Statistica Sinica Preprint No: SS-2024-0371

Title	On a Flexible Generalized Model Averaging Forecasting of Nonlinear Time Series
Manuscript ID	SS-2024-0371
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0371
Complete List of Authors	Rong Peng, Zudi Lu and Fangsheng Ge
Corresponding Authors	Zudi Lu
E-mails	zudilu@cityu.edu.hk
Notice: Accepted author version.	

On a Flexible Generalized Model Averaging Forecasting of Nonlinear Time Series

Rong Peng[1], Zudi Lu[2] and Fangsheng Ge[3]

[1]*Hunan University,* [2]*City University of Hong Kong*

[3]*University of Southampton*

Abstract: In nonlinear time series analysis, forecasting is fundamental but challenging with the curse of dimensionality for nonparametric regression of multiple lagged variables and the nonlinear/non-Gaussian features for response either continuous or discrete-valued. To address the challenges, we propose a unified framework of semiparametric Generalized MArginal Forecast Model Averaging (GMAFMA) under a flexible conditional exponential family of distributions for nonlinear forecasting of time series. This framework will not only overcome the curse of dimensionality with nonparametric forecasting but also flexibly adapt for both continuous- and discrete-valued non-Gaussian time series data, bridging the gap in existing methods for nonlinear forecasting. The GMAFMA procedure is developed by a semiparametric conditional likelihood method for estimation of the combining weights of the marginal forecasts, with asymptotic normality established under mild time series data generating conditions. Furthermore, an adaptively penalized GMAFMA (PGMAFMA) is suggested to find the most important marginal forecasts so that the forecasting is more interpretable and precise. The procedures are supported both by Monte Carlo simulations and various empirical applications, such as forecasting of the

Corresponding author: Zudi Lu, Department of Biostatistics, City University of Hong Kong, Hong Kong SAR, China. Email: zudilu@cityu.edu.hk

number of strike events in labor economics and the FTSE 100 index market moving direction in finance, and assessment of the causal effect of seatbelt law in reducing the road casualties.

Key words and phrases: Nonlinear/non-Gaussian time series, Generalized marginal forecast model averaging (GMAFMA), Penalized GMAFMA, Exponential family, Semi-parametric smoothing

1. Introduction

Forecasting is a fundamental but challenging task in many applications, particularly when it comes to large dimension for time series data that exhibit *non-linear and/or non-Gaussian dependence features* (Terasvirta, Tjøstheim and Granger, 2010). For *continuous-valued* time series data, non- and semi-parametric methods are found popular in literature (Fan and Yao, 2003; Gao, 2007). However, their counterparts of statistical inference for *discrete-valued* time series are still rare (c.f., Weiß (2018)), even though such data exist extensively. See, e.g., Cameron and Trivedi (1986), Winkelmann and Zimmermann (1994), Cameron and Trivedi (1996), Cameron and Trivedi (1998), Qaqish (2003), Cameron and Trivedi (2005), Weiß (2018), Fokianos et al. (2021) and Yu, Tang and Shi (2025), for the relevant developments under parametric frameworks. To bridge the gap, in this paper, we propose a *flexible framework of semiparametric Generalized MArginal Forecast Model Averaging* (GMAFMA) under a general conditional exponential family of distributions for nonlinear forecasting of dynamic series.

Specifically, our proposed framework will meet the challenges not only to overcome the curse of dimensionality but also adapt for both continuous- and discrete-valued non-Gaussian time series in semiparametrically dynamic nonlinear forecasting.

Under time series continuous-valued data, the literature on non- and semi-parametric approaches is very extensive. The reader is particularly referred to Fan and Yao (2003), Gao (2007), Terasvirta, Tjøstheim and Granger (2010), and more recent references such as Lu et al. (2009),bGao, Wang and Yin (2013), Gao et al. (2015), Dong, Gao and Linton (2023), Zhou et al. (2024), among others. However, for discrete-valued response time series, the studies still mostly focus on linear parametric models, such as the INAR, INARMA, and INGARCH models (Drost, Van den Akker and Werker, 2009; Fokianos, Rahbek and Tjøstheim, 2009; Weiß and Schnurr, 2023) that extend the AR, ARMA, and GARCH framework to integer-valued data. Further, in the field of forecasting aggregation or averaging, for example, Qaqish (2003) proposed a conditional linear family to generate correlated count data, and Yu, Tang and Shi (2025) proposed a class of aggregated forecast method for exponential family panel data. See also the recent review by Fokianos et al. (2021) and Weiß (2018). Unfortunately, investigations into nonparametric approaches for discrete valued time series data that we need in this paper are very rare (c.f., Li and Racine (2003); Zhang, Lu and Zou (2013)). To the best of our knowledge, even the popular local linear maximum likelihood estimation (Fan, Farmen and Gijbels, 1998) has only recently been developed under a general conditional exponential family of distributions for time series data by ourselves (Peng and Lu, 2023). Such nonparametric approaches are flexible both for discrete and continuous valued data, but suffer from severe dimensionality for nonlinear forecasting of time series data.

To get across the challenging dimensionality with nonparametric approach but fully utilizing its flexibility and adapt for discrete-valued response in dynamic nonlinear forecasting, this paper will make significant contributions in threefold.

Firstly, a flexible framework, namely the GMAFMA scheme, is proposed under a general conditional exponential family of distributions for dynamic forecasting of nonlinear time series. To our knowledge, it is the first attempt to flexibly combine both advantages of marginal regression model averaging (Li, Linton and Lu, 2015) and exponential family of distributions (McCullagh and Nelder, 1989) adapting to nonlinear forecasting for non-Gaussian continuous and/or discrete-valued time series data with identified nonlinear marginal features in a unified manner, extending the popular generalized linear model (GLM) for forecasting.

More specifically, on one hand, this scheme extends the idea of marginal regression model averaging (MARMA) of Li, Linton and Lu (2015) adapting to discrete-valued data. By applying the "separate and conquer" strategy, this scheme combines the nonparametric low-dimensional marginal regressions to approximate a large-dimensional regression. Li, Linton and Lu (2015) have developed a least squares based approach (hereafter denoted LS-MARMA), which has shown to be useful in many applications for time series with continuous-valued response data (c.f., Chen et al. (2016, 2018) and Chen, Li and Linton (2019)). Model averaging has become a useful tool to deal with model uncertainty (Hansen, 2007; Steel, 2020)) by combining different models, instead of model selection, among many working models, and has received increasing attention in recent years for i.i.d. data. See, for example, Li et al. (2018) for varying coefficient models, Li et al. (2022) for multinomial logistic models, and Racine et al. (2023) for i.i.d. mixed-data spline regression, among others. Further, Liao et al. (2021) extends it to linear infinite order autoregressive (AR(∞)) time series process. On the other hand, our scheme fully utilizes the flexible exponential family of distributions, which has been popular in generalized linear and additive modeling adapting for count-valued data (McCullagh and Nelder, 1989). This paper will also further extend Peng and Lu (2024) that is only for a binary valued time series classification and Zhang, Lu and Zou (2013) for a parametric model averaging under finite categories, both of which clearly cannot apply to the count-valued time series, such as Poisson autoregression analysis (Fokianos, Rahbek and Tjøstheim, 2009). Thus, by combining both advantages in flexible semiparametric model averaging and exponential family of distributions, our GMAFMA framework enjoys adapting to discrete/continuous-valued time series forecasting flexibly. It, therefore, bridges the gap in the existing literature.

