Chapter 1

Background Material

1.1 Preliminaries

1.1.1 Functional Data Analysis

Functional data analysis (FDA) is a field of statistics where it is assumed that the data observed at a given set of independent observation times (or coordinates etc.) represent noisy observations of some underlying function.

The approach taken here is to assume that an unknown differential equation can adquately - though not necessarily exactly - describe the process producing the data.

Specification of Function Spaces

The functions in question are generally all assumed to be members of some countably infinite dimensional vector space, such as the set of all functions $f(\cdot)$ such that $\int_0^T |f''(t)|^2 dt < \infty$ over some interval $[0, T]$.

This assumption implies that any given function can be represented as a countably infinite combination of basis elements, which are themselves functions. This means for a chosen set of basis elements $\{\phi_1(t), \phi_2(t), \dots\}$ and any given function $f(t)$, there is a set of coefficients $\{a_1, a_2, \dots\}$ such that:

$$f(t) = a_1\phi_1(t) + a_2\phi_2(t) + \dots$$

Functional Data Analysis can thus be regarded as a generalisation of multivariate statistics where the number of dimensions is potentially infinite.

Substantial complications are introduced into the statistical analysis because functions are generally much richer objects than real numbers or vectors. A function will generally have a different value for each input value, and the number of non-integer numbers on any interval - and hence potential inputs - is infinite. Functions cannot be trivially represented on paper or in computer memory in a similar fashion as real numbers or vectors.

For the purposes of this thesis, it is assumed that the functions in question are continous mappings.

In practice one attempts to resolve this difficulty by finding or otherwise constructing a discrete problem that resembles the functional problem, and then solve this approximate problem.

It might be that case that the approximate problem can itself only be solved approximately using numerical methods.

Statistical problems that involve differential equations are particularly difficult. More naive approaches force the practitioner to solve the ordinary differential equation (ODE) numerically everytime it is desired to evaluate the goodness of fit. For these situations, it is necesary by definition to use numerical analytic techniques to construct a proxy problem that resembles the original problem sufficiently well and that is sufficiently easy to work with.

For example, consider the problem of parametric estimation for a stochastic differential equation (SDE) of the form

$$dX = f(X;\theta)dt + \sigma dW.$$

Here $X(t)$ is the stochastic process being modelled, $f(\cdot;\theta)$ is a known function with a parameter θ to be estimated, σ is a volatility parameter, and $W(t)$ is a standard Brownian motion.

This SDE is equivalent to asserting for any time t and increment h that

$$X(t+h) = X(t) + \int_t^{t+h} f(X(s);\theta)ds + \sigma[W(t+h) - W(t)].$$

Suppose there are observations $X_1, X_2, \ldots, X_N$ of $X(t)$ at evenly spaced times, and that h is the distance between the time points. The integral formulation of the SDE suggests that if h is small enough, then

$$X_{k+1} \approx X_k + f(X_k;\theta) + \sigma h Z_{k+1}.$$

The Z_k here are i.i.d standard Normal random variables. This is known as the *Euler-Maruyama Approximation*.

Instead of attempting to estimate parameters for the SDE, we can fit parameters for a non-linear AR(1) process that acts as a proxy problem for the original SDE. This is a much more tractable problem than the original SDE.

In FDA, the assumption is usually made that all the functions can be represented as a linear combination from some choosen *finite* set of basis functions. Rather than discretise the differential operator as in the above example, the space of functions is discretised instead.

A differential equation (or a similar problem) over some finite dimesional space of functions with n dimensions can be represented as a problem over the Euclidean space R^n, this is a discrete problem.

The modelling process for functional data as described in Figure 1.1 can be more complex than standard statistical problems.

As is the case for typical stastical problems, the first step is to . Here one must only be certain that the model at used is sufficiently broad or well-specified to be able to actually capture the phenomona at hand.

Formulate a Model.

Construct a Discretised Model that Approximates the Orginal Model. Unless the statistical model is trivial, the next step is to construct a proxy model. This generally requires ideas from Numerical Analysis.

Conduct Statistical Analysis Using the Discretised Model. While the discretised model tends to be simpler than the original model, this task is not necesarily trivial as shall be seen.

Check the Approximation Error in Discretised Model. If the discretised model is too poor an approximation, then the results of any statistical analysis conducted could be substantially biased as a result of the approximation error introduced, even if the original model were perfectly sound. If the orginal model is biased, then the approximate one might be even more so.

Therefore, one should consider conducting post hoc checks. For example, running the analysis again with a more accurate approximate model and comparing the results with the original model. If both agree, it is evidence the approximate models are both reasonably accurate. Constructing a more accurate approximation is generally a straightforward and intuitive process, with the exact approach depending on the situation at hand.

In the context of FDA, this generally entails increasing the number of basis functions so that the associated approximation error is smaller. This is too complex a question to consider in sufficient detail at this point, so here is an illustrative example of how an analysis would be conducted, that only makes use of elementary ideas from Statistics and Numerical Analysis:

Suppose that one were attempting to estimate the parameters of an ODE by means of least squares, and one was using a finite-difference solver to compute the fitted values, and hence to determine the goodness-of-fit.

Once the fitting alogrithm had converged, one might run the solver again with a smaller stepsize and the same parameters and check if this has made a substantial change in the the sum of squared residuals.

If there has been a substantial change as a result of the stepsize reduction, then one would have to consider running the entire fitting procedure again starting from the previously computed parameter estimate, except with the smaller stepsize, taking the

new parameter estimates, and then checking again if decreasing the stepsize yet again produces substantial change in the goodness-of-fit statistic.

This procedure can even be automated. The Implicit Filtering alogrithm computes an approximate gradient using finite differnences and uses this to perform optimisation. If the alogorithm cannot produce a decrease in the objective function, or it cannot be certain that the true gradient isn't in fact zero, it reduces the stepsize. The algorithm terminates when the change in the objective function between changes in the stepsize has fallen below a chosen tolerance level.

If the fitting method used is slow however, then these such approaches can potentially be very slow due to the need to solve the same problem over and over again at increasing levels of precision.

Fortunately, Functional Data Analysis does not always require the recomputation of the curve in such a fashion whenever the parameters are changed. Instead of being implicitly represented as solutions of an ODE, functions are explicitly represented as elements in some finite dimesional vector space. As shall be seen, the objective function is generally a mapping from some vector space R^n to R that can often be evaluated reasonably easily, or at least more easily than having to run an ODE solver.

Check If Results of Statistical Analysis Are Consistent With Discretised Model. In the previous step, one checked that the approximate model was actually acting as a proxy for the original model. One must then check that the statistical analysis conducted using the approximate model is valid in its own right.

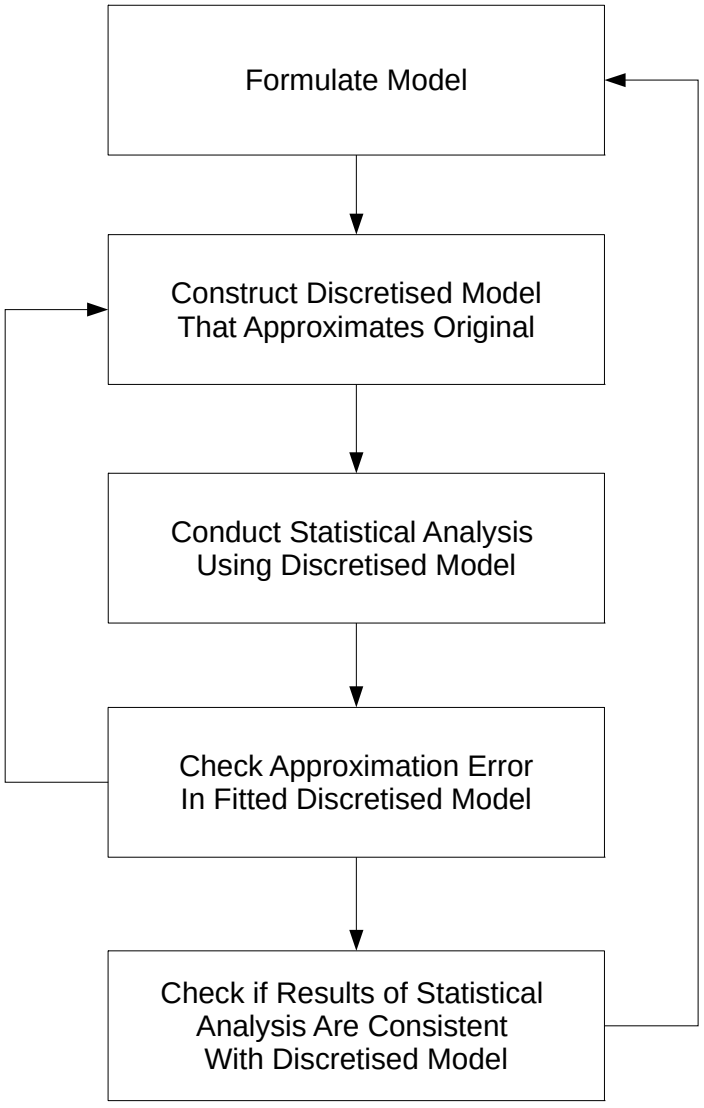

Figure 1.1: Statistcal Modelling Process For Functions

1.1.2 Penalised Regression

Suppose we have N noisy observations y_i at times t_i from some function $f(t)$, and we wish to estimate $f(t)$, from the data. A naive approach would be to estimate $f(t)$ by minimising a least squares criterion:

