Journal of Statistical Software
August 2018, Volume 86, Book Review 1. doi: 10.18637/jss.v086.b01

Reviewer: Virgilio Gómez-Rubio
Universidad de Castilla-La Mancha

Generalized Additive Models: An Introduction with R (2nd Edition)

Simon N. Wood
Chapman & Hall/CRC, Boca Raton, 2017.
ISBN 978-1-4987-2833-1. 476 pp. GBP 59.99 (H).
https://people.maths.bris.ac.uk/~sw15190/igam/

Generalized additive models (GAMs) are one of the main modeling tools for data analysis. GAMs can efficiently combine different types of fixed, random and smooth terms in the linear predictor of a regression model to account for different types of effects. Then this linear predictor can be conveniently linked to the mean of the observations, that are modeled using a distribution from the exponential family. As described in Wood's book, GAMs cover a wide range of statistical models used in practice, such as the general linear model, generalized linear models and mixed-effects models. This is stressed throughout the book with numerous examples.

The book starts by giving an overview of the general linear model in Chapter 1. Linear models are introduced with a very nice example on the estimation of the age of the universe and the estimation of Hubble's constant. Next, the author introduces the theory for linear models, which includes estimation methods and methods for model assessment and selection. In addition, a geometric interpretation of linear models is provided. The author also takes his time to discuss linear models with factors. This is important because this type of covariates may lead to unidentifiability of the model effects.

A section on practical linear modeling introduces the reader to the `lm()` function to fit linear models. Furthermore, a discussion is included on the possible problems (and how to tackle them) when fitting linear models to real data, for continuous and categorical (i.e., factors) covariates. Finally, the author describes several advanced topics on linear modeling. I have found particularly interesting his description of constraints and contrasts on the linear effects, as this is often a neglected topic when discussing linear regression models.

Chapter 2 introduces random effects for linear mixed models. It starts by providing a clear example of the main differences between fixed and random effects. Models with one and two random effects are discussed next. Different methods for the estimation of linear mixed models using maximum likelihood (ML) restricted maximum likelihood (REML) and the expectation-maximization (EM) algorithm are discussed in detail. The main theory is developed and computational details are described. A discussion of the different estimation methods is

given, which can be very helpful for practical data analysis. Selection of linear mixed effects models is also discussed here.

Although the author has developed his own packages for the estimation of some types of linear mixed models, there is a description of other packages, **nlme** (Pinheiro, Bates, DebRoy, Sarkar, and R Core Team 2018) and **lme4** (Bates, Mächler, Bolker, and Walker 2015), and how they compare to the **mgcv** package developed by the author. This is important as linear mixed effects models with complex random effects will be difficult to fit with some of these packages. Again, the author gives important advice on the analysis of data in practice using linear mixed models.

Generalized linear (mixed) models are included in Chapter 3. Here, the author extends the theory presented in the previous chapter to the case in which the response variable follows any distribution in the exponential family. Hence, the theory required for model fitting and the tools for model assessment and comparison are updated to cover this more general case. In particular, other estimation methods such as the iteratively re-weighted least square algorithm (IRLS) and penalized quasi-likelihood (PQL) are described.

The last part of this chapter is devoted to practical model fitting with R (R Core Team 2018), where different packages and estimation methods are compared. Here, the `gam()` function (in package **mgcv**) is introduced to fit generalized linear mixed models (GLMMs). Other functions described to fit GLMMs are `glmmPQL()` from the **MASS** package (Venables and Ripley 2002) and `glmer()` from package **lme4**.

GAMs are actually introduced in Chapter 4. Smooth functions defined as a linear combination of a basis of functions with compact support are used to create models with a smooth term in the linear predictor. This allows the author to introduce and discuss important topics such as how many knots to use, the choice of the basis functions to define the smooth term and how to set the smoothing parameter and the penalty term to control the degree of smoothing in the model. Package **mgcv** is described in more detail in this chapter, as well as the use of the `gam()` function for model fitting of GAMMs.

To me this is really where the book starts, as I was particularly eager to read about using smooth terms in the linear predictor of regression models. Furthermore, this chapter can serve as a gentle introduction to smoothers. A point that I have found very important is the description of how these models are regarded under a Bayesian perspective.

After this gentle introduction to smoothers, Chapter 5 deepens into smoothing splines. Several types of splines are described in detail, such as the widely known natural cubic splines, P-splines and thin plate splines. The book provides a thorough description (and discussion) of the underlying theory to implement the different types of smoothers described.

This chapter first develops one-dimensional splines and then extends these to consider more than one dimension and interactions between the covariates that define the smooth function. The last part of this chapter considers the relationship between smoothers and Gaussian Markov random fields (Rue and Held 2005), in particular, how the penalty term on the smoother parameters can be regarded as a Gaussian prior with sparse precision matrix. This provides another link between the theory of smoothers and Bayesian inference. This type of model is popular for disease mapping and geostatistics, to mention a few applications in spatial statistics.

Chapter 6 presents a general modeling framework for GAMs. This means that the general estimation procedures are also described here. These are, in particular, estimation methods

for the parameters of the fixed effects, the smoother and the smoothing parameter. In order to select the optimal degree of smoothing, several cross-validation criteria are discussed and compared. In addition, for those interested in a Bayesian approach to smoothing, computation of the marginal likelihood, for a given value of the smoothing parameter, is also discussed and an estimate using the Laplace approximation is provided.