Secondly, we propose an adaptive LASSO penalized scheme for an improved GMAFMA forecasting. In many situations of practice, a further challenge for multiple large-dimensional forecasting is how to improve interpretability and avoid "poor generalization ability", due to overfitting, especially when the data is limited. With irrelevant marginal forecasts included, the model will produce additional errors in estimation and forecasting, and it may make it hard to understand the truly important marginal forecasts. We have thus suggested a unified scheme for the GMAFMA framework by using our first-step estimation to adaptively select the important marginal forecasts to reduce the dimensionality via penalization. Zou (2006) and Zhao and Yu (2006) have proved that the conventional LASSO technique yields the consistent estimator under necessary conditions, while adaptive LASSO, as its extension, further enjoys the oracle property, with optimal convergence properties as if the true variables were known, in the case of independent data. Extending Al-Sulami et al. (2019) who studied the adaptive LASSO of Zou (2006) for spatial time series lag estimation, our proposed penalized GMAFMA (PGMAFMA) scheme hence extends the adaptive LASSO to nonlinear time series forecasting that allows for discrete-value data of unknown forms under the dependence of mild β-mixing condition. It will help to improve interpretability by extracting important information on marginal forecasts from multidimensional data, and, therefore, to provide potentially better predictions. We also further extend Chen et al. (2018) in penalized averaging, allowing both for continuous- and discrete-valued data.

Thirdly, we have investigated the performances of the proposed GMAFMA and PGMAFMA procedures for nonlinear time series forecasting. Numerical examples include Monte-Carlo simulations for binary data and count data, and real data implementations for the number of US manufacturing strike events in labor economics, FTSE100 market price moving direction in finance, and analysis of the causal effect of the UK seatbelt law in reducing the death of van drivers in health economics. Interestingly, our proposed GMAFMA outperforms the popular generalized linear model (McCullagh and

Nelder, 1989) in forecasting, and the PGMAFMA can even beat such popular machine learning methods as random forest (RF) and gradient boosting machines (GBM) in these examples.

The remainder of this paper is structured as follows. Section 2 first introduces the framework of generalized marginal forecasting model average (GMAFMA) in population for time series of conditional exponential family of distributions. Estimation by maximum conditional likelihood based model averaging and its asymptotic property are developed in Section 3. The penalised GMAFMA (PGMAFMA) is then suggested by applying an 'adaptive LASSO' penalty to penalize the unimportant marginal forecasts in Section 4, with asymptotic oracle properties developed. Further discussions on extension to covariates involving discrete-valued (lagged response) variables and to marginal forecasts of low-dimensional interactions are offered in Section 5. Numerical examples are demonstrated in Section 6 by simulations and applications to forecasting of strike events in labour economics, market moving direction in finance, and road casualty in health economics. Conclusions are put in Section 7. Technical details are given in the Supplementary Materials online.

2. A generalized marginal forecasts model averaging

In time series analysis, it is a usual practice to consider a stationary series with a non-stationary one turned by differencing operation (Box, Jenkins and Reinsel, 1994). We hence consider our model framework for a stationary time series process (Y_t, X_t^T)

that satisfies β-mixing dependence condition, with Y_t a response variable at time t, and $X_t = (x_{1t}, \ldots, x_{dt})^T$ a d-dimensional vector series representing the historical lagged information of Y_t and other relevant covariates available up to time $(t-1)$. Here X^T stands for the transpose of a vector (or matrix) X. Formally, the β-mixing property can be explicitly expressed as follows:

Definition 1. Let $Z_t = (Y_t, X_t)$ be a strictly stationary time series. The process Z_t is said to be β-mixing if

$$\beta(n) = E\left\{\sup_{B \in \mathcal{F}_{t+n}^{\infty}} |P(B) - P(B|Z_t, Z_{t-1}, \ldots)|\right\} \to 0$$

as $n \to \infty$, where $\mathcal{F}_{t+n}^{\infty}$ is the information field (also-called σ-algebra) generated by $\{Z_s, s \geq t+n\}$.

This β-mixing property is a useful concept in nonlinear time series analysis, which is easily satisfied when a process is of geometric ergodicity (Tjøstheim, 1990; An and Huang, 1996; Lu, 1998). We require this for the needed empirical process theory due to Doukhan, Massart and Rio (1995). Here the dimension d of X_t can be large because it may involve the long-lagged information of all the relevant variables in time series analysis. This setting is similar to that in our previous work of Li, Linton and Lu (2015) and Chen et al. (2018) for time series data with Y_t continuous-valued. But differently, in this paper, Y_t can be either discrete- or continuous-valued. To adapt for this generality, we consider that Y_t as a response random variable, given X_t, follows a general *conditional* exponential family of distributions.

2.1 Conditional exponential family of distributions

For generality, for any given set of past information available up to time $t-1$, denoted by I_{t-1}, we assume a generic form for a *conditional* exponential family (CEF) of distributions for Y_t given I_{t-1}, expressed by:

$$\mathfrak{m}_Y(y|I_{t-1}) = \mathfrak{m}_Y(y, \theta_t) \equiv \exp(y\theta_t - \psi(\theta_t) + \phi(y, \Theta_t)), \tag{2.1}$$

where $\psi(\cdot)$ and $\phi(\cdot)$ are known functions for a particular distribution family (c.f., McCullagh and Nelder (1989)). Here note that θ_t and Θ_t in (2.1) are the canonical and nuisance parameters, respectively, depending on (i.e., being functions of) the given information I_{t-1}, the functional forms of which are however unknown. As usual in the generalized linear model (GLM) literature (McCullagh and Nelder, 1989), the nuisance Θ_t in (2.1) is unnecessary to consider below, so we only specify $\mathfrak{m}_Y(y, \theta_t)$ relating to θ_t without containing Θ_t for the right hand side of (2.1). Further, note that in the continuous-valued case of Y_t, $\mathfrak{m}_Y(y|I_{t-1})$ is the conditional probability density function of Y_t given I_{t-1}. While in the discrete-valued case of Y_t, $\mathfrak{m}_Y(y|I_{t-1})$ becomes the conditional probability function of Y_t given I_{t-1}, that is $\mathfrak{m}_Y(y|I_{t-1}) = P(Y_t = y|I_{t-1})$. The widely applied Bernoulli, Binomial and Poisson distributions are just examples for the discrete valued case.

We make some remarks. Firstly, we highlight that for the information I_{t-1} in model family (2.1), we will specify it explicitly either in terms of the information of a certain component or the whole of the regressor vector $X_t = (X_{1t}, \cdots, X_{dt})^T$ in applications below. For example, if I_{t-1} represents the information of the partial, say the j-th component X_{jt}, of X_t, then it is a problem of marginal regression for forecasting of Y_t given X_{jt}, where $j = 1. \cdots, d$, while if I_{t-1} is for the whole of X_t, it is the forecasting that we are concerned with but may suffer from a curse of dimensionality for a large d. In this paper, we are interested in how to combine the marginal forecasts of Y_t given X_{jt}, for $j = 1. \cdots, d$, to optimally approximate the forecast of Y_t given X_t with the idea of model combination (averaging) applied. Secondly, now when we are based on X_t to represent I_{t-1}, our forecast model in (2.1) can be expressed via a link function $\eta(\cdot)$, with $\theta_t = \eta(\mu_t)$, where, letting $\dot{\psi}(\theta_t)$ be the derivative of $\psi(\theta_t)$, $\mu_t = E(Y_t|X_t) = \dot{\psi}(\theta_t)$ is the conditional mean that is to be estimated in forecasting. With $\psi(\cdot)$ and $\phi(\cdot, \cdot)$ known in (2.1), the function $\eta(\cdot) = (\dot{\psi})^{-1}(\cdot)$ is a known (canonical) link function as done in the traditional generalized linear model (GLM, McCullagh and Nelder (1989)), which is *not* a nonparametric link here. It is worth noting that under the GLM, the contribution of X_t to forecasting Y_t is modelled (via a known link $\eta(\cdot)$) with $\theta_t = \eta(E(Y_t|X_t))$ being a linear function of X_t, which may however be rather restrictive in many applications. In view of the fact that $\theta_t = \eta(E(Y_t|X_t))$ may be an (unknown) *nonlinear* function of X_t (though $\eta(\cdot)$ is a known link), we will model it *nonparametrically* (i.e., $\mu_t = E(Y_t|X_t)$ actually being nonparametric). Therefore, our nonparametric component lies in the *semiparametric* modelling of the forecast of Y_t given X_t with $\theta_t = \eta(\mu_t)$ being a nonparametric function of X_t, while the link function $\eta(\cdot) = (\dot{\psi})^{-1}(\cdot)$ itself is *known* for a particular (parametric) *conditional* distribution of Y_t given X_t with $\psi(\cdot)$ known under model family (2.1). Thirdly, in the case where $\psi(\cdot)$ is unknown in model (2.1), or the (parametric) conditional exponential family may be wrong, with a nonparametric link for $\eta(\cdot) = (\psi')^{-1}(\cdot)$ required, it is however not a simple matter, beyond the scope of this paper, but worth further investigation in future.

