Prof. Marco Avella
Scribe: Monica Yan

In this lecture we will introduce the additive models and see how GAM generalize the additive models as the same way GLM do to linear models.

1 Motivation

- Extend the linear model regression to allow for smooth unknown dependence on covariates
- Extend non-parametric regression to include multiple covariates.
- Extend the GLM framework to include smooth unknown dependence of response on covariates.
- Introduce main ideas with roughness penalty methods and spline basis.
- Discuss statistical inference ideas in this context.

2 Partially Linear Model (Spring Barley Data)

	Block 1		Block 2		Block 3	
Location t	Variety	Yield y	Variety	Yield y	Variety	Yield y
1	57	9.29	49	7.99	63	11.77
2	39	8.16	18	9.56	38	12.05
3	3	8.97	8	9.02	14	12.25
4	48	8.33	69	8.91	71	10.96
5	75	8.66	29	9.17	22	9.94
6	21	9.05	59	9.49	46	9.27
7	66	9.01	19	9.73	6	11.05
8	12	9.40	39	9.38	30	11.40
9	30	10.16	67	8.80	16	10.78
10	32	10.30	57	9.72	24	10.30
11	59	10.73	37	10.24	40	11.27
12	50	9.69	26	10.85	64	11.13
13	5	11.49	16	9.67	8	10.55
14	23	10.73	6	10.17	56	12.82
15	14	10.71	47	11.46	32	10.95
16	68	10.21	36	10.05	48	10.92
17	41	10.52	64	11.47	54	10.77
18	1	11.09	63	10.63	37	11.08
19	64	11.39	33	11.03	21	10.22
20	28	11.24	74	10.85	29	10.59
21	46	10.65	13	11.35	62	11.35
22	73	10.77	43	10.25	5	11.39
23	37	10.92	3	10.08	70	10.59
24	55	12.07	53	10.25	13	11.26
25	19	11.03	23	9.57	11	11.79

Figure 1: Spring Barley data (partial)

Figure 1 shows partial spring barley dataset[1]. It gives standardized yields from an agricultural field trial in which three blocks of long narrow plots were sown with 75 varieties of spring barley in a random order within each block. The yield from variety 27 in the third block is missing, but otherwise there are three replicates for each variety.

The simplest model that would accommodate the variety effects is a linear model with variety and block effects

$$y_{\nu b} = \tau_b + \beta_\nu + \varepsilon_{\nu b},\ \nu = 1, \ldots, 75,\ b = 1, 2, 3 \quad \text{where } \varepsilon_{\nu b} \overset{iid}{\sim} N(0, \sigma^2)$$

here τ_b is the yield for bth field; β_ν is the yield for νth barley variety.

This has residual sum of squares 94.87 on 147 degrees of freedom,giving $\hat{\sigma}^2 = 0.645$, while the standard error for a difference of variety effects $\hat{\beta}_{v1} - \hat{\beta}_{v2}$ is 0.655. As the model only consider two predictors: the block it belongs to τ_b, and the variety of the barley β_ν, it ignores the spatial variation, as the yield can be different due to the location within the block. So this model greatly overestimates σ^2, thereby decreasing the sensitivity of comparisons among the varieties of barley. Moreover the variety effect estimators may be biased if all three replicates of a particular variety happen to fall where the fertilities are higher than average.

A more flexible model that allows the yield for the νth variety in the bth block to depend on its location $t_{\nu b}$ is

$$y_{\nu b} = g_b(t_{\nu b}) + \beta_\nu + \varepsilon_{\nu b},\ \nu = 1, \ldots, 75,\ b = 1, 2, 3,$$

where $\varepsilon \overset{iid}{\sim} N(0, \sigma^2)$ and $g_b(\cdot)$ is smooth function that determines how the fertility pattern in block b depends on the location t.

Figure 2 gives some analysis of the spring data. Left panel shows fit of yield y as a function of location t for the three blocks. From bottom to the top are yields for blocks 1, 2 and 3 respectively (block 2 and 3 have been offset by adding 4 and 8). The smooth solid lines are the fits of polynomials of degree 20, 10 and 40 to the data from blocks 1, 2 and 3. Upper right shows the yields for block 1, with smoothing spline fit with 18 degrees of freedom. Lower right is cross-validation (solid) and generalized cross-validation (dots) criteria for smoothing spline fits to blocks 1, 2 and 3, with minima at roughly 20, 10 and 40 equivalent degrees of freedom.

Polynomial fittings of yields y and location t (left panel of Figure 2) show strong spatial patterns owing to fertility trends within each block, in addition to the variety effects. It also shows some of the disadvantages of polynomial fitting. The lower curve, for example, wiggles much more compared to a spline fit with the same degrees of freedom, which is shown in the upper right panel. The lower right panel shows how CV(λ) and GCV(λ) vary with the equivalent degrees of freedom for the three blocks. The fit to block 2 seems fairly reasonable, but block 3 is evidently overfitted with 40 degrees of freedom, and block 1 is probably also overfitted.

