Additive Models

Basics

- One approach to flexible modeling with multiple predictors is to use *additive models*:
$$Y = \beta_0 + f_1(x_1) + \cdots + f_p(x_p) + \varepsilon$$
where the f_j are assumed smooth.

- Variations include
  - some linear and some smooth terms
  $$Y = \beta_0 + \beta_1 x_1 + f_2(x_2) + \varepsilon$$
  - some bivariate smooth terms
  $$Y = f_1(x_1) + f_{23}(x_2, x_3) + \varepsilon$$

- A joint model using basis functions would be of the form
$$Y = X_0\beta + X_1\delta_1 + \cdots + X_p\delta_p + \varepsilon$$
with penalized objective function
$$\|Y - X_0\beta - \sum_{i=1}^{p} X_i\delta_i\|^2 + \sum_{i=1}^{p} \lambda_i \delta_i^T D_i \delta_1$$

- The model can be fit using the mixed model formulation with p independent variance components.

- An alternative is the *backfitting algorithm*.

Backfitting Algorithm

- For a model
$$f(x) = \beta_0 + \sum_{j=1}^{p} f_j(x_j)$$
with data y_i, x_{ij} and smoothers S_j
  - initialize $\widehat{\beta}_0 = \overline{y}$
  - repeat
$$\widehat{f}_j \leftarrow S_j \left[\{y_i - \widehat{\beta}_0 - \sum_{k \neq j} \widehat{f}_k(x_{ik})\}_1^n \right]$$
$$\widehat{f}_j \leftarrow \widehat{f}_j - \frac{1}{n} \sum_{i=1}^{n} \widehat{f}_j(x_{ij})$$
until the changes in the $\widehat{f}_j$ are below some threshold.
- A more complex linear term is handled analogously.
- For penalized linear smoothers with fixed smoothing parameters this can be viewed as solving the equations for the minimizer by a block Gauss-Seidel algorithm.
- Different smoothers can be used on each variable.
- Smoothing parameters can be adjusted during each pass or jointly.
  - `bruto` (Hastie and Tibshirani, 1990) uses a variable selection/smoothing parameter selection pass based on approximate GCV.
  - `gam` from package `mgcv` uses GCV.
- Backfitting may allow larger models to be fit.
- Backfitting can be viewed as one of several ways of fitting penalized/mixed models.

Some examples are available in

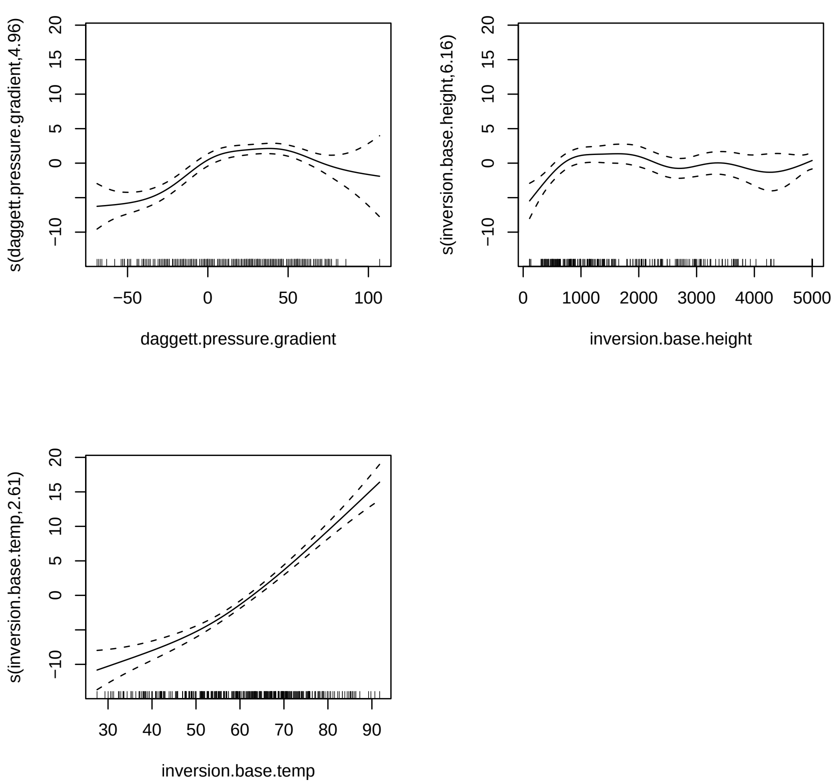
`http://www.stat.uiowa.edu/~luke/classes/STAT7400/examples/additive.Rmd`