Then, based on the given information I_{t-1} expressed by $X_t = (X_{1t}, \cdots, X_{dt})^T$, we have our generalized nonlinear regression:

$$\theta_t = \eta(\mu_t) = f(X_t), \tag{2.2}$$

where $f : R^d \to R$ is the unknown function we need to estimate, by which estimation of μ_t follows from (2.2). If f is a linear function, the model reduces to the popular time series generalized linear model (GLM) (Fokianos et al., 2021). However, when the functional form of f is unknown, it is more challenging. Non- and semi-parametric approaches to estimation of f in (2.2), though studied widely for independent and identically distributed (i.i.d.) data in the literature (c.f., Fan, Farmen and Gijbels (1998)), are, however, rare for time series count data, under conditional exponential family of distributions only done recently (c.f., Peng and Lu (2023)).

From the prediction perspective, we want to estimate the regression function $\mu_t = E(Y_t|X_t = (x_{1t}, \ldots, x_{dt})^T)$ as a forecast of Y_t. This is easily implemented by nonparametric methods via (2.2) if d is small (c.f., Fan, Farmen and Gijbels (1998) and Peng and Lu (2023)). However, a common scenario for X_t is that when considering more time series lag information in modeling and forecasting, the dimension d of X_t is large.

The accuracy of such estimation tends to deteriorate for a large dimension d, which was highlighted by Stock and Watson (2006) with the idea of model averaging applied based on a parametric GLM setting. This fact is well known with continuous-valued Y_t that the nonparametric estimation of μ_t suffers from curse of dimensionality (c.f., Fan and Yao (2003), Gao (2007), Terasvirta, Tjøstheim and Granger (2010)). Furthermore, it leads to an exponential increase in the costs for computation with large-dimensional covariate space. For instance, even for the popular generalized additive model (GAM), the computational cost may become large and it works poorly in prediction owing to overfitting if the dimension d of X_t is large; see the numerical performance of additive model forecasting in Li, Linton and Lu (2015), Chen et al. (2018) and Peng and Lu (2024).

2.2 Generalized marginal forecast model averaging approximation

When the information of I_{t-1} in (2.1) is expressed by a high dimensional $X_t = (X_{1t}, \cdots, X_{dt})^T$, a direct estimation of $\mu_t = E(Y_t|X_t)$ via (2.2) for forecasting of Y_t suffers from curse of dimensionality. In practice, a simple and easily applied idea is to use marginal information to do forecasting following the idea of "separate and conquer" in machine learning. For example, with I_{t-1} in (2.1) represented by X_{kt} only, that is, $\mathfrak{m}_Y(y, \theta_{kt}) \equiv \exp(y\theta_{kt} - \psi(\theta_{kt}) + \phi(y, \Theta_{kt}))$ with θ_{kt} and Θ_{kt} being functions of X_{kt}, estimation of $\mu_{kt} = E(Y_t|X_{kt})$ for forecasting of Y_t can be easily done by generalized marginal nonparametric regression (c.f., Peng and Lu (2023)), as given similarly in

(2.2):

$$\theta_{kt} = \eta(\mu_{kt}) = f_k(X_{kt}), \tag{2.3}$$

where $\eta(\cdot) = (\dot{\psi})^{-1}(\cdot)$ is known, and $f_k(X_{kt})$ is an unknown nonlinear function of X_{kt} that can be easily estimated (Peng and Lu, 2023), more precisely defined right after (3.2) below.

By the model averaging idea (c.f., Hansen (2007); Li, Linton and Lu (2015)), we propose combining the lower-dimensional marginal forecasts given in (2.3) for $k = 1, \cdots, d$, to approximate (2.2):

$$\begin{aligned}\theta_t = \eta(\mu_t) = f(X_t) &\approx \alpha_0 + \alpha_1 \eta(\mu_{1t}) + ... + \alpha_d \eta(\mu_{dt}) \\ &= \alpha_0 + \alpha_1 f_1(x_{1t}) + ... + \alpha_d f_d(x_{dt}) \equiv f_t^{MA} \equiv \theta_t(\boldsymbol{\alpha}),\end{aligned} \tag{2.4}$$

where $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_d)^T$ are the unknown model averaging coefficients to be defined. Note that we do not require $\sum_{j=1}^d \alpha_j = 1$ in (2.4) with α_0 for intercept, more conveniently in computation than the usual model average constraints that $\sum_{j=1}^d \alpha_j = 1$ with α_j's non-negative. Here the model averaging lies in the model combination of the marginal forecasts, $f_k(x_{kt})$'s, in (2.4) in a more general sense following Li, Linton and Lu (2015) that extends, while inspired by, Hansen (2007) on model average. See Li, Linton and Lu (2015) for more discussions.

We first give a proposition, showing that under the conditional independence of $X_t = (X_{1t}, \cdots, X_{dt})^T$ given Y_t for model (2.1), equation (2.4) can actually hold equally for suitably chosen α_0 and α_j, $j = 1, \cdots, d$. To simplify the discussion, we consider the most of the situations for exponential family where $\phi(y, \Theta_t) \equiv \phi(y)$ is independent of the nuisance parameter Θ_t in model (2.1), which include the popular Binomial and Poisson distributions in our simulation and real data examples in Section 6 below. A similar conclusion under a special case of Bernoulli distribution was established by Fang, Li and Xia (2022) with *i.i.d.* data. The proof of this proposition is relegated in Appendix A.2.1.

Proposition 1. Suppose that model (1) holds for any given information set I_{t-1} which may represent the information of $X_t = (X_{1t}, \cdots, X_{dt})^T$ and that of its component X_{jt}, $j = 1, \cdots, d$, respectively, and that $X_{1t}, \cdots, X_{dt}$ are conditionally independent given Y_t. Further assume that $\phi(y, \Theta_t) \equiv \phi(y)$ is independent of the nuisance parameter Θ_t in model (1). Then $\eta(\mu_t) = \alpha_0 + \alpha_1 \eta(\mu_{1t}) + \cdots + \alpha_d \eta(\mu_{dt})$ with $\eta(\cdot) = (\dot{\psi})^{-1}(\cdot)$ and $\alpha_1 = \cdots = \alpha_d = 1$ and $\alpha_0 = c$, where $\mu_t = E(Y_t|X_t)$ and $\mu_{jt} = E(Y_t|X_{jt})$ for $j = 1, \cdots, d$, and c is a constant.

