutions, San José, CA

Outline

- What is Functional Data Analysis?
- FDA and Differential Equations
- Examples:
  - Squid Neurons
  - ***Continuously Stirred Tank Reactor (CSTR)***
- Conclusions
- References

What is FDA?

- Functional data analysis is a collection of techniques to model data from dynamic systems
  - possibly governed by differential equations
  - in terms of some set of basis functions
- The ‘fda’ package supports the use of 8 different types of basis functions: constant, monomial, polynomial, polygonal, B-splines, power, exponential, and Fourier.

Observations of different lengths

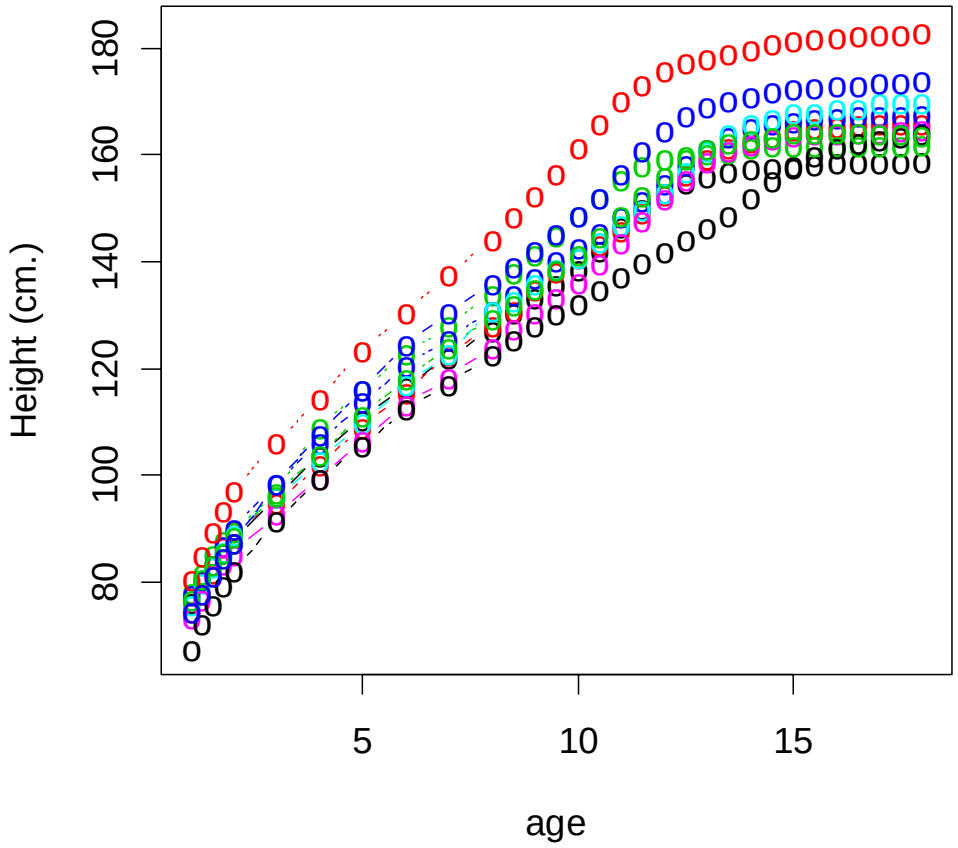
- Observation vectors of different lengths can be mapped to coordinates of a fixed basis set
- All examples in the ‘fda’ package have the same numbers of observations
- No conceptual obstacles to handling observation vectors of different lengths

Time Warping

- “start” and “stop” are sometimes determined by certain transitions
- Example: growth spurts in the life cycle of various species do not occur at exactly the same ages in different individuals (even within the same species)

