Generalized Additive Models

Simon Wood
School of Mathematics, University of Bristol, U.K.

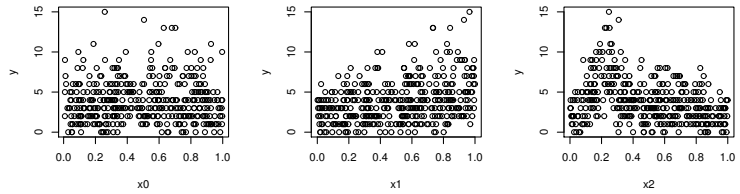
The model

- Response, y_i, predictors x_{ji}, model

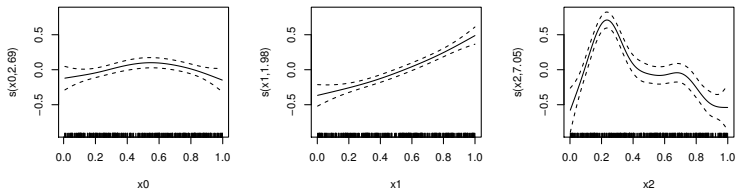
$$y_i \underset{\text{ind.}}{\sim} \pi(\mu_i, \boldsymbol{\theta}) \text{ where } g(\mu_i) = \mathbf{A}_i\boldsymbol{\gamma} + \sum_j f_j(x_{ji}).$$

  - π is a distribution: location parameter μ and other parameters $\boldsymbol{\theta}$.
  - The f_j are *smooth functions* to be estimated.
  - $\mathbf{A}$ is a known model matrix with associated parameters $\boldsymbol{\gamma}$ to be estimated.
  - g is a known *link function* (e.g. identity or log).
- If π is an exponential family distribution then this is a GLM with linear predictor dependent on smooth functions of predictors.

Example: Poisson regression

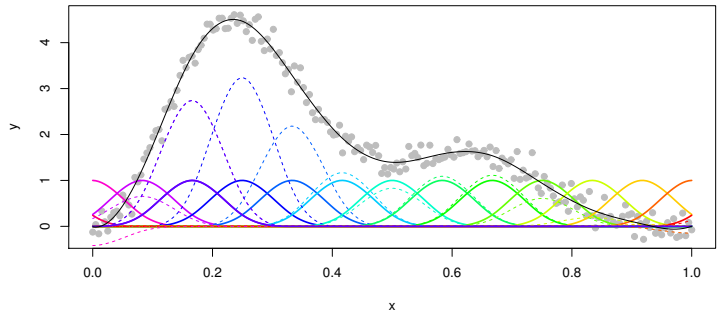


- $y_i \sim \text{Poi}(\mu_i)$ where $\log(\mu_i) = f_0(x_{0i}) + f_1(x_{1i}) + f_2(x_{2i})$.
- `gam(y~s(x0)+s(x1)+s(x2),family=poisson())`



Model representation and estimation

- Without $\sum f_j(x_{ji})$ the model is a standard regression model: use maximum likelihood estimation via Newton's method.
- With $\sum f_j(x_{ji})$ there are two problems:
  1. How to represent the f_j for estimation.
  2. How to control and estimate the degree of smoothness for the f_j.
- For 1 use a basis expansion $f_j(x) = \sum_k \beta_{jk} b_{jk}(x)$. $b_{jk}(x)$ is a known *basis function*, β_{jk} a coefficient to estimate.