In general, for time series data, the conditional independence of $X_t = (X_{1t}, \cdots, X_{dt})^T$ given Y_t may not hold and the conclusion in Proposition 1 may fail. We hence seek the approximation in (2.4) the *generalised marginal forecast model averaging* (GMAFMA) to optimally approximate the forecasting of Y_t by $\mu_t \approx \eta^{-1}(\theta_t(\boldsymbol{\alpha}))$ with $\theta_t(\boldsymbol{\alpha})$ defined in (2.4). The *optimal* approximation with $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_d)$ defined in population is in terms of minimizing the Kullback-Leibler (KL) distance, a natural distance function in distributions, from a "true" distribution $\mathfrak{m}_Y(y; \theta_t)$ in (2.1) with θ_t in (2.2), to a "working" approximate distribution $\mathfrak{m}_Y(y_t; \theta_t(\boldsymbol{\alpha}))$ with $\theta_t(\boldsymbol{\alpha})$ in (2.4):

$$KL(\boldsymbol{\alpha}) = 2E\{\log(\mathfrak{m}_Y(y_t; \theta_t)) - \log(\mathfrak{m}_Y(y_t; \theta_t(\boldsymbol{\alpha})))\}. \tag{2.5}$$

Here the KL distance is minimized with respect to $\boldsymbol{\alpha}$, which leads to the optimal approximation in (2.4) with the minimizer denoted by $\boldsymbol{\alpha}^*$ in population, that is

$$\boldsymbol{\alpha}^* = \arg\min_{\boldsymbol{\alpha}\in\mathfrak{A}} KL(\boldsymbol{\alpha}), \tag{2.6}$$

where $\mathfrak{A}$ is a compact subset of R^{d+1}, defined below.

Our proposed GMAFMA adapts to both discrete and continuous-valued non-Gaussian time series forecasting. It is different from the LS-MARMA (marginal regression model averaging) procedure of Li, Linton and Lu (2015) in that the LS-MARMA approximates $E(Y_t|I_{t-1})$ by $\alpha_0 + \sum_{k=1}^d \alpha_k E(Y_t|X_{kt})$ in terms of an $\mathcal{L}_2$ distance. Our GMAFMA extends the MARMA procedure by approximating the generalised regression in (2.2) by (2.4) via (2.6). In fact, in the special case of conditional Gaussianity for Y_t given X_t, a canonical link function of $\eta(\cdot)$ in (2.4) becomes an identity function, in which our GMAFMA reduces to the LS-MARMA.

Remark 1. The GMAFMA model (2.4) can be viewed as the combination of a series of marginal forecasting models. It does not mean that the model is true with equality holding in (2.4). In a more general setting, (2.4) is an approximation device that is defined in terms of (2.6). Such an approximation is essentially utilizing and combining the "weak learners" to become "a strong learner" in the sense of machine learning (c.f. Freund and Schapire (1997)).

Remark 2. Though, for simplicity of statement, we only put one-dimensional conditional forecasts $E(Y_t|X_{kt})$'s in (2.3) and (2.4), we can always add the interaction terms in the marginal forecasts, $E(Y_t|X_{jt}, X_{kt})$'s, for $1 \leq j \neq k \leq d$, into (2.4) , where Y_t is marginally forecasted by the information of X_{jt} and X_{kt}. Essentially, this marginal forecast can be any low-dimensional conditional forecast as addressed above, but due to the computational reason (say, "curse of dimensionality"), one may only consider one-dimensional or two-dimensional conditional marginal forecasts in practice.

Remark 3. As in Li, Linton and Lu (2015), though (2.4) may look like a special form of generalized additive models or GAMs at first glance, the marginal regressions f_k's combined here, as remarked above, can be replaced or added by other low-dimensional marginal regressions. For instance, if we consider a two-dimensional marginal regression $f(X_{jt}, X_{kt})$ in the approximation, it is then not of a usual GAM form. The motivation of affine combination used in (2.4) indeed comes from the model averaging by combining the easily estimable marginal forecasts to approximate the high-dimensional forecast that is hard to be well estimated directly. By doing so, it more easily avoids the shortcomings that the GAM suffers from; see Li, Linton and Lu (2015), Chen et al. (2018), and Peng and Lu (2024) for more discussions on this.

We finally point out that the KL distance (2.5) is used to measure the closeness of the approximation of a "working" distribution with equation (4) to the "true" one with equation (2), which is a useful idea applied in the literature (c.f., Zhang et al. (2016) and Yu, Zhang and Yau (2018) under the parametric model averaging setting).

Further, differently from Li, Linton and Lu (2015) with analytical solutions available for least squares estimation of LS-MARMA (see also Chen et al. (2018)), there are none under the maximum likelihood estimation in our GMAFMA framework and it is more challenging to establish the asymptotic properties for our procedures. We will apply the β-mixing empirical process theory due to Doukhan, Massart and Rio (1995) to get across this difficulty, which is the most helpful for this kind of problem (c.f., Lu, Tjøstheim and Yao (2007) and Peng and Lu (2023)).

3. Estimation for the GMAFMA

We now develop the estimation of the model averaging coefficients for the GMAFMA.

3.1 Estimation

A profile maximum likelihood plug-in method is suggested. First, for the KL-distance to be minimized with respect to $\boldsymbol{\alpha}$ with the (true) minimizer $\boldsymbol{\alpha}^*$ in (2.6), that is equivalently the maximizer of $E\{\log(\mathfrak{m}_Y(y_t;\theta_t(\boldsymbol{\alpha})))\}$ with respect to $\boldsymbol{\alpha}$, we can hence, given the observations $\{(y_t, X_t),\ t = 1, \cdots, n\}$, estimate the minimizer $\boldsymbol{\alpha}^*$ by $\widehat{\boldsymbol{\alpha}}^{*(n)}(\mathbf{f}) = \arg\max_{\boldsymbol{\alpha}} L_n(\boldsymbol{\alpha}, \mathbf{f})$ with

$$L_n(\boldsymbol{\alpha}, \mathbf{f}) = \frac{1}{n}\sum_{t=1}^{n}\left[y_t(\alpha_0 + \sum_{k=1}^{d}\alpha_k f_k(X_{kt})) - \psi(\alpha_0 + \sum_{k=1}^{d}\alpha_k f_k(X_{kt})) + \phi(y_t, \Theta_t)\right]. \tag{3.7}$$

This is a kind of maximum conditional likelihood estimation of $\boldsymbol{\alpha}^*$ given the initial information field I_0 for time series data, and $\widehat{\boldsymbol{\alpha}}^{*(n)}(\mathbf{f})$ is defined depending on $\mathbf{f}(\cdot) =$

$(f_1(\cdot), \ldots, f_d(\cdot))^T$, which we will estimate. Then, with estimated $\widehat{f}_k(X_{kt})$'s on hand, we can replace $f_k(X_{kt})$'s in (2.4) and (3.7) by $\widehat{f}_k(X_{kt})$'s, and get the final estimator of $\boldsymbol{\alpha}^*$ via plug-in by $\widehat{\boldsymbol{\alpha}}^{*(n)} = \widehat{\boldsymbol{\alpha}}^{*(n)}(\widehat{\mathbf{f}})$ with $\widehat{\mathbf{f}}(\cdot) = (\widehat{f}_1(\cdot), \ldots, \widehat{f}_d(\cdot))^T$.

