Computational Statistics Final

Daniel J. Park

March 20, 2020

1 The motivation for HGAMs

Scientists today using statistical techniques to analyze ecological data rely on two popular methods for modelling: generalized additive models (GAMs) and hierarchical generalized linear models (HGLMs). In their article, Pedersen, Miller, Simpson, and Ross show how incorporating a hierarchical structure to a GAM allows us to account for smooth functional relationships between predictor variables and the response differing across groups of observations while those relationships may still be pooled toward a similar shape.[1] We can investigate how functional relationships vary between groups of observations and whether a common relationship or shape models the data well.

One motivating hypothetical example for HGAMs in the paper is to estimate the relationship between the maximum size of a fish and temperature. Explanatory analysis of the sample might show that each species of fish displays its own smooth function of size and temperature, but that all species in the sample also share a similar trend in their responses over the temperature range. Fitting a smooth function to each species might ignore some common shape that they share, and we might overfit for the species with few observations. On the other hand, fitting one model for the entire data might be too general and do poorly in predicting any one species. An HGAM as a hierarchical model includes a global function relating size and temperature while incorporating species-specific functional relationships that can be penalized toward the global mean.

[1] Pedersen EJ, Miller DL, Simpson GL, Ross N. 2019. Hierarchical generalized additive models in ecology: an introduction with mgcv. PeerJ 7:e6876 https://doi.org/10.7717/peerj.6876

2 Overview of Generalized Additive Models

A GAM relates the expected response through an inverse link function to a linear combination of explanatory variables. The relationship between each covariate and the response can be modelled as a smooth function such as a spline, hence the $f_j\left(x^{(j)}\right)$, a smoother of the j-th variable. In equation form, a simple GAM looks like:

$$\mathbb{E}(Y) = g^{-1}\left(\beta_0 + \sum_{j=1}^{J} f_j\left(x^{(j)}\right)\right) \tag{1}$$

Each smoother is comprised of K basis functions multiplied by their respective coefficients and is represented by the equation:

$$f_j\left(x^{(j)}\right) = \sum_{k=1}^{K} \beta_{j,k} b_{j,k}\left(x^{(j)}\right) \tag{2}$$

To prevent overfitting, we estimate the coefficients $\beta_{j,k}$ subject to a constraint involving the model likelihood L, a smoothing parameter λ, and a penalty matrix S:

$$L - \lambda\beta^{\top} S\beta \tag{3}$$

where the contents of the matrix S depend on the type of smoother. The three types employed by Pedersen et al. are thin plate regression splines (TPRS), cyclic cubic smoothers, and random effects.

3 Types of HGAMs

Once we introduce hierarchical structure and are interested in variability between groups, we can choose between five different model structures:

1. A single common smoother for all observations.
2. A global smoother plus group-level smoothers that have the same wiggliness.
3. A global smoother plus group-level smoothers with differing wiggliness.
4. Group-specific smoothers without a global smoother, but with all smoothers having the same wiggliness.

5. Group-specific smoothers with different wiggliness.

Wiggliness here refers to how much a function changes over its range. Two functions can have a similar amount of wiggliness while taking different shapes, which is a measure of the average squared distance between them after being scaled to have a mean value of 0 across their ranges.

4 Reproducing the work in Pedersen et al.

I reproduce an example in Pedersen et al. concerning the `CO2` dataset available in R. As they explain, "This data is from an experimental study by Potvin, Lechowicz & Tardif (1990) of CO2 uptake in grasses under varying concentrations of CO2, measuring how concentration-uptake functions varied between plants from two locations (Mississippi and Quebec) and two temperature treatments (chilled and warm). Twelve plants were used and CO2 uptake measured at 7 CO2 concentrations for each plant (Figure 1). Here we will focus on how to use HGAMs to estimate inter-plant variation in functional responses."