Example: Ozone Levels

- Data set relates ozone levels to pressure gradient, temperature, and height of inversion.
- A `gam` fit is produced by

```
library(mgcv)
library(SemiPar)
data(calif.air.poll) # data are from SemiPar
fit <- gam(ozone.level ~ s(daggett.pressure.gradient)
                       + s(inversion.base.height)
                       + s(inversion.base.temp),
           data = calif.air.poll)
```

- The default plot method produces



Mixed Additive Models

- Mixed additive models can be written as

$$Y = X_0\beta + ZU + X_1\delta_1 + \cdots + X_p\delta_p + \varepsilon$$

  where U is a "traditional" random effects term with

$$U \sim N(0, \Sigma(\theta))$$

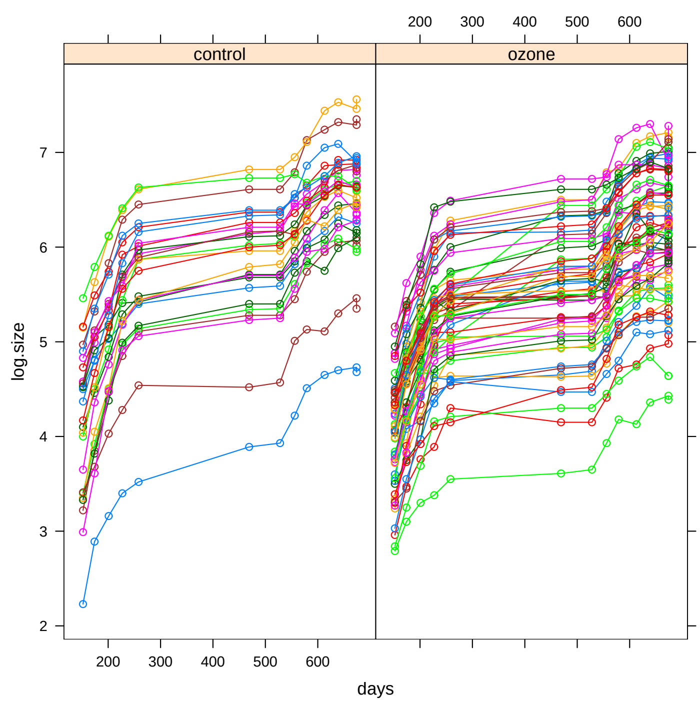
  for some parameter θ and the terms $X_1\delta_1 + \cdots + X_p\delta_p$ represent smooth additive terms.

- In principle these can be fit with ordinary penalized least squares or mixed models software.

Example: Sitka Pines Experiment

- An experiment on sitka pines measured size over time for 79 trees grown in an ozone-rich environment and a control environment. Measurements were taken at 13 time points.

```
data(sitka) # from SemiPar
library(lattice)
sitka$ozone.char <- ifelse(sitka$ozone, "ozone", "control")
xyplot(log.size ~ days|ozone.char, groups = id.num, type = "b",
       data = sitka)
```



- The plot suggests a model with
  - a smooth term for time
  - a mean shift for ozone
  - a random intercept for trees
  - perhaps also a random slope for trees

- The random intercept model can be fit with `spm` using

```
attach(sitka)
fit <- spm(log.size ~ ozone + f(days),
           random= ~ 1, group = id.num)
```

