A Literature Review on Generalized Additive Models

Yifang Zhang

Abstract: Generalized Additive Models (GAMs) are semi-parametric extensions of Generalized Linear Models (GLMs) where the only underlying assumption are additive and smooth components. Link functions are used in GAMs as well to construct a relationship between the mean of the response variables and some unknown smooth functions of some predictor variables. That leads to the principal advantage of GAMs that GAMs can model highly complex nonlinear relationships between the response and a set of predictor variables. GAMs can be applied for different kinds of data, including binary data, count data, or even longitudinal data. In additive, a class of GAMs called partially linear models have been paid a great attention since the models are more flexible than the standard linear models but are easier to illustrate the effect of each variables. In this literature review, we will focus on the PLM and also construct a simulaiton study on GAMs to test its abilitiy for prediction.

Key word: Generalized additive model; Partially linear models; Simulation

1. Introduction

1.1 Background of GAMs

Generalized Additive Models (GAMs) was produced by Hastie and Tibshirani[1] (1986,1990) by extending generalized linear models with the smoothing methods. They replaced the linear predictor $\eta = \sum_i^p \beta_j X_j$ with an additive predictor of the form $\eta = \sum_i^p s_j(X_j)$, and stated that a generalized additive model has the form[2]

$$g(\mu(x)) = \alpha + \sum_{j=1}^{p} f_j(x_j), \ E[f_j(x_j)] = 0,$$

As an extension of GLMs, GAMs can be applied to any of the data types that GLMs are used for including Gaussian, binary, multinomial, and Poisson data. Later on, Yee and Wild (1996) used vector smoothing to extend GAMs to a multivariate setting[3]. Furthermore, Rigby and Stasinopoulos developed a more general class called the generalized additive model for location, scale and shape (GAMLSS), where the data types is extended to a very general distribution family instead of just the exponential family[4].

GAMs have several advantages. First, GAMs is a powerful method for prediction and interpolation. Second, as an additive model, GAMs is easier to interpret but are more flexible than standard linear methods which can uncover hidden patterns in the data. And comparing to GLMs, GAMs shows lower AIC values and explained higher deviance. Later on, we will construct a simulation study to confirm that. And finally, regularization of GAMs allow us to avoid overfitting.

1.2 Statistical Methods

A general additive model has the following general form

$$\eta = g(\mu) = \alpha + f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p)$$

where $\mu \equiv E(Y)$ and Y has some exponential family distribution, and each of the x_i are predictor variables and the f_i are smooth functions of the predictors or terms.

According to Hastie in the book *Statistical Models in S*, several famous examples of the additive predicotr η are provided as following:

- simple additive model $y = f(x) + \epsilon$, having only one term;
- semi-parametric model $y = x^t\beta + f_1(z_1) + f_2(z_2) + \cdots + f_q(z_q) + \epsilon$, where several are linear and several are nonparametric terms. And it is actually the partially linear models;
- nonparametric logistic regression model $logitP(x) = log(\frac{P(x)}{1-P(x)}) = \eta(x)$.

To estimate the smooth functions f_j, Hastie and Tibshirani used the backfitting algorithmthe in the local scoring algorithm. And in fact there exist multiple methods for estimating the smooth functions f_j and two most popular and widely used methods are smoothing splines descirbed as Hastie and Tibshirani in their paper and penalized splines introduced by Simon N.Wood (2006) in his book *Generalized Additive Models: an introduction with R.*[56]

Here we use the **penalized regression method** as Wood suggested due to the availability of a variety of methods with efficient implementations. GAMs are fit to data by maximizing a penalized log-likelihood or a penalized log partial-likelihood as following:

- Step1: Initialize $\hat{\alpha} = g(\frac{1}{N\sum_{i=1}^{N} y_i})$, $f_j^0 = 0$ for $j = 1, 2, \cdots, p$.
- Step2: Construct an adjusted dependent variable z_ij as:

$$z_i = \eta_i^0 - (y_i - \mu_i^0)(\frac{\partial \eta_i}{\partial \mu_i})_0$$
$$\eta^0 = g(\mu_i)^0 = \alpha^0 + \sum_{i=1}^{p} f_j^0(x_{ij}) \ , \ \mu_i^0 = g^{-1}(\eta_i^0)$$