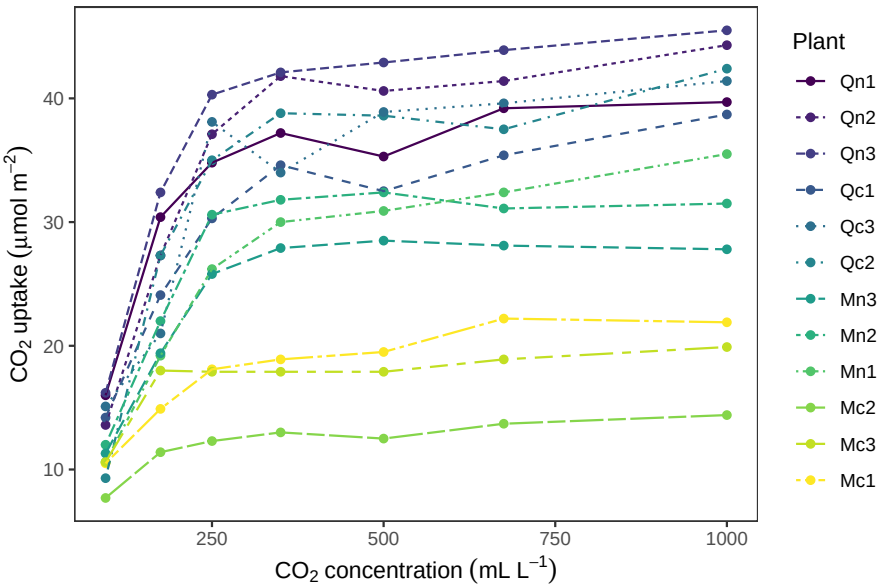

Figure 1: CO_2 Data by Plant.

The first model is a global smoother modelling the log of `uptake` as the sum of two smoothers—a thin plate regression spline of the log of `concentration`

and a random effect of `plant` for group-specific intercepts. In equation form, it looks like:

$$\ln(\texttt{uptake}_i) = f(\ln(\texttt{conc}_i)) + \zeta_{\texttt{plant}} + \epsilon_i \tag{4}$$

where $f(\cdot)$ is a smoother for concentration, ζ is a random effect for each plant-specific intercept, and ϵ_i is a normally-distributed error.

Figure 2 shows the predictions of the model described in Equation 4 at each level of observed CO_2 concentration for each species of plant surrounded by a shaded zone representing ± 2 standard errors. Because in Equation 4 we used a log-transformation of the data, we first exponentiate the predictions to bring them back to the original units for uptake and concentration. As expected, there is only one global smoother that is shifted vertically depending on the estimated intercept for each plant species. Although Model 1 seems to fit the data pretty well, it makes errors consistently toward one direction for a couple of the plants. For example, looking at the graph for Plant Qc2, the model underestimates uptake at most concentrations. For Plant Mc3, the model overestimates uptake for concentrations past 250mL/L.

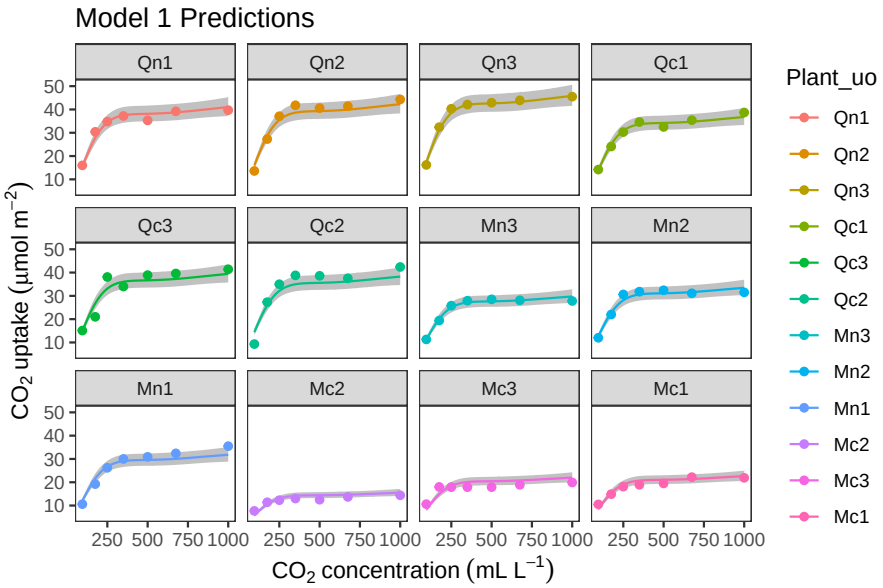

Figure 2: Model 1

Model 2 also incorporates a single global smoother, but now we allow for group-level smoothers that have similar wiggliness. We penalize large deviations

