Representing Penalties on Basis Expansions as a Polynomial In the Parameters (Working Title)

March 4, 2013

1 Basis Function Expansion Methods

Fitting sophisticated mathematical functions ot empirical data continues to be a challenge in statistical science. One approach, known as the basis function expansion method , offers a considerable degree of flexibility and mathematically tractables solutions. This approach involves repseting a function $y(t)$ as finite number of basis functions $\{\phi_k(t)\}$ through the conditional expectation

$$\mathbb{E}[y|t] = c_0\phi_0(t) + c_1\phi_1(t) + \cdots + c_N\phi_N(t),$$

where $\{c_k\}$ is a set of appropriately chosen coefficients. $\mathbb{E}[y|t]$ is the conditional expectation of our independent variable y, given our dependent data x. The conditional expectation $\mathbb{E}[y|t]$ is the estimate of y dependending exclusively on t that minimises the squared loss $\mathbb{E}[y(t) - E[y|t]]^2$. Unfortunately, this is a theoretical ideal, so me must make do with an estimate $\widehat{\mathbb{E}[y|t]}$ instead.

Example 1. Simple Linear Regression

Basis Function Methods might seem a little abstract, but they are ubiquitous. For example, Simple Linear Regression is an example of a basis function method. Take two basis functions $\{1, t\}$, so that $\hat{y}(t)$, our estimate of y given the value t, can be written as a linear combination of the two:

$$\widehat{\mathbb{E}[y|t]} = \hat{y}(t) = c_0 + c_1 t$$

This is the form of a simple linear regression model. If our basis consists of just the single constant function $\phi(t) = 1$ then we get a model of the form

$$\hat{y} = c_0.$$

In this case we would generally use the mean of the y values as our estimate of c_0. If we go in the other direction and add a quadratic function t^2 we get a quadratic regression model

$$\widehat{y(t)} = c_0 + c_1 t + c_2 t^2.$$

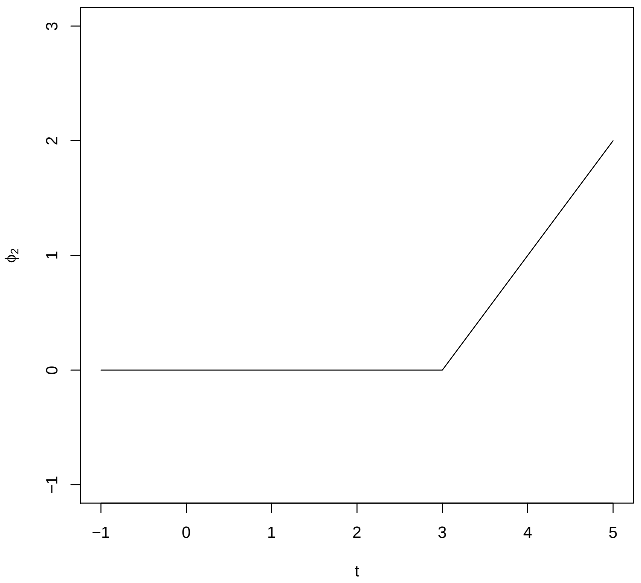

Figure 1: The ramp function ϕ_2.

Example 2. Ramp Functions

There are other choices of basis besides monomials we could use. We could have a basis of two functions $\{\phi_1, \phi_2\}$ Where the two functions are defined as follows

$$\phi_1(t) = 1$$
$$\phi_2(t) = \begin{cases} 0 & \text{if } t < \tau \\ t - \tau & \text{if } t \geq \tau \end{cases}.$$

The first function is a constant function, the second is a ramp function based at τ.

This basis would work very well for representing data which starts to increase linearly past a certain threshold. Ramp functions and constant functions are very cheap computationally, meaning we can spawn compartively a very large number of them. This could be used to cover situations where there data is unchanging over relatively long periods.