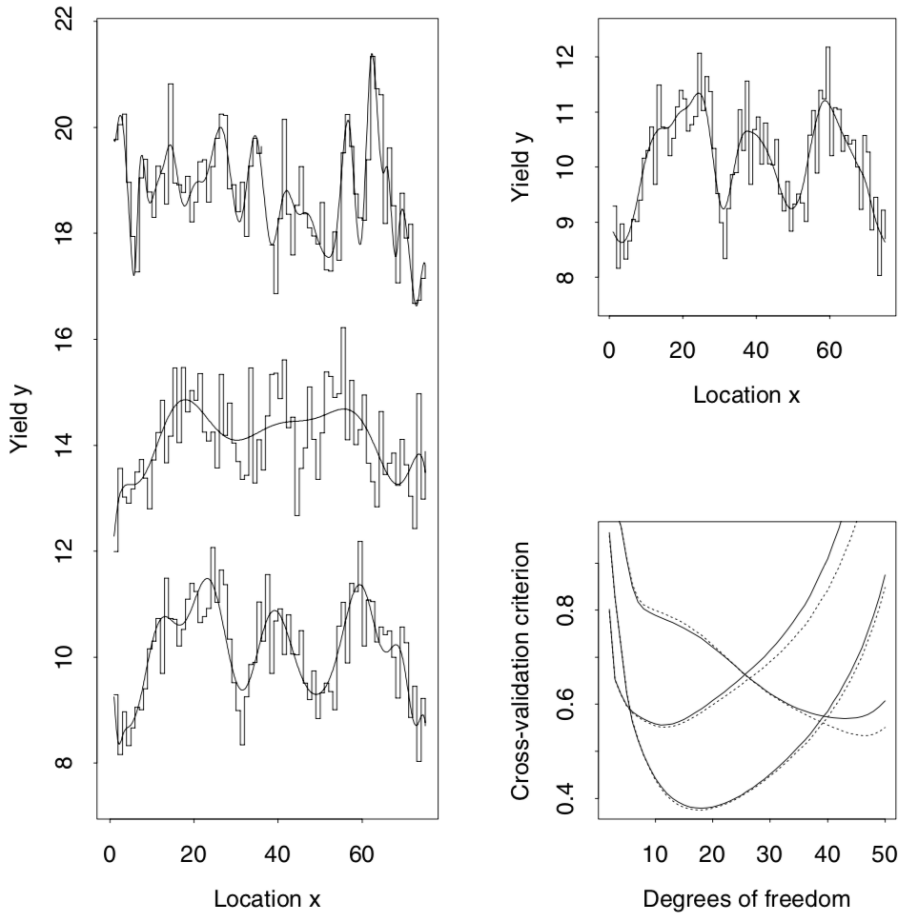

Figure 2: Spring barley data analysis ($g_b(t)$ estimation)

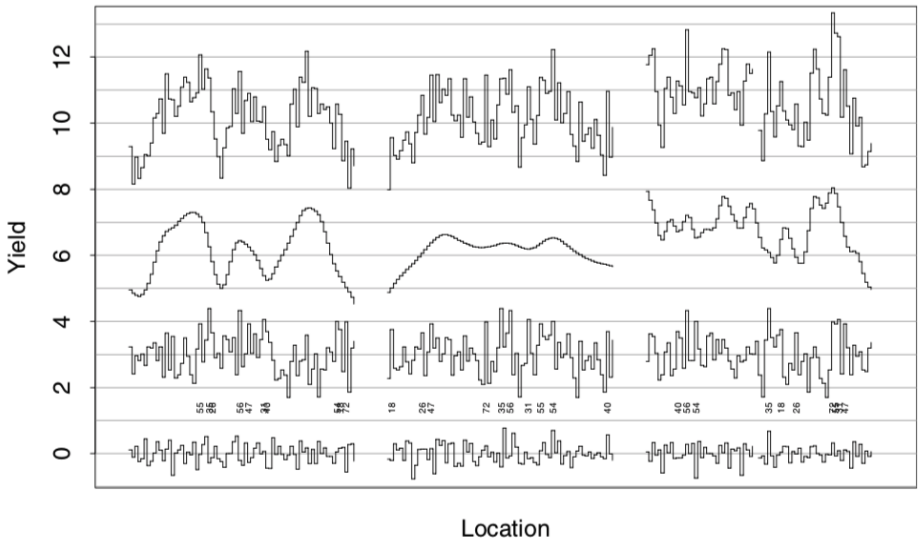

Figure 3: Spring barley data analysis

Figure 3 shows the each ingredients in model for all three blocks. Block 1 is shown on the left and block 3 on the right. Starting from the top, the panels are 1) the original yields y_b; 2) the fertility trend $\hat{g}_b(t)$; 3) variety effect estimates β_ν; and 4) the crude residuals. Both panels for the fertility trend $\hat{g}_b(t)$ and variety effect estimates β_ν are shifted above for display. The varieties with the ten largest β_ν are marked.

In this example, the model is partially linear as it is a linear model basic on predictors block fertility trend $g_b(t)$ and barley variety effect β_ν. However, the fertility trend $g_b(t)$ is estimated using non-parametric method (polynomial spline).

3 Additive Models

The partially linear model example above is using the idea of additive method. Besides, the non-parametric regression methods we introduced in previous lecture (polynomial expansions, spline smoothing) are also additive models. They share the same idea. In this section, we will introduce the additive models in systemically.