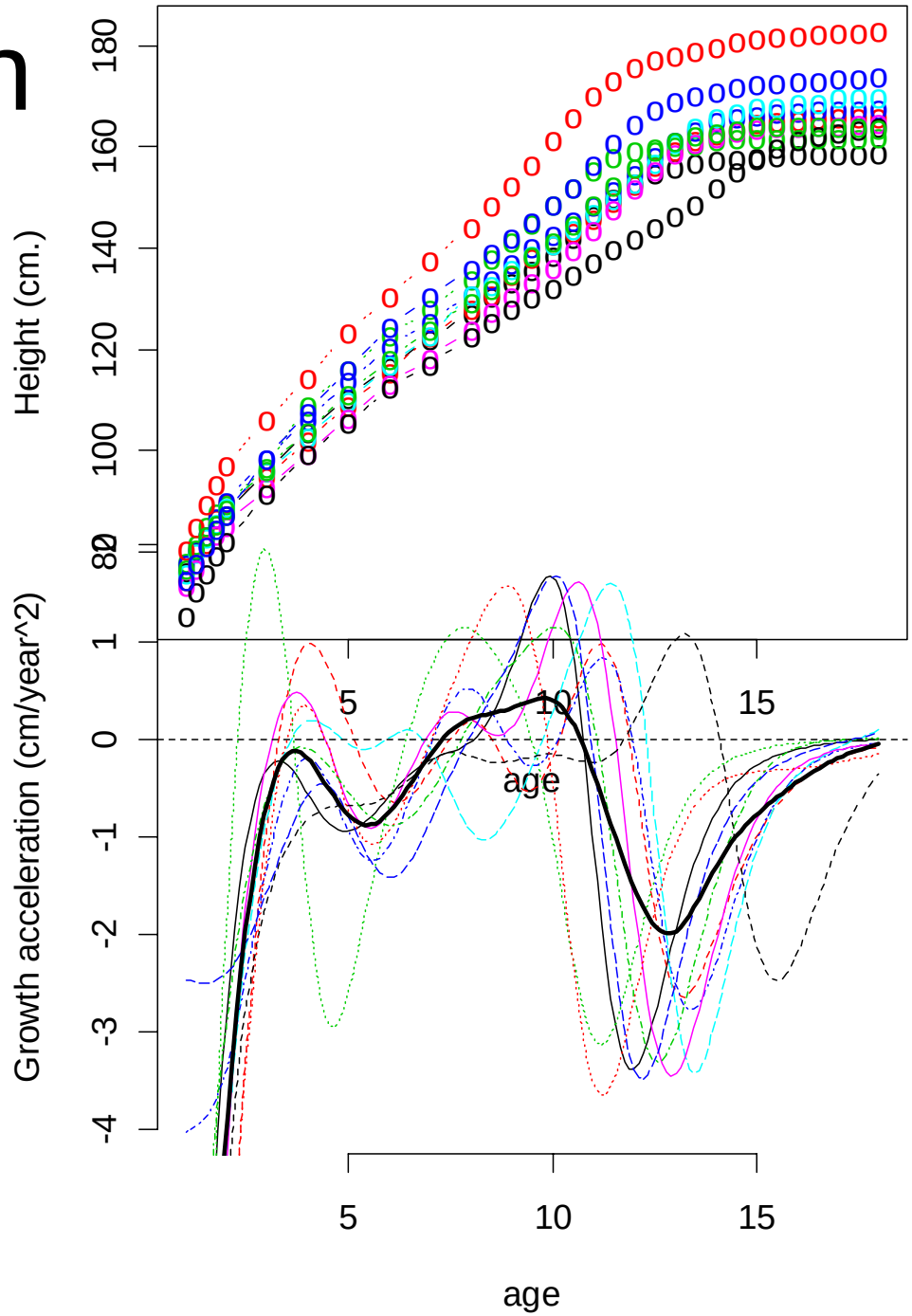
10 Girls: Berkeley Growth Study

- Tuddenham, R. D., and Snyder, M. M. (1954) "Physical growth of California boys and girls from birth to age 18", _University of California Publications in Child Development_, 1, 183-364.