Now we are to estimate the low-dimensional marginal nonparametric functions $f_k(X_{kt})$'s in (2.3), which are unknown (nonparametric) functions allowed to be nonlinear. We apply a nonparametric method to estimate $f_k(X_{kt})$ and the marginal conditional mean $\mu_{kt} = E(Y_t|X_{kt})$ through $\theta_{kt} = \eta(\mu_{kt}) = f_k(X_{kt})$ in (2.3). With Y_t, given X_{kt}, assumed to follow the conditional exponential family of distributions in the form of (2.1) with θ_t replaced by θ_{kt}, a maximum likelihood local linear fitting can be utilized for estimation of $f_k(\cdot)$ in (2.3) (c.f., Fan, Farmen and Gijbels (1998); Peng and Lu (2023)). By the Taylor's expansion of $f_k(X_{kt})$ at an (arbitrary) point x_{k0} where $f_k(\cdot)$ is differentiable and with X_{kt} in the neighbourhood of x_{k0},

$$\begin{aligned} f_k(X_{kt}) &\approx f_k(x_{k0}) + f_k'(x_{k0})(X_{kt} - x_{k0}) \\ &\equiv \beta_1 + \beta_2(X_{kt} - x_{k0}),\ if\ |X_{kt} - x_{k0}| \le h, \end{aligned} \tag{3.8}$$

with h a bandwidth to be appropriately selected. Then the time series local log conditional likelihood is thus given by

$$\ell_{h,x_{k0}}(\beta_1, \beta_2) = \sum_{t=1}^{n} \log \mathfrak{m}_Y(y_t, \beta_1 + \beta_2(X_{kt} - x_{k0}))K_h(X_{kt} - x_{k0}), \tag{3.9}$$

where $K_h(\cdot) = h^{-1}K(\cdot/h)$ with $K(\cdot)$ is a kernel function on $\mathbb{R}^1$. Then we have the estimator $\widehat{f}_k(x_{k0}) = \widehat{\beta}_1$ with $(\widehat{\beta}_1, \widehat{\beta}_2)$ being the minimizer of (3.9) with respect to (β_1, β_2). As usual in nonparametric estimation (Fan and Yao (2003)), the bandwidth h plays an important role for estimation of $f_k(\cdot)$. To simplify the computation, we could take h via, for example, *ThumbBw* function in R package **locpol** following Fan (2018). It appeared to work well for our numerical experiments. The theories for the estimated $f_k(\cdot)$'s we need in (2.4) were developed in Peng and Lu (2023).

Accordingly, we can have the marginal estimates, $\widehat{f}_k(X_{kt})$'s, by taking x_{k0} equal to X_{kt}. Then we can construct estimation of $\boldsymbol{\alpha}$ by minimizing the plug-in log-likelihood

$$\begin{aligned} L_n(\boldsymbol{\alpha}) = L_n(\boldsymbol{\alpha}, \widehat{\mathbf{f}}(\cdot)) = \frac{1}{n}\sum_{t=1}^{n}\{y_t\ (\alpha_0 + \sum_{k=1}^{d}\alpha_k \widehat{f}_k(X_{kt})) \\ - \psi(\alpha_0 + \sum_{k=1}^{d}\alpha_k \widehat{f}_k(X_{kt})) + \phi(y_t, \Theta_t)\}W(X_t). \qquad (3.10) \end{aligned}$$

Practically, to avoid the impact of the poor estimates of $f_k(\cdot)$'s, at extreme boundary values of X_{kt}, on the estimation of $\boldsymbol{\alpha}$ in (3.10), we may exclude the extreme boundary X_{kt}'s as usually done in time series semiparametric modelling (c.f., Masry and Tjøstheim (1995)) by adding a weight function $W(X_t) = \prod_{k=1}^{d} \mathbf{I}_{(c_{k0} \le X_{kt} \le c_{k1})}$ controlling the edge effects in (3.10), where $\mathbf{I}_{(\cdot)}$ is an indicator function and $c_{k0} < c_{k1}$ are appropriately chosen. Clearly, c_{k0} and c_{k1} can be taken sufficiently small and large, respectively, to cover all observations of X_{kt} if the edge effect on estimation is not serious. But, in general, to simplify the computation, we may take c_{k0} and c_{k1} as the 0.01-th and 0.99-th quantiles of the sample $\{X_{kt}, t = 1, 2, \cdots, n\}$, respectively, to avoid extreme edge effect, which seems to work well in our numerical experiments. The uniform convergence of $\widehat{f}_k(x_j)$'s to the true $f_k(x_j)$'s over the compact subset $[c_{k0}, c_{k1}]$ in R^1 (Peng and Lu, 2023) guarantees $\widehat{f}_k(X_{kt})$'s can consistently replace the $f_k(X_{kt})$'s in (3.7), as given in (3.10).

Our GMAFMA procedure is easy to implement in computation. In fact, from the computational perspective, with $\widehat{f}_k(X_{kt})$ made, maximization of (3.10) with equation (2.4) can be easily solved by an algorithm of weighted generalized linear model (GLM) estimation.

3.2 Asymptotic property

The estimator $\widehat{\boldsymbol{\alpha}}^{*(n)}$ maximizing (3.10) with respect to $\boldsymbol{\alpha} \in \mathfrak{A} \subset R^{1+d}$ enjoys nice asymptotic properties, where $\mathfrak{A}$ is a compact set of parameters such that $\boldsymbol{\alpha}^* \in \mathfrak{A}$, say, taking $\mathfrak{A} \equiv \prod_{k=0}^{d}[a_{0k}, a_{1k}]$ with a_{0k} and a_{1k} sufficiently small and large constants, respectively, which differ from the constants of c_{k0} and c_{k1} for $W(\cdot)$ to control the edge effects in (3.10), as discussed above. Correspondingly to (3.10), $\boldsymbol{\alpha}^*$ is the true parameter that maximizes with respect to $\boldsymbol{\alpha} \in \mathfrak{A}$:

$$L(\boldsymbol{\alpha}) = E\left[Y_t(\alpha_0 + \sum_{k=1}^{d} \alpha_k f_k^0(X_{kt})) - \psi(\alpha_0 + \sum_{k=1}^{d} \alpha_k f_k^0(X_{kt})) + \phi(y_t, \Theta_t)\right] W(X_t),$$

with $\mathbf{f}^0$ a d-dimensional vector of k-th component $f_k^0(\cdot)$'s the true functions in (2.3), defined more precisely by $f_k^0(x_k) = \arg\min_{\beta_1} E[\log \mathfrak{m}_Y(y_t, \beta_1)|X_{kt} = x_{k0}]$ correspondingly to (3.9).

As noted in Section 1, establishing the asymptotic properties of the likelihood based model averaging estimators is much more difficult than that of LS-MARMA in Li, Linton and Lu (2015). This is because both our maximum likelihood based model averaging and local linear marginal estimators, $\widehat{\boldsymbol{\alpha}}^{*(n)}$ and $\widehat{f}_k(x_k)$, defined by (3.10) and (3.9), respectively, do not own analytical solutions, while the LS-MARMA procedure has both analytical solutions.

We furthermore introduce some notations. Let

$$\mathbf{V} = \sum_{k=-\infty}^{\infty} Cov(\mathcal{V}_t, \mathcal{V}_{t-k}), \text{ with } \mathcal{V}_t = m^*(Z_t, \boldsymbol{\alpha}^*, \mathbf{f}^0) + \mathcal{D}_t, \tag{3.11}$$

where $m^*(Z_t, \boldsymbol{\alpha}^*, \mathbf{f}^0) = [Y_t - \dot{\psi}(\alpha_0^* + \sum_{k=1}^d \alpha_k^* f_k^0(X_{kt}))]\widetilde{\chi}_t(\mathbf{f}_0)W(X_t)$, and $\mathcal{D}_t$ is a $d \times 1$ vector whose k-th component is $\mathcal{D}_{t,k} = (\dot{\psi}'(f_k^0(X_{kt}))g_k(X_{kt}))^{-1}\omega(Y_t; f_k^0(X_{kt}))D_k(X_{kt})$, with $D_k(x_k) = \int_{R^d}[\dot{m}^*_{\mathbf{f},k}(y, x_{-k}, x_k)g_{Y,X_{-k},X_k}(y, x_{-k}, x_k)]dydx_{-k}$, and

$$\dot{m}^*_{\mathbf{f},k}(y,x) = -\dot{\psi}'(\alpha_0^* + \sum_{\ell=1}^d \alpha_\ell^* f_\ell^0(x_\ell))\alpha_k^*\widetilde{\chi}_t(\mathbf{f}^0)W(x) + [y - \dot{\psi}(\alpha_0^* + \sum_{\ell=1}^d \alpha_\ell^* f_\ell^0(x_\ell))]\widetilde{\boldsymbol{\gamma}}_k W(x), \mathit{and}$$