- Step3: Compute weights $w_i = (\frac{\partial \mu_i}{\partial \eta_i})_0(V_i^0)^{-1}$, where V_i^0 is the variance of y at μ_i^0.
- Step4: Penalized Spline Regression, i.e. minimize

$$||\sqrt(W)(z - X\beta)||^2 + \lambda\beta' S\beta$$

  w.r.t. β. Where X is the matrix of data on basis functions used to represent the regression function, W is a diagnoal matrix with i^{th} diagonal element be w_i, S is a matrix of known coefficients in the penalty function $\beta' S\beta$, and λ is a smoothing paramter. Compute f_j^1, η^1 and μ_i^1, which are the second stage estimates of f_j, η and μ_i.
- Step5: Repeat Step2 to Step4 replacing η^k by η^{k+1} until the difference between two successive values of η is less than a small prespecified vaule and convergence is obtained.

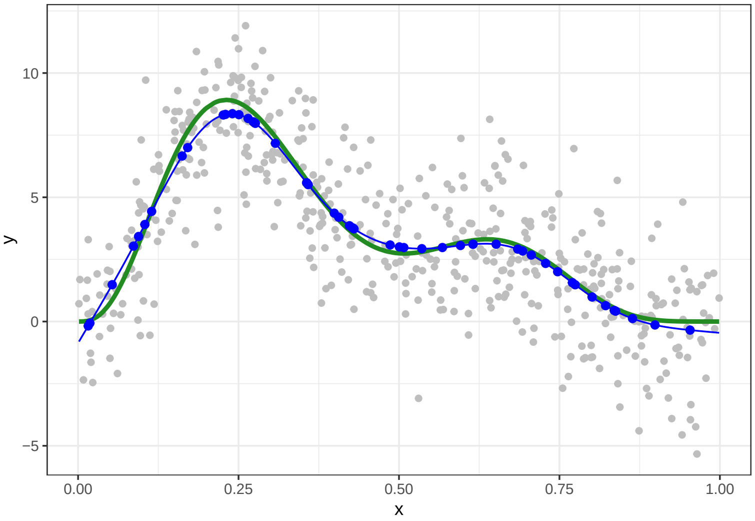
2. Simulation Study

2.1 The generalized additive Gaussian model

First we construct a simulation study on the generalized addtive model with the link function $g(\mu_i) = \mu_i$ to fit a Gaussian data.

```
## Loading required package: nlme
```

```
## This is mgcv 1.8-31. For overview type 'help("mgcv-package")'.
```

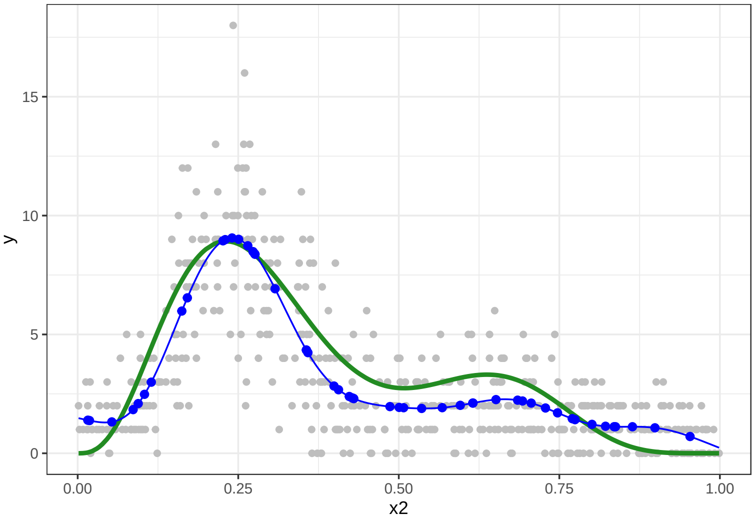


Here we simulate data from a strongly nonlinear term $f(x) = 0.2x^{11}(10(1-x))^6 + 10(10x)^3(1-x)^{10}$. The green curve is the true function f(x), while the blue curve is the fitted smooth estimate $\hat{f}(x)$.

We can find that the green curve and the blue curve basically coincide, thus GAMs is quite powerful for prediction. In addition, we get $AIC_{glm} = 2447.1$ while $AIC_{gam} = 2086.126$, and Deviance explained by GAM is 0.664 while deviance explained by GLM is just 0.289.

2.2 The generalized additive Poisson model

In addition, we construct a simulation study on the generalized addtive model with the link function $g(\mu_i) = \log(\mu_i)$ to fit a Poisson data.



