Generalized Additive Models

Simon Wood
Mathematical Sciences, University of Bath, U.K.

Introduction

- We have seen how to
    1. turn model $y_i = f(x_i) + \epsilon_i$ into $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ and a wiggliness penalty $\boldsymbol{\beta}^\mathsf{T}\mathbf{S}\boldsymbol{\beta}$.
    2. estimate $\boldsymbol{\beta}$ given $\boldsymbol{\lambda}$ as $\hat{\boldsymbol{\beta}} = \arg\ \min_\beta \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\boldsymbol{\beta}^\mathsf{T}\mathbf{S}\boldsymbol{\beta}$.
    3. estimate $\boldsymbol{\lambda}$ by GCV, AIC, REML etc.
    4. use $\boldsymbol{\beta}|\boldsymbol{\lambda} \sim N(\hat{\boldsymbol{\beta}}, (\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{S})^{-1}\sigma^2)$ for inference.
- . . . all this can be extended to models with multiple smooth terms, for exponential family response data . . .

Additive Models

- Consider the model

$$y_i = \mathbf{A}_i\boldsymbol{\theta} + \sum_j f_j(x_{ji}) + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2)$$

  - $\mathbf{A}_i$ is the i^{th} row of the model matrix for any parametric terms, with parameter vector $\boldsymbol{\theta}$. Assume it includes an intercept.
  - f_j is a smooth function of covariate x_j, which may vector valued.
- The f_j are confounded via the intercept, so that the model is only estimable under identifiability constraints on the f_j.
- The best constraints are $\sum_i f_j(x_i) = 0 \;\; \forall \; j$.
- If $\mathbf{f} = [f(x_1), f(x_2), \ldots]$ then the constraint is $\mathbf{1}^{\mathsf{T}}\mathbf{f} = 0$, i.e. $\mathbf{f}$ is orthogonal to the intercept. This results in minimum width CIs for the constrained f_j.[1]

[1] this fact is not often appreciated in the literature

Representing the model

- Choose a basis and penalty for each f_j.
- Let the model matrix for f_j be $\mathbf{X}$ and let $\lambda\boldsymbol{\beta}^\mathsf{T}\mathbf{S}\boldsymbol{\beta}$ be the penalty (more generally $\sum_j \lambda_j \boldsymbol{\beta}^\mathsf{T}\mathbf{S}_j\boldsymbol{\beta}$).
- Reparameterize to absorb the constraint $\mathbf{1}^\mathsf{T}\mathbf{X} = 0$ as follows
  1. Form QR decompostion
     $$\mathbf{Q}\begin{bmatrix} \mathbf{R} \\ \mathbf{0} \end{bmatrix} = \mathbf{X}^\mathsf{T}\mathbf{1} \text{ and partition } \mathbf{Q} = \begin{bmatrix} \mathbf{Y} & \mathbf{Z} \end{bmatrix}$$
  2. Setting $\boldsymbol{\beta} = \mathbf{Z}\boldsymbol{\beta}'$ then
     $$\mathbf{1}^\mathsf{T}\mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} \mathbf{R} & \mathbf{0} \end{bmatrix}\begin{bmatrix} \mathbf{Y}^\mathsf{T} \\ \mathbf{Z}^\mathsf{T} \end{bmatrix}\mathbf{Z}\boldsymbol{\beta}' = 0.$$
  3. So set $\mathbf{X}^{[j]} = \mathbf{XZ}$ and $\mathbf{S}_j = \mathbf{Z}^\mathsf{T}\mathbf{SZ}$...the constrained model and penalty matrices for f_j.

The estimable AM

- Now $y_i = \mathbf{A}_i\boldsymbol{\theta} + \sum_j f_j(x_{ji}) + \epsilon_i$ becomes $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where

$$\mathbf{X} = [\mathbf{A} : \mathbf{X}^{[1]} : \mathbf{X}^{[2]} : \cdots]$$

  and $\boldsymbol{\beta}$ contains $\boldsymbol{\theta}$ followed by the basis coefficients for the f_j.
- After suitable padding of the $\mathbf{S}_j$ with zeroes the penalty becomes $\sum_j \lambda_j \boldsymbol{\beta}^\mathsf{T}\mathbf{S}_j\boldsymbol{\beta}$.
- Now $\hat{\boldsymbol{\beta}} = \text{arg } \min_\beta \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sum_j \lambda_j \boldsymbol{\beta}^\mathsf{T}\mathbf{S}_j\boldsymbol{\beta}$.
- Again $\boldsymbol{\lambda}$ can be estimated by GCV, REML etc.