3.1 Formulation

Assume the sample of pairs (X_i, Y_i) is such that

$$Y_i = \alpha + f_1(X_{i1}) + \cdots + f_p(X_{ip}) + \varepsilon_i,\ i = 1, \ldots, n$$

- Response variable: Y_i
- Covariate: X_i
- Identifiability: unknown smooth regression functions $f_j(\cdot)$ such that $\mathbb{E}[f_j(X_{ij})] = 0$.
- Noise term: ε_i, assumed to be i.i.d. with $\mathbb{E}[\varepsilon_i] = 0$ and $\text{cov}(X_i, \varepsilon_i) = 0$.

3.2 Spline Approximation (refresher)

As we discussed in previous lecture, spline approximation tries to approximate the unknown function as a expansion of piecewise known functions

$$f_j(x_j) \approx f_{K_j}(x_j) = \sum_{k=1}^{K_j} \beta_{jk} b_k(x_j) = \boldsymbol{\beta}_j^T \mathbf{b}_j(x_j)$$

we can consider wigglingness penalties of the form

$$\int \{f''_{K_j}(x)\}^2 \mathrm{d}x = \int \boldsymbol{\beta}_j^T \mathbf{b}''_j(x) \mathbf{b}''_j(x)^T \boldsymbol{\beta}_j \mathrm{d}x = \boldsymbol{\beta}_j^T \mathbf{P}_j \boldsymbol{\beta}_j$$

[1] entire data can be found in A.C. Davison (2003). "Statistical Models" *Chapter 10.7* Table 10.21 pp.534-535

where

$$\mathbf{P}_j = \begin{bmatrix} \int b_1''(x)b_1''(x)\mathrm{d}x & \dots & \int b_1''(x)b_{K_j}''(x)\mathrm{d}x \\ \int b_2''(x)b_1''(x)\mathrm{d}x & \dots & \int b_2''(x)b_{K_j}''(x)\mathrm{d}x \\ \vdots & \ddots & \vdots \\ \int b_{K_j}''(x)b_1''(x)\mathrm{d}x & \dots & \int b_{K_j}''(x)b_{K_j}''(x)\mathrm{d}x \end{bmatrix}$$

Minimize the penalized least squares loss we get

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \mathrm{argmin}_{\boldsymbol{\beta}} \Big\{ \sum_{i=1} \{y_i - \sum_{j=1}^{p} f_{K_j}(x_{ij})\}^2 + \sum_{j=1}^{p} \lambda_j \int f_{K_j}''(x)^2 \mathrm{d}x \Big\} \\ &= \mathrm{argmin}_{\boldsymbol{\beta}} \Big\{ \|\mathbf{y} - \sum_{j=1}^{p} \mathbf{X}_j \boldsymbol{\beta}_j\|_2^2 + \sum_{j=1}^{p} \lambda_j \boldsymbol{\beta}_j^T \mathbf{P}_j \boldsymbol{\beta}_j \Big\} \\ &= \mathrm{argmin}_{\boldsymbol{\beta}} \Big\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \boldsymbol{\beta}^T \mathbf{P}(\boldsymbol{\lambda}) \boldsymbol{\beta} \Big\} \\ &= \mathrm{argmin}_{\boldsymbol{\beta}} \left\{ \left\| \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{X} \\ \mathbf{P}(\boldsymbol{\lambda})^{1/2} \end{bmatrix} \boldsymbol{\beta} \right\|_2^2 \right\} \end{aligned}$$

where $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_p)$ and $\mathbf{P}(\boldsymbol{\lambda}) = \mathrm{blockdiag}(\lambda_1 \mathbf{P}_1, \dots, \lambda_p \mathbf{P}_p)$.
Then the solution is again a ridge estimator.

$$\hat{\boldsymbol{\beta}} = \{\mathbf{X}^T\mathbf{X} + \mathbf{P}(\boldsymbol{\lambda})\}^{-1}\mathbf{X}^T\mathbf{y}$$

3.3 Choice of the smoothing parameter (cross-validation)

- Leave one out cross validation:

$$\mathrm{CV}(\lambda_1, \dots, \lambda_p) = \sum_{i=1}^{n} \Big\{ y_i - \sum_{j=1}^{p} \hat{f}_j^{(-i)}(x_i) \Big\}^2,$$

  where $\hat{f}_j^{(-i)}(x_{ij})$ indicates the fit at x_{ij} leaving out the ith observation with tuning parameters $\lambda_1, \dots, \lambda_p$. Since we have linear smoother of the form $\hat{\mathbf{f}} = \mathbf{S}_\lambda \mathbf{y}$, it simplifies to

$$\mathrm{CV}(\lambda_1, \dots, \lambda_p) = \sum_{i=1}^{n} \Big(\frac{y_i - \hat{f}(x_i)}{1 - (\mathbf{S}_\lambda)_{ii}} \Big)^2 = \sum_{i=1}^{n} \Big(\frac{y_i - \sum_{j=1}^{p} \hat{f}_j(x_i)}{1 - (\mathbf{S}_\lambda)_{ii}} \Big)^2,$$

