Modelling with smooth functions: GAMs, Covid and beyond.

Generalized additive models are generalized linear models in which the linear predictor is specified in terms of unknown smooth functions of predictor variables. For example, $y_i \sim \text{EF}(\mu_i, \phi)$, $g(\mu_i) = \alpha + \sum_j f_j(x_{ij})$, where EF is some exponential family distribution, g a known link function and the f_j are unknown functions. Inference with such model is facilitated by representing the f_j using intermediate rank basis expansions $f_j(x_{ij}) = \sum_k \beta_{kj} b_{kj}(x_{ij})$, where the β_{kj} are unknown coefficients and the b_{kj} known basis functions - often splines. The GAM is then simply a GLM, but over-parameterized, risking substantial overfit and poor inference. To avoid this, the usual GLM likelihood can be quadratically penalized using penalties and associated smoothing parameters that control the smoothness of the estimated f_j. The smoothing parameters can be selected to optimize prediction error. Alternatively an improper Gaussian smoothing prior can be specified, which implies posterior modes for the model coefficients corresponding to the penalized likelihood estimates. The Bayesian view also implies a full posterior distribution for the model coefficients, and opens the possibility of estimating the smoothing parameters by marginal likelihood maximisation. While theoretically straightforward, practical computation with this framework involves nested optimization methods that need to be carefully structured if stability and efficiency are to be maintained. However the resulting methods are rather general and extend beyond the simple GAM exponential family regression case to more general models, for example where location scale and shape parameters of a distribution all depend on multiple smooth functions of predictors, and to models dependent on linear functionals of smooth functions. An interesting example of the latter is the reconstruction of Covid-19 incidence (new infections per day), from daily deaths from Covid-19, and available information on the distribution of time from infection to death. The results suggest that in several European countries infections were in decline well before lockdowns were imposed. Such reconstructions rapidly lead to models with substantial non-linearity, somewhat different to regular regression models. The basic theoretical framework readily applies in these cases, but the numerical approach requires modifications if practical tractability is to be maintained. Such modifications are also discussed, and illustrated with the re-implementation of a large scale Covid statistical analysis based around a 700 state variable dynamic disease model fitted to multiple health service data streams. The key inferential target is a smooth function of time modifying transmission rates. Again the modelling strongly suggests that English Covid-19 infections were in sharp decline well before lockdown.