$$SSE(f) = \sum_{i=1}^{N} (y_i - f(t_i))^2$$

Here, $SSE(\cdot)$ is a function that assigns a real number to every real-valued function that is defined for all the t_i.

There is an obvious problem with this criterion - it does no t have a unique minimiser. Any function $g(t)$ such that $g(t_i) = y_i$ is will minimise $SSE(\cdot)$. There are an infinite number of degrees of freedom, but only a finite number of observations.

To ensure uniqueness, it is necessary to impose a further condition to discriminate between different candidates, a way to choose between different functions that interpolate a given set of points.

Smoothing Splines

One potential criterion is to introduce a second order penalty. If two functions fit the observed data equally well, the more regular or less "wiggly" function is chosen. There are several ways of translating this intuition into a formal fitting procedure.

A common choice is to measure the degree of irregularity by using the integral of the second derivative over a chosen interval $[0, T]$. The upper limit T should be choosen to allow for all observation times to be included.

$$\int_0^T |f''(t)|^2 dt.$$

For a given set of points, the smooth interpolating curve that minimises the energy integral above is given by an interpolating cubic spline.

Choosing the most regular interpolating curve is not necessarily a very good estimation strategy however because it strongly prioritises goodness-of-fit above all other considerations. If the data is noisy, there is a risk of overfitting and poor predictive power. There is a trade-off between bias and variance.

In practice, a joint estimation strategy is pursued that attempts to find a good balance between fidelity to the observed data and reasonably regular behavior. This involves minimising the following penalised least squares criterion:

$$PENSSE(f; \lambda) = \sum_{i=1}^{N} (y_i - f(t_i))^2 + \lambda \int_0^T |f''(t)|^2 dt$$

The λ term dictates the trade-off between fidelity to the data and regularity.

Suppose there were a candidate function $g(t)$, then by taking the cubic spline such that its value at t_i is equal to $g(t_i)$, we can produce a curve $s(t)$ that has the same least-squares error as $g(t)$, but with $\int [s''(t)]^2 dt \leq \int [g''(t)]^2 dt$. Thus, the curve that minimises $PENSSE$ can be assumed to be a cubic spline.

To find the minimiser of $PENSSE(\cdot; \lambda)$ first, assume that $f(t)$ can be represented as a linear combination of K cubic spline functions $\phi_i(t)$ that can represent any cubic spline with knots at the t_i. This implies that

$$f(t) = \sum_{i=1}^{K} c_i \phi_i(t).$$

Note that it is only required that the set of basis splines only possess enough resolution to represent the function that minimises $PENSSE$, it is not required that this set of splines is minimal.

A perfect basis is generally not needed, the basis need only be a reliable workhorse. This obviates what otherwise could prove to be a distracting nuiscence for a practioner who wishes to conduct data analysis instead of getting entangled in minutae.

Let the design matrix $\mathbf{\Phi}$ be defined by $\mathbf{\Phi}_{ij} = \phi_i(t_j)$, and let the weight matrix $\mathbf{R}$ be defined by $\mathbf{R}_{ij} = \int_0^T \phi_i''(t)\phi_j''(t)$. Then $PENSSE$ can be written in terms of the vector of coefficients $\mathbf{c}$ and observations $\mathbf{y}$ as:

$$PENSSE(\mathbf{c}; \lambda) = \|\mathbf{y} - \mathbf{\Phi c}\|^2 + \lambda \mathbf{c}'\mathbf{R}\mathbf{c}$$

The problem has been discretised into one on R^K.

The optimal value of $\mathbf{c}$ is given by

$$\hat{\mathbf{c}} = (\mathbf{\Phi}'\mathbf{\Phi} + \lambda \mathbf{R})^{-1}\mathbf{\Phi}'\mathbf{y}$$

This is an exact solution to the original problem because the span of the $\{\phi_i(t)\}$ contains the function that minimises $PENSSE$. The coefficient vector $\hat{\mathbf{c}}$ is the set of coordinates of the optimal function within this finite-dimensional vector space.

Piecewise Trigonometric Interpolation

Consider a more difficult penalised regression problem, that gives a sense of the limits of the approach employed so far.

$$PENSSE(f; \lambda) = \sum_{i=1}^{N} (y_i - f(t_i))^2 + \lambda \int_0^T |f''(t) - f(t)|^2 dt$$

$PENSSE$ can be minimised in this case taking by a piecewise function consisting of linear combinations of $\sin(t)$ and $\cos(t)$ over each interval, and matching them together.

Note that a function of the form $a_0 + a_1 \cos(t) + b_1 \sin(t) + a_2 \cos(2t) + b_2 \sin(2t) + \ldots$ can be written as a polynomial in e^{it} and e^{-it}. For this reason, such a piecewise trignometric function can also be referred to as a piecewise trignometric polynomial.

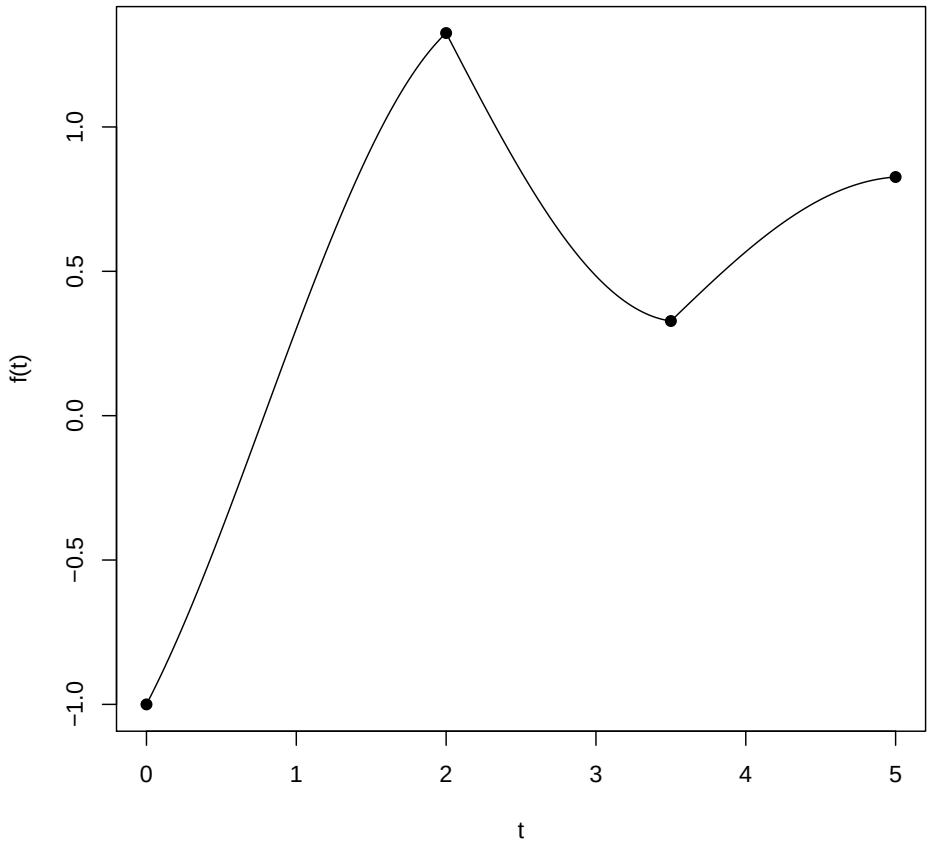

Figure 1.2: Plot of a Piecewise Trigonometic Curve. Note the kinks between segments.