- Generalized cross validation:
  The problem with ordinary cross validation arise as because, despite of the parameter estimate, the effective degree of freedom $tr(\mathbf{S}_\lambda)$ and expected prediction error being invariant to rotation of $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ by any orthogonal matrix $\mathbf{Q}$, while the elements $(\mathbf{S}_\lambda)_{ii}$ are not invariant. This sensitivity to an abitrary choice how fitting is done is unsatisfactory. One way out this, is to pick the a "good" rotation of $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$. For example we could pick a rotation that avoids as much a possible to have a few points with a high influence relative to others. This happens when we have very uneven values $(\mathbf{S}_\lambda)_{ii}$, which would entail that the CV score is dominated by a small proportion

of data. This suggests choosing the rotation $\mathbf{Q}$ to make $(\mathbf{S}_\lambda)_{ii}$ as even as possible.[2] This leads to the generalized cross validation score:

$$\mathrm{GCV}(\lambda_1, \ldots, \lambda_p) = \sum_{i=1}^{n} \Big(\frac{y_i - \hat{f}(x_i)}{1 - \mathrm{tr}(\mathbf{S}_\lambda)/n} \Big)^2 = \sum_{i=1}^{n} \Big(\frac{y_i - \sum_{j=1}^{p} \hat{f}_j(x_{ij})}{1 - \mathrm{tr}(\mathbf{S}_\lambda)/n} \Big)^2$$

3.4 An Example (Tree data)

	Girth	Height	Volume		Girth	Height	Volume
1	8.3	70	10.3	17	12.9	85	33.8
2	8.6	65	10.3	18	13.3	86	27.4
3	8.8	63	10.2	19	13.7	71	25.7
4	10.5	72	16.4	20	13.8	64	24.9
5	10.7	81	18.8	21	14.0	78	34.5
6	10.8	83	19.7	22	14.2	80	31.7
7	11.0	66	15.6	23	14.5	74	36.3
8	11.0	75	18.2	24	16.0	72	38.3
9	11.1	80	22.6	25	16.3	77	42.6
10	11.2	75	19.9	26	17.3	81	55.4
11	11.3	79	24.2	27	17.5	82	55.7
12	11.4	76	21.0	28	17.9	80	58.3
13	11.4	76	21.4	29	18.0	80	51.5
14	11.7	69	21.3	30	18.0	80	51.0
15	12.0	75	19.1	31	20.6	87	77.0
16	12.9	74	22.2				

Figure 4: Tree data

Figure 4 provides measurements of the girth, height and volume of timber in 31 felled black cherry trees. Note that girth is the diameter of the tree (in inches) measured at 4 ft 6 in above the ground. Figure 5 is the scatter plot of the tree data, which gets us some general idea about the data pattern : relation between girth and height, and height and volume seem not to follow the simple linear relation.

Let's consider the following two additive models

$$Volume_i = \alpha + f_1(Height_i) + f_2(Girth_i) + \varepsilon_i, \ i = 1, \ldots, 31 \tag{1}$$

$$\log(Volume_i) = \alpha + f_1(Height_i) + f_2(Girth_i) + \varepsilon_i, \ i = 1, \ldots, 31 \tag{2}$$

From Problem set 9, Exercise 3 (SMPracticals, Exercise 1 of Chapter 8), we know that fitting model (1) parametrically gives a very poor fit, while model (2) looks better. Figure 6 shows the components in model (1) (top) and model (2) (bottom). From the conic shape of threes

[2] More details in Wood's book p.258–260.

one could expect a quadratic *Girth* effect and linear *Height* effect on the volume of the trees.

Note that the unknown regression functions are centered at 0 in additive models in order to guarantee the identifiability of the estimated functions $\hat{f}_i$, hence $\int \hat{f}_i(t)\mathrm{d}t = 0$. From the right two panels of Figure 6 we can see that the lines are all centered around 0.

For model (2) we get the following summary

```
Family: gaussian
Link function: identity

Formula:
log(Volume) ~ s(Girth) + s(Height)

Parametric coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.27273    0.01499   218.3   <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(Girth)  2.41  3.029 218.72  < 2e-16 ***
s(Height) 1.00  1.000  30.14 5.45e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.975   Deviance explained = 97.8%
GCV = 0.0081245  Scale est. = 0.0069688  n = 31
```

Figure 7 gives model residual checking plots of fitting log additive model. The residuals plot (upper right panel) looks fine, and the response vs. fitted values scatter plot is also around the $y = x$ line.

4 Generalized Additive Models

4.1 Statistical Model

A GAM assumes that the response variables Y_i are drawn from independent distributions of the type

$$f(y_i; \theta_i, \phi) = \exp\Big[\big\{y_i\theta_i - b(\theta_i)\big\}/\phi + c(y_i, \phi)\Big],\ \forall i = 1, \ldots, n,$$

where b and c are some specific functions and ϕ a nuisance parameter. It corresponds to a *natural* exponential family with natural parameter θ_i/ϕ.
Furthermore it is assumed that

$$E[Y_i|\mathbf{X}_i = \mathbf{x}_i] = \mu_i \quad \text{and} \quad V[Y_i|\mathbf{X}_i = \mathbf{x}_i] = \phi V(\mu_i)$$