$\widetilde{\boldsymbol{\gamma}}_k$ is a $(d+1) \times 1$ vector with its $(k+1)$-th element 1 and 0 elsewhere, and $g_k(x_k)$ and $g_{Y,X_{-k},X_k}(y, x_{-k}, x_k) = g_{Y,X}(y,x)$ stand for the marginal and joint probability density functions (for continuous random variables) or probability functions (for discrete random variables) of X_{kt} and (Y_t, X_t), respectively, and x_{-k} is the vector of $x = (x_1, \cdots, x_d)^T$ with its k-th component x_k removed. We further define $\widetilde{\chi}_t(\mathbf{f}^0) = (1, f_1^0(x_{1t}), \ldots, f_d^0(x_{dt}))^T$ and

$$\boldsymbol{U} = \boldsymbol{U}(\boldsymbol{\alpha}^*, \mathbf{f}^0) = E[\psi''(\alpha_0^* + \sum_{k=1}^d \alpha_k^* f_k^0(X_{kt}))]\widetilde{\chi}_t(\mathbf{f}^0)\widetilde{\chi}_t(\mathbf{f}^0)^T W(X_t), \tag{3.12}$$

where $f_k^0(\cdot)$'s are the true functions in (2.3). In the discrete case, the relevant integration should be seen as a summation over the support of the involved random variable.

Under some mild conditions given in Assumption 1 (A1-A6) on the β-mixing time series (Y_t, X_t), the kernel function K, the weight function W and the bandwidth h, etc., detailed in Section A.2 of the Supplementary Material, the following theorem establishes the consistency and asymptotic normality of the maximum likelihood estimator $\widehat{\boldsymbol{\alpha}}^{*(n)} \to \boldsymbol{\alpha}^*$.

Theorem 1. Suppose that A1-A6 of Assumption 1 (in Section A.2 of the Supplementary Materials) hold, and $\mathbf{U}$ is positive define. If $nh^4 = O(1)$, then $\widehat{\boldsymbol{\alpha}}^{*(n)} - \boldsymbol{\alpha}^* = o_P(1)$ as $n \to \infty$. Furthermore, if $nh^4 = o(1)$, then $\sqrt{n}(\widehat{\boldsymbol{\alpha}}^{*(n)} - \boldsymbol{\alpha}^*) \xrightarrow{L} N(0, \mathbf{U}^{-1}\mathbf{V}\mathbf{U}^{-1})$,as $n \to \infty$, where $\xrightarrow{L}$ stands for convergence in distribution.

Remark 4. Theorem 1 indicates that the estimator $\widehat{\boldsymbol{\alpha}}^{*(n)}$ of our semiparametric model averaging forecasting procedure is asymptotically converging in probability to the true parameter at a root-n rate. Note that in our time series case, the matrix $\mathbf{V}$ in the asymptotic variance of Theorem 1 is much more complex than that under $i.i.d.$ data. In fact, under the $i.i.d.$ data, $\mathbf{V}$ defined in (3.11) reduces to $\mathbf{V} = Var(\mathcal{V}_t)$.

We finally comment that in Theorem 1, we allow the dimension d, the number of marginal forecasts, to be large, but it is fixed, independent of the sample size n. Extending Theorem 1 to the case of $d = d_n \to \infty$ as $n \to \infty$ is theoretically interesting (c.f., Theorem 4.3 of Li, Linton and Lu (2015)). However, the technical proof for such an extension is much more complex than that of Li, Linton and Lu (2015) with an easily available analytical solution because there is no analytical solution to the estimation in this paper. We conjecture that there is a similar result to Theorem 4.3 of Li, Linton and Lu (2015) for $d = d_n \to \infty$, extending Theorem 1 of this paper, but it needs a thorough revision of the proof in this paper as well as those in Peng and Lu (2023). So we put this theoretical investigation into $d = d_n \to \infty$ left for a future work. More significantly in practice, when d is large, we will develop a penalized procedure to select the most important marginal forecasts in next section. This will further improve the interpretability of our forecasts.

4. Penalized GMAFMA approach

A critical problem in practice is how to identify the important marginal forecasts to include and which are unimportant to exclude in the GMAFMA. Involving too many irrelevant marginal forecasts would add forecast errors and reduce the performance of prediction. When the low-dimensional interactions, as indicated in Remark 2 above, are taken into account, the number of the marginal forecasts involved increases greatly to $d + d(d-1)/2$. For instance, in the real data example of FTSE100 market moving direction in Section 6.2.2 with $d = 28$, the number of marginal forests is as high as 406 considering paired marginals. To get an improved prediction, we suggest removing the irrelevant marginal forecasts by applying an adaptive LASSO (Zou, 2006) penalty to the estimation in Section 3. Compared to other popular penalties including LASSO (Tibshirani, 1996) and SCAD (Fan and Li, 2001), the adaptive LASSO enjoys both easy implementation and good theory for time series from our experience (c.f., Chen et al. (2018) and Al-Sulami et al. (2019)).

4.1 An adaptive LASSO based estimation

Recall the log-likelihood function $L_n(\boldsymbol{\alpha}; \widehat{\mathbf{f}})$ defined in (3.10) with $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_d)^T$ to maximize L_n for $\boldsymbol{\alpha}$ in $\mathfrak{A}$ (a sufficiently large compact subset of R^{1+d}). We formulate a penalized GMAFMA (PGMAFMA) by the adaptive LASSO penalization on the coefficients, $\alpha_1, \ldots, \alpha_d$, using the Lagrange multiplier method to obtain the estimator $\widehat{\boldsymbol{\alpha}}$, viz:

$$\widehat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha} \in \mathfrak{A}} \widehat{R}(\boldsymbol{\alpha}) = -L_n(\boldsymbol{\alpha}; \widehat{\mathbf{f}}) + \lambda_n \sum_{k=1}^{d} \gamma_k |\alpha_k|, \tag{4.13}$$

where the tuning parameter vector $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_d)^T$ is taken with $\gamma_k = \frac{1}{|\widehat{\alpha}_k^{*(n)}|^{\iota}}$ component-wisely for $k = 1, \cdots, d$, for some $\iota > 0$ (ι typically taken as 1), following the adaptive LASSO penalization of Zou (2006) owing to its oracle properties (see Theorems 2–4). Here the term "adaptive" is just in the sense of adaptive LASSO of Zou (2006) simply for the penalization estimation that adapts to the true non-zero coefficients as if the non-zero coefficient feature variables (marginal forecasts) were known in advance. Note $\widehat{\boldsymbol{\alpha}}^{*(n)} = (\widehat{\alpha}_0^{*(n)}, \widehat{\alpha}_1^{*(n)}, \cdots, \widehat{\alpha}_d^{*(n)})^T$ is the maximizer of $L_n(\boldsymbol{\alpha}) = L_n(\boldsymbol{\alpha}; \widehat{\mathbf{f}})$ defined in (3.10) with respect to $\boldsymbol{\alpha}$ in $\mathfrak{A}$ under no penalization, which, according to Theorem 1, is a $\sqrt{n}$ consistent estimator to the true parameter $\boldsymbol{\alpha}^* = (\alpha_0^*, \alpha_1^*, \ldots, \alpha_d^*)^T$, the maximizer of $L(\boldsymbol{\alpha})$ defined in (3.2) with respect to $\boldsymbol{\alpha}$ in $\mathfrak{A}$. This follows from Zou (2006) that one can use any consistent estimator of $\boldsymbol{\alpha}^*$ as the initial estimator $\widehat{\boldsymbol{\alpha}}^{*(n)}$ used in adaptive LASSO estimation. The Lagrange multiplier λ_n in (4.13) is a tuning parameter that varies with n and will be selected.