The above basis has two disadvantages though. Firstly the second basis function is only continuous, not differentiable. It would not be wise to use this basis if we wanted to estimate $\mathrm{d}y/\mathrm{d}t$. Secondly it is very arbitrary. It is not obvious why it would be useful compared to a quadratic model for example; the latter also allows to us estimate derivatives of all orders and can still acommodate data which varies in its rate of change.

This underscores an important point. There are many types of basis besides the ones usually encountered, such as polynomial bases. Many of them are fungible from the pure mathematical point of view in terms of how well they can fit a function. The statistician should make a sensible choice of basis. If our choice of basis is good, then it will be able to fit the data with only a few terms, and we might be able to avoid estimating many coefficients. This is one of the reasons simple linear regression and quadratic regression are useful; they can capture much variation in the observed data in spite of being very simple.

Example 3. Fourier Basis Functions

If our data has a periodic component to it, such as the observed temperature over the course of the year, or a time series of annual sales, then it would be wise to use a basis consisting of periodic functions. This suggests that we should use a Fourier basis consisting of the set of functions $\{\cos(n\omega t), \sin(n\omega t)\}$ where $0 \leq n \leq N$ for some N and ω is the frequency. The frequency depends on our time scale, they are related by the formula $\omega = \frac{2\pi}{T}$, where T is the period.

Furthermore Fourier basis functions have several other desirable properties. They are smooth, meaning that they can be used to estimate any derivivative of the data, at leastin theory. They are orthogonal, which can make certain problems more convienient and they are closed under differentiation, meaning that the derivative of a combination of Fourier basis functions, is itself a combination of Fourier basis functions. The latter two properties will prove very useful later, as we shall see.

A Fourier basis can represent a massive class of functions; any square integrable function on some interval can be represented by them. They cannot detect how frequencies change in space however, only their global behaviour. They can also be somewhat mundane - almost certain to work, but unlikely to provide anything too interesting.

Example 4. B-spline Basis Functions

Roughly speaking, B-splines are compactly supported polynomial functions, or more practically they are nonzero only inside of a given interval. More formally a B-Spline basis consists of a degree n, which determines the degree of the basis functions, and a set of knots, that is a set of K time points $\{t_0, \dots, t_K\}$.

Since they are compact they generally can only indvidually capture local information about the data. One of the main advantages of B Splines is that they can represent any other spline of the same degree and smoothness with the same knots. This makes them useful for Statistics since they assume less about the form of the data, they can help us avoid bias.

2 Least Squares Fitting

Least Squares is one of the most well known statistical estimation techniques. If we have a set of n observations y_i, measured at times t_i, then the least squares criterion, chooses the estimated function $\hat{y}$ from some set of functions S that minimises the sum of the squares of the error

$$\widehat{y(t)} = \underset{f \in S}{\operatorname{argmin}} \sum_{i=1}^{n} [y_i - f(t_i)]^2.$$

We assume S is the span of some set of m basis functions i.e $S = \{\sum_{i=1}^{m} c_i \phi_i(t) | c_i \in \mathbb{R}\}$. This suggests that to find $\hat{y}$ we only need to estimate to coefficients $\hat{c}_i$. Then we have completely deterimined $\hat{y}$. We can then write the least squares criterion as:

$$(\hat{c}_1, \ldots, \hat{c}_m) = \underset{c_1 \ldots c_K}{\operatorname{argmin}} \sum_{i=1}^{n} [y_i - \sum_{j=1}^{m} c_j \phi_j(t_i)]^2$$

The above expression is cumbersome, we can use vector notation to simplify it. Firstly, we have $f(t) = \mathbf{c}'\phi(\mathbf{t})$, where $\mathbf{c} = (c_1, ..., c_m)$ and $\phi = (\phi_1(t), \ldots, \phi_2(t))'$. If we construct a matrix $\mathbf{\Phi}$, where the ith row of $\mathbf{\Phi}$ is $\phi(t_i)$, and let $\mathbf{y} = (y_1, \ldots, y_n)$ we can write the least squares problem as

$$\hat{\mathbf{c}} = \underset{\mathbf{c} \in \mathbb{R}^m}{\operatorname{argmin}} (\mathbf{y} - \mathbf{\Phi c})'(\mathbf{y} - \mathbf{\Phi c}).$$

The expression on the right hand side is an example of a *Quadratic Form*, they are the generalisation of quadratic functions to finite dimensional vector spaces. We can have quadratic forms on infinite dimensional vectors spaces too, but that is not relevant here.