As can be seen in Figure 1.2, a piecewise trignometric polynomial of second degree generally fails to be smooth at the boundary points, and thus has a kinked appearance. For the purposes of statistical modelling, it is strongly desireable impose the additional constraint that $f(t)$ must at least everywhere differentiable.

1.1.3 Finte Dimensionalisation: the General Case

To find an exact solution to the two problems in the previous section, it was necessary to construct a finite dimensional function space that contained the minimal function. However it is not guaranteed that this is always possible. In practice, one would hope that the optimal function can approximated sufficiently well by taking a linear combination from some choosen set of functions. Spline bases tend to be a reliable workhorse that are effectively the default choice. They provide a good balance between being well behaved as objects for regression and having good approximating power.

For comparison, Chebyshev Polynomials can often provide better approximation power for a given number of basis functions and are used in Computational Fluid Dynamics for this purpose. Unfortunately, they can be poorly behaved statistically because they consist of high order polynomials that are difficult to fit to data.

Functional Data Analysis thus consists of the following steps:

1. Formulate a model for $f(t)$. Usually, this takes the form of a penalised regression model, where $f(t)$ is defined as the function that minimises some kind of penalised error

2. Assume that $f(t)$ can be written as a finite combination of chosen basis functions. In practice, this is only approximately true, so it is important to ensure that our basis can actually approximate the optimal $f(t)$ sufficently well. The function $f(t)$ can thus be written:

$$\begin{aligned} f(t) &= \sum_{i=1}^{K} c_i \phi_i(t) \\ &= [c_1, \ldots, c_K]'[\phi_1(t), \ldots, \phi_K(t)] \\ &= \mathbf{c}'\phi(t) \end{aligned}$$

   Note that $f(t)$ is now defined by the coefficent vector $\mathbf{c}$.

3. Formulate the model in terms of the coefficient vector $\mathbf{c}$. A statistical problem over some given functional space has been transformed into a statistical problem over R^K.

As is done in some texts, we will immediately provide an example with a very small basis to illustrate these steps [1]. Consider the following penalised regression problem:

$$PENSSE(f;\lambda) = \sum_{i=1}^{N}(y_i - f(t_i))^2 + \lambda \int_0^1 |t^2 f'' - 0.5f|^2 dt$$

The differential equation associated with the penalty term is known as an Euler's Equation, and can be regarded here as a toy equation representative of equations such

as Bessel's equation. The solution is given by $f(t) = at^{r_1} + bt^{r_2}$, where r_1 and r_2 are the roots of the quadratic equation $r^2 - r - 0.5 = 0$. Thus, $r_1 \approx -0.36$ and $r_2 \approx 1.36$.

Some texts in the field, for the sake of illustration it will be assumed that that $f(t)$ can be written as a quadratic - a linear combination of the basis functions $\{1, t, t^2\}$:

$$f(t) = at^2 + bt + c$$

Then:

$$\begin{aligned}
\int_0^1 |t^2 f'' - 0.5f| dt &= \int_0^1 \left| at^2 - \frac{1}{2}(at^2 + bt + c) \right|^2 dt \\
&= \int_0^1 \left| \frac{1}{2}\left(at^2 - bt - c\right) \right|^2 dt \\
&= \frac{1}{4}\int_0^1 |at^2 - bt - c|^2 dt \\
&= \frac{1}{4}[a \ -b \ -c]'\mathbf{H}[a \ -b \ -c] \\
&= \frac{1}{4}[a \ b \ c]'(\mathbf{A}'\mathbf{H}\mathbf{A})[a \ b \ c] \\
&= [a \ b \ c]'\mathbf{K}[a \ b \ c]
\end{aligned}$$

Here $\mathbf{K} = \frac{1}{4}\mathbf{A}'\mathbf{H}\mathbf{A}$, the elements of the matrix $\mathbf{H}$ are defined by $\mathbf{H}_{ij} = \int_0^1 t^i t^j dt = 1/(i + j + 1)$, and elements of the matrix $\mathbf{A}$ are given by:

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

Thus, the penalised error is given by:

$$PENSSE(a, b, c; \lambda) = \sum_{i=1}^{N} (y_i - at_i^2 - bt_i - c)^2 + \lambda[a \ b \ c]'\mathbf{K}[a \ b \ c] \tag{1.1}$$

Thus, we now gone from a problem specified in terms of functions, to a penalised least squares problem in the three coefficents a, b and c. The quality of this approximate model as λ gets larger and larger depends on how well the functions $t^{-0.36}$ and $t^{1.36}$ can be respectively approximated by quadratics over the interval $[0, 1]$.

To illustrate this example further, the method was fitted to simulated data. A solution to the ODE $t^2 f'' - f = 0$ was generated over the interval $[0, 1]$, samples were taken at various points before being corrupted by Gaussian noise. The quadratic that minimised (1.1) with $\lambda = 100$ was then found. For comparison, the data was also fitted to a quadratic using ordinary least squares. The original function $f(t)$, the perturbed data, and the two fitted functions are all shown in Figure 1.3

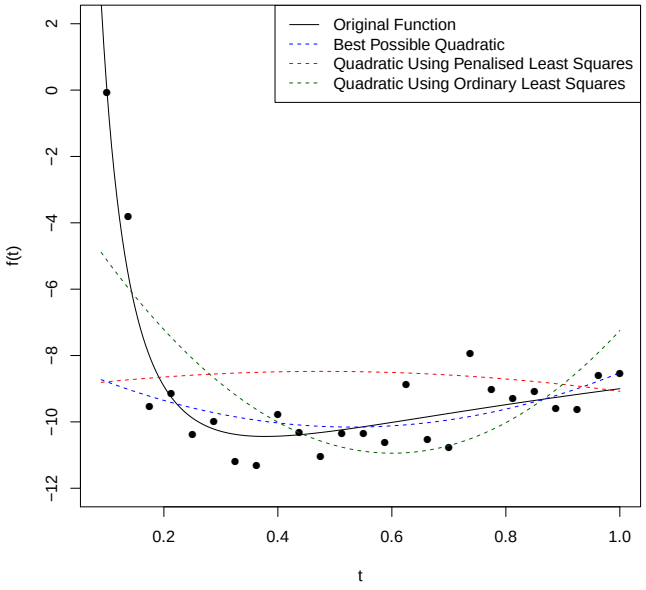

Figure 1.3: Performing FDA with the differential operator $Lf = t^2 f'' - 0.5f$ and the basis set $\{1, t, t^2\}$.

It's already been noted that the quality of the model depends partially on how well $f(t)$ can ever be approximated by a quadratic over $[0, 1]$ in the first place. Therefore, the quadratic $q(t)$ that minimises $\int_0^1 |f(t) - q(t)|dt$ was found numerically and also plotted in Figure 1.3.

Figure 1.3 suggests that $f(t)$ can be approximated reasonably well by quadratics for so long as one stays away from the point $t = 0$. This is consistent with theory. The ODE $t^2 f'' - f = 0$ behaves degenerately at the origin. When $t = 0$, the ODE has what is known as a singular point, the term in front of f'' becomes zero so that the ODE reduces to $(0)^2 f'' - f = 0$. Additionally, it is always the case that the second derivative diverges to infinity at 0 if $f(t)$ is of the form $at^{-0.36} + bt^{1.36}$. As a result of both the singular point and infinite curvature at $t = 0$, polynomial approximation is predicted to be execptionally tricky around this point.[12, 30]

Comparing the two fits in Figure 1.3, it is fair to argue that the penalised regression model captures the shape of $f(t)$ better than ordinary least squares away from $t = 0$. Both models seem to have similar predictive power on average. The penalised fit is being heavily influenced by the singularity at $t = 0$ and probably would have performed better if a more robust loss function than least squares were used.

