Smooth additive models for very large datasets

S. Wood

Abtract:

Motivated by trying to model 4 decades worth of daily air pollution data from 1000+ monitoring stations across the UK, I will discuss the production of scalable statistical computation methods for smooth additive models with up to 10000 coefficients, for 10s of millions of data. The methods rest on three simple ideas: a fitting iteration that avoids computationally awkward log determinant terms that feature in the smoothing fitting objective, the use of scalable parallel block Cholesky methods as the main matrix decomposition required, and a marginal covariate discretization approach that saves both storage and FLOPS. I will discuss what went wrong computationally to motivate what went right, and present the results of the air pollution modelling exercise.