3 Roughness Penalties

It is well known that the Least Squares gives us the *Best Linear Unbiased Estimator* for $\mathbf{y}$ the function we assume to be generating the data. Nonetheless it is often useful to employ a form of *regularisation* which constrains how much $\hat{y}$ is allowed to vary. Intuitively, this reduces the *variance* of $\hat{y}$ and helps guard against overfitting. In the context of basis function expansions, the most commonly used penalty is the curvature penalty

$$PEN(\hat{y}) = \int_D [\widehat{y(t)''}]^2 \mathrm{d}t.$$

Here D is our domain of interest and f'' stands for the second derivative of the function f. This criterion for fitting a curve is very appealing in many regards. It can be shown using the theory of Finite Elements that if we constrain $\hat{y}$ to pass through two points, then the $\hat{y}$ minimises the penalty is the best approximation to the straight line going through the points that we can form with this basis.

This penalty can also be represented as a polynomial. Let $\langle f, g \rangle = \int_D f(t)g(t)\mathrm{d}t$. We can see that $\langle \cdot, \cdot \rangle$ defines an inner product on $L^2(D)$, the set of square integrable functions on D. The penalty can be written in the form $PEN(f) = \langle f'', f'' \rangle$. Substituting in the expansion for f we get

$$\langle f'', f'' \rangle = \langle \sum_{i=1}^{m} c_i \phi_i'', \sum_{i=1}^{m} c_i \phi_i'' \rangle \quad = \sum_{i=1}^{m} \sum_{j=1}^{m} \langle c_i \phi_i'', c_j \phi_j'' \rangle = \sum_{i=1}^{m} \sum_{j=1}^{m} c_i c_j \langle \phi_i'', \phi_j'' \rangle.$$

We can see that the terms $\langle \phi_i'', \phi_j'' \rangle$ depend only on our choice of basis and so are "fixed" for our purposes. This implies that the penalty can be represented as a polynomial in the c_i. We can do better however and represent the penalty as a quadratic form. If we define an $m \times m$ matrix $\mathbf{K}$ by $\mathbf{K}_{ij} = \langle \phi_i'', \phi_j'' \rangle$ it can be seen that $\langle f'', f'' \rangle = \mathbf{c}'\mathbf{K}\mathbf{c}$.

Note that this relies on the fact that the second derivative operator maps a linear combination of ϕ_i into a linear combination ϕ_i''. We can always represent an inner product on some finite vector space as $\mathbf{b}'\mathbf{A}\mathbf{b}$, where the terms depend on the problem at hand. We can simply construct the matrix $\mathbf{A}$ for the ϕ_i'' and have the $\mathbf{c}$ in place of the $\mathbf{b}$ because differentiation is linear.

4 Penalised Least Squares

We have two distinct, but compelling objectives. We want a fit $\hat{y}$ that maintains fidelity to the data, as represented by the goodness of fit, but simultaneously adheres to the requirement of a smooth description of the data. The solution is to use a combination of the two penalties, the penalised sum of squared errors

$$PENSSE(f) = \sum_{i=0}^{n} (y_i - f(t_i)^2 + \lambda \int_D [f(t)'']^2 \mathrm{d}t.$$

If we assume again that f is the sum of basis functions, depending on a vector of coefficients $\hat{\mathbf{c}}$; we can use the previous results to show that the Penalised Sum of Squares can be writen as the sum of two quadratic forms:

$$PENSSE(c) = (\mathbf{y} - \mathbf{\Phi c})'(\mathbf{y} - \mathbf{\Phi c}) + \lambda \mathbf{c}'\mathbf{K}\mathbf{c}.$$

Where $\mathbf{y}$, $\mathbf{K}$ and $\mathbf{\Phi}$ have the same definitions as they did before.