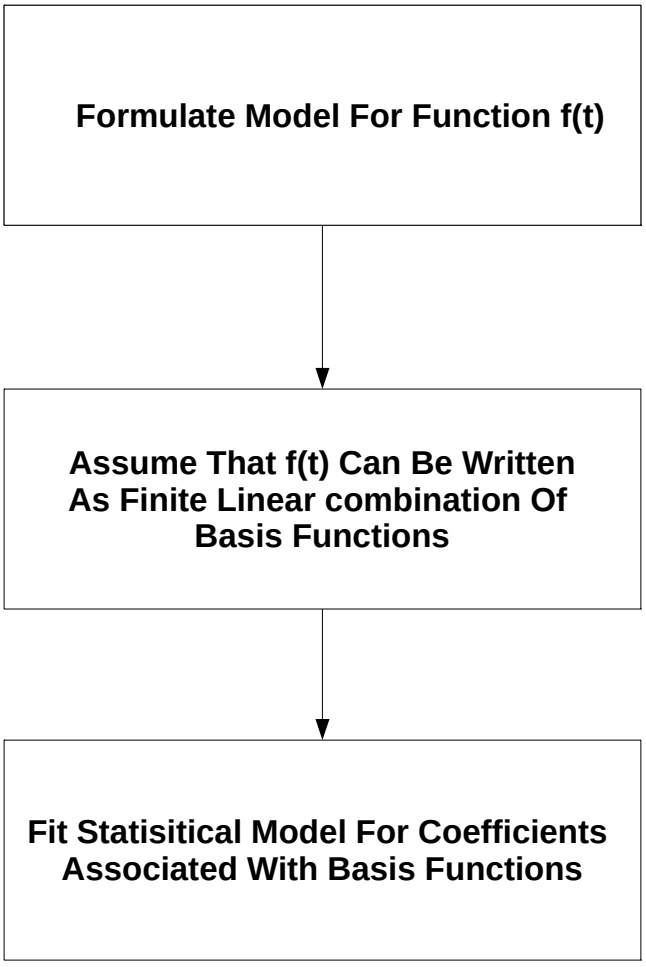

Figure 1.4: Statistical Modelling Process For Functional Data Analysis

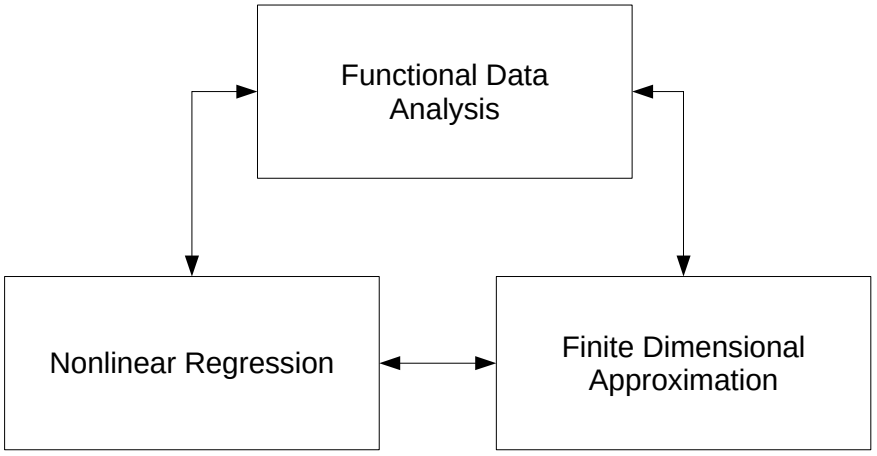

Figure 1.5: Elements of Functional Data Analysis

1.2 The Two-Stage Parameter Cascade

Consider the following penalised regression problem:

$$PENSSE(f,\theta) = \sum_{i=1}^{N}(y_i - f(t_i))^2 + \lambda \int_0^1 |T_\theta f|^2 dt.$$

Here T_θ is some differential operator, that is parameterised by an unknown θ that is to be estimated.

T_θ can be an ordinary differntial operator or a partial differential operator; and whether it is linear, quasi-linear, or nonlinear.

There are two statistical objects to be estimated here: the parameter θ, and the function $f(t)$.

Ramsay and Cao propose the following hierarchial approach to estimation[4]:

Given a fixed value of θ, let $f(t|\theta)$ denote the function that minimises $PENSSE(f,\theta)$

For a given value of θ, it's associated mean square error is defined by:

$$MSE(\theta) = \sum_{i=1}^{N}[y_i - f(t_i|\theta)]^2$$

By making $f(t)$ dependent on θ, the fitting problem has been reduced to a problem in non-linear least squares.

This leaves the issue of estimating the optimal value of θ - Ramsay and Cao propose the use of gradient descent.

For a given value of θ, $f(t|\theta)$ is found. These two values together are then used to compute $MSE(\theta)$ and $\nabla MSE(\theta)$. Finally, a new value of θ is computed by perturbing θ in the direction of the gradient.This scheme is sketched out in Figure 1.6.

It is assumed that $f(t)$ can be represented by a finite vector $\mathbf{c}$ associated with an appropriate basis. This leads to a pair of nested optimisation problems: the *Inner Optimisation* involves finding the value of $\mathbf{c}$ that minimises the penalised least squares criterion given θ, and the *Middle Optimisation* entails finding the of value of θ that minimises $MSE(\theta)$.

There is thus a "cascade" of estimation problems, where the results of the lower level estimation problem feeds back in to the higher level one.

Note that every time a new value of θ is introduced, the associated function $f(t|\theta)$ must be computed from scratch. The middle optimisation can thus generate many inner optimisation subproblems as the parameter space is explored, and these in turn could require multple iterations to complete if no explicit formula for $\mathbf{c}$ given θ is available.

Figure 1.6 is a high level overview of the Parameter Cascade, in particular, the step of computing $f(t|\theta)$ is presented as a single atomic and organic step, even though it could be a complex process in its own right. This risks masking some of the comptuational work that is happening. A more complete description is provided in Figure 1.7.

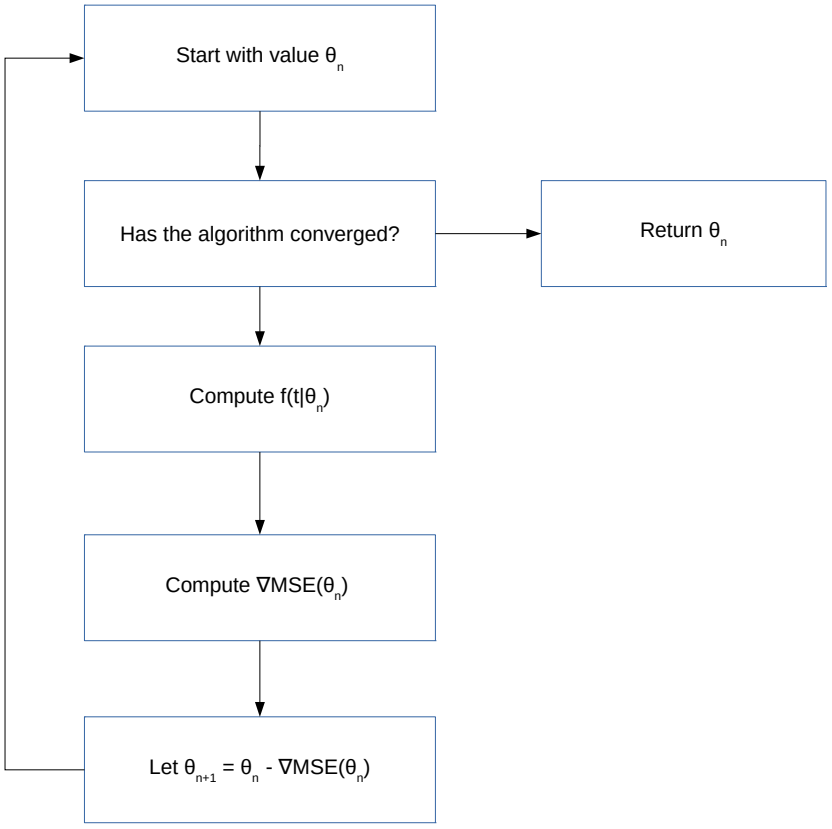

Figure 1.6: Two Stage Parameter Cascade (Simplified)

