AMs, GAMs, and SpAMs

Zachary Weller

Department of Statistics, Colorado State University
wellerz@stat.colostate.edu

November 18, 2014

Motivation

- Regression plays a fundamental role in many data analyses
- In reality, effects may not be nonlinear
- *Additive models* (AMs) allow a more flexible form for the regression function:

$$E(Y|X_1, \ldots, X_p) = \alpha + f_1(X_1) + \ldots + f_p(X_p) = \sum_{j=1}^{p} f_j(X_j) \tag{1}$$

- The $f_j's$ give above are general smooth, nonparametric functions
- Non-parametric form for the f_j functions makes model more flexible
- Additivity is retained allowing for easy interpretation

GAMs

- Analogous to generalized linear models, we can model the conditional mean, $\mu(X)$, of Y via a link function g to create *generalized additive models* (GAMs):

$$g(\mu(X)) = \alpha + f_1(X_1) + \ldots + f_p(X_p) \tag{2}$$

  - $g(\mu) = logit(\mu)$ for modeling binomial probabilities
  - $g(\mu) = log(\mu)$ for log linear models for Poisson counts

- Can mix in linear and other non-parametric form with the nonlinear terms (e.g. if one of the feature variables is a factor)
  - $g(\mu) = X^T\beta + \alpha_k + f(Z)$
  - $g(\mu) = f(X) + h_k(Z)$

Fitting AMs: population level

- Letting P denote the joint distribution of (X_i, Y_i), $\mathcal{H}_j$ be a Hilbert subspace of $L_2(P)$, and $\mathbb{E}$ be the expectation w.r.t. $\mathbf{X}$ and ϵ, at the population level, we seek the solution to:

$$\min_{f_j \in \mathcal{H}_j, 1 \leq j \leq p} \mathbb{E}(Y - \sum_{j=1}^{p} f_j(X_j))^2 \tag{3}$$

- We seek to estimate all p functions, $f_j(X_j)$, simultaneously $\rightarrow$ given sample data, this is done via backfitting

Fitting AMs: sample version

- To employ the backfitting algorithm, first choose a nonparametric smoothers, S_j (e.g. cubic splines)
- Backfitting algorithm
  - set $\widehat{\alpha} = \frac{1}{N}\sum_{i=1}^{N} y_i$
  - set $\hat{f}_j = 0, \forall j$
  - Iterate until convergence (i.e. $\Delta(\hat{f}_j) < \delta$):

    For each $j = 1, \ldots, p$

    1. Compute residual: $R_j = Y - \widehat{\alpha} - \sum_{k \neq j} \hat{f}_k(X_k)$;
    2. Smooth residuals: $\hat{f}_j \leftarrow S_j R_j$
    3. Center: $\hat{f}_j \leftarrow \hat{f}_j - \frac{1}{N}\sum_{i=1}^{N} \hat{f}_j(x_{ij})$

- Implemented in the R package 'gam'
- Backfitting algorithm must be modified for GAMs

R examples

1. AMs: synthetic data
2. GAMs: predicting email spam
   - Data set containing information from 4601 email messages sent to George Forman from Hewlett-Packard laboratories
   - Response: 1 = spam, 0 = not spam (39.4% spam messages)
   - Features: percentage of words in a message that match given word (e.g. money), character frequency (e.g. ;), capital letters (e.g. length of longest uninterrupted sequence of capital letters)
   - Sensitivity and Specificity: may be more serious to classify a genuine email as spam than vice-versa. We can alter the balance between the two types of classification errors in two ways:
     1. change the losses (i.e. change the posterior probability needed to classify a message as spam)
     2. re-weight the data to encourage model to better fit data in 'email' class (i.e. upweight the '0' instances in the data)