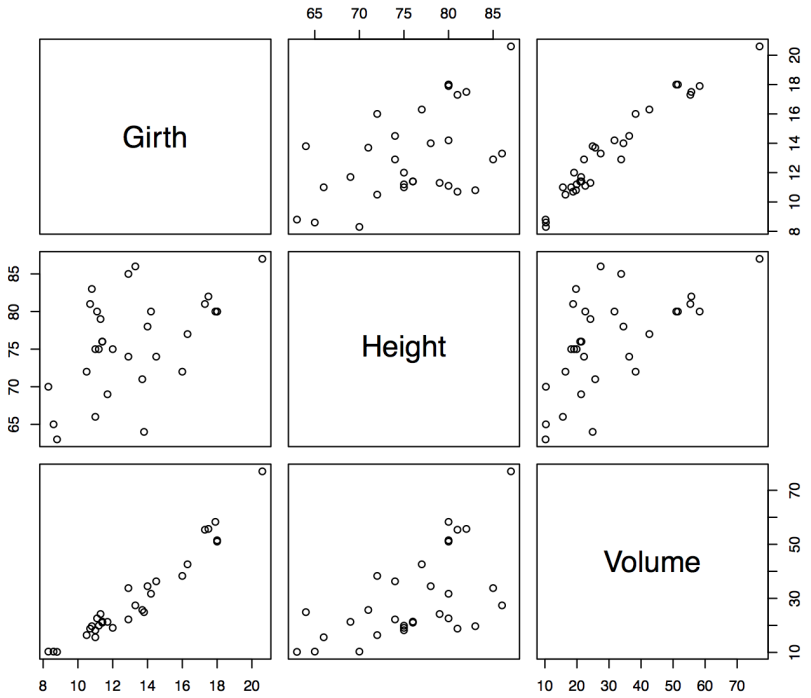

Figure 5: Scatter plot of Tree data

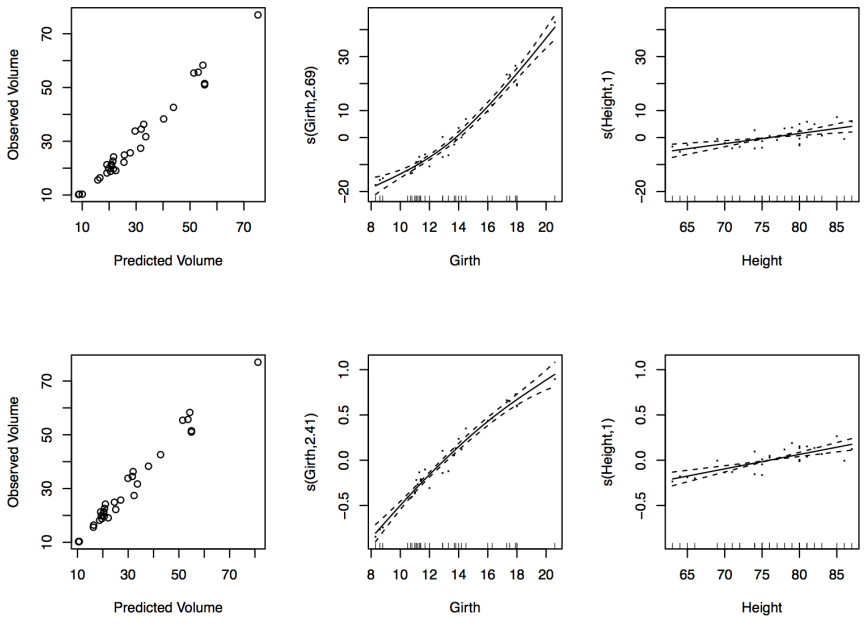

Figure 6: Additive models fitting of Tree data

and that

$$g(\mu_i) = \alpha + \sum_{j=1}^{p} f_j(x_{ij})$$

where $f_1, \ldots, f_p$ are smooth unknown functions.

4.2 Key Ideas

- GAM extend the additive models framework in the same way GLM extend the linear model framework
- Approximate the unknown functions using splines basis

$$f_j(x_j) \approx f_{K_j}(x_j) = \sum_{k=1}^{K_j} \beta_{jk} b_k(x_j) = \boldsymbol{\beta}_j^T \mathbf{b}_j(x_j)$$

- Consider penalties of the form

$$\int \{f''_{K_j}(x)\}^2 \mathrm{d}x = \boldsymbol{\beta}_j^T \mathbf{P}_j \boldsymbol{\beta}_j$$

- Estimate $\boldsymbol{\beta}$ by penalized MLE. Defining $D(\boldsymbol{\beta}) = 2\ell_{\max} - \ell(\boldsymbol{\beta})$ ($\ell_{\max}$ is saturated log likelihood) we get $\hat{\boldsymbol{\beta}}$ by minimizing the penalized deviance

$$\hat{\boldsymbol{\beta}} = \operatorname{argmin}\left\{ D(\boldsymbol{\beta}) + \boldsymbol{\beta}^T \mathbf{P}(\boldsymbol{\lambda}) \boldsymbol{\beta} \right\}$$

- Estimation of $\lambda_1, \ldots, \lambda_p$ by generalizations of CV, GCV, etc.