Linear functional generalization

- Occasionally we may want a model that depends on an f_j in some way other than simple evaluation. So let L_{ij} be a linear operator and consider an extended model

$$y_i = \mathbf{A}_i \boldsymbol{\theta} + \sum_j L_{ij} f_j(x_j) + \epsilon_i$$

  e.g. $L_{ij} f_j = \int k_i(x) f_j(x) dx$ (k_i known), or just $L_{ij} f_j = f(x_{ji})$.
- Dropping j for now, we can discretize $L_i f(x) \simeq \sum_k \tilde{L}_{ik} f(x_k)$.
- So $L_i f(x) \simeq \sum_k \tilde{L}_{ik} \tilde{\mathbf{X}}_k \boldsymbol{\beta}$, where $\tilde{\mathbf{X}}_k$ is k^{th} row of model matrix evaluating $f(x)$ at the points x_k.
- Then the model matrix for $L_i f(x)$ is $\tilde{\mathbf{L}} \tilde{\mathbf{X}}$. The penalties are just those for f.
- Hence the extended model can be written in the same general form as the simple AM.

Generalized Additive Models

- Generalizing again, we have

$$g(\mu_i) = \mathbf{A}_i\boldsymbol{\theta} + \sum_j L_{ij} f_j(x_j), \quad y_i \sim \text{EF}(\mu_i, \phi)$$

  where g is a known smooth monotonic link function and EF an exponential family distribution.
- Set up model matrix and penalties as before.
- Estimate $\boldsymbol{\beta}$ by penalized MLE. Defining the *Deviance*. $D(\boldsymbol{\beta}) = 2\{l_{\text{max}} - l(\boldsymbol{\beta})\}$ (l_{max} is saturated log likelihood)...

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} D(\boldsymbol{\beta}) + \sum_j \lambda_j \boldsymbol{\beta}^{\mathsf{T}} \mathbf{S}_j \boldsymbol{\beta}$$

- $\boldsymbol{\lambda}$ estimation is by generalizations of GCV, REML etc.

GAM computation: $\hat{\boldsymbol{\beta}}|\mathbf{y}$

- Penalized likelihood maximization is by Penalized IRLS.
- Initialize $\hat{\boldsymbol{\eta}} = g(\mathbf{y})$ and iterate the following to convergence.
  1. Compute z_i and w_i from $\hat{\eta}_i$ (and $\hat{\mu}_i$) as for any GLM.
  2. Compute a revised $\boldsymbol{\beta}$ estimate

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_i w_i(z_i - \mathbf{X}_i\boldsymbol{\beta})^2 + \sum \lambda_j \boldsymbol{\beta}^\mathsf{T}\mathbf{S}_j\boldsymbol{\beta}$$

  and hence revised estimates $\hat{\boldsymbol{\eta}}$ and $\hat{\boldsymbol{\mu}}$.
- Newton based versions of w_i and z_i are best here, as it makes $\boldsymbol{\lambda}$ estimation easier.

EDF, $\boldsymbol{\beta}|\mathbf{y}$ and $\hat{\phi}$

- Let $\mathbf{S} = \sum_j \lambda_j \mathbf{S}_j$ and $\mathbf{W} = \text{diag}\{E(w_i)\}$.
- The Effective Degrees of Freedom matrix becomes

$$\mathbf{F} = (\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X} + \mathbf{S})^{-1}\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}$$

- Then the EDF is $\text{tr}(\mathbf{F})$. EDFs for individual smooths are found by summing the F_{ii} values for their coefficients.
- In the $n \to \infty$ limit

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

- The scale parameter can be estimated by

$$\hat{\phi} = \sum_i w_i(z_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})^2/\{n - \text{tr}(\mathbf{F})\}.$$

$\boldsymbol{\lambda}$ estimation

- There are 2 basic computational strategies for $\boldsymbol{\lambda}$ selection.
  1. Single iteration schemes estimate $\boldsymbol{\lambda}$ at each PIRLS iteration step, by applying GCV, REML or whatever to the working penalized linear model. This approach need not converge.
  2. Nested iteration, defines a $\boldsymbol{\lambda}$ selection criterion in terms of the model deviance and optimizes it directly. Each evaluation of the criterion requires an 'inner' PIRLS to obtain $\hat{\boldsymbol{\beta}}_\lambda$. This converges, since a properly defined function of $\boldsymbol{\lambda}$ is optimized.
- The second option is usually preferable on grounds of reliability, but the first option can be made very memory efficient with very large datasets.
- The first option simply uses the smoothness selection criteria for the linear model case, but the second requires that these be extended...