from the average by making the group-level smoothers share a penalty term. The equation form of this model is slightly modified from before:

$$\ln(\texttt{uptake}_i) = f(\ln(\texttt{conc}_i)) + f_{\texttt{plant}}(\ln(\texttt{conc}_i)) + \epsilon_i \tag{5}$$

We still keep the term representing the global smoother, but now we add plant-specific smoothers of the log of concentration for a given plant. Since each $f_{\texttt{plant}}(\cdot)$ contains its own intercept, we don't need to encode the $\zeta_{\texttt{plant}}$ term from Model 1.

Figure 3 shows the fitted global smoother and group-level smoothers. From the plot on the right side, we notice first that the plants differ in their average log uptakes. For instance, the range of the purple curve on the bottom doesn't overlap with the range of the pink curve above it, and the average value of the purple curve is quite far from that of the gold curve on the very top. Secondly, we can see that the shape of each plant's response function differs. Some slope down and then up while others do vice versa over the domain of log concentration. Some are flatter curves while others have steeper 'bumps'.

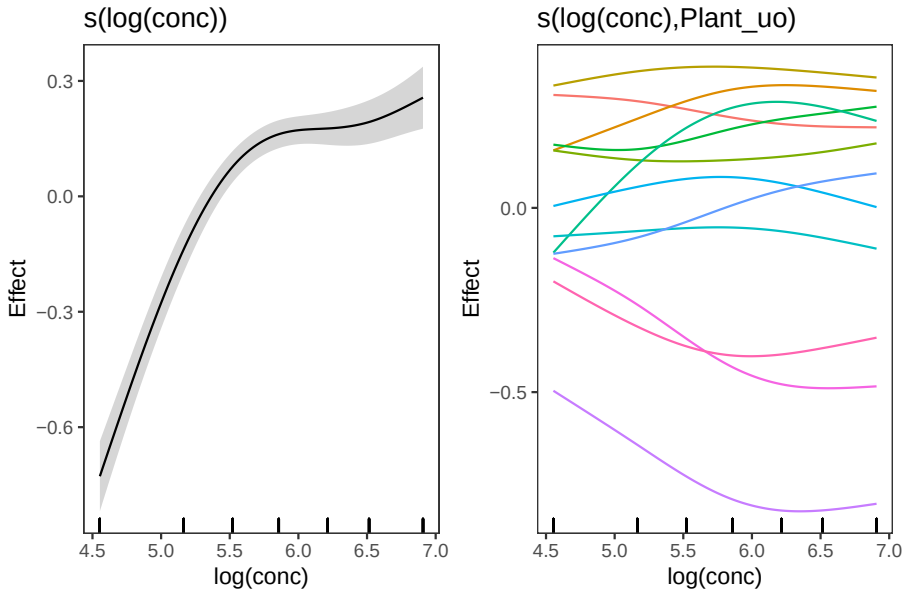

Figure 3: Global smoother (left) and fitted plant-level smoothers (right) for Model 2

The predictions from Model 2 shown in Figure 4 display a better fit to the observed data. Now that we added plant-specific smoothers that are allowed to differ

in shape from the global average while maintaining similar levels of wiggliness, we see that the estimates no longer consistently over- nor under-estimate the true uptake for any of the plants.