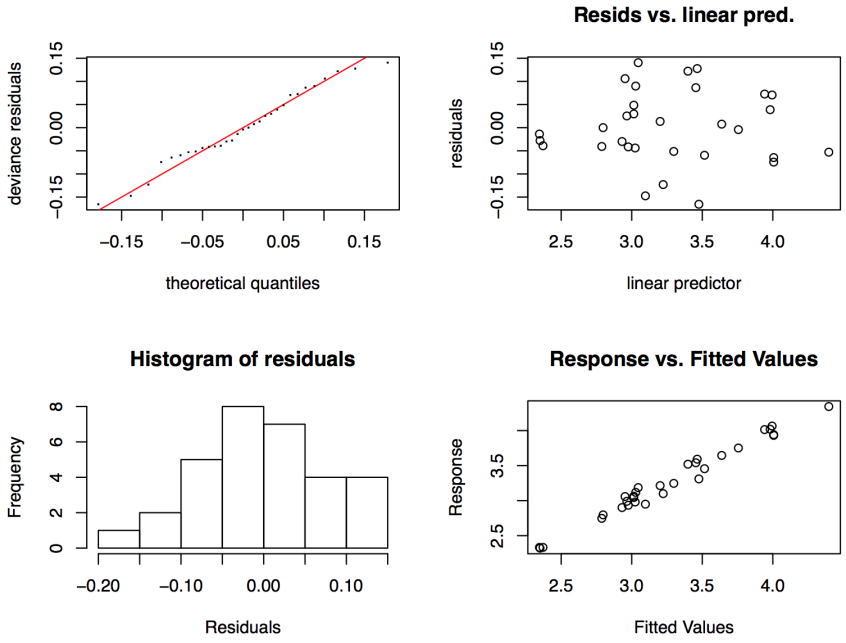

Figure 7: Model check of Tree data (log additive model)

4.3 Computational Aspect

- Penalized Iteratively Reweighted Least Squares (IRLS) algorithm: Newton based pseudo responses z_i and w_i used as in the IRLS for GLM.
- Initialize $\hat{\boldsymbol{\mu}}$ and iterate until convergence
  1. Compute z_i and w_i from $\hat{\mu}_i$ as for any GLM.
  2. Compute a new estimate of $\boldsymbol{\beta}$ as

$$\hat{\boldsymbol{\beta}} = \text{argmin} \sum_{i=1}^{n} w_i (z_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \boldsymbol{\beta}^T \mathbf{P}(\lambda) \boldsymbol{\beta}$$

  and revise estimate $\hat{\boldsymbol{\mu}}$

- Effective degrees of freedom:

$$\mathbf{S}_\lambda = \{\mathbf{X}^T \mathbf{W} \mathbf{X} + \mathbf{P}(\boldsymbol{\lambda})\}^{-1} \mathbf{X}^T \mathbf{W} \mathbf{X}$$

- GCV generalization:

$$GCV(\boldsymbol{\lambda}) = \frac{n D(\hat{\boldsymbol{\beta}}_\lambda)}{\{n - \text{tr}(\mathbf{S}_\lambda)\}^2}$$

4.4 Statistical Inference

- Bayesian confidence intervals widely used for splines! It is default in `mgcv` package in R.
- In particular, we have the Bayesian approximation

$$\boldsymbol{\beta} | \mathbf{y} \sim N(\hat{\boldsymbol{\beta}}, (\mathbf{X}^T \mathbf{W} \mathbf{X} + \mathbf{P})^{-1} \phi)$$

- The scale parameter can be estimated by

$$\hat{\phi} = \frac{1}{n - \text{tr}(\mathbf{S}_\lambda)} \sum_{i=1}^{n} w_i (z_i - \mathbf{x}_i^T \boldsymbol{\beta})^2$$

  where
- If we have partial linear parts in the model i.e.

$$g(\mathbb{E}[Y | \mathbf{X} = \mathbf{x}]) = \sum_{j=1}^{p_1} f_j(x_{ij}) + \mathbf{x}_{[2]}^T \alpha,$$

  where we treat in $\mathbf{x}^T = (\mathbf{x}_{[1]}^T, \mathbf{x}_{[2]}^T)$, $\mathbf{x}_{[1]}^T$ is the non-linear part, and $\mathbf{x}_{[2]}^T$ is the linear part. then under regularity conditions

$$\sqrt{n}(\hat{\alpha} - \alpha) \xrightarrow[n \to \infty]{\mathcal{D}} N(\mathbf{0}, \mathbf{V}).$$

  So we can see that even though we estimate some components with non-parametric techniques, the parametric parts converge at usual $O(\sqrt{n})$ rate

4.5 An Example (Air Pollution Data)

The air pollution data are from Peng and Welty (2004) and are contained in a data frame `chicago`. The response of interest is the daily death rate in Chicago, `death`, over a number of years. Possible explanatory variables for the observed death rate are levels of ozone, `o3median`, levels of sulpher dioxide, `so2median`, mean daily temperature, `tmpd`, and levels of particulate matter, `pm10median` (as generated by diesel exhaust, for example). In addition to these air quality variables, the underlying death rate tends to vary with `time` (in particular throughout the year), for reasons having little or nothing to do with air quality. The data set consists of 5114 observations. Figure 8 shows the histogram and time series plot for first 200 death rate observations.

