پانزدهمین سمینار احتمال
و فرآیندهای تصادفی
۸ و ۹ شهریور ۱۴۰۴
دانشگاه کردستان

15th Seminar On Probability and Stochastic Processes
August 30-31, 2025
University of Kurdistan

انجمن آمار ایران

Improved Generalized Additive Models in Time Series

Ahad Malekzadeh, Fatemeh Rajabi Naraki *

Department of Computer Science and Statistics, K.N. Toosi University of Technology, Tehran, Iran

Abstract

Generalized additive models (GAMs) have become a powerful and well-established tool in the field of statistics, which demonstrating remarkable performance for determining the model governing the changes of a random variable over time. By implementing GAM on time series data, it is possible to examine the behavior of nonlinear data and predict the future. One of the most important issues in these models is determining the appropriate link function because choosing it incorrectly will lead to unbiased results. In this paper, we intend to present an improved interpretation of the GAM that determines the appropriate link function for a time series dataset. Furthermore, the efficiency and performance of this method are illustrated with real-world examples.

Keywords: Generalized additive model, Link function, Improved generalized additive model, Smooth function, Splines, Penalized likelihood.
Mathematics Subject Classification (2020): 62M10, 91B84

1 Introduction

Data collection and analysis has become a principal foundation of decision-making and planning in various fields in the past decade. With the advancement of technology and increased access to data, organizations and individuals can better utilize this information. Data analysis is now used in most parts of industry and science, including engineering, economics and finance, marketing, medicine and health, etc. Ultimately, this approach has been influential not only in the business world but also in the daily lives of people. In general, data collection and analysis is known as a powerful tool in today's world that has a profound impact on all aspects of human life.

In the meanwhile, time series data performs a significant obligation in data analysis and refers to a set of data that is collected over time and arranged in chronological order. This data usually consists of repeated observations of a particular variable at regular intervals. Time series data allows analysts to identify long-term trends and seasonal or periodic patterns. Using various models such as Autoregressive Integrated Moving Average (ARIMA), Segmented Linear Regression Models, neural networks, etc. more accurate predictions can be achieved,

*Corresponding author: Ftm23rjb@gmail.com

nevertheless a more appropriate approach for this type of data is to use Generalized Additive Models (GAMs).

One of the first approaches that proposed to model nonlinear relationships was the additive model (AM), which was presented by Friedman and Stuetzle (1981), and then Hastie and Tibshirani (1986) invented its generalized form is called GAM and presented it to the world of science, which can fit smooth relationships between variables without knowing the non-linear relationship in advance. Consequently, several researchers, including Wood (2017), Yee (2015), James et al. (2013), etc., investigated and developed the infrastructure of GAMs. Initially, the additive model was defined for observations y_i and predictor variables $x_1, ..., x_p$ as follows,

$$AM : E(y_i|x_j) = \sum_{j=1}^{p} f_j(x_{ij}), \tag{1}$$

where the terms f_j denote non-parametric, smooth functions for the independent variable that are fitted to the data. As indicated, after defining the additive model, a generalized version of the model was obtained by applying a function to the expected value $E(y_i|x_j)$. The following model determines the structure of a nonparametric GAM in general,

$$GAM : S[E(y_i|x_j)] = S(\mu_i) = \sum_{j=1}^{p} f_j(x_{ij}), \tag{2}$$

where $\mu = E(Y)$ denotes the expected value, S represents a function that links the expected value to the predictor variables $x_1, ..., x_p$, which is known in most contexts as the canonical link function with the symbol $g(.)$ and note that an exponential family distribution for y_i (e.g. normal, binomial, or poisson distributions) is specified.

Mainly, GAMs endure in both parametric and nonparametric methods and are utilized in various data scenarios such as when the variable is independent, time series, or has a lag. GAMs were introduced in time series epidemiology studies by Schwartz (1993) and have since been used and developed by many as a standard approach to solving the control variable problem (Schwartz (1994b), Bremner et al. (1999), Dominici et al. (2000)). To illustrate the importance of using GAMs in time series data, we can cite studies that have been conducted in this field, for instance, the impact of air pollutants on human health has been investigated by using GAM by Ravindraa (2019), or to describe the physiology of heart and lung interaction, using GAM to decomposition time series and waveforms has been discussed by Enevoldsen et al. (2023), and after the COVID-19 pandemic spread around the world, many studies have been conducted on the diagnosis of disease incidence, mortality rates, and vaccination, such as: A extension of the GAM in spatial-temporal way for modeling COVID-19 mortality risk in Toronto was investigated by Feng (2022), and Eckardt et al. (2024) wrote their paper on the use of generalized functional additive mixed models with compositional variables for COVID-19 prevalence curves.