Again, the differece between the true curve (green one) and the blue curve (fitted one) is not big. But the fitting is not as great as what we get in Gaussion case. Furthermore, $AIC_{glm} = 2305.7$ while $AIC_{gam} = 2017.455$. Deviance explained by GAM is 0.667, while deviacne explained by GLM is only 0.25.

So, in general we can say GAMs provide a more powerful prediction when it comes to complex nonlinearity between predictors and the response. And comparing to GLMs, GAMs provides both lower AIC values and higher deviance explanation.

3. Problems and Solutions for GAMs with Multiple Smoothing Parameter Estimation

3.1 Problems: Bias estimates, underestimate variances when concurvity exists

GAMs have been widely and effective applied in a variety of research areas including Business and Economics, epidemiology, genetics, and medicine as a more flexible apporach then fully parametric methods. And the gam function in R stadio is a powerful method for researchers to construct GAMs.

During the literature reviews, we find that GAMs has been widely used in many time-series analyses, and one typical topic is Air Pollution Study. While when it comes to multiple nonparametric smooth functions been included in the model, Dominici, McDermott, Zeger and Samet (2002) shown that the default convergence parameters in the S-Plus statisticl function gam could be too lax to assure convergence of the backfitting algorithm and could lead to biased estimates[7]. They found that for time-series studies of air pollution and health, where mostly more than one nonparametric smooth function would be included in the model, the use of GAMs should be careful. Default convergence parameters need to be modified, but at the meantime, the penalized likelihood GAM optimized would lead to an increasing of bias in pollution effect estimates according to their study. Furthermore, GAMS failed to detect concurvity (*the nonparametric analogue of multicollinearity*) even though the convergence of the backfitting algorithm is guaranteed.

Ramsay, Burnett and Krewski (2003) studied the effect of concurvity in GAMs and found that when concurvity was present in the data, statistical software such as gam functions in R can seriously underestimate the variance of fitted model parameters and lead to significance tests with inflated Type I error[8].

3.2 Several Solutions

Even thougth GAMs would lead to bias when multiple smooth parameters are considered in the model, GAMs is still a powerful technique in epidemiologic or other research because by using nonparametric regression, epidemiologists don't have to rely on difficult-to-verity assumptions mentioned by Dominici et al. (2002). So researchers find possilbe solutions for such situations.

- He Shui (2004) explored an alternate classof models, generalized linear models with natural cubic splines which might not be affected as much as GAMs by concurvity[9]. But since GLM with natural cubic splines resluted in loss of flexibility, the author also investigated an alternative approach to fit a GAM, a partial regression approach described by Speckman[10] (1988) with kernel smoothing to estimate the partially linear model

  $$Y = X^T\beta + g(T) + \epsilon$$

  Shui shown that the partial regression approach performs better than the standard approach when concurvity exists in the data.

- Wood (2004) proposed that GAMs with a ridge penalty could provide a practical solution[11]. The mothod is based on the pivoted QR decomposition and the singular value decomposition. Wood also shown that this new method enhanced numerical stability of the basic penzlized least squares mothds.

- Souza et al.(2017) proposed a hybrid generalized additive model-principal component anaylsis-vector auto-regressive (GAM-PCA-VAR) model for a study to time series of respiratory disease and air pollution data[12]. They combined PCA and GAMs along with a VAR process where PCA is used to eliminate the multicollinearity existing in the data and VAR model is used to handle the serial correlation of the data to produce white noise processes as covariates in the GAM.

4. Partially Linear Models - A Class of GAMs

4.1 Introduction

Parially linear models (PLM) are special cases of the generalized additive models, which assume the realtion between the response variables and the covariates can be represented as[13]:

$$Y = X^T \beta + g(T) + \epsilon$$

where X, T are d-mensional and scalar regressors, β is a vector of unknown parameters, $g(\cdot)$ is an unknown smooth function, and ϵ is an error term with mean zero conditional on X and T.

PLMs allow easier interpretation of the effect of each variable and are more flexible than the standard linear model since they are semi-parametric models, as combining both parametric and nonparametric components. Another advantage of PLM is a easier computations comparing with additive models[14].

4.2 Estimations

There are several different methods been proposed for the estimation of PLMs.