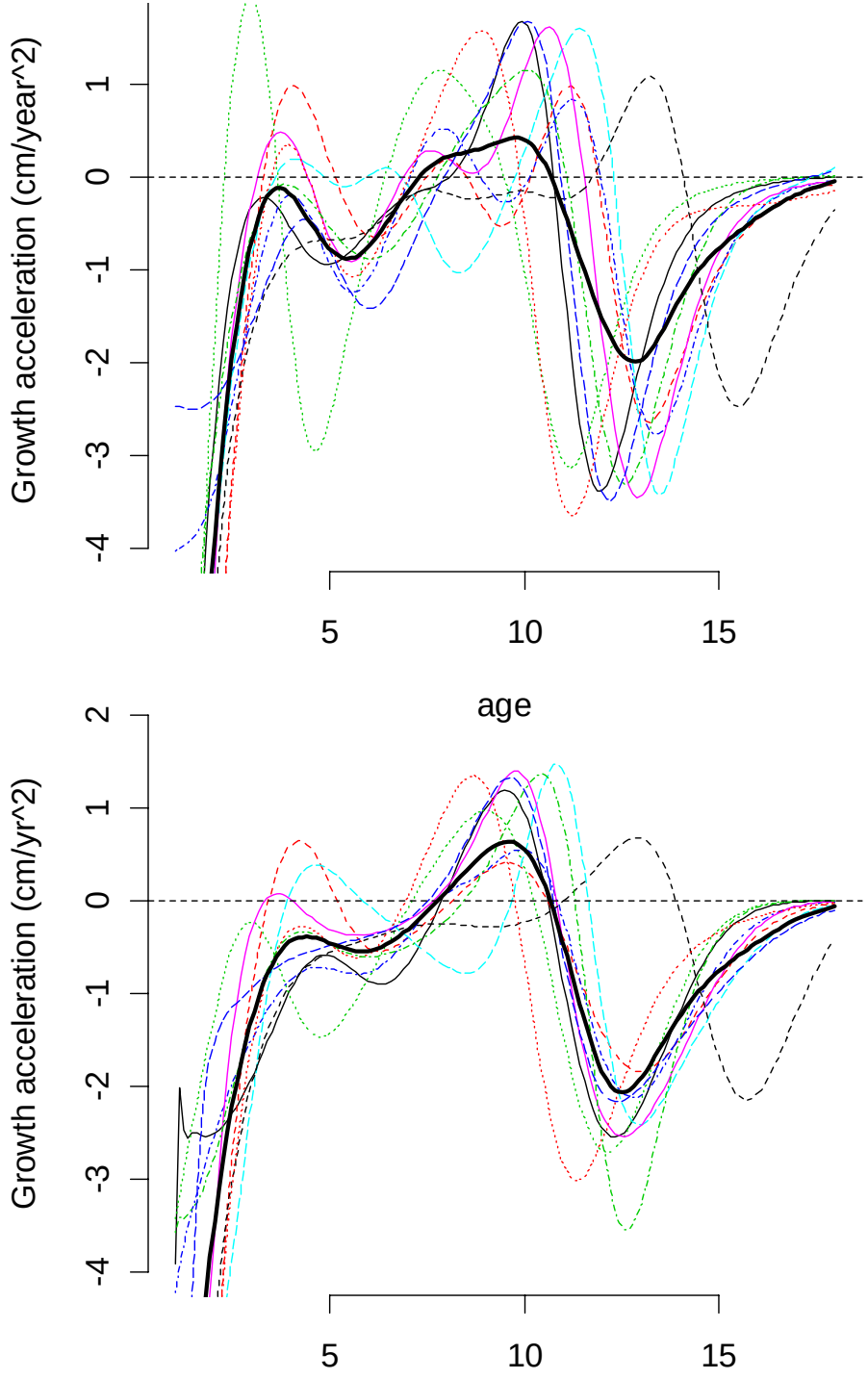
Acceleration

- Growth spurts occur at different ages
- Average shows the basic trend, but features are damped by improper registration



Registration

- register.fd all to the mean
- Not perfect, but better



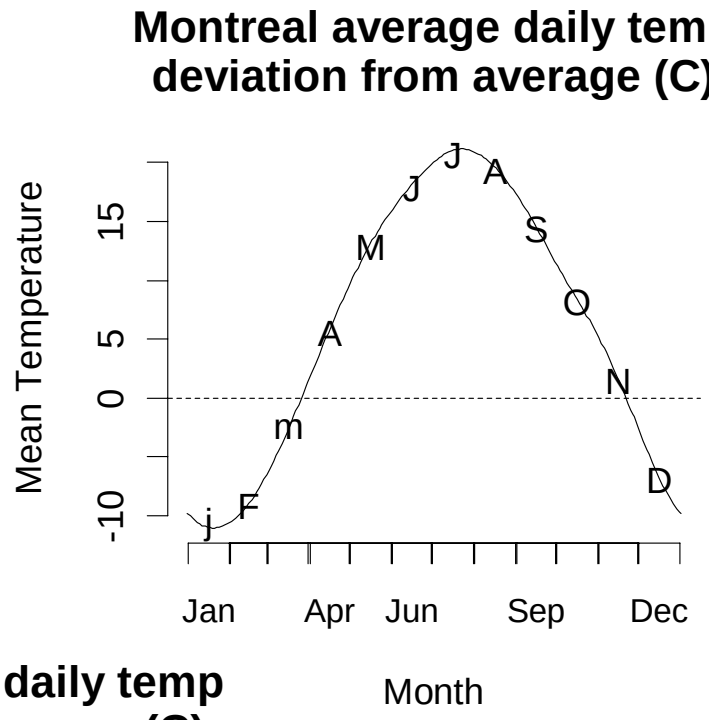
A Stroll Along the Beach

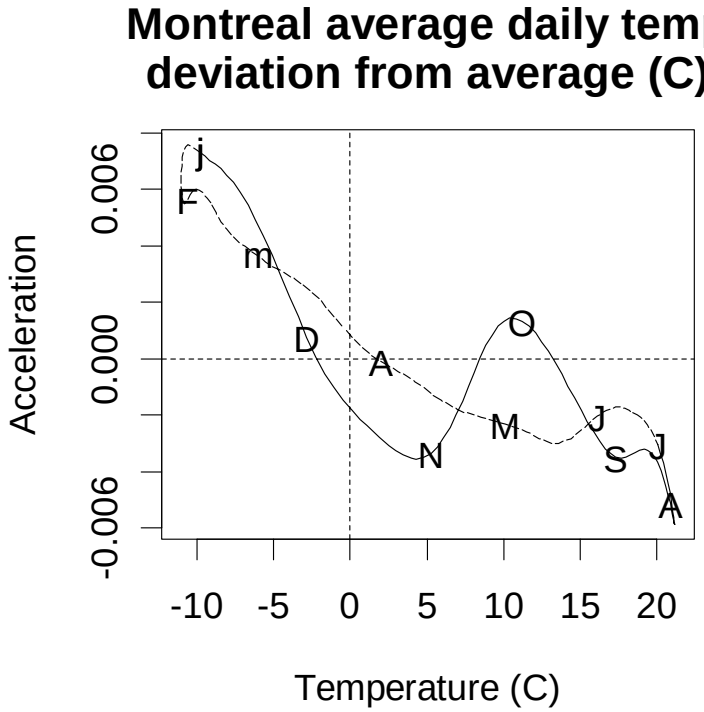
- Light intensity over 365 days at each of 190*143 = 27140 pixels was
  - smoothed
  - functional principal components

