Additive Smooth Models: basic theory

Simon N. Wood

University of Bath

Additive smooth models

- Consider a univariate response y and corresponding predictors x_1, x_2 ,x_3 ... (possibly vectors).
- An additive smooth model has a structure like

$$y_i = \mathbf{X}_i\boldsymbol{\beta} + f_1(x_{1i}) + f_2(x_{2i}) + \cdots + \epsilon_i$$

  — the f_j are smooth functions and $\mathbf{X}_i\boldsymbol{\beta}$ is the linear predictor for any *strictly parametric model components*.
- The ϵ_i are 0 mean r.v.s, with variance σ^2. Normality is assumed if CIs or H_0 tests are required.
- The additive smooth structure offers a nice balance of flexibility and structure.

ASM representation

- How can the f_j be estimated? Use a basis expansion for each smooth term. The model becomes

$$y_i = \mathbf{X}_i\boldsymbol{\beta} + \sum_{k=1}^{K_1} \beta^*_{1k} b_{1k}(x_{1i}) + \sum_{k=1}^{K_2} \beta^*_{2k} b_{2k}(x_{2i}) + \cdots + \epsilon_i$$

- Ignorance of parametric form of f_j OK, if K_j large enough and *basis functions*, $b_{jk}(x)$, chosen carefully.

- Only unknowns are now parameters. Absorbing $b_{jk}(x_{ji})$ into $\mathbf{X}_i$ and β^*_{jk} into $\boldsymbol{\beta}$, the model becomes

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \text{(a linear model!)}$$

Not over-fitting

- To avoid model mis-specification bias we made K_js fairly large (not too large or computational efficiency is lost). $\Rightarrow$ There is a danger of overfitting.
- To avoid overfitting penalize lack of smoothness. e.g. estimate $\boldsymbol{\beta}$ by minimizing not just RSS, but

$$\mathcal{S}(\boldsymbol{\beta}) = \text{RSS} + \sum_j \lambda_j \times [\text{wiggliness of } f_j]$$

- Smoothing parameters λ_j, assumed known for moment, while e.g. $[\text{wiggliness of } f_j] = \int f_j''(x)^2 dx$.

Model estimation

- Given basis functions, most wiggliness measures (including $\int f_j''(x)^2 dx$) can be written as $\boldsymbol{\beta}^{\mathsf{T}}\mathbf{S}_j\boldsymbol{\beta}$ where $\mathbf{S}_j$ contains known coefficients.
- f_j only identifiable to within an additive constant $\Rightarrow$ impose sum-to-zero constraints, and absorb into basis.
- Fitting objective becomes, minimize

$$\mathcal{S}(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sum \lambda_j \boldsymbol{\beta}^{\mathsf{T}}\mathbf{S}_j\boldsymbol{\beta} \text{ w.r.t. } \boldsymbol{\beta}.$$

- Resulting estimate is $\hat{\boldsymbol{\beta}} = \mathbf{B}\mathbf{y}$ where $\mathbf{B} = (\mathbf{X}^{\mathsf{T}}\mathbf{X} + \sum_j \mathbf{S}_j)^{-1}\mathbf{X}^{\mathsf{T}}$ (not for computational use!)

Inference based on $\hat{\boldsymbol{\beta}}$

- Covariance matrix of $\hat{\boldsymbol{\beta}}$ easily derived from that of $\mathbf{y}$

$$\mathbf{V}_{\hat{\boldsymbol{\beta}}} = \mathbf{B}\mathbf{B}^{\mathsf{T}}\sigma^2.$$

- If we also assume normality (i.e. $\mathbf{y} \sim N(\mathbf{0}, \mathbf{I}\sigma^2)$) then

$$\hat{\boldsymbol{\beta}} \sim N(\mathbb{E}(\hat{\boldsymbol{\beta}}), \mathbf{B}\mathbf{B}^{\mathsf{T}}\sigma^2) \tag{1}$$

- But $\mathbb{E}(\hat{\boldsymbol{\beta}}) \neq \boldsymbol{\beta}$ unless $\boldsymbol{\beta} = \mathbf{0}$ (strictly *in null space of penalty*)...so (1) sometimes useful for hypothesis testing, but probably not otherwise.