- Engle et al.[15] (1986) and Heckman[16] (1986) used spline smoothing and defined estimators of β and g as the solution of
$$\arg\min_{\beta,\ g} \frac{1}{n}\sum_{i=1}^{n}\{Y_i - X_i^T\beta - g(T_i)^2\} + \lambda\int\{g^{''}(u)\}^2 du$$
- Speckman (1988) estimated the nonparametric component by $W\gamma$, where $W_{n\times q}$ is full rank and γ is an additional parameter. The estimator of β based on $Y = X\beta + W\gamma + \epsilon$ is
$$\hat{\beta_S} = \{X^T(I - P_W)X\}^{-1}\{X^T(I - P_W)Y\}\ ,\ P_W = W(W^TW)^{-1}W^T$$
- Hamiltonand Truong[17] (1997) developed the local linear method. For PLM $Y = \alpha B + s(X) + \epsilon$, then estimators for α and $m(x)$ are
$$\hat{\alpha} = \{((I - S)B)'((I - S)B)\}^{-1}((I - S)B)'(I - S)Y$$
and
$$\hat{m}(x) = S'(x)(Y - \hat{\alpha}B)$$
where
$$\begin{aligned} S'(x) &= a'[X'(x)W(x)X(x)]^{-1}X'(x)W(x),\ a^{'} = (1, 0, 0) \\ W(x) &= diag[K_H(X_1 - x), \cdots, K_H(X_n - x)],\ K_H(x) = |H|^{-1}K(H^{-1}x) \\ X_i'(x) &= [1, \{H^{-1}(X_i - x)\}'] \end{aligned}$$
where $K(\cdot)$ is a kernel function on R^2 and H a 2×2 symmetric and positive-defined matrix.

4.3 Variable Selection

Partially linear models are effectively used for analyzing high-dimensional data, which lead to variable selection an important procedure when we apply PLMs. For parametric methods, we have plenty of variable selection procedures. The Akaike infromation criterion (AIC) proposed by Akaike (1973) is one widely used criterion.Schwarz (1986) proposed the bayesian information criterion (BIC) which is another widely used criterion in our daily works. Depending on AIC or BIC, regressions like stepwise regression and best suset selection can be used in practical questions.

However, when it comes to nonparametric methods, Breiman (1996) found that such proceduers suffer from several weknesses, one is the lack of stability. A small change on data will provide a very different model.

In attempt to overcome these weaknesses, Fan and Li (2004) proposed an effective kernel estimator for nonparametric function estimation and use the smoothly clipped absolute deviation (SCAD) penalty for variable selection. Where Bunea (2004) used a penalized least squares criterion for selectoin and established the consistency property of their estimator. Li and Liang (2009) proposed two classes of variable selection procedures, penalized least squares and penalized quantile regression, using the nonconvex penalized principle[18]. In the context of partially smoothing spline models, Ni, Zhang and Helen Zhang[19] (2009) proposed a new regularization method for simultaneous variable selection and model estimation in partially linear models via double-penalized least squares.

In addition, there are two recent development on variable selection for PLMs. Yang, Fang, Wang and Shao[20] (2017) developed a novel variable selection based on the idea of gradient learning for general PLMs. This procedure used the reproducing-kernel-Hilbert-space tool to learn the gradients and the group-lasso to select variables.

Wang, Cai and Li[21] (2021) mentioned that most existing variable selection procedures in PLMs are based partial residuals which involve a two-step estimation procedure. They proposed a new Bayesian subset selection procedure for PLMs by modifying the Bayesian shrinking and diffusing priors.

4.4 Detect the model structure

Since Partially Linear Models (PLMs) is a class of models between linear and nonparametric models, one natural question except for variable selection is, given a set of covariates, how should we decides which covariates have linear effects while others have nonlinear effects. To achieve a highly interpretable model to show how each covariate associated with the response variable and predict the response, the structure selection problem fundamentally important.

While there is quite little literature in this area and one paper I found useful for detecting the model structure is *Linear or Nonlinear? Automatic Structure Discovery for Partially Linear Models* by Zhang, Cheng and Liu[22] (2011). They proposed a new approach with a new regularization framework in the context of smoothing spline ANOVA models, called the Linear and Nonlinear Discoverer (LAND) to identify model structure annd estimate the regression function simultaneously.

The underlying true regression model has the form

$$y_i = b + \sum_{j \in I_L} x_{ij}\beta_j + \sum_{j \in I_N} f_j(x_{ij}) + \sum_{j \in I_O} 0(x_{ij}) + \epsilon_j$$

where I_L, I_N, I_O are the index sets for nonzero linear effects, nonzero nonlinear effects and null effects. Clearly, $I_L \cup I_N \cup I_O = I$ the whole index set.