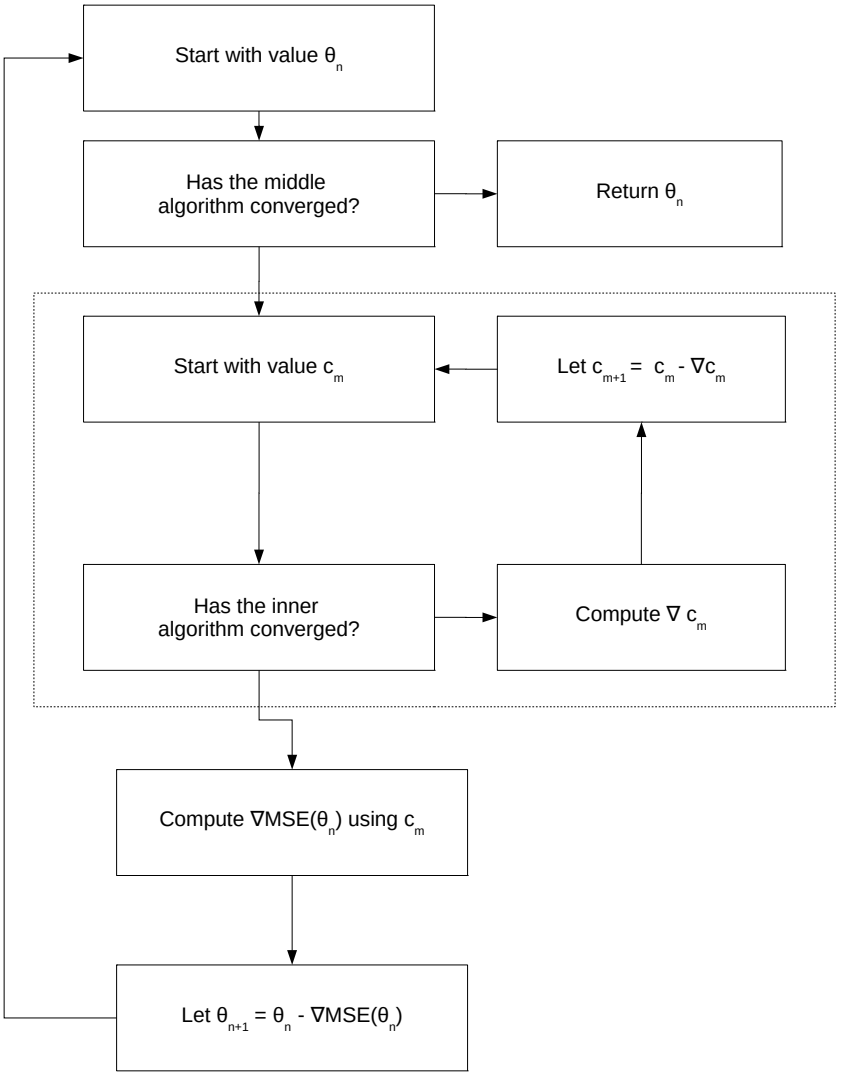

Figure 1.7: Schematic of the Two Stage Parameter Cascade With the Inner Optimisation Visible

1.3 The Data2LD Package

The Data2LD package is an R package intended to perform smoothing using the Paramter Cascade to fit linear differential operators with a forcing function, that is, ODEs of the form:

$$\sum \beta_i(t) D^i f(t) = u(t)$$

1.3.1 Reflux Data

The Reflux data, plotted in Figure 1.8, describes the output of an oil refining system. A given fraction of oil is being distilled into a specific tray, at which point it flows out through a valve. At a given time, the valve is switched off, and distillate starts to accumulate in the tray [26].

ODE Fit

$$\begin{cases} y'(t) = -\beta y(t) & t \leq t_0 \\ y'(t) = -\beta y(t) + u_0 & t \geq t_0 \\ y(0) = 0 \\ y'(0) = 0 \end{cases}$$

This ODE admits an exact solution. Letting $\gamma = u_0/\beta$ and C be an arbitray constant, then the solution is given by

$$y(t) = \begin{cases} 0 & t \leq t_0 \\ \gamma + Ce^{-\beta(t-t_0)} & t \geq t_0 \end{cases}$$

Without loss of generality the exponential term $Ce^{-\beta(t-t_0)}$ can be replaced with one of the is of the form $Ce^{-\beta t}$. This is the case because $Ce^{-\beta(t-t_0)} = Ce^{-\beta t}e^{-\beta t_0} = [Ce^{-\beta t_0}e^{-\beta t}]$, the $e^{-\beta t_0}$ term is thus absorbed into the constant term.

In order to ensure that $y(t)$ is continous at t_0 and monotone increasing, we require that $\gamma + C = 0$ and that $\beta > 0$

This model is perfectly adaquate, but it turns out that the constraint $C = -\gamma$ is unsuitable from the point of view of numerical parameter estimation.

However, if we allow t_0 to vary, we can allow C to assume any negative value while preserving monotonicity and continuity.

Assume that $y(t)$ is instead given by:

$$\tilde{y}(t) = \max(0, \gamma + Ce^{-\beta(t-t_0)})$$

The function $\tilde{y}(t)$ satisfies the same ODE and initial conditions as $y(t)$ except that the change point t_0 is shifted to t_0' defined by:

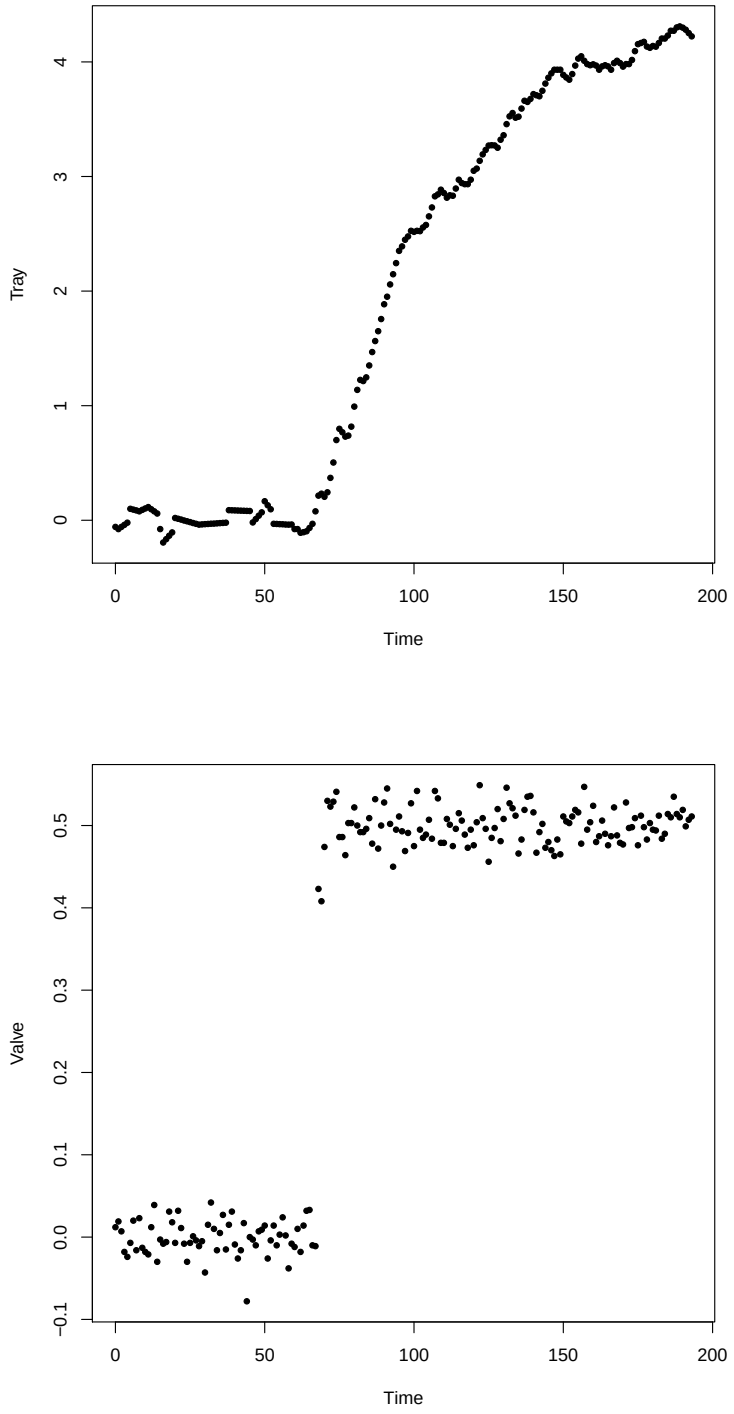

Figure 1.8: Reflux Data

$$t_0' = \max\left(t_0, t_0 - \frac{1}{\beta}\ln\left(\frac{-\gamma}{C}\right)\right)$$

The function $\tilde{y}(t)$ is a combination of simpler functions, joined together using the maximum operator instead of the addition operator, see Figure 1.9.