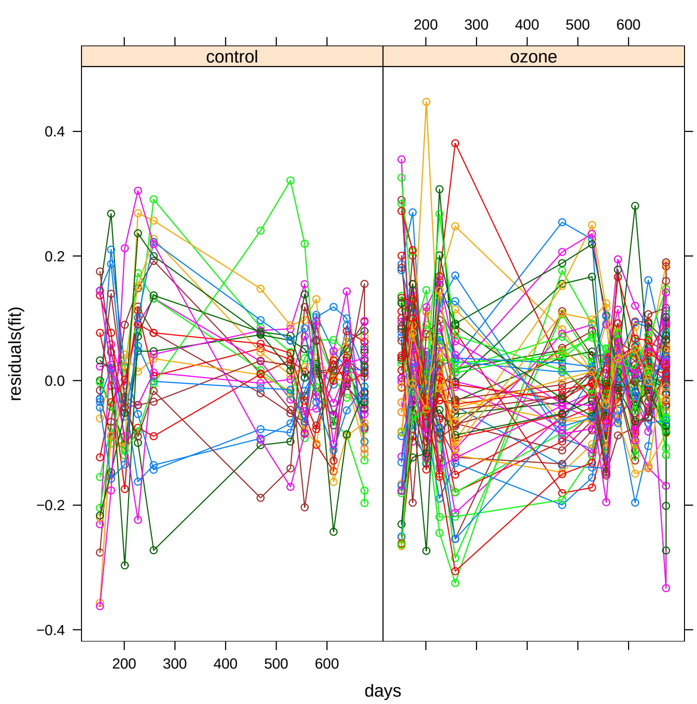
  and by `gamm` with

```
trees <- as.factor(id.num)
fit <- gamm(log.size ~ ozone + s(days),
            random = list(trees = ~ 1))
```

- `spm` cannot fit a more complex random effects structure at this point. Using `gamm` we can fit random slope and intercept with

```
fit <- gamm(log.size ~ ozone + s(days),
            random = list(trees = ~ 1 + days))
```

- Residuals don't show any further obvious pattern.



- Autocorrelated errors over time might be worth considering.

Generalized Additive Models

- Standard generalized linear models include

$$y_i \sim \text{Bernoulli}\left(\frac{\exp\{(X\beta)_i\}}{1+\exp\{(X\beta)_i\}}\right)$$

  and

$$y_i \sim \text{Poisson}\left(\exp\{(X\beta)_i\}\right)$$

- Maximum likelihood estimates can be computed by *iteratively reweighted least squares (IRWLS)*
- Penalized maximum likelihood estimates maximize

$$\text{Loglik}(y, X_0\beta + X_i\delta) - \frac{1}{2}\lambda\delta^T D\delta$$

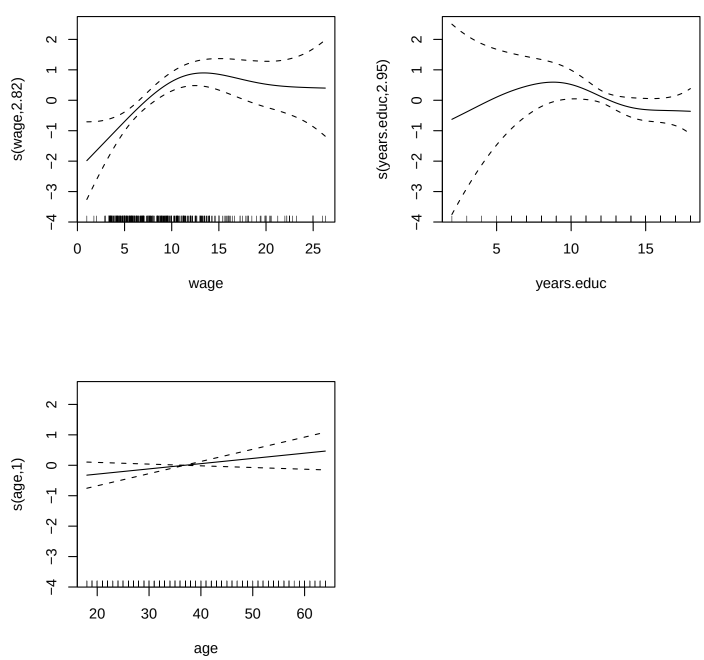
- This has a mixed model/Bayesian interpretation.
- GLMM (genelarized linear mixed model) software can be used.
- The IRWLS algorithm can be merged with backfitting.