We offer three remarks on the penalty term in (4.13). First, the use of a LASSO-type method follows the logic of variable selection as in linear regression (cf. Zou (2006)). Even when $f_k(\cdot)$ is nearly zero in population, its estimate from (3.9) may not be, especially with large d, leading to error accumulation in prediction. The adaptive LASSO helps by shrinking such α_k estimates to zero in n (4.13). Second, while Mallows-type penalties (cf. Hansen (2007); Zhang et al. (2016)) are common in averaging of parametric models, our method differs by combining weak marginal forecasts using nonparametric flexibility to capture non-linearities and avoid the curse of dimensionality (cf. Li, Linton and Lu (2015); Chen et al. (2018)). Third, existing Mallows-type methods basically focus on parametric OLS or likelihood. As suggested by a referee, we have added Section A.6 in Appendix to test an MLE-based Mallows-type averaging for nonparametric marginal regressions on the US Strike dataset (c.f., Section 6.2.1). Its performance appeared to be worse when compared with our penalized GMAFMA (see Table A.4 in Appendix), likely due to poor approximation of the effective degrees of freedom needed for Mallows-type penalty, not as simple as that for parametric models. Any further investigation into it is hence beyond the scope of this paper. See Section A.6 for more details and discussions.

Notice that both the log-likelihood function plus penalty in (4.13) and the equations of its gradient function equal to 0 are hard to solve in a closed form. An algorithm is particularly presented in Section A.3 of the online Supplementary Material to tackle the computing problem. Clearly, one can first use the estimated coefficients ($\widehat{\alpha}_k$) from

GMAFMA procedure to construct the penalty weight (γ_k). Then, for any λ_n given, (4.13) can be treated as an optimization problem with the $\mathfrak{L}_1$ norm penalty. The computation of using adaptive LASSO in our PGMAFMA can be done via, say, the popular **glmnet** package in R, in a similar manner. The optimal selection of tuning parameter λ_n in the large-dimension case can be dealt with, say, by a popular idea of cross-validation on data.

In applications, we need to tune two parameters: (i) the bandwidth h for the marginal nonparametric regressions, and (ii) the penalty λ_n, in our PGMAFMA. The choice of h has been discussed in Section 3.1. For λ_n, we may apply cross-validation, say by direct 10-fold cross-validation using *cv.glmnet* from the **glmnet** package, or by a grid search to more carefully minimize the deviance of the fitted values, in the data examples of Section 6. Moreover, as suggested by a referee, we have additionally discussed a forward cross-validation for selection of λ_n in Appendix A.7 by using either fixed-length or expanding window, with results reported in Appendix A.7 (Table A.4) for the US Strike dataset of Section 6.2.1. Our finding indicates that the cv.glmnet seems to work robustly, with the forward CV actually getting the same variables selected as the cv.glmnet did for the Strike dataset, after penalization. On the other hand, for the forward CV, the chosen optimal λ_n may vary somehow across the window types and sizes. Consistent with Zhang and Zhang (2023), the forward CV introduces an additional hyperparameter of window size to tune at a higher computational cost. See Appendix A.7 for more details and discussions.

4.2 Asymptotic properties

Consider a setting where only a part of the marginal forecasts are useful in prediction. The penalized procedure, therefore, forces the non-relevant marginal forecast weight to (near) zero in our estimation, with $\boldsymbol{\alpha}^*$ denoting d-dimensional vector of true coefficients α_k^*, $k = 1, \cdots, d$. We denote by $\boldsymbol{\alpha}^{*1}$ the d_0-dimensional vector of non-zero true parameters and $\boldsymbol{\alpha}^{*2}$ the $(d - d_0)$-dimensional vector of zero true parameters, that is

$$\boldsymbol{\alpha}^* = \begin{bmatrix} \boldsymbol{\alpha}^{*1} \\ \boldsymbol{\alpha}^{*2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\alpha}^{*1} \\ \mathbf{0} \end{bmatrix}.$$

For simplicity, let $\mathcal{A} = \{j : \alpha_j^* \neq 0\} = \{1, ..., d_0\}$, and $d_0 < d$.

In addition to Assumption 1, we need some additional mild regularity conditions (B1-B2) in Assumption 2, given in Section A.2 of the online Supplementary Material. The following theorems show that the proposed estimator enjoys the asymptotic normality and consistency (which is called *Oracle Property*) as if the true non-zero coefficients were known. See Section A.2 of the online Supplementary Materials for more technical details.

Theorem 2. Let Assumptions 1 and 2 (in Section A.2 of the Supplementary Materials) hold. Suppose $\frac{\lambda_n}{\sqrt{n}} \to 0$ and $\lambda_n n^{(\iota-1)/2} \to \infty$. Then there exists a global minimizer $\widehat{\boldsymbol{\alpha}}$ of the $\widehat{R}(\boldsymbol{\alpha})$ defined in (4.13) such that $\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\| = O_P(\frac{1}{\sqrt{n}})$, where $\boldsymbol{\alpha}^*$ is the true parameter.

Now let $\widehat{\boldsymbol{\alpha}}^1$ be the d_0-dimensional vector of all $\widehat{\alpha}_j$'s for $j \in \mathcal{A}$, and $\widehat{\boldsymbol{\alpha}}^2$ the $(d - d_0)$-

dimensional vector of all $\widehat{\alpha}_j$'s for $j \in \mathcal{A}^c$, where $\widehat{\alpha}_j$ denotes the j-th component of $\widehat{\boldsymbol{\alpha}}$.

Theorem 3. (Consistency) Let Assumptions 1 and 2 (in Section A.2) hold. Suppose $\frac{\lambda_n}{\sqrt{n}} \to 0$, and $\lambda_n n^{(\iota-1)/2} \to \infty$ for some $\iota > 0$. Then the irrelevant marginal forecast model averaging weights are estimated to be zero with probability tending to one: $P(\widehat{\boldsymbol{\alpha}}^2 = \boldsymbol{\alpha}^{*2} = 0) \to 1 \;\; as \;\; n \to \infty$.

Theorem 4. (Asymptotic Normality) Let Assumptions 1 and 2 (in Section A.2) hold. Suppose $\frac{\lambda_n}{\sqrt{n}} \to 0$ and $\lambda_n n^{(\iota-1)/2} \to \infty$. Then $\sqrt{n}(\widehat{\boldsymbol{\alpha}}^1 - \boldsymbol{\alpha}^{*1}) \xrightarrow{\text{L}} N(0, \boldsymbol{U}_1^{-1}\boldsymbol{V}_1\boldsymbol{U}_1^{-1})$, where $\boldsymbol{U}_1$ and $\boldsymbol{V}_1$ are the $d_0 \times d_0$ sub-matrices of $\boldsymbol{U}$ and $\boldsymbol{V}$ corresponding to (i,j)-th components with $i, j \in \mathcal{A}$, respectively.

As discussed at the end of Section 3, extending Theorems 2–4 to the case of $d = d_n \to \infty$ as $n \to \infty$ is theoretically interesting, but it is still an open question as the technical proof for such an extension becomes much more challenging than in Chen et al. (2018). So we put this theoretical extension of $d = d_n \to \infty$ also left for a further work in particular in view of the length of this paper. But the practical implications from our above theorems are sufficient for applications as the dimension d is fixed but allowed to be large, as illustrated in our numerical data examples below.

5. Further discussions

We discuss an extension of the GMAFMA and PGMAFMA to some more general settings.