Some might be sceptical at the ad hoc fashion in which we are proceeding, but this often proves necessary in the course of mathematical modelling with ODEs. Often, one can only make an educated or inspired guess about the coarse behaviour in advance, and then refine the model depending on how the proposed ODE behaves. This introduces issues with 'Researcher Degrees of Freedom'.

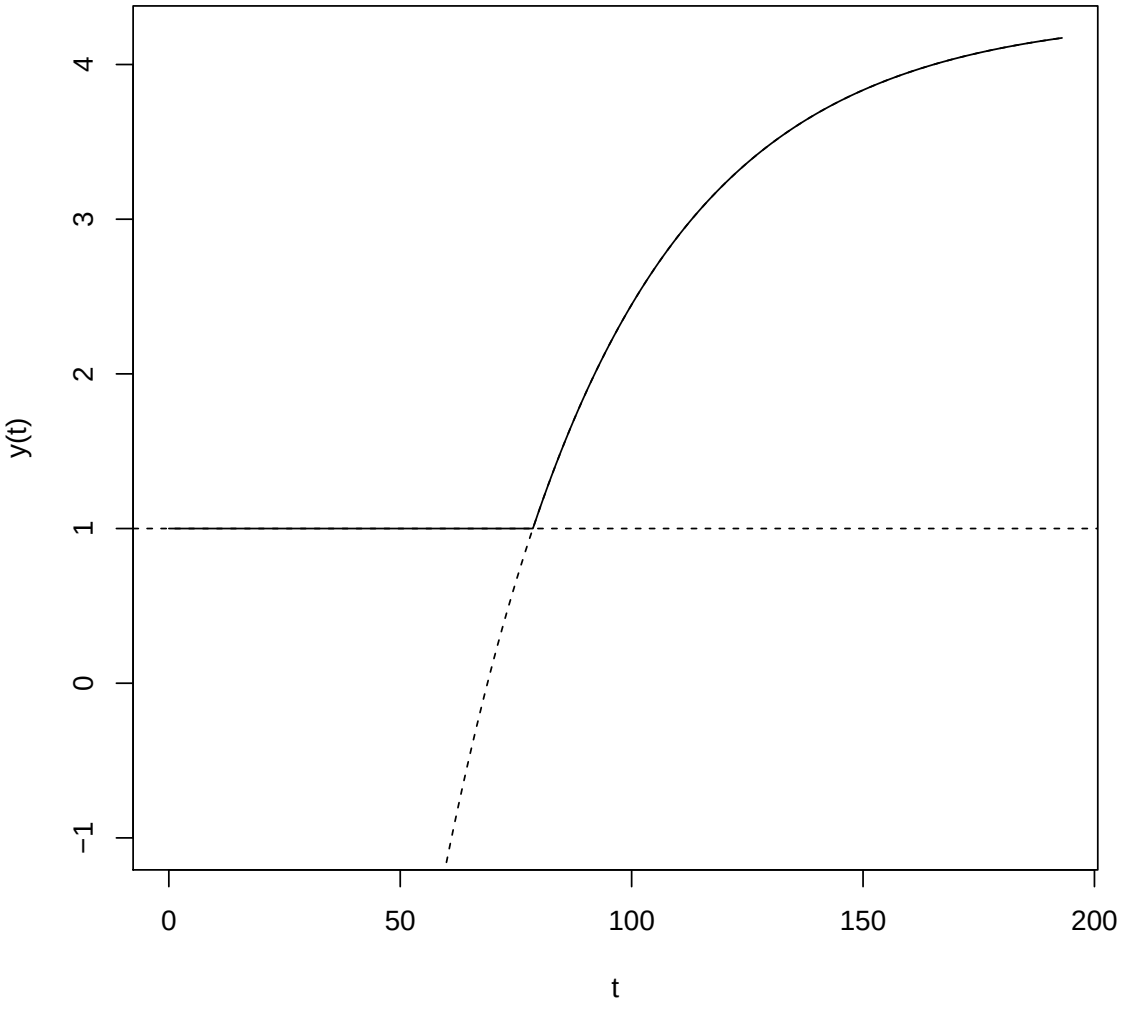

Figure 1.9: Plot of $\tilde{y}(t)$ and its constituent functions

Parameter Estimation

We assume that the breakpoint t_0 is known in advance. Then our model for $y(t)$

$$y(t) = \begin{cases} 0 & t \leq t_0 \\ \beta_0 + \beta_1 e^{\beta_2 t} & t \geq t_0 \end{cases}$$

Note that this function might not be well defined at t_0, we will adress the question of matching later on. We must estimate the three unkown coefficients $\beta_0, \beta_1, \beta_2$.

Estimating β_0 from the data The plot suggests that $\beta_2 < 0$, and $\beta_1 < 0$, under this assumption, we have that:

$$\lim_{t \to \infty} y(t) = \beta_0$$

Where the convergence happens monotonically from below

So an initial estimate for β_0 is given by $\hat{\beta}_0 = \max(y_i)$

Estimating β_1 and β_2 given β_0 The model can be rearranged so that

$$\log(\beta_0 - y(t)) = \log(\beta_1) - \beta_2 t$$

This equation is only valid so long as the left hand side is well defined however. For our intial estimate of β_0, it is only necessary to exclude the largest observed value of y.

Simulataneous Estimation Once we have reasonable estimates for β_0, β_1 and β_2, we can use non linear regression to estimate all three.

Matching

We have now produced seperate estimates for $y(t)$ at $t \leq t_0$ and $t \geq t_0$, these distinct functions do not necessarily agree at $t = t_0$.

To stich the two functions together, we let $\hat{y}(t) = \max(0, \beta_0 + \beta_1 e^{\beta_2 t})$. This is a continous function that entirely satisfies the orginal ODE, except for the precise location of the transition point.

This seems very "ad hoc", but is consistent with how ODE's are stiched togther in Applied Maths texts.

1.3.2 Partial Differential Equation - the Transport Equation

A linear PDE that would be analogous to the linear ODE above would be the Transport Equation:

$$\frac{\partial u(x,t)}{\partial t} + \beta \frac{\partial u(x,t)}{\partial x} = 0$$

The ODE $y'(t) + \beta y(t) = 0$ can be thought of as a simplifcation of the Transport Equation, where it is assumed that $u(x,t)$ only varies with time, and not with space.

A general solution to the Transport Equation is given by:

$$u(x,t) = f(x - \beta t)$$

The function $f(\cdot)$ is unspecified. The solution $u(x,t)$ is constant along the rays $x = \beta t + C$

The solution is in effect an animation of the shape $f(x)$ moving to the right at fixed speed β.

Statistically speaking, fitting the Transport Equation to observed data is a semi-parametric problem because one of the parameters to be estimated is a function. This is also a transformation model, since the plot of $u(x,t)$ with respect to x at a fixed time t is a transformed version of $f(x)$, the curve at $t = 0$.

If the parameter governing the transformation process - β - is known, $f(\cdot)$ is reasonably easy to estimate. If β is unknown, one might attempt to maximise a profiled objective function in β.

Suppose there are n observed values y_i at time t_i and location x_i.

The value observed at a point x at time t depends only on $x - \beta t$. The function $f(\cdot)$ could thus be estimated by non-parametrically regressing the observed values at y_i against $x_i - \beta t_i$

What if β were unknown? The above discussion suggests a hierarchial approach to estimation: for a given choice of β, to fit an associated function $f(\cdot|\beta)$ using an appropriate non-parametric estimation method, and compute the associated least squares error. Define the function that associates each β with its sum of squared error:

$$H(\beta) = \sum_{i=1}^{n} [y_i - f(x_i - \beta t_i|\beta)]^2$$

(In case the left hand side might be slightly unclear - for the ith observation, the associated function $f(\cdot|\beta)$ is evaluated at $x_i - \beta t_i$.)