At the present time, we intend to address the issue of selecting the appropriate link function utilizing GAM on time series data. Since GAMs can work with different types of statistical distributions (e.g., normal, binary, Poisson, etc.), the link function is chosen depending on the type of distribution. For example, for binary data (e.g., success/failure), the logit function is usually used, or for count data, the Poisson (log) link function may be used. Using these functions, GAMs provide the ability to model complex, nonlinear relationships, helping analysts to derive more accurate results from the data. The correct choice of link function is of great importance depending on the type of data and the type of distribution of the dependent variable and can have a significant impact on the accuracy and interpretability of the model. In addition, if the link function is misspecified, it will provide biased and uncertain results. In this article, we intend to address an improved version of GAM (IGAMs) on time series data that considering the issue of selecting the appropriate link function.

The rest of the paper is as follows. In Section 2, we explain the theory of IGAMs when the data are time series, and discuss the select of the link function and parameter estimation.

In Section 3, we present an example from the COVID-19 study which we fit these models and measure the model selection criteria for them. Eventually, in Section 4, we conclude and discuss.

2 Model specification

The GAM presented by Hastie and Tibshirani (1986) is the classical addendum of general linear models (GLMs) (introduced by Nelder and Wedderburn (1972)) that replaces the linear predictor $\sum \beta_j x_{ij}$ with the additive predictor term $\sum f_j(x_{ij})$. In this section, we explain the use of GAMs in time series studies and attempt to provide explanations about introducing the model, estimating parameters, and choosing the appropriate link function.

2.1 Time Series by using GAM

GAM has been introduced as an effective and flexible approach for performing nonlinear regression analysis in time series studies. In time series studies, GAM has been the most widely used method because it allows nonparametric adjustments for nonlinear seasonal confounding effects, trends, and other covariate variables. These models have been broadly used in many analyses that are time-dependent, including climate change, health and medical data, product price changes, and inflation, etc. The GAM for a trend in a series of T observations y_t at observation times x_t $(t = 1, 2, ..., T)$ is defined as follows,

$$E(y_t|x_t) = \mu_t = S(\sum_{j=1}^{p} f_j(x_{t-j})), \tag{3}$$

where S is true unknown response function that estimated non-parametrically and f_j is smooth function. Solutions for estimating the true response function have been introduced, including the use of kernel functions by Hardle et al. (1993) and the use of P-spline approach by Muggeo and Ferrara (2008). To ensure the uniqueness and identifiability of the model, several constraints need to be imposed on the model. We assume that for a fixed value of $c > 0$ the following constraints are required:

$$\sum_{j=1}^{p} \int_{min(x_j)}^{max(x_j)} f_j(t)^2 dt = c, \tag{4}$$

furthermore, each predictor function is postulated to be (the usual conditions used to obtain identifiability in GAMs):

$$\int_{min(x_j)}^{max(x_j)} f_j(t) dt = 0, \tag{5}$$

in addition, we suppose that the response function is monotonically increasing, i.e.

$$\frac{\partial h_T(t)}{\partial t} \geq 0. \tag{6}$$

The main reason for imposing these restrictions is to achieve interpretable results. For non-monotonic response functions, the interpretation of the functions $fj()$ becomes vastly difficult, because an increasing (decreasing) function $fj()$ does not translate into an increasing (decreasing) effect on the response.

2.2 Link function and Estimation procedure

As previously mentioned, within the context of GLMs and GAMs, the inverse of the response function is generally called the link function $(g = S^{-1}(.))$, which is approximated by a combination of functions of the form $S = h_0(h(.))$, where $h_0()$ is a constant response function (usually the canonical response function). In this manner, the inner unknown function $h()$

must be estimated, which is characterized by expanding the basis into the $h(\eta_i) = \mathbf{\Phi}^T(\eta_i)\boldsymbol{\alpha} = \sum_{k=1}^{m_k} \phi_k(\eta_i)\alpha_k$, where η_i is the vector of the m_h B-spline basis functions evaluated at η_i and α is the basis coefficient vector corresponding to the inner function. We also use a basis expansion of B-spline written by $f_j(x) = \boldsymbol{\psi}_j^T(x)\boldsymbol{\beta}_j, j = 1, \ldots, p$, to estimate the unknown functions $f_j()$ where $\psi_j(x)$ is the vector of the m_j basis functions evaluated at x and β_j is the corresponding coefficient vector. Finally, by substituting the above expressions into equation 3, we obtain the following model estimation as IGAM,

$$IGAM : E(y_t|x_t) = \mu_t = h_0(\mathbf{\Phi^T}(\sum_{j=1}^{p} \psi_{ij}^T(x)\beta_j)\boldsymbol{\alpha}). \tag{7}$$