Model representation with basis

- The basis expansions for the f_j turn the model into

$$y_i \underset{\text{ind.}}{\sim} \pi(\mu_i, \boldsymbol{\theta}) \text{ where } g(\mu_i) = \mathbf{X}_i\boldsymbol{\beta},$$

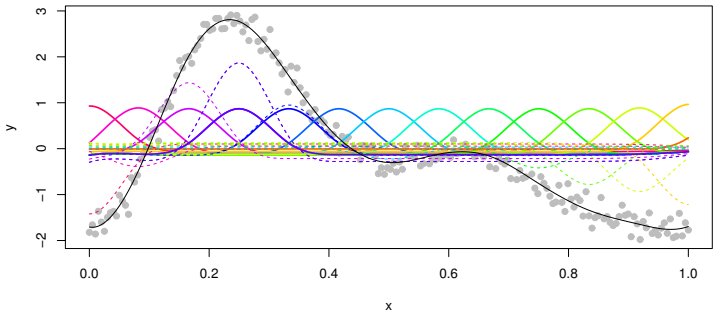
$\boldsymbol{\beta}^\mathsf{T} = (\boldsymbol{\gamma}^\mathsf{T}, \boldsymbol{\beta}_1^\mathsf{T}, \boldsymbol{\beta}_2^\mathsf{T} \ldots)$ and

$$\mathbf{X} = \begin{bmatrix} A_{11} & A_{12} & \cdots & b_{11}(x_{11}) & b_{12}(x_{11}) & \cdots & b_{21}(x_{21}) & \cdots \\ A_{21} & A_{22} & \cdots & b_{11}(x_{12}) & b_{12}(x_{12}) & \cdots & b_{21}(x_{22}) & \cdots \\ \cdot & \cdot & \cdots & \cdot & \cdot & \cdots & \cdot & \cdots \\ \cdot & \cdot & \cdots & \cdot & \cdot & \cdots & \cdot & \cdots \end{bmatrix}$$

- If π is an exponential family distribution this is just a richly parameterized GLM.

Identifiability

- One nuisance: the f_j in $\sum_j f_j(x_{ji})$ are only identifiable to within an additive constant.
- Impose identifiability constraints $\sum_i f_j(x_{ji}) = 0$, for all j.
- Conveniently handled by absorbing into the basis (modifies basis functions and loses one, but easily automated)...



Controlling smoothness

- We could control smoothness via the number of basis functions, but this is computationally awkward to optimize.
- Instead define a smoothing penalty to impose in fitting, e.g.

$$\int f_j''(x)^2 dx$$

- Given $f_j(x) = \boldsymbol{\beta}_j^\mathsf{T}\mathbf{b}(x)$ where $\mathbf{b}(x)^\mathsf{T} = (b_{j1}(x), b_{j2}(x), \ldots)$ then $f_j(x)'' = \boldsymbol{\beta}_j^\mathsf{T}\mathbf{b}''(x)$ so that, by definition of $\boldsymbol{\mathcal{S}}_j$,

$$\int f_j''(x)^2 dx = \int \boldsymbol{\beta}_j^\mathsf{T}\mathbf{b}''(x)\mathbf{b}''(x)^\mathsf{T}\boldsymbol{\beta}_j dx = \boldsymbol{\beta}_j^\mathsf{T}\boldsymbol{\mathcal{S}}_j\boldsymbol{\beta}_j.$$

- Penalty is 0 for linear functions of x ($\boldsymbol{\mathcal{S}}_j$ rank 2 deficient).
- So f_j is represented by a basis and a quadratic penalty.

Penalized model fitting

- $l(\boldsymbol{\beta})$ is the log likelihood, l_{sat} the saturated log likelihood.
- Let the model *deviance* be $D(\boldsymbol{\beta}) = 2(l_{\text{sat}} - l(\boldsymbol{\beta}))$.
- For notational convenience let $\mathbf{S}_j$ is a zero padded version of $\boldsymbol{\mathcal{S}}_j$, such that $\boldsymbol{\beta}_j^\mathsf{T}\boldsymbol{\mathcal{S}}_j\boldsymbol{\beta}_j \equiv \boldsymbol{\beta}^\mathsf{T}\mathbf{S}_j\boldsymbol{\beta}$.
- Model fitting amounts to finding

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmin}}\, D(\boldsymbol{\beta}) + \sum_j \lambda_j \boldsymbol{\beta}^\mathsf{T}\mathbf{S}_j\boldsymbol{\beta}.$$

- The λ_j are *smoothing parameters* controlling the trade-off between fitting the data closely and having a smooth model.
- We'll need to select the λ_j somehow, but they allow continuous fine control of the smoothness of the f_j.