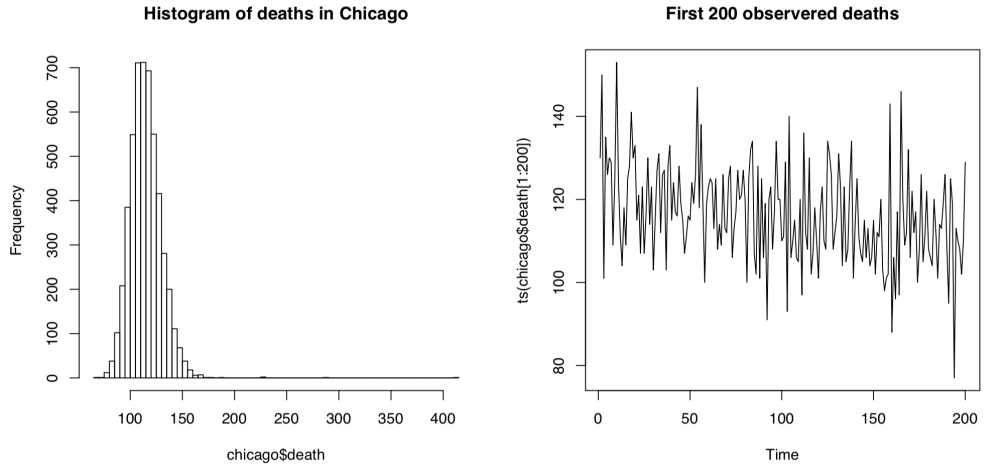

Figure 8: Air Pollution data

Consider the following Poisson regression model with canonical link, where all predictors are estimated using spline smoothing

$$\log\{\mathbb{E}[\texttt{death}_i]\} = f_1(\texttt{time}_i) + f_2(\texttt{pm10median}_i) + f_3(\texttt{ao2median}_i) + f_4(\texttt{o3median}_i) + f_5(\texttt{tmpd}_i)$$

The estimated smooths are shown in Figure 9, and indicts a problem with the distribution of `pm10median` values, in particular, which might be expected to cause leverage problems. Figure 10 gives the estimate of the smooth of `time` with (top) and without partial residuals (bottom).

Check the model using the command `gam.check`, result is shown in Figure 11. For Possion data with moderately high means, the distribution of the standardized residuals should be quite close to normal, so that the QQ-plot is obviously problematic, there is a large deviance at the tail. As all the plots make clear, there are a few gross outliers that are very problematic in this fit.

After a more detailed examination of the data, it shows that the highest temperatures in the temperature record `tmpd` were recorded in the few days preceding the high mortalities, when

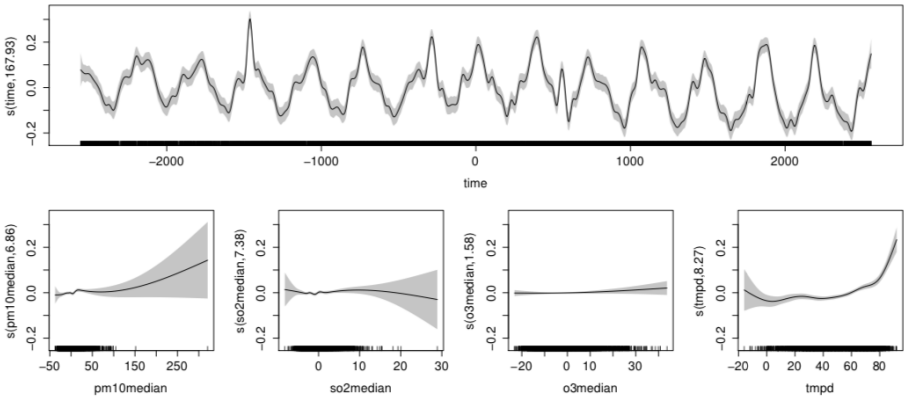

Figure 9: GAM fitting of Air Pollution data

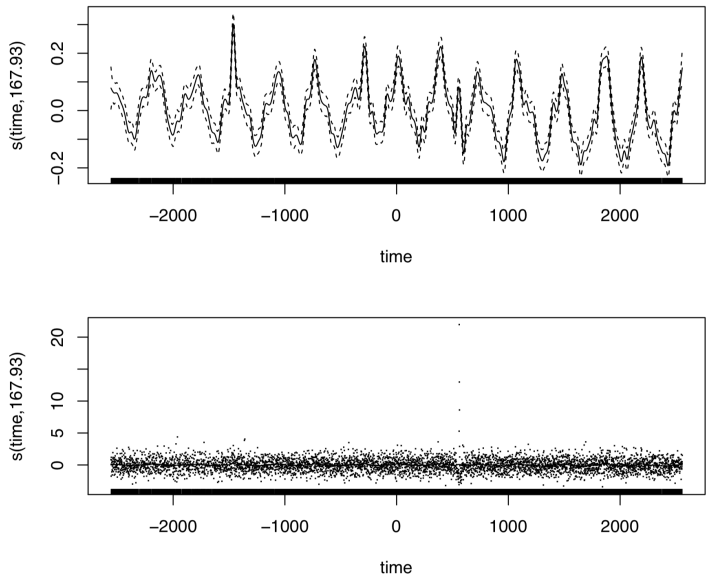

Figure 10: Time smooth estimation (Air pollution data)

there were also high ozone levels recorded. This suggests that the effects of temperature and pollution levels have lag on the mortalities, so it might be better to predict it other than only on the day itself. Besides, such model including lag might be more sensible on biological/medical aspects: the pollution levels and temperatures recorded in the date are not high enough to cause immediate disease and mortality. It seems more plausible that any effects should take some time to manifest themselves via aggravation of existing medical conditions.