Various methods have been proposed for estimating GAM parameters, initially Hastie and Tibshirani (1986) used the local scoring algorithm which is Fisher score update using local score estimation and then, Wood (2004) showed that parameters can be estimated using a penalized likelihood approach. Here, for estimating IGAM parameters, based on what was obtained in Tutz and Petry (2016), present an algorithm based on boosting techniques, where each boosting iteration consists of two steps (1) updating the response function with a fixed prediction and (2) updating the prediction with a fixed response function. Both updates are based on penalized and constrained Fisher scoring. (See Tutz and Petry (2016) for more detailed)

3 Application to the data

Following the global pandemic of the coronavirus disease (COVID-19), some scientists and researchers have been able to gather a lot of evidence about the virus and learn more about how it is transmitted and how to treat it more effectively. With the number of hospitalizations for the coronavirus increasing daily, especially in 2021, it was necessary to study this data to detect and predict the peak of the disease in order to maintain global health around the world. It should be considered that vaccination has saved millions of lives during the COVID-19 pandemic. Within a year of the start of the pandemic, several vaccines against the virus were developed. This speed was possible due to previous research on coronavirus vaccines, increased attention and funding for research and development, and a high interest in volunteering for clinical trials. Here, we intend to examine the doses of Covid-19 vaccine administered in several countries such as China, Iran, France, etc., based on the data available on the Global Vaccination website OurWorldinData. Here we are going to examine the doses of COVID-

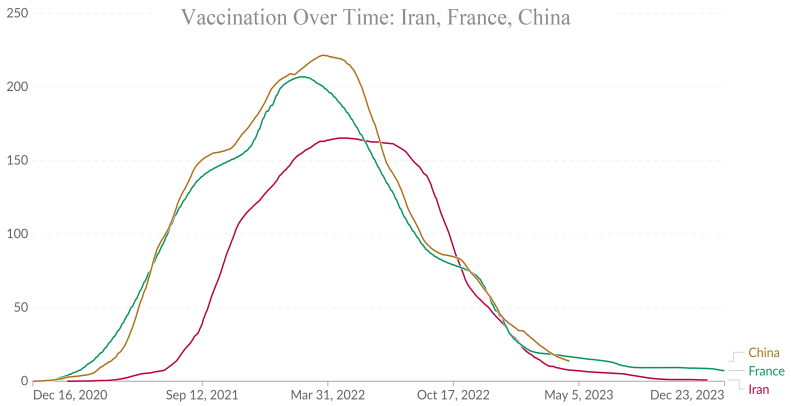


Figure 1: Per 100 population. The value shown for each date is the total number of vaccine doses administered in the 12 months prior to that date for the three countries China, Iran, and France.

19 vaccine administered in each days per 100 people in some countries such as China, Iran,

Italy, etc. in a two-year period from 2021 to 2023 based on the data available on the Global Vaccination Website.The data for these three regions are given in Figure 1. We have fitted the Additive model(AM), the generalized Additive model(GAM) and the improved generalized Additive model(IGAM) on the data and measured the Root Mean Squared Error (RMSE), the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC) and the R-squared (R^2) criteria and we construct its values separately for each region in Table 1.

Table 1: Criteria for three processes AM, GAM and IGAM

	AM	GAM	IGAM
China	RMSE = 74.507463	RMSE = 4.982501	RMSE = 4.480922
	AIC = 9815.482	AIC = 5200.566	AIC = -2500.448
	BIC = 9829.739	BIC = 5252.811	BIC = -2448.233
	R^2 = 0.01122333	R^2 = 0.9955313	R^2 = 0.9983965
Iran	RMSE = 63.250842	RMSE = 4.031698	RMSE = 2.771918
	AIC = 11371.87	AIC = 5766.393	AIC = -1831.283
	BIC = 11386.65	BIC = 5820.581	BIC = -1777.16
	R^2 = 0.03341473	R^2 = 0.9960379	R^2 = 0.9967077
France	RMSE = 64.178143	RMSE = 5.775090	RMSE = 5.155585
	AIC = 12182.87	AIC = 6944.363	AIC = -2134.83
	BIC = 12197.86	BIC = 6999.267	BIC = -2079.966
	R^2 = 0.1346133	R^2 = 0.9929344	R^2 = 0.9950166