Fitting algorithm given the λ_j

- Use *Penalized Iteratively Re-weighted Least Squares* (PIRLS).
- Iteratively solve penalized linear model fitting problem

$$\tilde{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_i W_i(z_i - \mathbf{X}_i\boldsymbol{\beta})^2 + \sum_j \lambda_j \boldsymbol{\beta}^\mathsf{T}\mathbf{S}_j\boldsymbol{\beta},$$

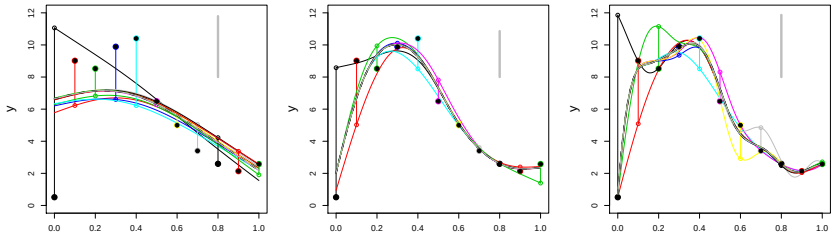
  - z_i is pseudodata depending on y_i and the previous $\hat{\mu}_i$.
  - W_i depends on $\hat{\mu}_i$ and is related to the variance of y_i.
  - Exact forms depend on π and the link function g.
- $\tilde{\boldsymbol{\beta}}$ eventually converges on required $\hat{\boldsymbol{\beta}}$.
- Notice how each step is fitting a *working linear model*.

Degrees of freedom

- $\dim(\boldsymbol{\beta})$ is now only a good measure of model degree of freedom if all the smoothing parameters are zero!
- e.g. if all $\lambda_j \to \infty$ then each smooth is a linear function of x with just 2 degrees of freedom, irrespective of $\dim(\boldsymbol{\beta})$.
- To characterize *effective degrees of freedom* consider the shrinkage of parameters by the smoothing penalties.
- At PIRLS convergence $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X} + \sum_j \lambda_j \mathbf{S}_j)^{-1}\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{z}$.
- But removing all the penalization gives $\tilde{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{z}$.
- So $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X} + \sum_j \lambda_j \mathbf{S}_j)^{-1}\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}\tilde{\boldsymbol{\beta}}$. i.e. $\hat{\boldsymbol{\beta}}$ is a shrunken version of the unpenalized $\tilde{\boldsymbol{\beta}}$, with shrinkage matrix $\mathbf{F} = (\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X} + \sum_j \lambda_j \mathbf{S}_j)^{-1}\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}$.
- $F_{ii} = \partial\hat{\beta}_i/\partial\tilde{\beta}_i$ are shrinkage factors and their sum, $\operatorname{trace}(\mathbf{F})$, is a measure of *effective degrees of freedom*.

Smoothing parameter selection

- One option is leave-one-out cross-validation.
  - Leave out each data point in turn, and then predict it using a model fitted only to the data not left out.
  - The best model is the one with lowest average error in predicting the left out data



- Each panel shows predictions of data left out of spline fits - the prediction error and the corresponding spline have the same colour. The grey bar is the mean error.
- Left is too smooth, right is too wiggly, middle is better.

Generalized cross validation

- The average leave-one-out cross validation error can actually be computed from a single fit to all the data!
- But it lacks some invariance properties that might be desirable. It is also awkward to optimize for multiple smoothing parameters.
- Generalized cross validation removes theses problems. For the Gaussian-identity link case, the averaged error becomes

$$n\sum_i (y_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})^2/(n - \operatorname{trace}(\mathbf{F}))^2$$

  — residual variance per residual degree of freedom.
- In the general non Gaussian case this becomes

$$nD(\hat{\boldsymbol{\beta}})/(n - \operatorname{trace}(\mathbf{F}))^2$$

- Prediction error criteria like GCV are not the only possibility ...