Example: Trade Union Membership

- Data relating union membership and various characteristics are available.
- A Bernoulli generalized additive model relates the probability of union membership to the available predictor variables.
- One possible model is fit by

```
data(trade.union) # from SemiPar
fit <- gam(union.member ~ s(wage) + s(years.educ) + s(age)
                        + female + race + south,
           family=binomial,
           subset=wage < 40,    # remove high leverage point
           data=trade.union)
```

- The estimated smooth terms are



- Some summary information on the smooth terms:

```
Approximate significance of smooth terms:
                edf Est.rank Chi.sq p-value
s(wage)       2.818    9.000  27.26 0.00127 **
```

```
s(years.educ) 2.951     9.000  11.43 0.24710
s(age)        1.000     1.000   2.30 0.12940
```

Alternative Penalties

- Bases for function spaces are infinite dimensional
- Some form of penalty or *regularization* is needed.
- Penalties often have a useful Bayesian interpretation.
- Most common penalties on coefficients δ
  - quadratic, $\sum \delta_i^2$ or, more generally, $\delta^T D \delta$
  - absolute value, L_1, LASSO: $\sum |\delta_i|$

Ridge Regression

- *Ridge regression* uses the L_2 penalty $\lambda \sum \delta_i^2$.
- Using a quadratic penalty $\delta^T D \delta$ with strictly positive definite D is sometimes called *generalized ridge regression*.
- The minimizer of

  $$\min_{\delta} \{ \|Y - X\delta\|^2 + \lambda \delta^T D \delta \}$$

  is

  $$\widehat{\delta_\lambda} = (X^T X + \lambda D)^{-1} X^T Y$$

  which shrinks the OLS estimate towards zero as $\lambda \to \infty$.
- If $X^T X = D = I$ then the ridge regression estimate is

  $$\widehat{\delta_\lambda} = \frac{1}{1+\lambda} \widehat{\delta}_{\text{OLS}}$$

LASSO

- The LASSO (Least Absolute Shrinkage and Selection Operator) or L_1-penalized minimization problem

$$\min_{\delta}\{\|Y - X\delta\|^2 + 2\lambda \sum |\delta_i|\}$$

does not in general have a closed form solution, but if $X^T X = I$ then

$$\widehat{\delta}_{i,\lambda} = \text{sign}(\widehat{\delta}_{i,\text{OLS}})(|\widehat{\delta}_{i,\text{OLS}}| - \lambda)_+$$

The OLS estimates are shifted towards zero and truncated at zero.

- The L_1 penalty approach has a Bayesian interpretation as a posterior mode for a Laplace or double exponential prior.

- The variable selection property of the L_1 penalty is particularly appealing when the number of regressors is large, possibly larger than the number of observations.

- For least squares regression with the LASSO penalty
  - the *solution path* as λ varies is piece-wise linear
  - there are algorithms for computing the entire solution path efficiently
  - Common practice is to plot the coefficients $\beta_j(\lambda)$ against the *shrinkage factor* $s = \|\beta(\lambda)\|_1/\|\beta(\infty)\|_1$

- R Packages implementing general L_1-penalized regression include `lars`, `lasso2`, and `glmnet`.

- A paper, talk slides, and R package present a significance test for coefficients entering the model.

Elastic Net

- The *elastic net* penalty is a combination of the LASSO and Ridge penalties:
$$\lambda\left[(1-\alpha)\sum \delta_i^2 + 2\alpha \sum |\delta_i|\right]$$
  - Ridge regression corresponds to $\alpha = 0$.
  - LASSO corresponds to $\alpha = 1$.
- λ and α can be estimated by cross-validation.
- Elastic net was introduced to address some shortcomings of LASSO, including
  - inability to select more than n predictors in $p > n$ problems;
  - tendency to select only one of correlated predictors.
- The `glmnet` package implements elastic net regression.
- Scaling of predictors is important; by default `glmnet` standardizes before fitting.

