Curve fitting using monotone smoothing splines

Malte Larsson

It is a common problem to fit a curve to data. When it is known that the function should be increasing, this could be done using monotone smoothing splines. This thesis investigates an algorithm for finding this curve. The algorithm seems to work well for most parts, although there are cases when it is very slow.

The problem of fitting a curve to a data set is one that arises in many different fields, for example statistics and biology. One way to fit a curve is by using smoothing splines. Smoothing splines are curves that both should fit a data set consisting of points fairly well, but also be sufficiently smooth. The smoothing helps to remove the effects of noise in the data, and give a reasonable shape to the fitted curve. We consider the problem of finding monotone smoothing splines, which is useful when we know that the curve should be increasing. An example of this can be to find the average length of children given their age, as this is something that can be expected to be increasing.

Mathematically we want to find an increasing function $y(t)$ such that given a data set with points (t_i, α_i), the value

$$\int_0^T (y''(t))^2 dt + \lambda \sum_{i=1}^{n} (y(t_i) - \alpha_i)^2$$

is as small as possible. The parameter λ decides the balance between how important it is that the curve is close to the data points and how smooth it should be. It depends on the data set what value this should have.

One can show that on every interval between two data points the optimal curve can be one of two types of functions. This can be used to determine the curve exactly, given the values of some variables. Then it is possible to reformulate the problem as an optimization problem over these variables. Now we don't need to find an entire function anymore, but only the values of some variables. This can be used to construct an algorithm that makes the problem solvable for computers.

The algorithm is tested on some created data sets of different shapes and sizes. When the parameter λ is too large, the algorithm can be very slow and for data sets with more than 15 points it can run for hours or more. However when λ is close to the best choice or smaller, the algorithm is generally very fast and it can fit monotone smoothing splines to large data sets with hundreds of points. This is good since we are mostly interested in solutions when λ is chosen well.