GAM course proposal

Title: Generalized additive models and their extensions: the penalized regression spline approach in R.

Simon N Wood
Mathematical Science
University of Bath
Bath BA2 7AY, UK.
s.wood@bath.ac.uk
01225 386603

1 Summary

This course provides an overview of the theory of generalized additive models represented by reduced rank penalized splines, and their practical use with the mgcv package in R. Here generalized additive models include generalized additive mixed models, varying coefficient/geographic regression models, structured additive regression models, generalized linear additive smooth structure models, signal regression models etc, since all of these fit into the same inferential and computational framework (quadratically penalized GLMs). The course will give a compact overview of the essential theory of penalized regression splines and GAMs, focussing on the key theoretical concepts that underpin the more detailed literature: bases, penalties, the Bayesian model of smoothing, and smoothing parameter selection. It will then cover the various types of smooth (one dimensional, isotropic and tensor product interactions) that form the basic toolkit for model construction. Model checking, building and selection will be discussed, including practical exercises with the mgcv package in R. The course will finish with a look at some more advanced GAM topics: beyond exponential family distributions, spatial and temporal autocorrelation, functional data analysis, and inference via posterior simulation. Participants should preferably bring a laptop, with the latest version of R installed.

2 Contents

1. Introduction: demonstration of some GAM-type models. Live data analysis presentation.
2. Penalized regression smoothers: the basics. Slide based presentation: visual/mathematical.
   - Splines. Bases, penalties, estimation.
   - Reduced rank splines.
   - The Bayesian, mixed model connection.
   - Effective degrees of freedom.
   - Smoothness selection (GCV, REML etc.).
3. Lab exercise: building a simple basis penalty smooth in R.
4. Extending the model. Slide based presentation: mostly mathematical.
   - Several smoothers in a model, and identifiability constraints.
   - Linear functionals of smooths (e.g. signal regression, varying coefficient models).
   - Generalization to exponential family response variables: estimation and smoothing parameter estimation.
5. Some software: R package mgcv. Live demo, working through handout.
   - Specifying and estimating models.
   - Examining the results (summaries and plots).

- Prediction.

6. A toolkit of smoothers. Slide based presentation: heavily visual.
   - Zero dimensional smooths: simple Gaussian random effects.
   - One dimensional smoothers, including cyclic and adaptive.
   - Isotropic smooths of several variables: thin plate splines, more general Duchon splines, splines on the sphere, Gaussian Markov random fields and finite region smoothing.
   - Smooth interactions via tensor product smoothing (including spatio-temporal smoothing).
   - Smooth - factor interactions.
7. Model Checking and selection. Slide based presentation + live demo.
   - Residual checking, as GLMs.
   - More checking: basis dimension, concurvity, partial residuals.
   - Selection tools: generalized AIC, approximate p-values, approximate GLRT.
   - Backwards/forwards selection.
   - Selection by further penalization.
8. Computer exercises. Participants working through exercises on their own laptops.
   - Basic GAM specification.
   - Model building and checking.
9. Some more advanced topics. Slide presentation: visual/mathematical.
   - Models beyond exponential family: ordered categorical, survival data, multivariate models etc.
   - Spatial and temporal autocorrelation.
   - Functional data analysis. Function on scalar and scalar on function regression.
   - Posterior simulation, for inference about any quantity derived from a fitted GAM.
10. Advanced topic exercises.

3 Learning outcomes

The overall aims of the course are:

1. To provide the broad-brush theoretical overview of these models and methods that allows them to be used appropriately and reliably, and makes the literature reasonably accessible.
2. To provide an overview of what is readily computationally possible, and the basic knowledge that allows participants to progress rapidly after the course to using more advanced models.

More detailed intended outcomes are:

1. To understand the notion of a penalized regression spline type smoother and its mathematical equivalence to a Gaussian random effect (and Gaussian random field).
2. To understand the concepts of Marginal Likelihood (e.g. REML) and Prediction error (e.g. GCV) smoothness estimation.
3. To understands the generalization from simple smoothing to models involving multiple smooth functions and exponential family responses.

4. To appreciate that any (bounded) linear functional of a smooth can replace the smooth in a model without fundamentally changing the GAM inferential problem.
5. To obtain an overview of the types of smooth term that can be included in a GAM.
6. To appreciate how to check a fitted GAM.
7. To be aware of some alternative model selection strategies that can be employed with GAMs
8. To be able to use R package mgcv to fit simple GAMs, and to be aware of its facilities for fitting more elaborate models.

4 Target audience & why attend

Statisticians familiar with regression modelling, but wanting to know more about practical semi-parametric modelling and what is possible. Some familiarity with R assumed.

Semi-parametric regression is now sufficiently reliable and well developed theoretically to be part of the statistician's standard toolkit. This course gives a compact overview of the key concepts, and their implementation in R.

5 Assumed Knowledge & Preparatory reading

1. The course assumes that you are familiar with the theory and use of Generalized Linear Models, up to about the level of Chapter 2 of Wood (2006).
2. General familiarity with R and its help system (including browsing html help) is assumed.
3. It is also assumed that you are familiar with the use of `glm` in R for the fitting of GLMs, and have used R's `predict`, `summary`, `anova`, `residuals`, `plot` and `AIC` commands to examine fitted GLMs.
4. To complete the R practicals in the course you will need to bring a laptop with the latest version of R and `mgcv` installed. Installing package `gamair` and `gamm4` might also be a good idea. A reasonable capacity laptop battery, fully charged, would be a good idea.

Some familiarity with Wood SN, (2006) Generalized Additive Models: An introduction with R, would be useful.

6 Room set up needed

A data projector and black/whiteboard + desks for participants to work at. A room layout that allows me to get round to participants to help with exercises is ideal.

Participants are encouraged to bring their own laptops with the most recent version of R pre-installed, *so it would be helpful to have power points for participants.* Datasets and scripts for the exercises will be distributed by USB memory stick.