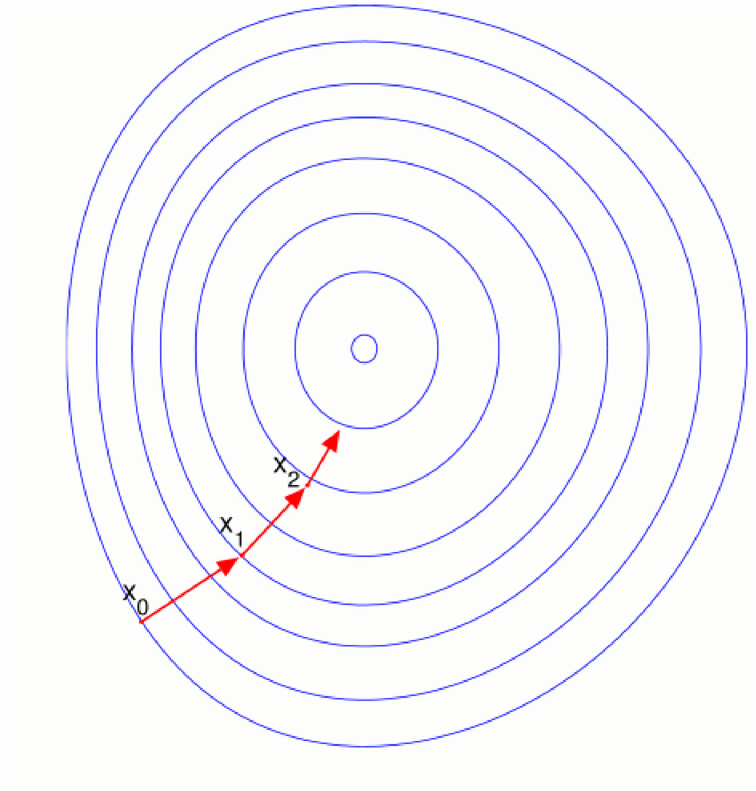
Curve Fitting with the Least Square Method

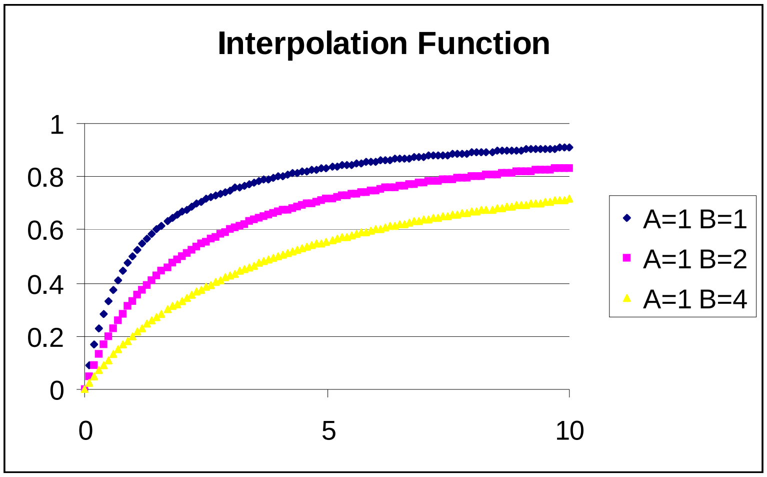


Matthieu Bultelle
Department of Bio-Engineering
Imperial College, London

Context

We wish to model the positive feedback of the production of AHL. For this purpose we need to interpolate a set of N experimental measurements $(X_i, Y_i)_{i=1...N}$ with the function

$$x \to f(x) = \frac{Ax}{B+x}$$



There are several ways to do the interpolation, some are more robust than others. We chose to use the least square methods that is to minimize the expression

$$\varepsilon(A,B)) = \sum_i \left(Y_i - \frac{AX_i}{B+X_i} \right)^2$$

The minimum of the function is obtained for the point (A,B) such as $\nabla \varepsilon = 0$

We have therefore the following necessary conditions

$$(1) \qquad \partial_A \varepsilon = 2\sum_i \left(\frac{AX_i}{X_i+B} - Y_i \right) \left(\frac{X_i}{X_i+B} \right) = 0$$

$$(2) \qquad \partial_B \varepsilon = 2\sum_i \left(\frac{AX_i}{X_i+B} - Y_i \right) \left(\frac{-AX_i}{(X_i+B)^2} \right) = 0$$

Equation (1) can be simplified to yield the condition

$$(1) \qquad A = \frac{\sum_i \frac{X_i Y_i}{X_i + B}}{\sum_i \frac{{X_i}^2}{(X_i + B)^2}} = F(B)$$

Likewise (2) can be modified into

$$(2) \qquad A = \frac{\sum_i \frac{X_i Y_i}{(X_i + B)^2}}{\sum_i \frac{{X_i}^2}{(X_i + B)^3}} = G(B)$$

The intersections of the curves $(A = F(B))$ and $(A = G(B))$ are potential extrema of the $\varepsilon(A, B)) = \sum_i \left(Y_i - \frac{AX_i}{B + X_i} \right)^2$. It is easy to prove that they actually are local minima. If the data are kind to us there is only one intersection. In the general case we have more than one intersections. To determine which local minimum is the absolute minimum (the point we are after), we need to compute $\varepsilon(A, B))$ for all the candidates – the overall minimum is of course the point that returns the lowest value.