- http://www.stat.berkeley.edu/~wickham/userposter.pdf

Other fda capabilities

- **Correlations**
  - even with series of different lengths!
- **Phase plane plots**
  - good estimates of derivatives





afda-ch03.R
fda-ch01.R
fda-ch02.R

Script files for fda books

- Ramsay and Silverman
  - (2002) Applied Functional Data Analysis (Springer)
  - (2006) Functional Data Analysis, 2nd ed. (Springer)
- ~R\library\fda\scripts
  - Some but not all data sets discussed in the books are in the ‘fda’ package
  - Script files are available to reproduce some but not all of the analyses in the books.
  - plus CSTR demo

FDA and Differential Equations

- Many dynamic systems are believed to follow processes where output changes are a function of the outputs, **x**, and inputs, **u** (and unknown parameters θ):

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

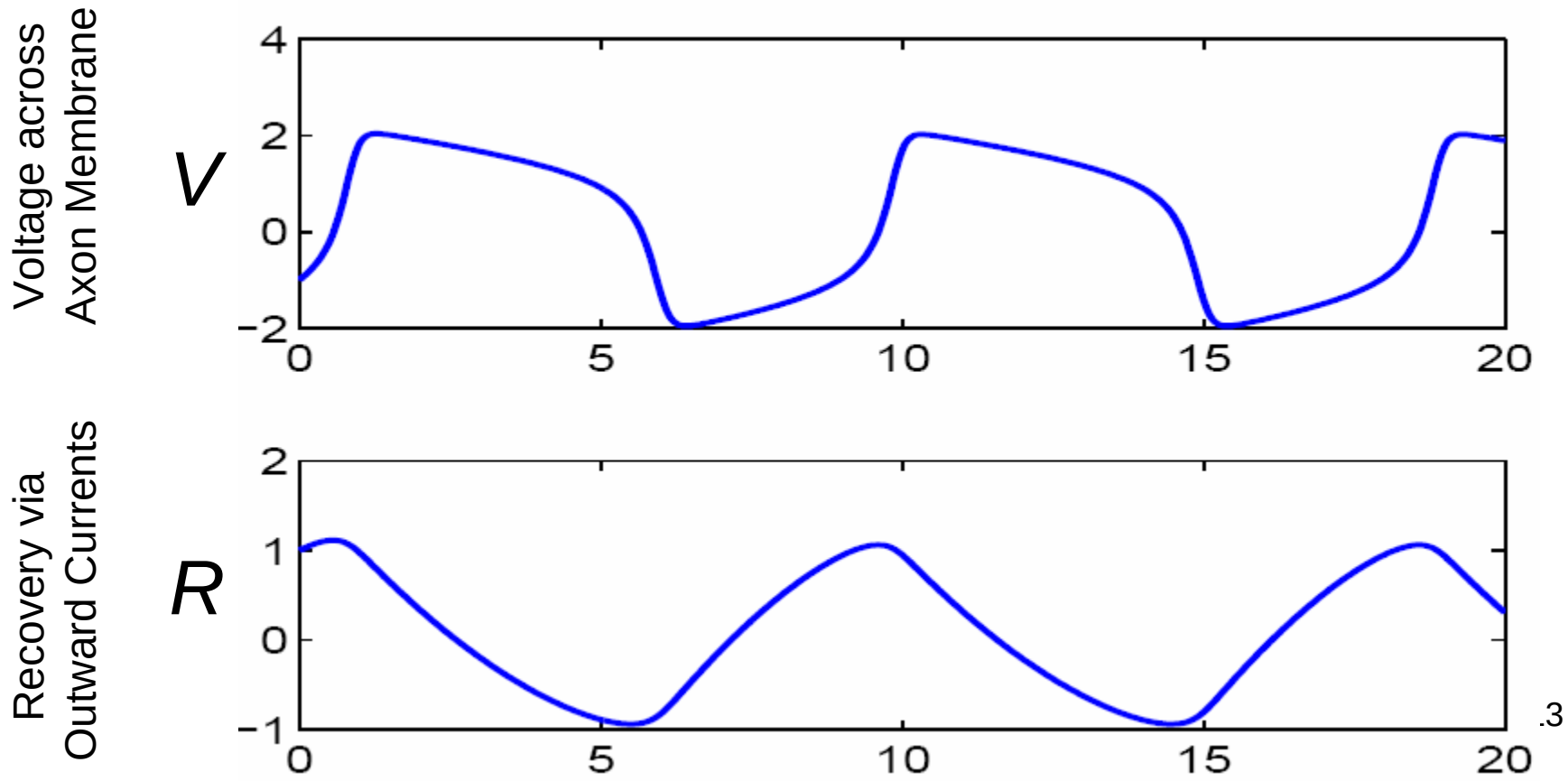
- Matlab was designed in part for these types of models

Squid Neurons

- FitzHugh (1961) - Nagumo et al. (1962) Equations: Estimate *a*, *b* and *c* in:

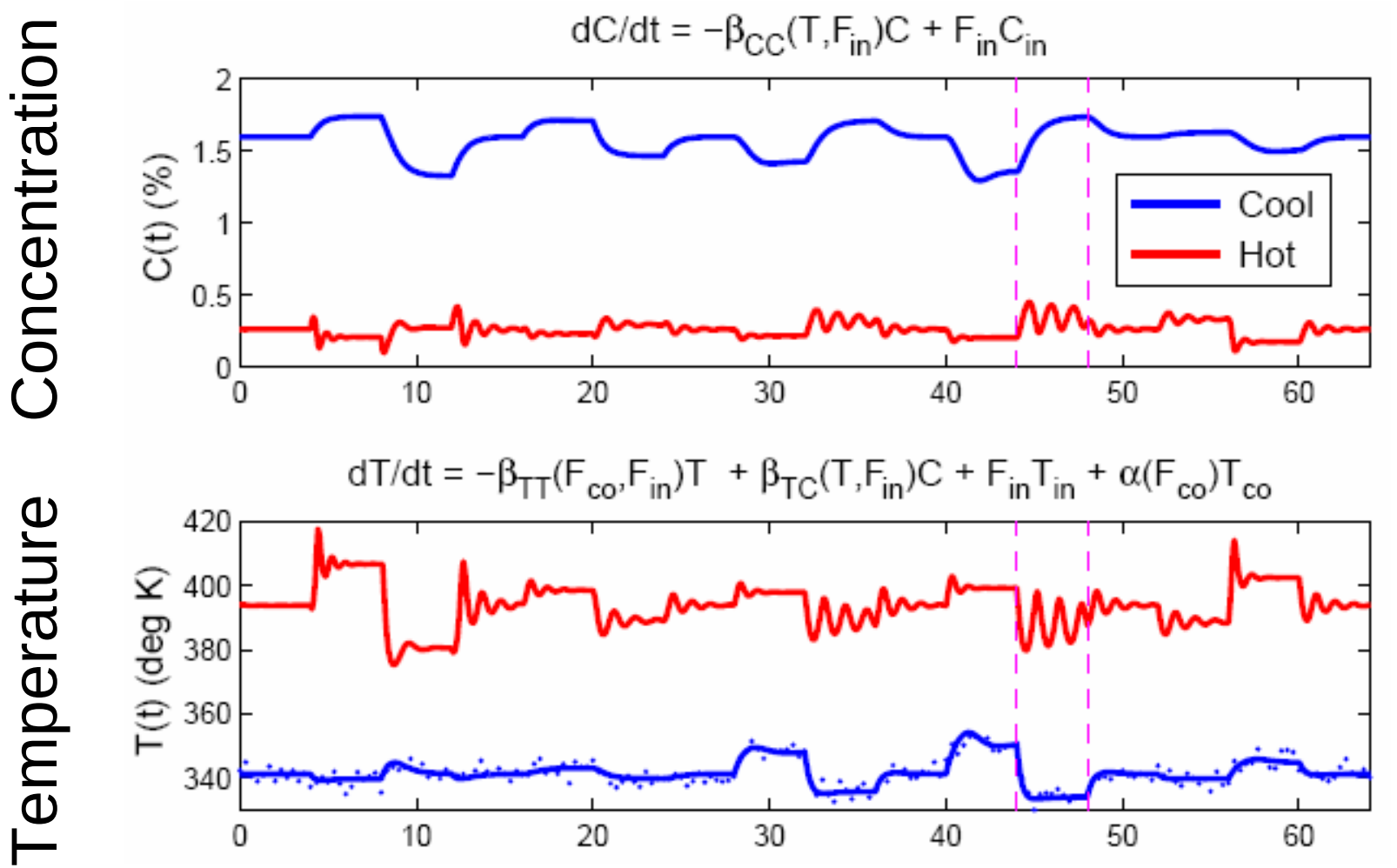
$$\dot{V} = c\left[V - \left(V^3/3\right) + R\right]$$