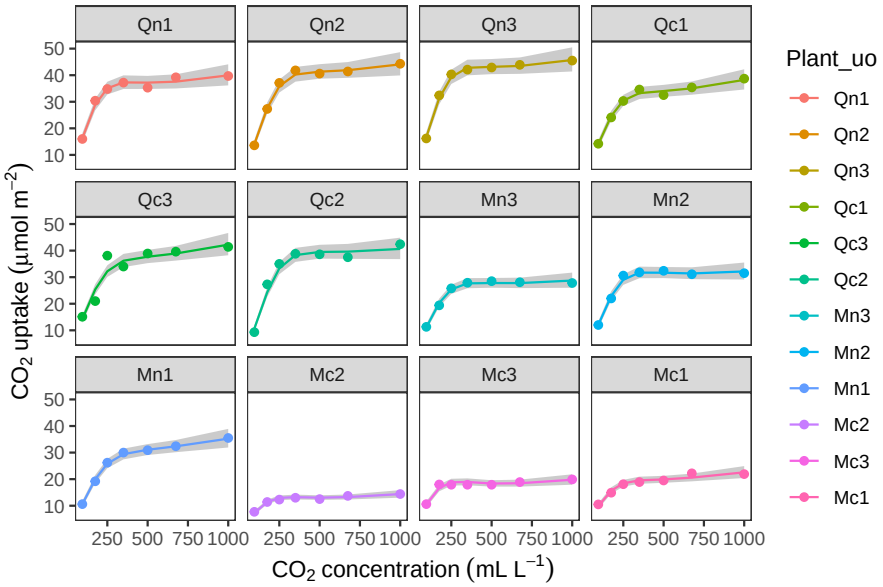

Figure 4: Model 2

Model 3 combines a single global smoother with group-level smoothers, except now we allow each group smoother to have its own smoothing parameter and potentially differing levels of wiggliness. As evident from Figure 5, the smoother for Plant Mc1 displays greater wiggliness than the smoother for Plant Qc1.

Model 3 (predictions shown in Figure 6) is useful if the Plant groups differ substantially in their wiggliness, but it seems from the plot that Model 3 does not improve the fit much beyond Model 2 because the data does not wiggle or vary so differently among plant species.

Model 4 takes the form of Model 2, except we remove the global smoother $f(\ln(\texttt{conc}_i))$ and use only the group-level smoothers that share a common smoothing parameter: "This model assumes all groups have the same smoothness, but that the individual shapes of the smooth terms are not related."[2] In this CO_2 example, the model and plots for Model 4 are very similar to those of Model 2, but this may not always be the case. If there are only a few observations in each group

[2]Ibid., 21.

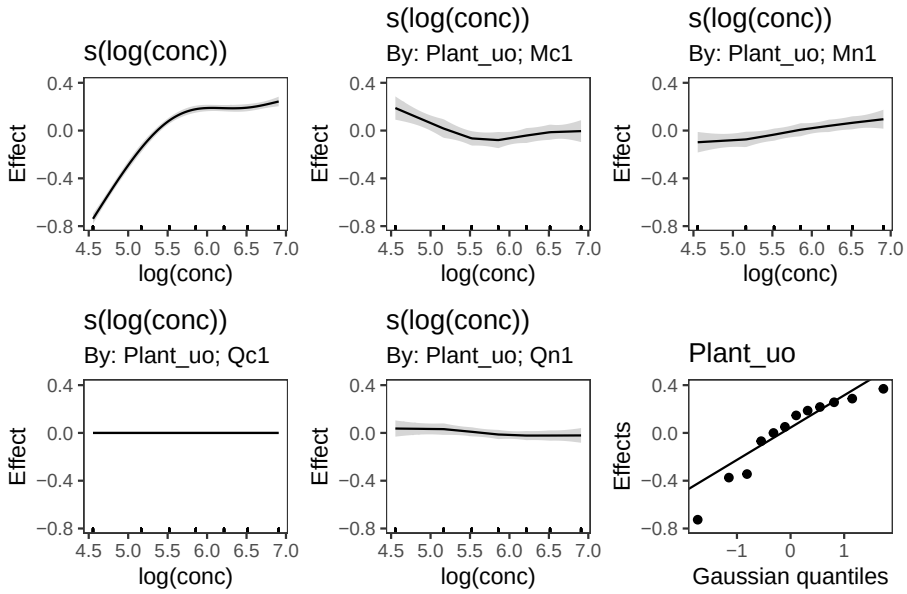

Figure 5: Global smoother (top left), random effect intercepts for Plant (bottom right), and a selection of group-level smoothers

in some training dataset, estimates from Model 4 will likely be more variable than those from Model 2, since without the global smoother the model cannot assume some shared functional shape among groups.