Bayesian Inference

- We penalize wiggliness because we think that a smooth truth is ‘more probable’ than a wiggly one.
- So let ‘model wiggliness’ have a negative exponential prior distribution (Wahba, ’83; Silverman ’85),

$$\pi(\boldsymbol{\beta}) \propto e^{-\frac{1}{2}\sum \lambda_j \boldsymbol{\beta}^{\mathsf{T}} \mathbf{S}_j \boldsymbol{\beta}}$$

- The model says $\mathbf{y}|\boldsymbol{\beta} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{I}\sigma^2)$, so Bayes rule

$$\Rightarrow \boldsymbol{\beta}|\mathbf{y} \sim N(\hat{\boldsymbol{\beta}}, (\mathbf{X}^{\mathsf{T}}\mathbf{X} + \sum_j \mathbf{S}_j)^{-1}\sigma^2)$$

- Credible intervals for any function of $\boldsymbol{\beta}$ follow.

Inference based on fit

- There's always an unpenalized model with fitted value behaviour that is 'close' to a penalized model with given smoothing parameters.
- Hence, conditional on smoothing parameters, the distributions of F-ratio statistics for comparison of penalized models should be approximately the distribution of the equivalent un-penalized analogues.
- So F-ratio testing can be used for model comparison, provided we can work out the approximate 'equivalent degrees of freedom' of the penalized models.

Degrees of freedom

- The point of penalizing the fit is to reduce the model's freedom to vary. How many degrees of freedom does the penalized fit have?
- Without penalization, $\tilde{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$.
- With penalization, $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{X} + \sum_j \lambda_j \mathbf{S}_j)^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$.
- So $\hat{\boldsymbol{\beta}} = \mathbf{F}\tilde{\boldsymbol{\beta}}$ where $\mathbf{F} = (\mathbf{X}^\mathsf{T}\mathbf{X} + \sum_j \lambda_j \mathbf{S}_j)^{-1}\mathbf{X}^\mathsf{T}\mathbf{X}$.
- $\mathbf{F}_{ii} = \partial\hat{\beta}_i/\partial\tilde{\beta}_i$ is a measure of the DoF of the i^{th} parameter, while $\mathsf{tr}(\mathbf{F}) \equiv$ *the whole model DoF*.
- $\hat{\sigma}^2 = \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2/(n - \mathsf{tr}(\mathbf{F}))$.

Smoothness selection

- So far everything is conditional on λ: how should λ values be chosen?
- There are two main choices:
  1. Use prediction error criteria, such as cross validation, GCV, AIC, BIC etc.
  2. Use the Bayes smoothing model to decompose each smooth into fixed effect and random effect components: the model can then be estimated by maximum likelihood or REML, with $\boldsymbol{\lambda}$ treated as variance parameters.

`mgcv::gam` does 1, while `mgcv::gamm` does 2. Concentrate on 1, here.

Ordinary cross validation

- Leave one out cross validation seeks to minimize estimated mean square prediction error to select $\boldsymbol{\lambda}$.

$$\text{minimize } V_o(\boldsymbol{\lambda}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu}_i^{[-i]})^2 \text{ w.r.t. } \boldsymbol{\lambda}$$

where $\hat{\mu}_i^{[-i]}$ is predicted y_i from fit to all data except y_i.

- If $\mathbf{A} = \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X} + \sum_j \lambda_j \mathbf{S}_j)^{-1}\mathbf{X}^\mathsf{T}$, so $\hat{\boldsymbol{\mu}} = \mathbf{A}\mathbf{y}$, then

$$V_g(\boldsymbol{\lambda}) = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{(1 - A_{ii})^2}$$