How many Local Minima are there?

There is no way to know how many local minima there are but it is easy to know how many there are in the worst case scenario

It is easy to prove that the equation $F(B) = G(B)$ can be turned into a polynomial equation of degree 5N-5. So there cannot be more the 5N-5 intersection points – which can still be many.

Do we know where they are located?

Up to a point. We are only interested in the positive values of B, so we have 0 as a lower bound for B. Unfortunately we do not have a simple upper bound for B.

An easy pragmatic solution is available to us however. Just plot $B \rightarrow G(B) - F(B)$!
It will be easy to identify a value of B (call it B_{lim}) which is sure to be an upper bound (the estimation does not have to be that precise!!!).

A little physical sense also helps: if you have done your experiments properly you have acquired some data in the saturation phase. If this is the case you can be sure that X_{max} is larger than B, and X_{max} can therefore be sued as an upper bound for B in the search for the overall minimum for $\varepsilon(A, B))$.

How do I find the solutions of G(B) = F(B)

We now have a lower and an upper bound for B.
For complex equations like the one we are interested in I would recommend the following strategy (which is not brute force but is still computationally intensive). Implementation will require a few programming skills (I recommend Matlab or C as language).

1) Cut the segment $[0, B_{\lim}]$ into a large number (p+1) of equally-spaced points $\beta_j = jB_{\lim} / p$. Ideally p is large (1000 or even better).

2) Compute $G(B) - F(B)$ for all the points $\beta_j = jB_{\lim} / p$

3) For j=0 to j= p, Compare the sign of $G(\beta_j) - F(\beta_j)$ with $G(\beta_{j+1}) - F(\beta_{j+1})$

If there has been a change of sign between then there is a zero of the function between β_j and β_{j+1}. Find this zero of the function by dichotomy.
Providing the initial sampling of $[0, B_{\lim}]$ has been fine enough (p large enough) we have found all the solutions of the function.

Reminder : Finding zeros of a function by dichotomy

Dichotomy = ‘cutting in two’

Let us assume we have a function f continuous on $[a,b]$ such as $f(a) < 0$ and $f(b) > 0$.
Note if we have $f(a) > 0$ and $f(b) < 0$ instead it does not matter just switch –f for f!!

It can be proven that f has a zero between a and b (f being continuous). Please note that there may be more than one zero between a and b. The method detailed below is going to yield one of them only, not all of them. To get all the zeros between a and b you need to resample $[a,b]$ more finely.

Let us call ε the precision of the estimation of this zero.
We want to return a value x such as $|x - x_0| < \varepsilon$ where $f(x_0) = 0$.
For this purpose we build two series $(U_j)_{j\geq 0}$ and $(V_j)_{j\geq 0}$ with the following rules

Initialization: $U_0 = a, V_0 = b$

Computing Rank j+1: Compute $F\left(\frac{U_j + V_j}{2}\right)$.

If it is positive then $\begin{cases} U_{j+1} = (U_j + V_j)/2 \\ V_{j+1} = V_j \end{cases}$ else $\begin{cases} V_{j+1} = (U_j + V_j)/2 \\ U_{j+1} = U_j \end{cases}$

Repeat the operation while $|U_j - V_j| \geq \varepsilon$

A less complicated way to find the overall Minimum

Excel offers a way to implement the least square method with any interpolation function.
All it needs is

- the expression of the interpolation function
- the data $(X_i, Y_i)_{i=1...N}$
- a starting point for the search (A_0,B_0)

However, the optimization algorithm is not as robust as we could hope and therefore nothing ensures that the software will not return a local minimum instead of the overall minimum.
We can use the previous results to get a more robust interpolation. The idea is to use a (potentially) large number of starting points and let Excel do the rest.

We assume that the upper bound B_{lim} has been found.

1) Cut the segment $[0, B_{\lim}]$ into a large number (p+1) of equally-spaced points $\beta_j = jB_{\lim} / p$. Ideally for every value $\beta_j = jB_{\lim} / p$ we would associate a value of α_j of A such as (α_j, β_j) is a good starting point. If your experiments were done properly then you did some measures in the saturated phase of the curve. you can therefore use α_j = Y_{max} = Max (Y_I) all the time.

2) For every value of β_j , run the Excel Simulation with (Y_{max}, β_j) as starting point.
We call the result (A_j, B_j).

3) Compute the error function $\varepsilon(A, B)) = \sum_i \left(Y_i - \frac{AX_i}{B + X_i} \right)^2$ for (A_j, B_j)

The point (A_j, B_j) that achieves the lowest value of $\varepsilon(A, B))$ will be a good approximation of the minimum of $\varepsilon(A, B)) = \sum_i \left(Y_i - \frac{AX_i}{B + X_i} \right)^2$ if you have used enough points (p is large enough).