Lastly, Model 5 is Model 3 after dropping the first term representing the global smoother. Model 5 assumes the group smoothers may take on differing wiggliness and that their shapes do not share a common trend.

Since Pedersen et al. do not show plots comparing the observed data and the predictions from Models 4 and 5, I do not reproduce them here.

5 Comparing a GAM and HGAMs on an example

I now compare the performance of HGAMs on a dataset presented at the American Statistical Association's 1983 Exposition that contains data about 398 cars and their miles per gallon (mpg).[3] The dataset also contains the following predictor variables: number of cylinders, engine displacement in cubic inches, horsepower,

[3] http://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/auto-mpg.data

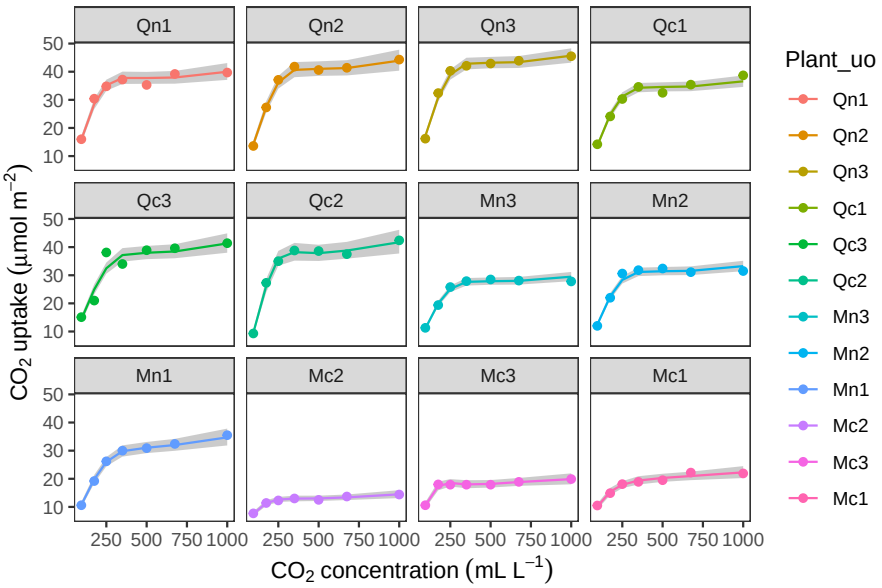

Figure 6: Model 3

weight in pounds, time it takes to accelerate from 0 to 60mph in seconds, years released after 1900, and region of origin. I am interested in the variation in functional relationship between mpg and covariates among cars from three different regions (1. American 2. European 3. Japanese).

From a pairwise correlation plot, I noticed that cylinders, displacement, and weight are strongly correlated. To avoid issues of multicollinearity, simplify the model, and reduce computational costs, I chose not to include `cylinders` and `weight` in the model. I also shifted the `year` column by the minimum year so that the covariate values started at 0 instead of 70. The observed mpg data look approximately normally distributed, so I assume that $\mathbb{E}[\texttt{mpg}]$ follows a Gaussian distribution. After splitting the dataset into training and test sets according whether the car model year was even or odd, I fit Model 1, the global smoother for all observations. In equation form, it looks like:

$$\texttt{mpg}_i = f(\texttt{disp}_i) + f(\texttt{hp}_i) + f(\texttt{accel}_i) + f(\texttt{year}_i) + \zeta_{\texttt{origin}_i} + \epsilon_i \quad (6)$$

Then I proceeded to incorporate the other four hierarchical structures onto GAMs as outlined in Section 3, fitting Models 2, 3, 4, and 5. The `gam.check` function in R produces a QQ-plot of the residuals, a histogram of residuals, and a

graph of response vs. fitted values. Diagnostic plots for all 5 models show that the normality assumption of the residuals still approximately holds, which means that the GAM and HGAMs are appropriate. As a reference, I show some diagnostic plots for Model 5 (Figure 7).