$$\dot{R} = -\left(V - a + bR\right)/c$$



Tank Reactions

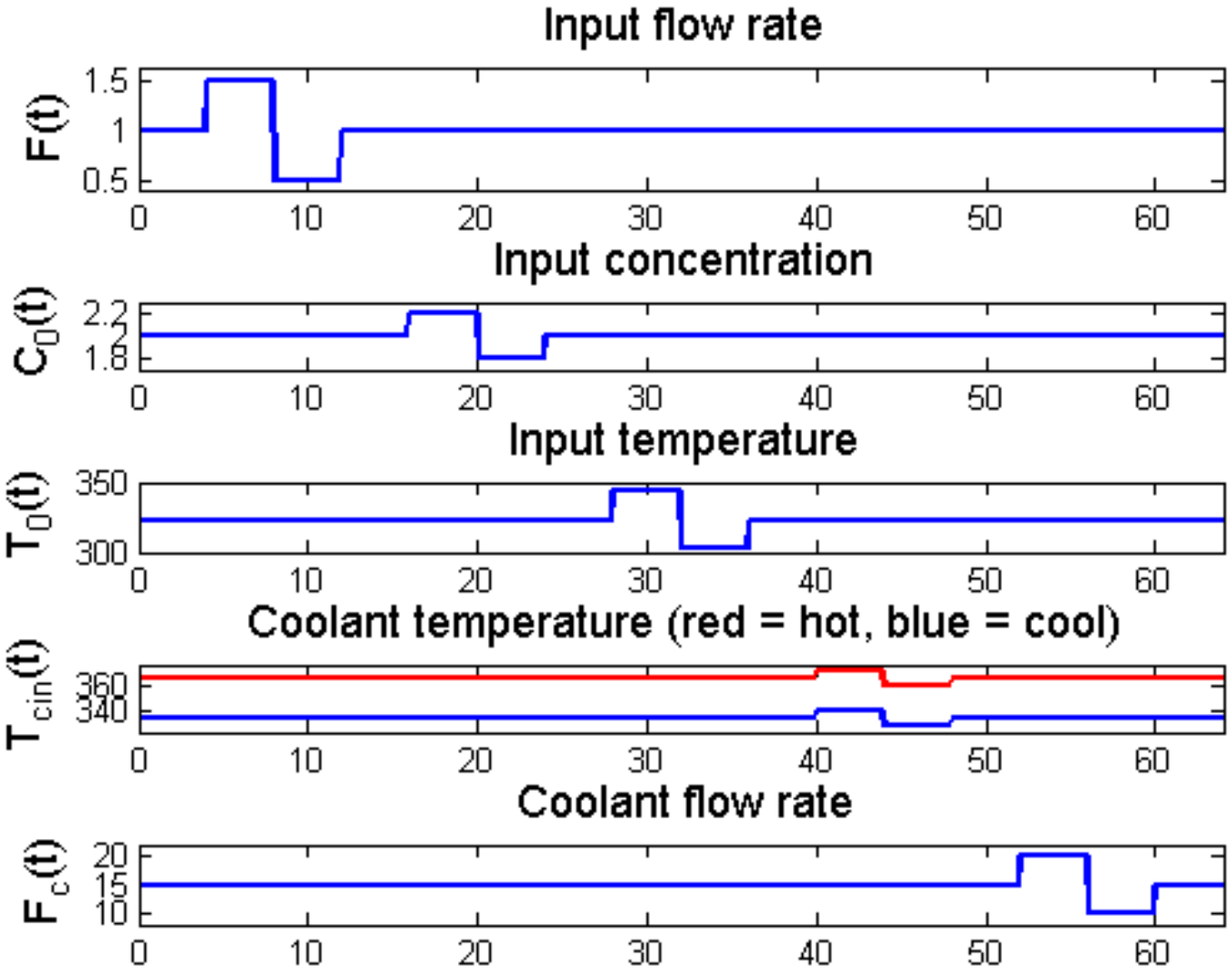
- Continuously Stirred Tank Reactor (CSTR)



Functional Data Analysis Process

1. Select Basis Set
2. Select Smoothing Operator
   - e.g., differential equation
   - equivalent to a Bayesian prior over coefficients to estimate
3. Estimate coefficients to optimize some objective function
4. Model criticism, residual plots, etc.
5. Hypothesis testing

Inputs to Tank Reaction Simulation



Computations: Nonlinear ODE

$$dC/dt = -\beta_{CC}(T, F_{\text{in}})C + F_{\text{in}}C_{\text{in}}$$

$$dT/dt = -\beta_{TT}(F_{\text{co}}, F_{\text{in}})T + \beta_{TC}(T, F_{\text{in}})C + F_{\text{in}}T_{\text{in}} + \alpha(F_{\text{co}})T_{\text{co}}$$

$$\beta_{CC}(T, F_{\text{in}}) = \kappa \exp\left\{-10^4 \tau\left(1/T - 1/T_{\text{ref}}\right)\right\} + F_{\text{in}}$$

$$\beta_{TT}(F_{\text{co}}, F_{\text{in}}) = \alpha(F_{\text{co}}) + F_{\text{in}}$$

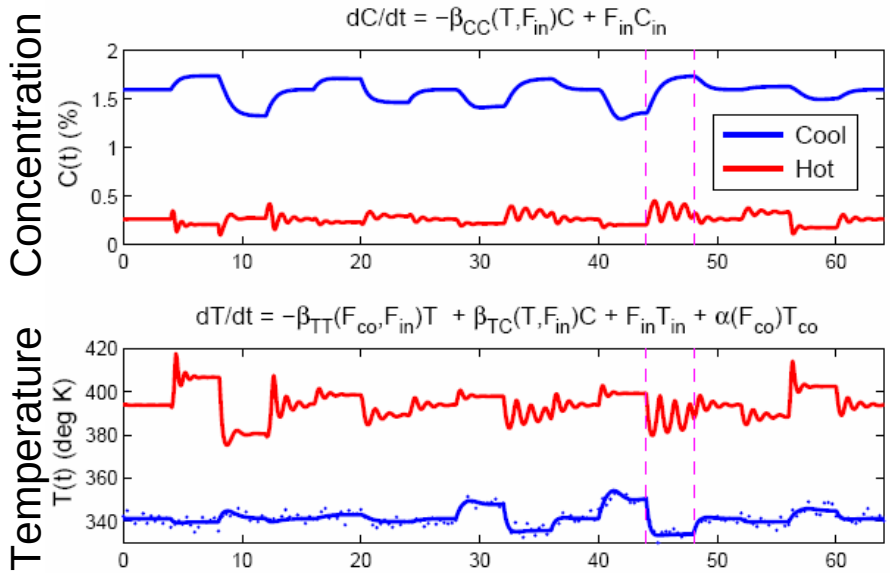
$$\beta_{TC}(T, F_{\text{in}}) = 130\beta_{CC}(T, F_{\text{in}})$$

$$\alpha(F_{\text{co}}) = aF_{\text{co}}^{\,b+1} / \left(F_{\text{co}} + aF_{\text{co}}^{\,b}/2\right)$$