Sparse Additive Models

- Sparse additive models (SpAMs): "Additive models only have good properties when ... p is not large relative to the sample size n"
- May want/need to constrain the index set $\{j : f_j \not\equiv 0\}$
- 2008 paper by Ravikumar, Lafferty, Liu, and Wasserman develops theory and implementation of SpAMs for both AMs and GAMs
- Derive what is effectively a hybrid of backfitting and LASSO
- Show that their estimator satisfies both *sparsistence* ("sparsity pattern consistent") and *persistence* (form of risk consistent)

SpAMs Theory

Modifying (3) to include a scaling parameter for each function and an additional constraint provides:

$$\min_{\beta \in \mathbb{R}^p, f_j \in \mathcal{H}_j} \mathbb{E}(Y - \sum_{j=1}^{p} \beta_j f_j(X_j))^2 \tag{4}$$

$$\text{subject to:} \qquad \sum_{j=1}^{p} |\beta_j| \leq L, \tag{5}$$

$$\mathbb{E}(f_j^2) = 1, \quad j = 1, \ldots, p \tag{6}$$

In penalized Lagrangian form:

$$\begin{aligned} \mathcal{L}(f, \lambda) &= \frac{1}{2}\mathbb{E}(Y - \sum_{j=1}^{p} f_j(X_j))^2 + \lambda \sum_{j=1}^{p} \sqrt{\mathbb{E}(f_j^2(X_j))} \\ &= \frac{1}{2}\mathbb{E}(Y - \sum_{j=1}^{p} f_j(X_j))^2 + \lambda \sum_{j=1}^{p} \|f_j(X_j)\| \end{aligned} \tag{7}$$

SpAMs Theory

Theorem

The functions, $f_j \in \mathcal{H}_j$, that minimize (7) are given by:

$$f_j = \left[1 - \frac{\lambda}{\sqrt{\mathbb{E}(P_j^2)}}\right]_+ P_j \quad a.s. \tag{8}$$

where $[\cdot]_+$ denotes the positive part, and $P_j = \mathbb{E}[R_j | X_j]$ is the projection of the residual $R_j = Y - \sum_{k \neq j} f_k(X_k)$ onto $\mathcal{H}_j$

Again, we can implement the sample version via backfitting and soft thresholding algorithm. We can estimate the projection $P_j = \mathbb{E}(R_j | X_j)$ by applying a smoother to the residuals: $\hat{P}_j = S_j R_j$. Furthermore, estimate $\sqrt{\mathbb{E}(P_j^2)}$ by $\hat{s}_j = \frac{1}{\sqrt{n}} \|\hat{P}_j\| = \sqrt{\text{mean}(\hat{P}_j^2)}$.

SpAM Backfitting Algorithm

- Choose regularization parameter λ (can be chosen by GCV or C_p)
- Initialize $\hat{f}_j = 0, \forall j$
- Iterate until convergence:
  For each $j = 1, \ldots, p$
  1. Compute residual: $R_j = Y - \widehat{\alpha} - \sum_{k \neq j} \hat{f}_k(X_k)$;
  2. Estimate $P_j = \mathbb{E}(R_j | X_j)$ by smoothing: $\hat{P}_j = S_j R_j$;
  3. Estimate norm: $\hat{s}_j^2 = \frac{1}{n} \sum_{i=1}^n \hat{P}_j^2(x_{ij})$;
  4. Soft threshold: $\hat{f}_j = [1 - \lambda / \hat{s}_j]_+ \hat{P}_j$;
  5. Center: $\hat{f}_j \leftarrow \hat{f}_j - \frac{1}{N} \sum_{i=1}^N \hat{f}_j(x_{ij})$

- Not currently implemented in R
- A related algorithm, *sparse additive machine*, is implemented in the ‘SAM’ package $\leftarrow$ can be viewed as a functional version of support vector machines

The End

References:

- Hastie, Trevor, et al. *The elements of statistical learning*. Vol. 2. No. 1. New York: Springer, 2009.
- Ravikumar, Pradeep, et al. "Sparse additive models." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71.5 (2009): 1009-1030.

Questions?