The Journal of Open Source Software

gratia: An R package for working with generalized additive models

Gavin L. Simpson[1]

1 Department of Animal and Veterinary Science, Aarhus University, Denmark

DOI: DOIunavailable

Software

- Review
- Repository
- Archive

Editor: Pending Editor

Reviewers:

- @Pending Reviewers

Submitted: N/A
Published: N/A

Summary

Generalized additive models (GAMs) are an extension of the generalized linear model (GLM) to allow the effects of one or more covariates on the response to be modelled as a smooth function of the covariate. GAMs are increasingly being used in many applied science subjects because the smooth functions of covariates allow for flexible relationships between covariates and the response to be learned from the data through the use of penalized splines. Within the R ecosystem, Simon Wood's `mgcv` package is widely used to fit GAMs to data as it is a *Recommended* package that ships with R as part of the default desktop installation. Additionally, a growing number of other R packages build upon `mgcv`, for example as an engine to fit specialised models not handled by `mgcv` itself, or to make use of the wide range of splines available in `mgcv`.

The `gratia` package builds upon `mgcv`, providing functions that make working with GAMs fitted using `mgcv` easier. At its core, `gratia` takes a *tidy* approach providing `ggplot2`-based replacements for `mgcv`'s base graphics-based plotting capabilities, functions for model diagnostics and exploration of fitted models, as well as a family of functions for drawing samples from the posterior distribution of a fitted GAM. Additional functionality is provided to facilitate the teaching of GAMs.

In this short paper, I briefly introduce GAMs, before providing an overview of the niche filled by `gratia`. Finally, I provide a brief example of some of the main features of the `gratia`.

Generalized additive models

A GAM has the general form

$$
\begin{aligned}
y_i &\sim \mathcal{D}(\mu_i, \boldsymbol{\phi}) \\
g(\mu_i) &= \mathbf{A}_i \boldsymbol{\gamma} + \sum_{j=1} f_j(x_{ji})
\end{aligned}
$$

where observations y_i are assumed to be conditionally distributed $\mathcal{D}$ with expectation $\mathbb{E}(y_i) = \mu_i$ and dispersion parameter ϕ. The expectation of y_i is given by a linear predictor of strictly parametric terms, whose model matrix is $\mathbf{A}_i$ with parameters γ, plus a sum of smooth functions of covariates $f_j()$. $g()$ is a link function mapping values on the linear predictor to the scale of the response.

Statement of need

`mgcv` is state-of-the-art R-based software for fitting GAMs are their extensions to data sets on the order of millions of observations. The package is continually maintained and ships with the standard R installation as a *recommended* package. `mgcv` provides functions for plotting the estimated smooth functions of a model, as well as for producing model diagnostic plots. These functions produce plots using base graphics, the original plotting system for R. One of the original motivations driving the development of `gratia` was to provide equivalent plotting capabilities for GAMs fitted by `mgcv` using the `ggplot2` package and the grammar of graphics. To facilitate this, `gratia` provides functions for representing the model terms using *tidy* principles that are suited to plotting via `ggplot2` or manipulation within the *tidyverse* ecosystem of packages. This functionality allows for high-level plotting using the `draw()` method, as well as easily customisable plot generation using lower-level functionality.

Taking a Bayesian approach to smoothing with penalized splines, it can be shown that GAMs fitted by `mgcv` are an empirical Bayes model with improper multivariate normal priors on the basis function coeficients.

```r
library("mgcv")
library("gamair")
library("gratia")
library("ggplot2")
library("dplyr")
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:nlme':
##
##     collapse
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
data(chl, package = "gamair")
chl <- chl |>
  as_tibble()
chl
```

```
## # A tibble: 13,840 x 6
##        lon   lat jul.day  bath   chl chl.sw
##      <dbl> <dbl>   <int> <int> <dbl>  <dbl>
##  1 -0.0018  60.0     148   120  1.5   1.01
##  2 -0.002   60.4     267   110  0.9   1.23
##  3 -0.0058  72.7     204  2739  0.4   0.391
##  4 -0.0173  50.6     110    44  0.48  1.76
##  5 -0.0227  53.9     212     4  0.4   2.06
##  6 -0.0277  53.9     212     4  0.6   2.06
##  7 -0.0327  50.3     124    50  1.05  1.73
##  8 -0.0671  55.5     273    62  1.1   1.17
##  9 -0.0704  50.6     109    44  1.5   1.75
## 10 -0.0707  53.6     137     6  2.65 12.3
## # i 13,830 more rows
```

```r
m <- gam(chl ~ s(lon, lat, bs = "sos"), data = chl, method = "REML")
summary(m)
```

```
##
## Family: gaussian
## Link function: identity
##
```

```
## Formula:
## chl ~ s(lon, lat, bs = "sos")
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.59802    0.01766   90.51   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##              edf Ref.df     F p-value
## s(lon,lat) 45.42     49 65.38  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.188   Deviance explained = 19.1%
## -REML =  29847  Scale est. = 4.3144    n = 13840
```