Boosting Geoadditive Regression Models

Thomas Kneib[1], Torsten Hothorn[1] and Gerhard Tutz[1]

[1] Department of Statistics, Ludwig-Maximilians-University Munich, Germany

Geoadditive regression models provide a comprehensive model class for analysing complex regression data combining nonlinear functional relationships, interaction surfaces, varying coefficient terms, and random effects with temporal and spatial structures [1]. While model choice and variable selection are still issues of major concern even in simple linear regression models, they become even more demanding in geoadditive regression: Should a continuous covariate be included into the model at all and if so as a linear effect or as a nonparametric, flexible effect? Is a spatial effect required in the model, i.e., is spatial correlation present beyond the spatial variation accounted for by spatially varying covariates? Are some of the covariate effects spatially varying?

To answer these and related questions, we propose a systematic, fully automated componentwise boosting procedure [2]. The general idea is to specify a set of candidate terms for the model and to iteratively apply the best-fitting candidate term to an updated residual vector. Each candidate term is associated with a base-learning procedure that, in our modelling framework, can be derived based on penalised splines for univariate nonparametric effects and varying coefficients, bivariate tensor product penalised splines for spatial effects and interactions, and cluster-specific random effects. All base-learning procedures can be cast into a general modelling framework leading to simple penalised least-squares fits. This, in turn, allows to devise a generic componentwise boosting procedure for a comprehensive model class.

When applying a suitable stopping rule to the iterative process, the boosting algorithm implements a means of model choice and variable selection. Base-learners formed by covariates with negligible effects will never be selected, leading to variable selection. Moreover, concurring base-learners can be defined for the same covariate, such as linear versus non-linear modelling. The boosting procedure compares both alternatives in each iteration and only selects the one with the better fit. Thereby it provides an automated check for whether nonlinear modelling is actually required, i.e., a means of model choice.

One major difficulty is to obtain base-learners that are comparable in complexity to avoid biased selection towards more flexible effects. The trace of the hat matrix, i.e. the equivalent degrees of freedom of a flexible effect will be used as a general measure of complexity for the base-learners. A suitable reparametrisation allows us to specify any desired degree of freedom for a base-learner and to answer additional model choice questions.

We demonstrate the versatility of our approach with two case studies. In the first example, an analysis of habitat suitability based on species abundance data, we demonstrate the impact of spatial correlation on variable selection and address further model choice problems in geoadditive models and models with space-varying coefficients. In the second example, forest health data are analyzed based on a complex model including nonparametric, spatial, interaction and random effects.

References

[1] Fahrmeir L, Kneib T, Lang S (2004) Penalised Structured Additive Regression for Space-Time Data: A Bayesian Perspective. *Statistica Sinica*, **14**, 731–761.

[2] Kneib T, Hothorn T, Tutz G (2007) Model Choice and Variable Selection in Geoadditive Regression Models. Department of Statistics Technical Report No. 3. Available from `http://www.statistik.lmu.de/~kneib`.