4 parameters: κ, τ, a, b

estimate parameters (κ, τ, a, b)

- Compute Input vectors
- Define functions
- Call differential equation solver
- Summarize, plot



Three problems

- Estimate (κ, τ, a, b) to minimize SSE in Temperature only

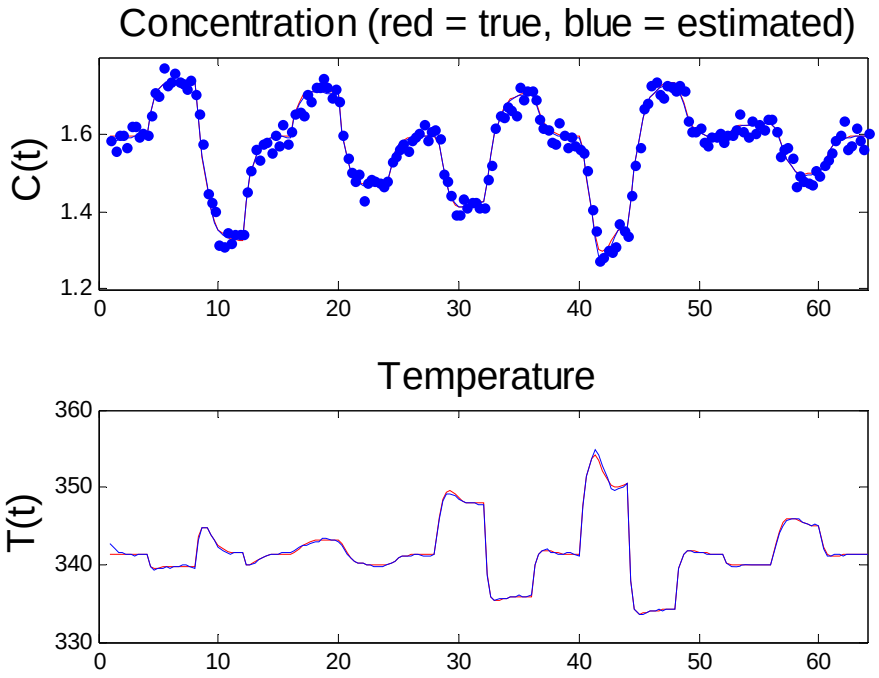
	function		SSE	SSE-min
Matlab	lsqnonlin		5.09888	0.00236
R	nls		5.09652	0
	optim	Nelder-Mead	5.09652	0
		BFGS	5.09652	0
		CG	5.09900	0.00248
		SANN	5.17504	0.07852
	nlminb		5.09652	0

SSE(Temp, Conc)