Non-Convex Penalties

- The elastic net penalties are convex for all α.
- This greatly simplifies the optimization to be solved.
- LASSO and other elastic net fits tend to select more variables than needed.
- Some non-convex penalties have the theoretical property of consistently estimating the set of covariates with non-zero coefficients under some asymptotic formulations.
- Some also reduce the bias for the non-zero coefficient estimates.
- Some examples are
  - smoothly clipped absolute deviation (SCAD);
  - minimax concave penalty (MCP).
- MCP is of the form $\sum \rho(\delta_i, \lambda, \gamma)$ with

$$\rho(x,\lambda,\gamma) = \begin{cases} \lambda|x| - \frac{x^2}{2\gamma} & \text{if } |x| \leq \gamma\lambda \\ \frac{1}{2}\gamma\lambda^2 & \text{otherwise} \end{cases}$$

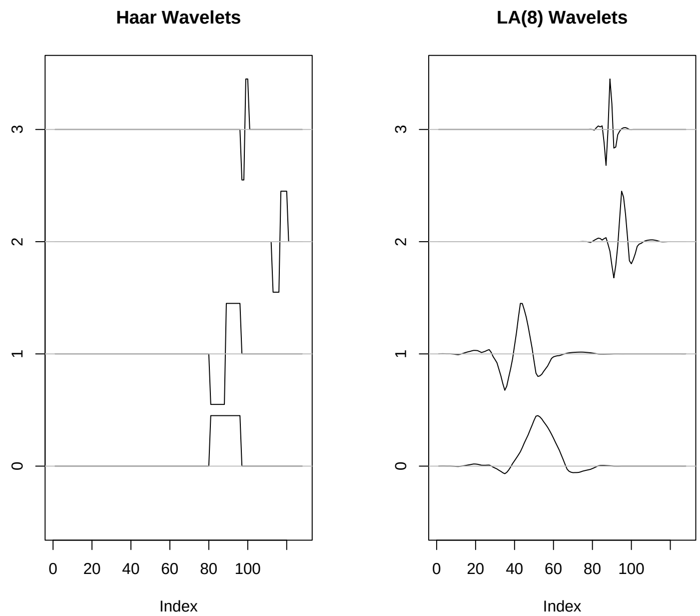
  for $\gamma > 1$.
- This behaves like $\lambda|x|$ for small $|x|$ and smoothly transitions to a constant for large $|x|$. SCAD is similar in shape.
- Jian Huang and Patrick Breheny have worked extensively on these.

Alternative Bases

- Many other bases are used, including
  - polynomials
  - trigonometric polynomials (Fourier basis)
  - wavelet bases
- Different bases are more suitable for modeling different functions
- General idea: choose a basis in which the target can be approximated well with a small number of basis elements.

Wavelets

- Wavelet smoothing often assumes observations at $N = 2^J$ equally spaced points and uses an orthonormal basis of N vectors organized in J levels.



- A common approach for wavelet smoothing is to use L_1 shrinkage with

$$\lambda = \widehat{\sigma}\sqrt{2\log N}$$

  A variant is to use different levels of smoothing at each level of detail.

- $\widehat{\sigma}$ is usually estimated by assuming the highest frequency lavel is pure noise.
- Several R packages are available for wavelet modeling, including `waveslim`, `rwt`, `wavethresh`, and `wavelets`
- Matlab has very good wavelet toolboxes.
- S-Plus also has a good wavelet library.

Other Approaches

- MARS, multiple adaptive regression splines. Available in the `mda` package.
- `polymars` in package `polyspline`.
- Smoothing spline ANOVA.
- Projection pursuit regression.
- Single and multiple index models.
- Neural networks.
- Tree models.