This is a non-linear least squares problem in β. To estimate β, one would attempt to find the value of β that minimises $H(\beta)$.

The Parameter Cascade Algorithm entails a similar hierarchial approach.

1.3.3 Quasi-linear Differential Equations

Up until this point, it has been possible to use techniques from Applied Mathematics to construct solution strategies on a case by case basis. As differential equations become more complex, this approach begins to rapidly become non-viable. In this section, we will introduce quasi-linear variations of the previous two examples.

The difference between a quasi-linear and a linear differential equation is that the coefficients in a quasi-linear equation are allowed to depend on the unknown function. Instead of an ODE such as $y' = \beta(t)y$, one would have an ODE such as $y' = \beta(y,t)y$. As

we shall see, though quasi-linear problems tend to be reminiscent of linear ones, they are nonetheless substanially more complicated, and require more technical knowledge, and even ingenuity to tackle.

For a quasi-linear variation of a linear ODE, consider the Van Der Pol Equation:

$$y''(t) + \beta(1 - y(t))^2 y'(t) + y(t)$$

This ODE has no obvious solution.

Even if a solution exists, an estimation strategy might be difficult to derive. Consider the inviscid Burger's Equation:

$$\frac{\partial u(x,t)}{\partial t} + \beta u(x,t)\frac{\partial u(x,t)}{\partial x} = 0$$

This equation is identical to the Transport Equation except that the rate term is equal to $\beta u(x,t)$. The solution is given by:

$$u(x,t) = f(s)$$

Here $f(\cdot)$ is some arbitary function as before, and s is implicitly defined as the solution of the equation $x = \beta f(s)t + s$. Since $s = x - \beta ut$, this can be written as:

$$u(x,t) = f(x - \beta ut)$$

Fitting this model is substantially trickier than the Transport Equation. There is no clean seperation between the problem of estimating $f(\cdot)$ and β since $u(x,t)$ appears on the righthand side and scales β.

A further complication is that $u(x,t)$ might only define a *relation*, instead of a function. There might be multiple values of u associated with a given (x,t) that satisfy the solution equation. Physically speaking, multiple values correspond to shock waves.

Discussion We see that the level of knowledge required to devise fitting strategies can increase substantially even with seemingly modest increases in the complexity of the differential equation.

Consider the following quasi-linear model of genetic drift in a population proposed by R.A. Fisher (in the *Annals of Eugenics*) [8]:

$$\frac{\partial u(x,t)}{\partial t} + \beta_1\frac{\partial u(x,t)}{\partial x} = \beta_2 u(x,t)(1 - u(x,t))$$

This problem is similar to the previous two PDEs we discussed, it even admits travelling wave solutions of the form $f(x+Ct)$ as Fisher himself noted. Nonetheless, it is a much more difficult problem than the previous two despite the apparently modest increase in complexity. One would likely have to consult a textbook that covers nonlinear PDEs that can generate waves in fair degree of detail to be able to devise a fitting strategy. This is quite a specialised subject!

As the complexity increases, a practioner will find themselves spending more and more time doing Applied Mathematics, and less and less time doing Statistics.

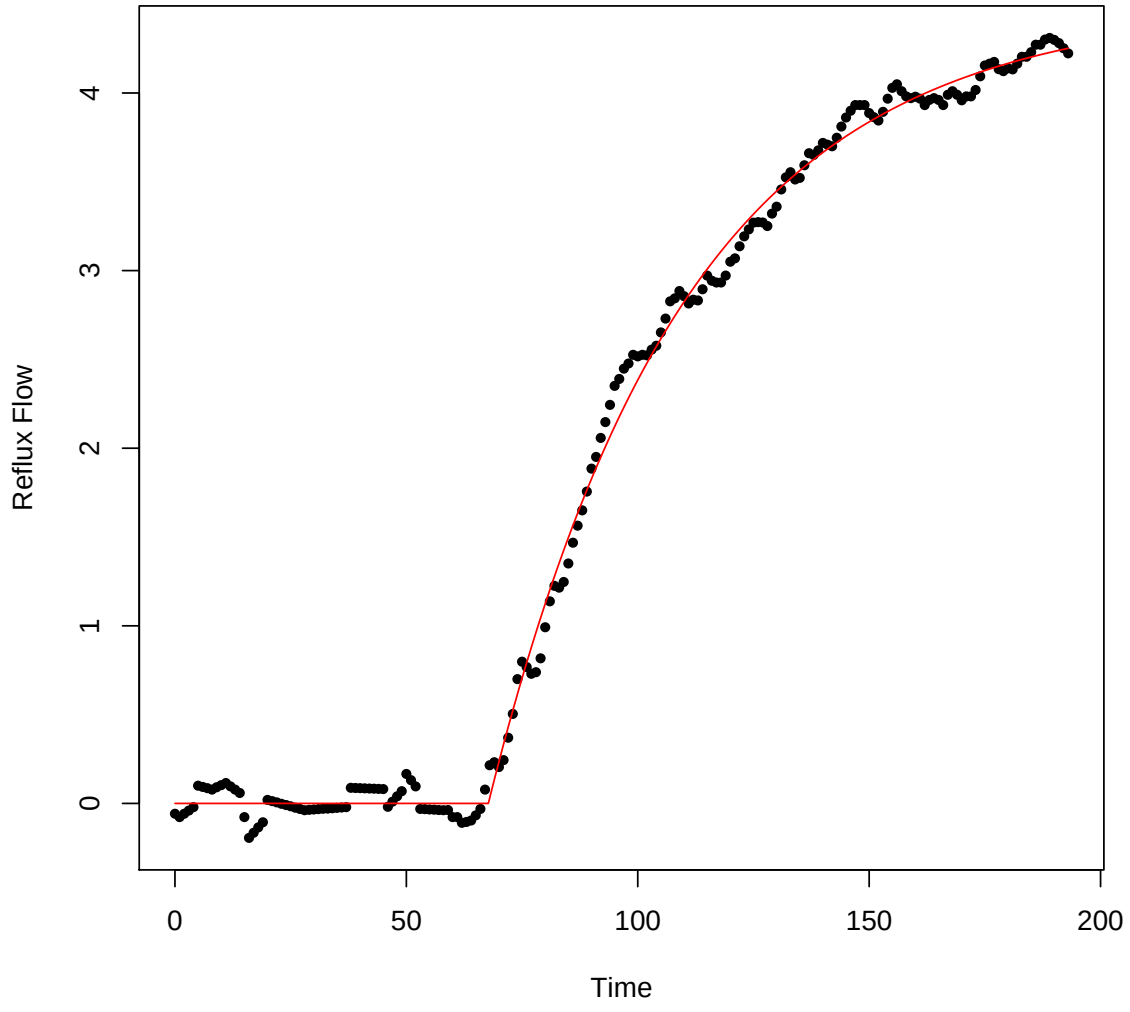

Figure 1.10: Fitted curve using NLS

Functional Model

Instead of the previous approach, which requires a considerable amount of domain-specific knowledge, the functional model is generally employed by Statisticians instead. It has the advantage of being more broadly applicable.

The functional model asserts that

$$y'(t) \approx -\beta y(t) + u(t)$$

Where $y(\cdot)$ and $u(\cdot)$ are functions to be estimated, and β is a single scalar parameter.

It is assumed that $u(t)$ is a step function of the form

$$u(t) = a\mathbb{I}_{[0,t_0)}(t) + b\mathbb{I}_{[t_0,\infty)}(t)$$

And that $y(t)$ can be expanded as a linear combination of B-Splines. The knots are duplicated at t_0 so that the first derivative is discontinous. This model was fitted using the Data2LD package.

1.4 Three Stage Parameter Cascade

Up to this point, the structural parameter λ has been treated as fixed. But it is possible to extend the Parameter Cascade to estimate λ.

It is necessary to an *Outer Criterion* $F(\lambda)$ that determines how good a given choice of λ is.

A common choice of outer criterion is the so-called Generalised Cross Validation.

Just as the problem of fitting a function $f(\cdot|\theta)$ can generate an optimisation subproblem, so that of fitting a third level in the cascade can generate a series of subproblems to find the best parameter choice associated with a given value of λ, which in turn generates a series of subproblems to find the fitted function as the parameter space is explored.