Deviance based $\boldsymbol{\lambda}$ selection criteria

- Mallows' C_p/ UBRE generalizes to

$$\mathcal{V}_a = D(\hat{\boldsymbol{\beta}}_\lambda) + 2\phi\text{tr}(\mathbf{F}_\lambda)$$

- GCV generalizes to

$$\mathcal{V}_g = nD(\hat{\boldsymbol{\beta}}_\lambda)/\{n - \text{tr}(\mathbf{F})\}^2$$

- Laplace approximate (negative twice) REML is

$$\mathcal{V}_r = \frac{D(\hat{\boldsymbol{\beta}}) + \hat{\boldsymbol{\beta}}^\mathsf{T}\mathbf{S}\hat{\boldsymbol{\beta}}}{\phi} - 2l_s(\phi) + (\log|\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X} + \mathbf{S}| - \log|\mathbf{S}|_+) - M_p\log(2\pi\phi).$$

Nested iteration computational strategy

- Optimization wrt $\boldsymbol{\rho} = \log \boldsymbol{\lambda}$ is by Newton's method, using analytic derivatives.
- For each trial $\boldsymbol{\lambda}$ used by Newton's method...
  1. Re-parameterize for maximum numerical stability in computing $\hat{\boldsymbol{\beta}}$ and terms like $\log |\mathbf{S}|_+$.
  2. Compute $\hat{\boldsymbol{\beta}}$ by PIRLS (full Newton version).
  3. Calculate derivatives of $\hat{\boldsymbol{\beta}}$ wrt $\boldsymbol{\rho}$ by implicit differentiation.
  4. Evaluate the $\boldsymbol{\lambda}$ selection criterion and its derivatives wrt $\boldsymbol{\rho}$
- ...after which all the ingredients are in place for Newton's method to propose a new $\boldsymbol{\lambda}$ value.
- As usual with Newton's method, some step halving may be needed, and the Hessian will have to be peturbed if it is not positive definite.

One last generalization: GAMM

- A generalized additive mixed model has the form

$$g(\mu_i) = \mathbf{A}_i\boldsymbol{\theta} + \sum_j L_{ij} f_j(x_j) + \mathbf{Z}\mathbf{b}, \quad \mathbf{b} \sim N(\mathbf{0}, \boldsymbol{\psi}), \quad y_i \sim \text{EF}(\mu_i, \phi)$$

- ... actually this is not much different to a GAM. The random effects term $\mathbf{Zb}$ is just like a smooth with penalty $\mathbf{b}^\mathsf{T}\boldsymbol{\psi}^{-1}\mathbf{b}$.
- If $\boldsymbol{\psi}^{-1}$ can be written in the form $\sum_k \lambda_k \mathbf{S}_k$ then the GAMM can be treated *exactly* like a GAM. (`gam`).
- Alternatively, using the mixed model representation of the smooths, the GAMM can be written in standard GLMM form and estimated as a GLMM. (`gamm`/`gamm4`).
- The latter option is often preferable when there are many random effects, and the former when there are fewer.

Inference for GAMMs

- For many GAMMs we are interested in making inferences about the smooths, but are using the other random effects to model 'nuisance' randomness.
- In this case we often want to use the large sample result

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

  for inference, where $\tilde{\mathbf{W}}^{-1} = \mathbf{W}^{-1} + \mathbf{Z}^\mathsf{T}\psi\mathbf{Z}/\phi$.
- The point here is that inference about the smooths and other fixed effects takes account of the uncertainty induced by both random effects and residual variability.
- Note that $\tilde{\mathbf{W}}$ usually has exploitable sparse structure, so that its inverse is not too expensive.

Summary

- A GAM is simply a GLM in which the linear predictor partly depends linearly on some unknown smooth functions.
- GAMs are estimated by a penalized version of the method used to fit GLMs.
- An extra criterion has to be optimized to find the smoothing parameters.
- A GAMM is simply a GLMM in which the linear predictor partly depends linearly on some unknown smooth functions.
- From the mixed model representation of smooths, GAMMs can be estimated as GAMs or GLMMs.
- Inference for GAMs and GAMMs is really Bayesian, but without any need to simulate.