Based on what is given in Table 1, for data of all regions, it is clear that the AM is not a good fit and not suitable at all and should not be used for such data. While both GAM and IGAM approaches perform relatively well, their R^2 value is close to 1, which proofs this. However, the IGAM is a better fit and the lower values of the RMSE, AIC and BIC confirm it.

Discussion and Results

As has been shown in numerous papers over the past decades, GAM is a significant method for modeling nonlinear process data that can be predicted without the need for an analyst in the form of a nonlinear relationship. Here, we improve GAM (IGAM) to provide a more appropriate approach by properly selecting the link function, which is one of the main pillars of the model that needs to be estimated. We also analyzed a real-world application, data on the dose of the vaccine administered for COVID-19, and we saw the positive performance of IGAM in the example results given.

References

Bremner S. A., Anderson H. R., Atkinson R. W., McMichael A. S., Strachan D. P., Bland J. M. and Bower J. S. (1999), Short term association between outdoor air pollution and mortality in London 1992–4, *Occupational and Environmental Medicine*, 56, 237–244.

Dominici F., Samet J. and Zeger S. L. (2000), Combining evidence on air pollution and daily mortality from the largest 20 US cities: a hierarchical modeling strategy, *Journal of the Royal Statistical Society*, 163, 263–84.

Eckardt M., Mateu J. and Greven S. (2024), Generalized functional additive mixed models with (functional) compositional covariates for areal Covid-19 incidence curves, *Journal of the Royal Statistical Society Series C*, Applied Statistics, 2024, 73, 880–901.

Enevoldsen J., Simpson G.L. and Vistisen S.T. (2023), Using generalized additive models to decompose time series and waveforms, and dissect heart–lung interaction physiology, *Journal of Clinical Monitoring and Computing*, 37:165–177

Feng C. (2022), Spatial-temporal generalized additive model for modeling COVID-19 mortality risk in Toronto, Canada, *ScienceDirect*, Spatial Statistics 49 (2022) 100526.

Friedman J.H. and Stuetzle W. (1981), "Projection Pursuit Regression", *Journal of the American Statistical Association* 76:817–823.

Hardle W., Hall P. and Ichimura H., (1993), *Optimal smoothing in single-index models*, Ann. Statist. 21, pp. 157–178.

Hastie T.J. and Tibshirani R.J. (1986), Generalized Additive Models, *Statistical Science*, 3(1), 297-318.

Hastie T.J. and Tibshirani R.J. (1990), *Generalized Additive Models*, Chapman and Hall, London.

James G., Witten D., Hastie T. and Tibshirani R. (2013), *An Introduction to Statistical Learning*, springer, New York, 112.

Muggeo V.M.R. and Ferrara G., (2008), Fitting generalized linear models with unspecified link function: A p-spline approach, Comput. Statist. Data Anal. 52 , pp. 2529–2537.

Nelder J. and Wedderburn R. (1972), Generalized Linear Models, *Journal of the Royal Statistical Society*, Series A (General), Blackwell Publishing, 135(3), 370–384.

Ravindraa K., Rattana P., Morb S. and Aggarwal A.N. (2019), Generalized additive models: Building evidence of air pollution, climate change and human health, *ScienceDirect* Environment International 132 104987.

Schwartz J. (1993), Air pollution and mortality in Birmingham, AL, *American Journal of Epidemiology*, 137, 1136–1147.

Schwartz J. (1994), Nonparametric smoothing in the analysis of air pollution and respiratory illness, *The Canadian Journal of Statistic*, 22, 471–487.

Tutz G. and Petry S. (2016), Generalized additive models with unknown link function including variable selection, *Journal of Applied Statistics*

Wood, S. N. (2004), *Generalized Additive Models: an introduction with R*, first ed, Chapman and Hall/CRC, New York.

Wood, S. N. (2017), *Generalized Additive Models: an introduction with R*, 2nd ed, CRC Press.

Yee T.W. (2015), *Vector Generalized Linear and Additive Models with an Implementation in R*, Springer.