Bayesian smoothing

- Why smooth? Because we think the truth is more likely to be smooth than wiggly.
- We could formalize this belief with a prior on wiggliness

$$\pi(\boldsymbol{\beta}) \propto \exp\left(-\sum_j \lambda_j \boldsymbol{\beta}^{\mathsf{T}} \mathbf{S}_j \boldsymbol{\beta}/2\right).$$

  ...recognisable as $\boldsymbol{\beta} \sim N(\mathbf{0}, \{\sum_j \lambda_j \mathbf{S}_j\}^-)$ (improper Gaussian).
- Our model specifies the likelihood. Applying Bayes' rule

$$\boldsymbol{\beta}|\mathbf{y} \underset{n\to\infty}{\sim} N\left(\hat{\boldsymbol{\beta}}, \{\mathbf{X}^{\mathsf{T}}\mathbf{W}\mathbf{X} + \sum_j \lambda_j \mathbf{S}_j\}^{-1}\right)$$

  where $\hat{\boldsymbol{\beta}}$ is penalized MLE from earlier[1].

[1] any scale parameters absorbed in $\mathbf{W}$

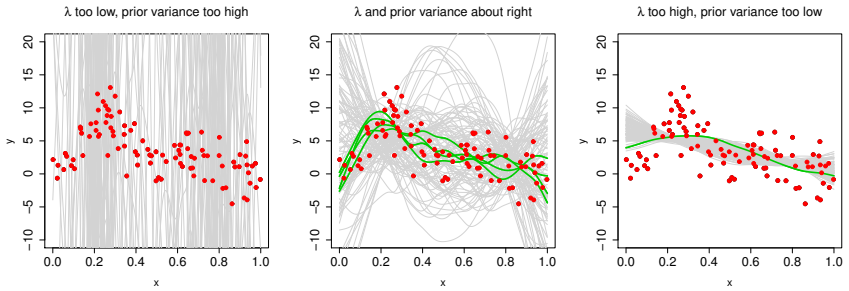
Consequences of Bayesian Model

- Smooths are Gaussian random fields!
- Can produce credible intervals for f_j — well calibrated.
- Can do inference via MCMC (e.g. `mgcv:jagam`).
- Structure is like a mixed model with Gaussian random effects
  - Can estimate as mixed model (e.g. `gamm` or `gamm4`).
- We can estimate smoothing parameters to maximize the *marginal likelihood*

$$\pi(\mathbf{y}|\boldsymbol{\lambda}, \boldsymbol{\theta}) = \int \pi(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\theta})\pi(\boldsymbol{\beta}|\boldsymbol{\lambda}) d\boldsymbol{\beta}$$

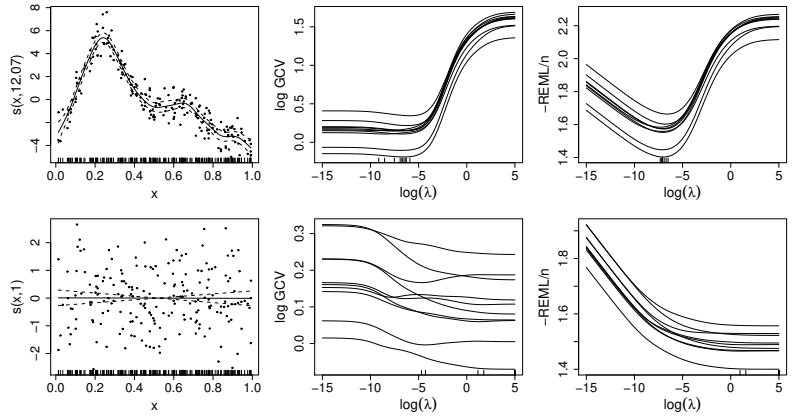
- Integral is intractable, but we can use Laplace approximation. i.e. replace integrand with exponential of second order Taylor expansion of its log about $\hat{\boldsymbol{\beta}}$. The approximation is proportional to a Gaussian density and is tractable.