5 Multivariate Splines

What if our data is spatial in nature? In this case our data will often be at a series of points $\mathbf{x}_i = (x_i, y_i); i = 1, \ldots, n$. It is actually not too difficult to extend our results; it is almost as easy as replacing the t_i with $\mathbf{x_i}$.

As before we expand our function y as a basis function expansion

$$y(\mathbf{x}) = \sum_{i=1}^{m} c_i \phi(\mathbf{x}).$$

Notice that even though we are working in more than one dimension, our sum is still "one dimensional" in that it has only one index. This is deliberate as it makes our life much easier.

Finding a least squares estimate is identical, so we will not cover it here.

Example 5. Fitting the Displacement of a Membrance Using a Laplacian Penalty

On model of the mechanics of a membrane is the emphwave equation. Let $u(x, y, t)$ be the displacement of the membrane at position (x, y) at time t. Then u is assumed to satisfy the partial differential equation:

$$u_{tt} = c\Delta u.$$

Here $u_{tt} = \frac{\partial^2 u}{\partial t^2}$, c is a parameter known as the wave speed and Δ is a differential operator known as the Laplacian. In two dimensions it is defined as

$$\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}.$$

If the membrane doesn't move with time so $u_{tt} = 0$ then the wave equation reduces to Laplace's equation

$$\Delta u = 0.$$

This is what is known as a steady state model, notice the coefficient c has disappeared.

If we have measurements of the displacements at various points, this suggests that we use a penalised model of the form

$$PENSSE(f) = \sum_{i=0}^{n}(y_i - f(\mathbf{x}_i))^2 + \lambda \int_{\Omega} |\Delta f(\mathbf{x})|^2 \mathbf{dx}.$$

As before we can express this penalty as a quadratic form. Define an inner product on our functions by

$$\langle f, g \rangle = \int_{\Omega} f(\mathbf{x}) g(\mathbf{x}) \mathbf{dx}.$$

Then the laplacian penalty can be written as

$$\int_{\Omega} |\Delta f(\mathbf{x})|^2 \mathbf{dx} = \langle \Delta f, \Delta f \rangle.$$

If $f = \sum c_i \phi_i$ then $\Delta f = \sum c_i \Delta \phi_i$. In the same manner as the previous roughness penalty, we can see that if f is a basis expansion then we can write the penalty as a polynomial in the coefficients

$$\langle \Delta f, \Delta f \rangle = \sum_{i=1}^{m} \sum_{j=1}^{m} c_i c_j \langle \Delta \phi_i, \Delta \phi_j \rangle.$$

This can of course be written as $\mathbf{c}'\mathbf{Kc}$ where $\mathbf{K}_{ij} = \langle \Delta \phi_i, \Delta \phi_j \rangle$.

We define the matrix $\mathbf{\Phi}$ as before with $\mathbf{\Phi}_{ij} = \phi_j(\mathbf{x}_i)$, $\mathbf{c} = (c_1, \ldots, c_m)$. We can then write the penalty in terms of the coefficients

$$PENSSE(c) = (\mathbf{y} - \mathbf{\Phi c})'(\mathbf{y} - \mathbf{\Phi c}) + \lambda \mathbf{c}'\mathbf{Kc}.$$

It is worth noting that this is identical to the expression we derived above. This is because both $\frac{d^2}{dt^2}$ and Δ are linear operators.