Therefore, the model we suggested here is definite not the best model we can try. This example only try to give an idea how GAM work in real data[3].

[3]check S. Wood (2017). "Generalized Additive Models" *Chapters 7.4* for more models one could try.

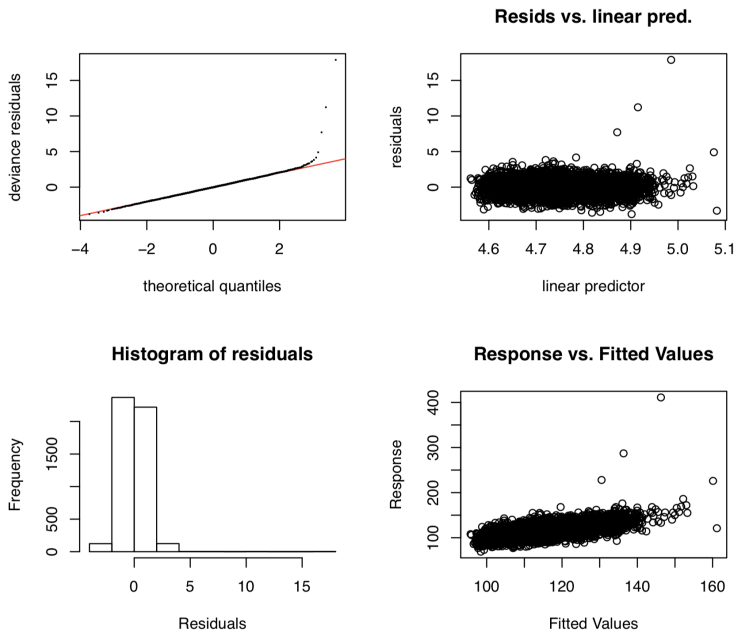

Figure 11: GAM fitting residuals check (Air pollution data)

4.6 Curse of Dimensionality

Why do we only consider univariate smooth functions? Figure 12 in "Elements of Statistical Learning" by Hastie, Tibshirani and Friedman (2008) can give us the idea. In a nutshell, we need to much more data points to make a fair "local" method estimate in high dimensions.

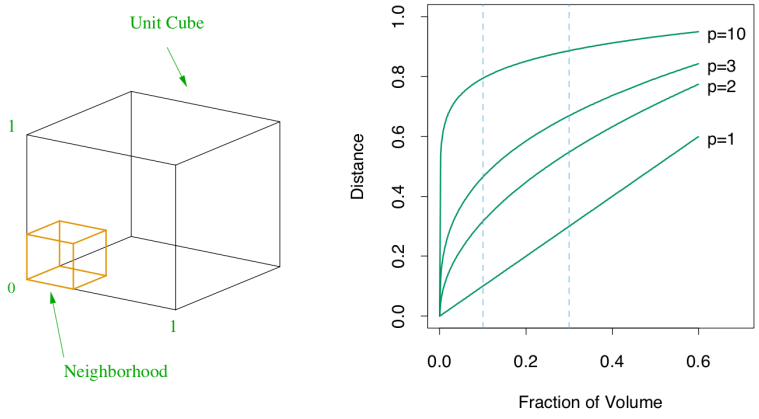

Figure 12: Curse of Dimensionality

The figure on the right shows the side-length of the sub-cube needed to capture a fraction r of the volume of the data, for different dimensions p. If we want our local estimate based on r percent of the space, then our estimation will based on the local subcell with edge equals to $r^{1/p}$.

For example, in one dimension where $p = 1$, if we want to make local average estimation each time based on 10% of the whole coordinate, then we need to compute based on data points exactly in $r^{1/p} = 0.1^{1/1} = 10\%$ of the range. When $p = 2$, it goes to $r^{1/p} = 0.1^{1/2} = 32\%$ of the coordinate range. If $p = 10$, it becomes $r^{1/p} = 0.1^{1/10} = 80\%$, which means to capture 10% of the data to form a local average, we must cover 80% of the range of each input variable. it is no longer a "local" problem. And we can also see, even with reducing r dramatically doesn't help much, as the fewer observations we average, the higher is the variance of our fit.

5 Remarks

- There are also kernel smoothing variants to the GAM framework. Splines are more popular probably because of speed and good implementations such as the `mgcv` R package.
- We did not discuss how to choose the number of parameters and knots in the splines. There are no widely accepted theoretical guidelines.
- Tuning parameters are critical and much effort has been devoted to this since the introduction of GAM in the late 80's.
- A different extension of GLM is to consider

  $$g(\mathbb{E}[Y|\mathbf{X} = \mathbf{x}]) = \mathbf{x}^T\boldsymbol{\beta},$$

  where the link function is left unspecified and is estimated nonparametrically. These models are called Single Index Models.

6 Key Takeaways

- We extended the GLM machinery to include smooth unspecified dependence on covariates
- Spline smoothing as a general approach that allows us to introduce nonparametric approach
- Key idea is introduce linear combination of spline basis functions and roughness penalty functions.
- Ideas for smoothing parameter selection in the univariate case are straightforwardly extended at the expense of an increases computational cost
- Bayesian approach can be useful for statistical inference with splines!