How marginal likelihood smoothness selection works



1. Choose λ to maximize the average likelihood of random draws from the prior implied by λ.
2. If λ too low, then almost all draws are too variable to have high likelihood. If λ too high, then draws all underfit and have low likelihood. The right λ maximizes the proportion of draws close enough to data to give high likelihood.

Prediction error vs. likelihood $\boldsymbol{\lambda}$ estimation



1. Pictures show GCV and REML scores for different replicates from same truth.
2. Compared to REML, GCV penalizes overfit only weakly, and so is more likely to occasionally undersmooth.

Applying the λ estimation methods

- There are 2 possibilities, for both we work with $\rho = \log(\lambda)$:
  1. Apply smoothness selection to the working penalized model at each PIRLS step.
  2. Optimize GCV/REML for the model itself using a Newton method.
     - Each trial ρ requires an inner iteration to find the corresponding $\hat{\beta}$.
     - Use implicit differentiation to find $\partial\hat{\beta}/\partial\rho$, so that derivatives required by outer Newton method can be computed.
- Option 1 is easier to code and adapt to big data situations.
- Option 2 gives better convergence guarantees.

Model selection

- We need means for comparing models/deciding what terms to include. In many cases the gold standard might be prediction of hold-out data, but other approaches are also helpful.

1. Null space penalization: add a penalty (and smoothing parameter) for each f_j which allows it to be penalized to zero during smoothing parameter estimation.
2. P-values: by 'inverting' the Bayesian CI for f_j, compute a p-value for $H_0 : f_j = 0$.
3. Akaike's Information Criterion: this becomes

$$-2l(\hat{\beta}) + 2 \times (\text{Effective Degrees of Freedom})$$

   - Actually the derivation arrives at the EDF as $\text{trace}(\mathbf{V}_\beta \mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X})$ where $\mathbf{V}_\beta$ is the Bayesian covariance matrix for β.
   - Decent performance of the AIC requires that we correct $\mathbf{V}_\beta$ for smoothing parameter uncertainty, but a simple correction seems to suffice.

Model checking

- As for any regression model, examine standardised residuals to check for violations of mean-variance and independence assumptions.
- As for any regression model, details of the distribution beyond these properties are less important (consider quasi-likelihood theory), but violation may have some influence on smoothness selection.
- The basis dimension used for each smooth should be checked. Is it overly restrictive?
  - EDF close to its upper limit is suspicious.
  - Simple informal randomization tests can be used to try and detect residual pattern with respect to x_j which might indicate that the basis for f_j is too small.
- See `gam.check` in `mgcv` to get started.

Extensions

- Simple independent Gaussian random effects can be included as 0-dimensional smooths, using same methods.
- $y_i \underset{\text{ind.}}{\sim} \pi(\mu_i, \boldsymbol{\theta})$ does not cover all interesting regression models!
- $\mathbf{y} \sim \pi(f_1, f_2, f_3, \ldots)$ is much more general, and for regular enough π general methods are possible. This covers e.g. multivariate responses and Cox Proportional Hazards models.
- $y_i \underset{\text{ind.}}{\sim} \pi(\theta_{1i}, \theta_{2i}, \ldots)$ where $g_j(\theta_{ji}) = \sum f_j$. Refereed to as *distributional regression* or GAMs for location scale and shape (GAMLSS).
- Models can depend on linear functional of smooth functions: e.g. scalar on function regression.

Summary

- GAMs allow a response to depend on smooth functions of predictor variables.
- The smooth functions are represented using a basis expansion and quadratic smoothing penalty.
- Basis coefficients are estimated by penalized MLE.
- Penalization implies a notion of effective degrees of freedom.
- Cross validation can be used to select the degree of penalization.
- The quadratic penalties are equivalent to Gaussian priors on the coefficients, providing a Bayesian interpretation.
- The Bayesian approach provides useful confidence intervals, and an alternative approach to smoothness estimation via marginal likelihood maximization.
- Model selection and checking are similar to any regression model (but check the basis dimension).