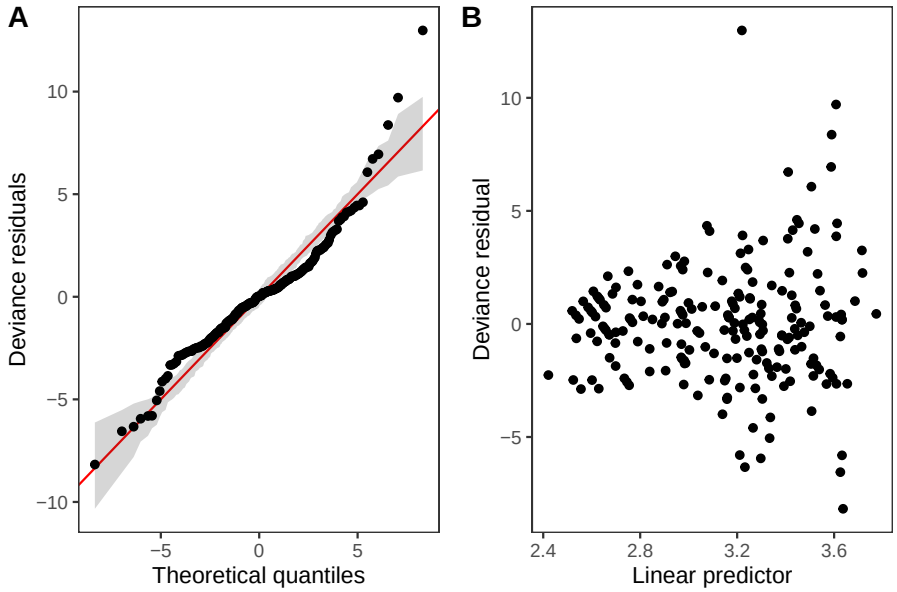

Figure 7: Model 5 QQ-plot and Fitted vs. Residual plot

To compare how the models perform against each other and relative to a null model containing only intercepts for a car's region of origin, I produced a table of total deviance by origin and model. Models 1-5 are all much more accurate in predicting mpg from the test data than the null model. What we see among Models 1-5 is that for American cars, allowing covariate smoothers to have differing wiggliness and not including global smoothers (Model 5) performs best. However, for European and Japanese cars, the global smoothers (Model 1) predicts with the least deviance from the test responses.

Why does Model 5 perform best on American cars and Model 1 on European and Japanese ones? This may have to do with unbalanced group sizes in the training data - we used 126 American, 41 European, and 47 Japanese cars to train the models. Since the European and Japanese groups are relatively small, global smoothers would help share information about common shapes and wiggliness from the American group. As Pedersen et al. discuss, the decision over which

	Total Deviance of held out data					
Origin	Intercept only	Model 1	Model 2	Model 3	Model 4	Model 5
American	4825	873	711	724	722	692
European	604	161	206	226	220	211
Japanese	1166	384	408	459	390	396

Table 1: Out-of-sample predictive ability for Models 1-5 applied to the 1983 Cars mpg dataset. Deviance values represent the total deviance of model predictions from observations for test data. 'Intercept only' results are for a null model with only origin-level random effect intercepts included.

model to use should be influenced in part by the goal of the study. Are we interested in the shape of the global average mpg? Then Models 1-3 would be best suited for the task. Or are we only interested in how the relationship between mgp and covariates varies depending on a car's region of origin? In which case Models 4-5 would be appropriate.

Model	df	AIC	ΔAIC
Model 1	21	1154	61
Model 2	35	1101	8
Model 3	35	1093	0
Model 4	31	1104	12
Model 5	37	1093	1

Table 2: Degrees of Freedom and AIC for HGAMs

The original intention of the study and expert knowledge of the automobile domain is important to keep in mind during model selection. While AIC may be a robust measure of how well a model balances predictive power and parameter complexity (comparisons shown in Table 2, Pedersen et al. stress that it should not be the sole factor: "Instead, model selection should be based on expert subject knowledge about the system, computational time, and most importantly, the inferential goals of the study."

References

[1] Pedersen EJ, Miller DL, Simpson GL, Ross N. 2019. Hierarchical generalized additive models in ecology: an introduction with mgcv. PeerJ 7:e6876 https://doi.org/10.7717/peerj.6876