So further study should be done in this area, to develop if there exists more methods for researchers to detect the model structure when considering partially linear models to achieve a highly interpretable model.

Reference

[1] Trevor Hastie and Robert Tibshirani (1986), “Generalized Additive Models”, *Statistical Science*, Vol.1, No.3, 297-318.

[2] Trevor Hastie and Robert Tibshirani (1987), “Generalized Additive Models: Some Applications”, *Journal of the American Statistical Association*, Vol.82, No.398, pp.371-386.

[3] T.W.Yee and C.J. Wild (1996), “Vector Generalized Additive Models”, *Journal of the Royal Statistical Society*, Vol.58, Issue 3, pp.481-493.

[4] R.A.Rigby and D.M.Stasinopoulos (2005), “Generalized additive models for location, scale and shape”, *Journal of the Royal Statistics*, Vol.54, Issue 3, pp.507-554.

[5] Suneel Babu Chatla and Galit Shmueli, “Efficient estimation of COM–Poisson regression and a generalized additive model”, *Computational Statistics and Data Analysis*, 121(2018)71-88.

[6] Simon N. Wood (2006), Generalized Additive Models: an introduction with R.

[7] Francesca Dominici, Aidan McDermott, Scott L.Zeger, and Jonathan M.Samet (2002), “On the Use of Generalized Additive Models in Time-Series Studies of Air Pollution and Health”, *American Journal of Epidemiology.*

[8] Timothy O. Ramsay, Richard T.Burnett and Daniel Krewski (2003), “The Effect of Concurvity in Generalized Additive Models Linking Mortality to Ambient Particulate Matter”, *Epidemiology*, Vol.14, No.1, pp.18-23.

[9] He Shui (2004), “Generalized additive models for data with concurvity: Statistical issues and a novel model fitting approach”.

[10] Paul Speckman (1988), “Kernel Smoothing in Partial Linear Models”, *Royal Ststistical Society*, Vol.50, No.3, pp/413-436.

[11] Simon N.Wood (2004), “Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models”, *Journal of the American Statistical Association*, Vol.99, No.467, pp.673-686.

[12] Juliana B.de Souza, Valderio A. Reisen, Glaura C. Franco, Marton Ispany, Pascal Bondon and Jane Meri Santos (2017), “Generalized additive models with principal component analysis: an application to time series of respiratory disease and air pollution data”, *Applied Statistics Series C.*

[13] Wolfgang Hardle, Yuichi Mori and Philippe Vieu, Statistical Methods for Biostatistics and Related Fields.

[14] Hua Liang, “Estimation in partially linear models and numerical comparisons”, *Computational Statistics and Data Analysis*, 50(2006)675-687.

[15] Robert F.Engle, C.W.J.Granger, John Rice and Andrew Weiss (1986), “Semiparametric Estimates of the Relation Between Weather and Electricity Sales”, *Journal of the American Statistical Association*, Vol.81, N.394, pp.310-320.

[16] Nancy E. Heckman (1986), “Spline Smoothing in a Partly Linear Model”, *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol.48, No.2, pp.244-248.

[17] Scott A.Hamilton and Young K.Truong (1997), “Local Linear Estimation in Partly Linear Models”, *journal of multivariate analysis*, 60,1-19.

[18] Hua Liang and Runze Li (2009), “Variable Selection for Partially Linear Models With Measurement Errors”, *Journal of the American Statistical Association*, 104:485,234-248.

[19] Xiao Ni, Hao Helen Zhang, and Daowen Zhang, “Automatic model selection for partially linear models”, *Journal of Multivariate Analysis*, 100(2009)2100-2111.

[20] Lei Yang, Yixin Fang, Junhui Wang and Yongzhou Shao (2017), “Variable selection for partially linear models via learning gradients”, *Electronic Journal of Statistics*, Vol.11, 2907-2930.

[21] Jia Wang, Xizhen Cai and Runze Li (2021), "Variable selection for partially linear models via Bayesian subset modeling with diffusing prior", *Journal of Multivariate Analysis*, Vol.183.

[22] Hao Helen Zhang, Guang Cheng and Yufeng Liu (2011), "Linear or Nonlinear? Automatic Structure Discovery for Partially Linear Models", *Journal of the American Statistical Association*, Vol.106, No.495, pp.1099-1112.