Once the underlying theory for GAM fitting has been laid out, the extension to include random effects is presented. These can be easily estimated by noting the link between smoothers and random effects (i.e., random effects can be represented as certain types of smoothers).

This chapter also discusses general methods for model comparison, such as the AIC, for GAMMs. The author describes the main differences between marginal and conditional AIC, and when they should be used. General theory for hypothesis testing on the parameters of the smoothers is also provided. A detailed discussion on the computation of p-values for random effects terms, a case in which the effective number of parameters may be difficult to obtain, is also included here.

Finally, Chapter 7 focuses on fitting GAMMs with R. The main package described here is **mgcv** (developed by the author of the book), but for some examples other software packages are also used. For instance, the **JAGS** software (Plummer 2003) for Bayesian inference is sometimes used (Wood 2016). Although the chapter starts by summarizing how smooth terms are defined using a formula in the R language, this only takes a few pages and most of the chapter is spent on the practical and detailed analysis of different datasets.

The examples discussed start with the analysis of brain images from magnetic resonance scanning (to measure brain activity). Smooth terms are used to display the (spatial) distribution of brain activity and assess any similarities between the two brain hemispheres.

The next example models the probability of suffering from diabetic retinopathy in a set of patients depending on several covariates, such as body mass index. Smoothers are interesting here because the probability of suffering from the disease (in the logit scale) is modeled to depend on several covariates and their effects do not need to be necessarily linear.

The third example in Chapter 7 describes a Poisson regression to model the number of deaths in Chicago using covariates that measure air pollution. Here, smoothers are introduced into the model to account for seasonal patterns, for example.

Smooth spatial terms are considered for the fourth example to model fish stock in some parts of the Atlantic ocean. Here, egg counts are modeled using the sample locations and different covariates measured in the ocean, such as surface temperature and salinity. In this particular case, Tweedy and negative binomial models are considered to account for over-dispersion of the data.

Spatial smoothing is also employed in an example to map the spatial distribution of a bird species in Portugal using a grid of pixels. Presence-absence in each pixel is modeled using a logistic regression on the pixel coordinates. Here, different types of two dimensional smoothers are used to model the spatial distribution of bird species.

Examples based on GAMMs are discussed next. A spatio-temporal model, using a smooth term in three dimensions plus random effects, is used to model the number of sole eggs spawned in the Bristol channel. Note that in this case random effects are used to model overdispersion in a Poisson distribution.

Similarly, a temporal model using random effects is used to show an example on the analysis

of daily temperature in Cairo over nearly a decade. Short term autocorrelation in the data is modeled using autocorrelated errors, included as random effects in the linear predictor.

An example on fully Bayesian inference is developed to show the use of function `jagam()`, which provides an interface to fit GAMMs with **JAGS**, in the analysis of growth trajectories of a number of trees under different environmental conditions.

In addition to these examples, this chapter includes some advanced ones on survival analysis, multivariate additive models with a bivariate response and functional data analysis. The chapter ends with a summary of other R packages to fit GAMMs.

The book includes two appendices with the main results for maximum likelihood estimation and matrix algebra. The first appendix is useful as a reference on maximum likelihood estimation whilst the second is useful to check some important properties of matrix decompositions which are used for efficient computation and model fitting of GAMMs.

To sum up, Wood's book is an excellent text on additive models which, I hope, will become a classic reference book. It provides a detailed description of the different types of models and computational details to implement these models, as well as a description of different R packages to fit GAMMs. Furthermore, the author provides a critical comparison of the different methods and approaches described in the book, as well as useful advice on the practicalities of data analysis with GAMMs and model building, assessment and selection.

References

Bates D, Mächler M, Bolker B, Walker S (2015). "Fitting Linear Mixed-Effects Models Using **lme4**." *Journal of Statistical Software*, **67**(1), 1–48. doi:10.18637/jss.v067.i01.

Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team (2018). ***nlme**: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-137, URL https://CRAN.R-project.org/package=nlme.

Plummer M (2003). "**JAGS**: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling." In K Hornik, F Leisch, A Zeileis (eds.), *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*. Technische Universität Wien, Vienna, Austria. URL https://www.R-project.org/conferences/DSC-2003/Proceedings/Plummer.pdf.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rue H, Held L (2005). *Gaussian Markov Random Fields. Theory and Applications*. Chapman & Hall/CRC, Boca Raton. doi:10.1201/9780203492024.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*. 4th edition. Springer-Verlag, New York. URL http://www.stats.ox.ac.uk/pub/MASS4.

Wood SN (2016). "Just Another Gibbs Additive Modeler: Interfacing **JAGS** and **mgcv**." *Journal of Statistical Software*, **75**(7), 1–15. doi:10.18637/jss.v075.i07.

Reviewer:

Virgilio Gómez-Rubio
Department of Mathematics
Universidad de Castilla-La Mancha
Avda. España s/n
02071 Albacete, Spain
E-mail: Virgilio.Gomez@uclm.es
URL: https://becarioprecario.github.io/

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
August 2018, Volume 86, Book Review 1 *Published:* 2018-09-04
doi:10.18637/jss.v086.b01