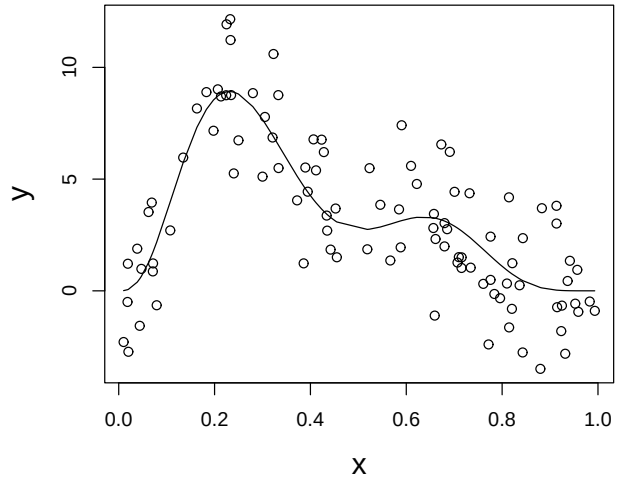
OCV not invariant

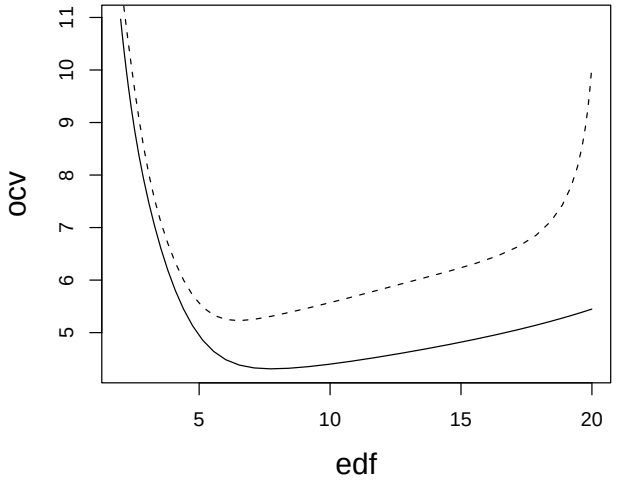
- For any $\perp$ matrix $\mathbf{Q}$, fitting objective

$$\mathcal{S}_Q(\boldsymbol{\beta}) = \|\mathbf{Q}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\|^2 + \sum \lambda_j \boldsymbol{\beta}^{\mathsf{T}} \mathbf{S}_j \boldsymbol{\beta}$$

  yields identical inferences about $\boldsymbol{\beta}$ as $\mathcal{S}(\boldsymbol{\beta})$.

- But $\mathcal{S}_Q$ and $\mathcal{S}$ yield different OCV functions!





Generalized cross validation

- If we choose $\mathbf{Q}$ so that each rotated observation counts equally in the OCV function, then we get GCV (e.g. Craven and Wahba, '79).
- The resulting GCV function,

$$V_g(\boldsymbol{\lambda}) = \frac{n\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2}{[n - \mathsf{tr}(\mathbf{A})]^2}$$

  is invariant. Note that $\mathsf{tr}(\mathbf{A}) = \mathsf{tr}(\mathbf{F})$.
- Main problem solved by `mgcv::gam` is to minimize functions like V_g efficiently w.r.t. $\boldsymbol{\lambda}$.

Other λ selection criteria

- Can increase 'cost' per DoF in GCV function

$$V_g(\boldsymbol{\lambda}) = \frac{n\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2}{[n - \gamma \mathsf{tr}(\mathbf{F})]^2} \quad \gamma \geq 1$$

- If σ^2 is known, can estimate $\mathbb{E}(\|\boldsymbol{\mu} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2)/\sigma^2$ by

$$V_u(\boldsymbol{\lambda}) = \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2/\sigma^2 + 2\mathsf{tr}(\mathbf{F})$$

which is basically Mallows' statistic, Craven and Wahba's UBRE or AIC. The penalty per DoF can be increased to give, e.g. BIC.

Generalized Additive Models

- The model is generalized to

$$g\{\mathbb{E}(y_i)\} = \mathbf{X}_i\boldsymbol{\beta} + f_1(x_{1i}) + f_2(x_{2i}) + \cdots$$

- g is a known 'link function'.
- y_i are independent and either
  1. assume y_i follow an exponential family distribution (use maximum likelihood), or
  2. assume only $\mathrm{var}(y_i) = \phi V(\mathbb{E}(y_i))$, where V is a known function (use quasi-likelihood).