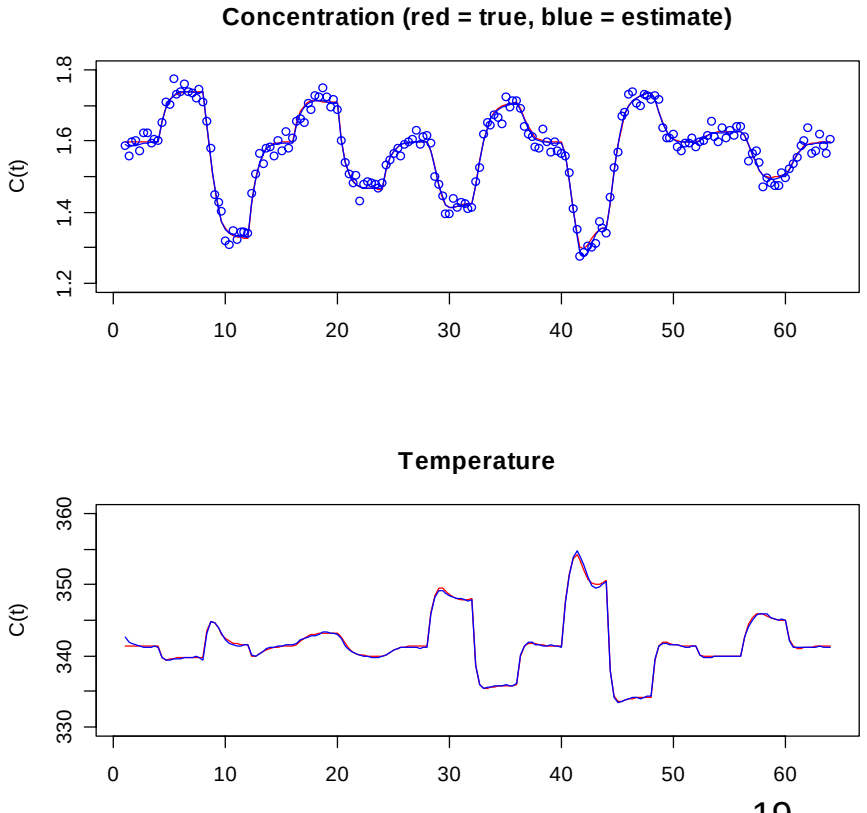
Median absolute relative error

	Matlab	R
Concentration	1.149E-03	1.145E-03
Temperature	2.640E-04	2.636E-04

- Matlab: lsqnonlin



- R: nls



R vs. Matlab

- Gave comparable answers
- R code for CSTR slightly more accurate but requires much more compute time
  - coded by different people
- R has helper functions not so easily replicated in Matlab
  - *summary.nls* →
  - confint.nls
  - profile.nls

```
      Estimate StdErr  t    Pr(>|t|)
kref   0.466  0.004 113.0  < 2e-16 ***
EoverR 0.840  0.009  94.7  < 2e-16 ***
a      1.720  0.232   7.4  8.2e-13 ***
b      0.496  0.050  10.0  < 2e-16 ***
```

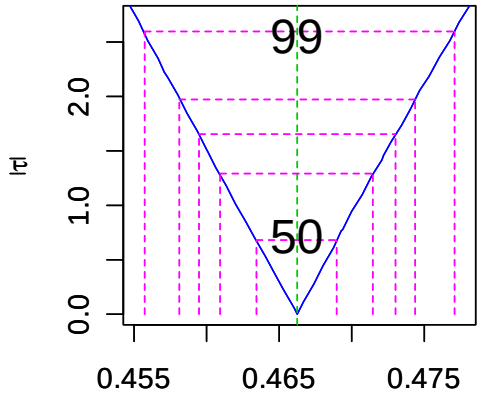
confint.nls

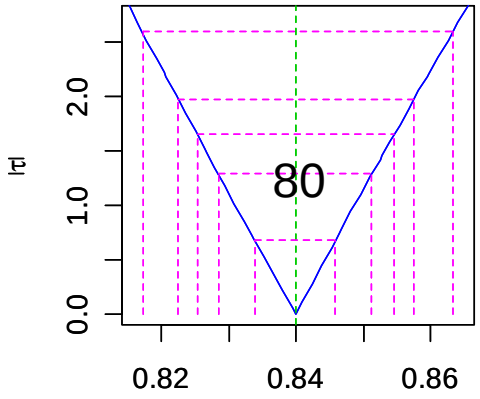
- Likelihood-based confidence intervals: generally more accurate than Wald intervals
  - Wald subject to parameter effects curvature
  - Likelihood: only affected by intrinsic curvature

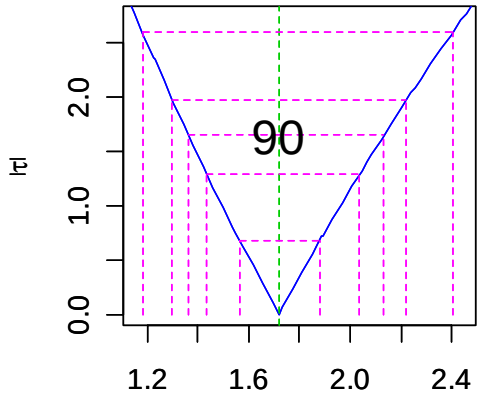
```
> confintNlsFit
            2.5%   97.5%
kref       0.458   0.474
EoverR     0.823   0.858
a          1.300   2.222
b          0.401   0.599
```

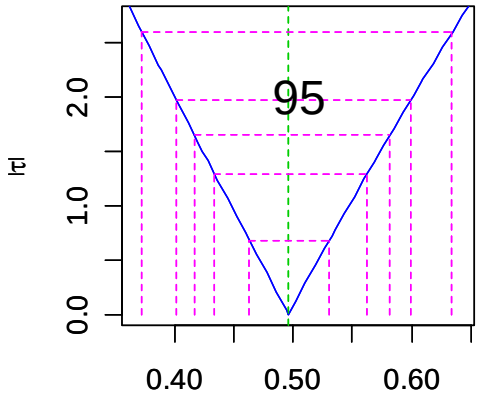
plot.profile.nls

- for a plot showing the sqrt(log(LR))









Conclusions

- R and Matlab give comparable answers
- R:nls has helper functions absent from Matlab:lsqnonlin

- Functional data analysis tools are key for
  - estimating derivatives and
  - working with differential operators

References

- www.functionaldata.org
- Ramsay and Silverman (2006) *Functional Data Analysis*, 2nd ed. (Springer)
- _____(2002) *Applied Functional Data Analysis* (Springer)
- Ramsay, J. O., Hooker, G., Cao, J. and Campbell, D. (2007) Parameter estimation for differential equations: A generalized smoothing approach (with discussion). *Journal of the Royal Statistical Society, Series B*. To appear.

NOT free-knot splines

- For this, see
  - **`DierckxSpline package`**
  - Companion to Dierckx, P. (1993). Curve and Surface Fitting with Splines. Oxford Science Publications, New York.
- R package by Sundar Dorai-Raj
  - links to Fortran code by Dierckx available from www.netlib.org/dierckx
- soon to appear on CRAN