Many packages do not implement the three state parameter cascade. They instead expect practioners to find the best choice of λ by cycling through a set of predetermined values or even just employing manual adjustment.

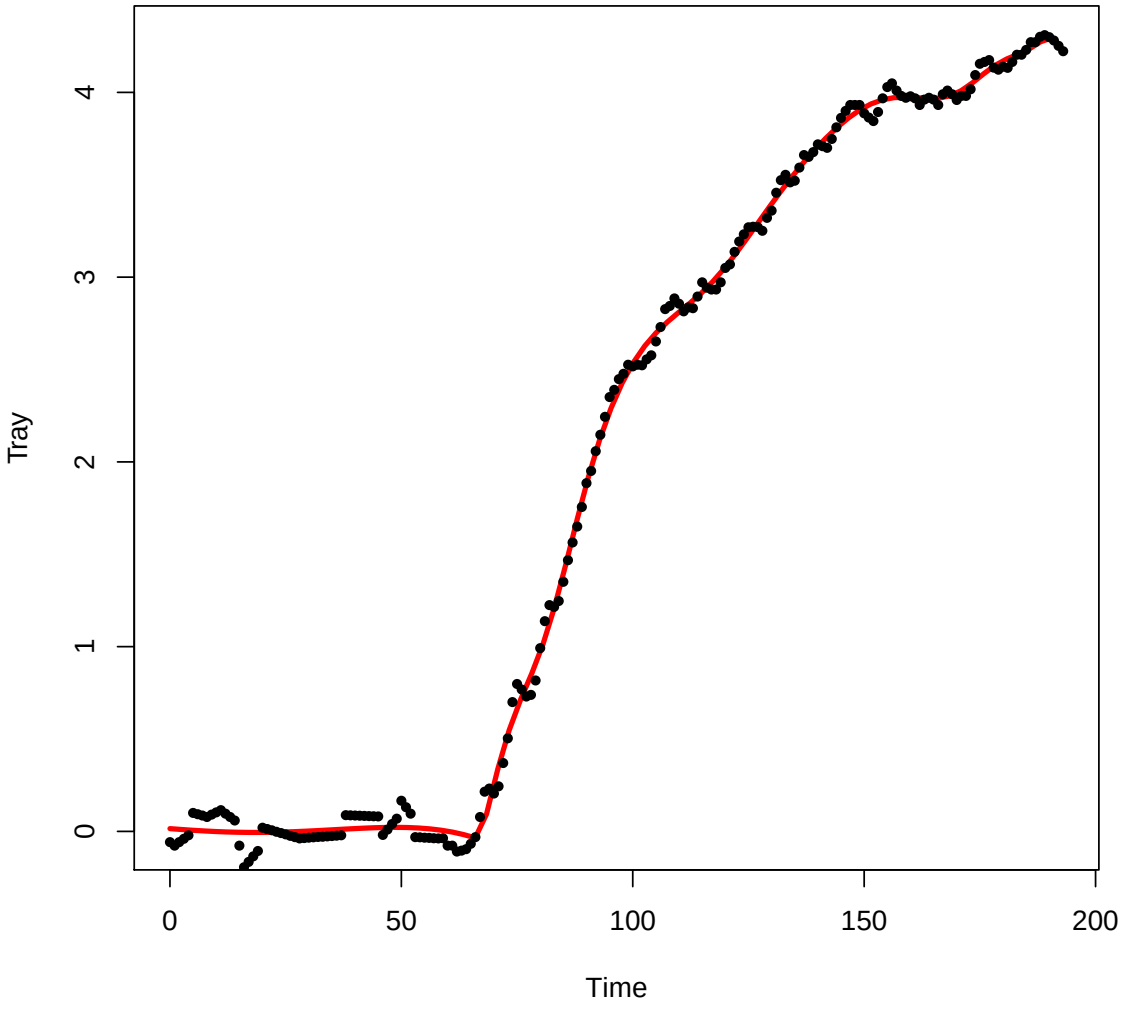

Figure 1.11: Fitted curve using FDA

Bibliography

[1] John P Boyd. *Chebyshev and Fourier spectral methods*. Courier Corporation, 2001.

[2] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[3] Richard P Brent. *Algorithms for minimization without derivatives*. Courier Corporation, 2013.

[4] Jiguo Cao and James O Ramsay. Parameter cascades and profiling in functional data analysis. *Computational Statistics*, 22(3):335–351, 2007.

[5] Kwun Chuen Gary Chan. Acceleration of expectation-maximization algorithm for length-biased right-censored data. *Lifetime data analysis*, 23(1):102–112, 2017.

[6] Edwin KP Chong and Stanislaw H Zak. *An introduction to optimization*, volume 76. John Wiley & Sons, 2013.

[7] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[8] Ronald Aylmer Fisher. The wave of advance of advantageous genes. *Annals of eugenics*, 7(4):355–369, 1937.

[9] Bengt Fornberg. Generation of finite difference formulas on arbitrarily spaced grids. *Mathematics of computation*, 51(184):699–706, 1988.

[10] PR Graves-Morris, DE Roberts, and A Salam. The epsilon algorithm and related topics. *Journal of Computational and Applied Mathematics*, 122(1-2):51–80, 2000.

[11] David R Hunter and Kenneth Lange. A tutorial on MM algorithms. *The American Statistician*, 58(1):30–37, 2004.

[12] Eugene Isaacson and Herbert Bishop Keller. *Analysis of numerical methods*. Courier Corporation, 2012.

[13] Carl T Kelley. *Implicit filtering*, volume 23. SIAM, 2011.

[14] C.T. Kelley. A brief introduction to implicit filtering. `https://projects.ncsu.edu/crsc/reports/ftp/pdf/crsc-tr02-28.pdf`, 2002. [Online; accessed 12-October-2019].

[15] Jack Kiefer. Sequential minimax search for a maximum. *Proceedings of the American mathematical society*, 4(3):502–506, 1953.

[16] Kenneth Lange. *Optimization*. Springer, 2004.

[17] Kenneth Lange. *Numerical analysis for statisticians*. Springer Science & Business Media, 2010.

[18] Kenneth Lange. The MM algorithm. `https://www.stat.berkeley.edu/~aldous/Colloq/lange-talk.pdf`, April 2007. [Online, accessed 18-September-2019].

[19] Randall J LeVeque. *Finite difference methods for ordinary and partial differential equations: steady-state and time-dependent problems*, volume 98. Siam, 2007.

[20] Steve McConnell. *Code complete*. Pearson Education, 2004.

[21] Geoffrey McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.

[22] J Nocedal and SJ Wright. *Numerical Optimisation*. Springer verlag, 1999.

[23] Naoki Osada. *Acceleration methods for slowly convergent sequences and their applications*. PhD thesis, PhD thesis, Nagoya University, 1993.

[24] Yudi Pawitan. *In all likelihood: statistical modelling and inference using likelihood*. Oxford University Press, 2001.

[25] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. ISBN 3-900051-07-0.

[26] James Ramsay. Functional data analysis. *Encyclopedia of Statistics in Behavioral Science*, 2005.

[27] Walter Rudin et al. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1964.

[28] Christopher G Small. A survey of multidimensional medians. *International Statistical Review/Revue Internationale de Statistique*, pages 263–277, 1990.

[29] Karline Soetaert, Thomas Petzoldt, and R. Woodrow Setzer. Solving differential equations in R: Package deSolve. *Journal of Statistical Software*, 33(9):1–25, 2010.

[30] Gerald Teschl. *Ordinary differential equations and dynamical systems*, volume 140. American Mathematical Soc., 2012.

[31] Keller Vandebogart. Method of quadratic interpolation. `http://people.math.sc.edu/kellerlv/Quadratic_Interpolation.pdf`, September 2017. [Online; accessed 13-September-2019].

[32] Jet Wimp. *Sequence transformations and their applications*. Elsevier, 1981.

[33] Stephen Wright. Optimization for data analysis. In Michael W. Mahoney, John C. Duchi, and John C. Duchi, editors, *The Mathematics of Data*, chapter 2, pages 49–98. American Mathematical Society and IAS/Park City Mathematics Institute and Society for Industrial and Applied Mathematics, 2018.

[34] Tong Tong Wu, Kenneth Lange, et al. The MM alternative to EM. *Statistical Science*, 25(4):492–505, 2010.