GAM representation

- As for ASM, basis expansion allows model to be written as a GLM $g\{\mathbb{E}(y_i)\} = \mathbf{X}_i\boldsymbol{\beta}$.
- Fit is measured by log-likelihood ($l(\boldsymbol{\beta})$ or equivalently $l(\mu)$), or more conventiently *Deviance*

$$D(\boldsymbol{\mu}) = 2\{l(\mathbf{y}) - l(\boldsymbol{\mu})\}$$

  which takes on the role of the RSS.
- Again penalization avoids overfitting. $\hat{\boldsymbol{\beta}}$

$$\text{minimizes } D(\boldsymbol{\beta}) + \sum \lambda_j \boldsymbol{\beta}^{\mathsf{T}} \mathbf{S}_j \boldsymbol{\beta} \text{ w.r.t. } \boldsymbol{\beta}$$

GAM estimation

- Penalized MLE is performed by Penalized IRLS.
- Let $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\mu}}$ be current best estimates. Define

$$z_i = g'(\hat{\mu}_i)(y_i - \hat{\mu}_i) + \mathbf{X}_i\hat{\boldsymbol{\beta}} \text{ and } W_{ii} = \{V(\hat{\mu}_i)g'(\hat{\mu}_i)^2\}^{-1}$$

- The $\boldsymbol{\beta}$ minimizing

$$\|\sqrt{\mathbf{W}}(\mathbf{z} - \mathbf{X}\boldsymbol{\beta})\|^2 + \sum \lambda_j \boldsymbol{\beta}^{\mathsf{T}}\mathbf{S}_j\boldsymbol{\beta}$$

  is (almost always) an improved $\hat{\boldsymbol{\beta}}$ estimate.

- Iteratively repeating the steps finds the penalized likelihood maximizing estimates.

GAM Inference

- The ASM results have large sample equivalents for GAMs.
- If $\mathbf{B} = (\mathbf{X}^{\mathsf{T}}\mathbf{W}\mathbf{X} + \sum \lambda_j \mathbf{S}_j)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{W}$ then $\hat{\boldsymbol{\beta}} \dot{\sim} N(\mathbb{E}(\hat{\boldsymbol{\beta}}), \mathbf{B}\mathbf{B}^{\mathsf{T}}\phi)$ [frequentist].
- $\boldsymbol{\beta} \dot{\sim} N(\hat{\boldsymbol{\beta}}, (\mathbf{X}^{\mathsf{T}}\mathbf{W}\mathbf{X} + \sum \lambda_j \mathbf{S}_j)^{-1}\phi)$ [Bayesian].
- Approximate GLRT and F-ratio tests are possible, as if using a GLM, provided the effective degrees of freedom ($\mathrm{tr}(\mathbf{F})$) is used as the model DoF.

Smoothness selection

- There are two approaches.
- Performance iteration (Gu '92). Select $\boldsymbol{\lambda}$ by GCV/UBRE applied to the working linear model at each P-IRLS step.
  + Fast. - Does not always converge.
- Outer iteration. Use Deviance, $D(\hat{\boldsymbol{\beta}})$, in place of the RSS, $\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2$, in the GCV or UBRE score definitions, and optimize directly w.r.t. $\boldsymbol{\lambda}$.
  + Good convergence. - Slower.
  Note: UBRE becomes generalized AIC here.

Prediction

- Because model has a parametric representation as a GLM, prediction from a fitted model is just like prediction from a GLM.
- Given new covariate values a 'prediction matrix', $\mathbf{X}^p$, is created, exactly as if producing a new model matrix, except that any covariate dependent details of basis function form depend on the *original* fit covariates.
- The predicted values are then $\mu_i^p = \mathbf{X}_i^p \hat{\boldsymbol{\beta}}$.
- By simulating from the posterior of $\boldsymbol{\beta}$, samples from the posterior of any function of $\boldsymbol{\beta}$ can be obtained!

Open problems

- All the distributional results are conditional on the smoothing parameters.
- This reduces the reliability of p-values, and causes problems with the component wise performance of credible intervals/confidence intervals.
- Work on smoothing parameter unconditional confidence intervals is promising but incomplete (see Wood 2006 *GAMs: An Intro with R*).

Other approaches

There are several R packages providing GAMs or similar models. See the package docs for full references.

- `gam` provides Trevor Hasties original backfit GAMs.
- `gss` provides full smoothing spline based models (including computationally efficient versions).
- `assist` is an alternative for full spline smoothing.
- `gamlss` moves beyond smooth models of the mean of a response to model variance skew and kurtosis as smooth functions of covariates.
- `vgam` looks at multivariate extensions and much more.