6 Non Linear Penalties

Suppose we believed that our data could be approximately modelled by and ODE of the form $f'' = f^2$ and we wished to incorporate this data into our model. It would be reasonable to include a penalty of the form:

$$\int_D (f''(t) - f(t)^2)^2 \mathrm{dt}$$

We can also represent this penalty as a quadratic form. As usual $f = \sum c_i \phi_i$. Expanding out the two terms in the penalty we get

$$f''(t) = \sum_{i=0}^{m} c_i \phi_i(t)''$$

$$f(t)^2 = \sum_{i=0}^{m} \sum_{j=0}^{m} c_i c_j \phi_i(t) \phi_j(t).$$

Hence,

$$f''(t) - f(t)^2 = \sum_{i=0}^{m} c_i \phi(t)'' - \sum_{i=0}^{m} \sum_{j=0}^{m} c_i c_j \phi_i(t) \phi_j(t).$$

Notice the second term is a two dimensional, finite sum. We now need to find norm. As before we can write this as an inner prouct: $\langle f'' - f^2, f'' - f^2 \rangle$. The above expression is in the form $a_1 \phi_1'' + \cdots + a_m \phi_m'' + b_{11} \phi_1 \phi_1 + \ldots b_{mm} \phi_m \phi_m$. This is a linear combination of a finite set of functions. So we can represent it as a quadratic form. The form above is a little awkward to work with. We would like to have it vary with one index only, or combine the ϕ_i''' and the $\phi_i \phi_j$ together. We would define a function $\pi(n)$ that returns the appropriate function either some ϕ_i or $\phi_i \phi_j$ depending on the value. Defining such a function is tricky however. One function is as follows

$$\pi(k) = \begin{cases} \phi_k'' & \text{if } k \leq m \\ \phi_{(k-1)|m} \phi_{(k-1) \bmod m} & \text{if } k > m \end{cases}$$

Here $a|b$ is integer division, i.e $a|b = \text{floor}(a/b)$.

If we define $\psi_k(t)$ to be $\pi_k(t)$ and define a similar function $\sigma(i)$ for the coeffients we can define $f'' - f^2$ we get:

$$f'' - f^2 = \sum \sigma(i) \psi_i$$

We now can express the penalty as a quadratic form:

$$\int_D (f(t)'' - f(t)^2)^2 dt = \sigma(\mathbf{c})' \mathbf{K} \sigma(\mathbf{c})$$

Here $\mathbf{K}_{ij} = \langle \psi_i, \psi_j \rangle$ and $\sigma(\mathbf{c}) = (\sigma(1), \ldots, \sigma(m+m^2)) = (c_1, \ldots, c_m, c_1 c_1, \ldots, c_m c_m)$. The inner product is $\langle f, g \rangle = \int_D f(t) g(t) dt$. Since $\sigma(\mathbf{c})$ is a second order polynomial in the c_i, the penalty is actually as fourth order polynomial in the c_i.

We can write the Penalised Sum of Squared Errors in terms of the c_i:

$$PENSSE(\mathbf{c}) = (\mathbf{y} - \mathbf{\Phi c})'(\mathbf{y} - \mathbf{\Phi c}) + \lambda\sigma(\mathbf{c})'\mathbf{K}\sigma(\mathbf{c})$$

Remark. Alternative Representation Of The Penalty

It is difficult to deal with the indexing above. An alternative is to break the function into two parts $f = f' + f^2$ where $f' = \sum c_k \phi'_k$ and $f^2 = \sum c_k c_l \phi_k \phi_l$. Since $\langle f', f^2 \rangle = \langle f^2, f'$. We can represent the penalty in the form of the sum of three parts

$$\langle f' + f^2, f' + f^2 \rangle = \mathbf{c}'\mathbf{Kc} + 2\mathbf{c}'\mathbf{Lb} + \mathbf{b}'\mathbf{Mb}.$$

Here $\mathbf{K}_{ij} = \langle \phi'_i, \phi'_j \rangle$. $\mathbf{L}$ and $\mathbf{M}$ similarly represent $\langle f', f^2 \rangle$ and $\langle f^2, f^2 \rangle$.