5.1 Discrete-valued covariates

In the case of discrete-valued response, considering the useful lags of the response variable in our GMAFMA and PGMAFMA procedures leads to discrete-valued covariates of X_t. We extend the idea of nonparametric kernel regression for discrete covariates in Li and Racine (2003) and Racine and Li (2004) to the time series setting. That is, for example, in (2.3), if X_{kt} is discrete-valued, we then consider the local linear fitting in (3.9) by applying the idea with the usual kernel function $K_h(X_{kt} - x_{k0})$ for continuous value in (3.9) replaced by a discrete-valued random variable kernel function $K_\lambda(X_{kt}, x_{k0})$. Racine and Li (2004) defined $K_\lambda(X_{kt}, x_{k0}) = 1$ for $X_{kt} = x_{k0}$ and $K_\lambda(X_{kt}, x_{k0}) = \lambda$ for $X_{kt} \neq x_{k0}$. Note that when $\lambda = 0$, the kernel function $K_\lambda(X_{kt}, x_{k0})$ becomes an indicator function which takes value 1 if $X_{kt} = x_{k0}$, and 0 otherwise. If $\lambda = 1$, $K_\lambda(X_{kt}, x_{k0}) = 1$ becomes a constant. The range of λ is $[0, 1]$. For details, the reader is referred to Li and Racine (2003) and Racine and Li (2004).

5.2 Marginal covariate interactions

As discussed in Remark 2, we can further consider the marginal covariate interaction effects which are useful in our GMAFMA and PGMAFMA procedures. Because of the curse of dimensionality, we usually do not consider interactions of too many covariates (but the pairs of covariates are reasonable) in nonparametric estimation. For example, in (2.3), if X_{kt} is replaced by (X_{jt}, X_{kt}), $j \neq k$, with X_{jt} and X_{kt} being either discrete or continuous valued, we can extend the local linear fitting in (3.9) by replacing X_{kt} and x_{k0} with (X_{jt}, X_{kt}) and (x_{j0}, x_{k0}), respectively, and the kernel function $K_h(X_{kt} - x_{k0})$ by the product of two marginal kernel functions, depending on X_{jt} and X_{kt} being continuous or discrete valued, say $K_{h_1}(X_{jt} - x_{j0})K_{h_2}(X_{kt} - x_{k0})$ if both X_{jt} and X_{kt} are continuous-valued, and similarly for other cases by combining the idea discussed in Section 5.1. Note that by taking Taylor's expansion of $f_{jk}(X_{jt}, X_{kt})$ at an arbitrary point (x_{j0}, x_{k0}) at which it is differentiable and that (X_{jt}, X_{kt}) is in the neighborhood of (x_{j0}, x_{k0}), we then have

$$\begin{aligned} f_{jk}(X_{jt}, X_{kt}) &\approx f_{jk}(x_{j0}, x_{k0}) + f'_{jk,1}(x_{j0}, x_{k0})(X_{jt} - x_{j0}) + f'_{jk,2}(x_{j0}, x_{k0})(X_{kt} - x_{k0}) \\ &\equiv \beta_1 + \beta_2(X_{jt} - x_{j0}) + \beta_3(X_{kt} - x_{k0}). \end{aligned} \tag{5.14}$$

Then the time series local log conditional likelihood for estimation of $\beta_1 = f_{jk}(x_{j0}, x_{k0})$ is

$$\ell(\beta_1, \beta_2, \beta_3) = \sum_{t=1}^{n} [\log \mathfrak{m}_Y(y_t, \beta_1 + \beta_2(X_{jt} - x_{j0}) + \beta_3(X_{kt} - x_{k0}))K_{h_1,h_2,jk,t}(x_{j0}, x_{k0})], \tag{5.15}$$

with $K_{h_1,h_2,jk,t}(x_{k0}, x_{k0}) = K_{h_1}(X_{jt} - x_{j0})K_{h_2}(X_{kt} - x_{k0})$, if both X_{jt} and X_{kt} are continuous-valued, and similarly for other cases with $K_{h_1}(X_{jt} - x_{j0})K_{h_2}(X_{kt} - x_{k0})$ replaced by other suitable kernel functions as discussed above. Here $\mathfrak{m}_Y(\cdot, \cdot)$ is given in

(2.1).

We can add the interaction terms $f_{jk}(X_{jt}, X_{kt})$'s into (2.4) to extend the GMAFMA in Section 2 and the PGMAFMA in Section 4, which will be applied in the data examples.

6. Numerical examples

Both simulations and real data examples are demonstrated to support our procedures.

6.1 Monte-Carlo simulations

To save space, details of simulation settings in this subsection are provided in Section A.4 of the online Supplementary Materials for the reference of interested readers.

We consider two non-Gaussian data generating processes (DGPs) for Monte-Carlo simulation to demonstrate the performance of our proposed GMAFMA framework:

- DGP 1 (Conditional binomial distribution) - the binary classification forecasting problem with two-dimensional interaction marginals (a model of a non-GAM form);
- DGP 2 (Conditional Poisson distribution) - the count data prediction problem, with discrete-valued lagged and other exogenous information accounted for.

For DGP 1, the data is generated by $Y_t = I(y_t > 0)$ with

$$y_t = \sum_{k=1}^{3} g_{0k}(y_{t-k}) + \cos(2x_{t1}x_{t2}) + \log(1 + (x_{t3}x_{t4})^2) + x_{t5}x_{t6} + \varepsilon_t, \tag{6.16}$$

where $g_{0k}(y_{t-k}) = -\sin(2y_{t-k})$, and the model involves lags up to 3 of time series y_t and two-way interactions of 6 covariates $(x_{t1}, \cdots, x_{t6})$ independently following $N(0, 1)$. For prediction of Y_t, comparisons of our PGMAFMA and GMAFMA with GLM (generalized linear model), LASSO-based GLM (GLMNET), random forest (RF) and linear autoregression (AR) of y_t based classifier (see Figure 1) are detailed in Section A.4.1 of the Supplementary Materials, with working variables of 15 Gaussian $N(0, 1)$ covariates, x_{tk}, $k = 1, \cdots, 15$, and lags up to 15 of y_t (in total $d = 30$ predictors). For easy presentation, we consider only a setting of combining 29 marginal forecasts (c.f., a form similar to (2.4)) including 15 lagged forecasts $f_k(Y_{t-k})$ for $k = 1, \ldots, 15$ and 14 paired forecasts $f_{j,j+1}(x_{tj}, x_{t,j+1})$ for $j = 1, \ldots, 14$. The testing AUC (area under curve) values of 100 replications are depicted in boxplot in Figure 1 for the six methods, with AUC the larger the better.

Similarly, for DGP 2 with $Y_t|I_{t-1} \sim Poisson(\mu_t)$, we considers a log-nonlinear structure for modelling μ_t, involving 3 covariates together with lag order 3 of Y_t:

$$\log \mu_t = \frac{1}{4}\sum_{k=1}^{3} g_{0k}(Y_{t-k}) + 3\cos(x_{t1}) + 2e^{2x_{t2}} + 6x_{t3}^2 + \epsilon_t, \text{with } \ g_{0k}(Y_{t-k}) = -\sin(Y_{t-k}), \tag{6.17}$$

where the 3 covariates (x_{t1}, x_{t2}, x_{t3}) independently follow the $N(0, 1)$ distribution. Again, we consider a working setting of 15 Gaussian covariates, x_{tk}, $k = 1, \cdots, 15$, and long lags of order up to 15 of Y_t for six methods (with GBM instead of RF) as indicated in Figure 2. This leads to a total of $d = 30$ marginals in (2.4), considering only one-dimensional marginals, $f_k(Y_{t-k})$'s and $f_{15+j}(x_{tj})$'s for $k, j = 1, \ldots, 15$, used for pre-

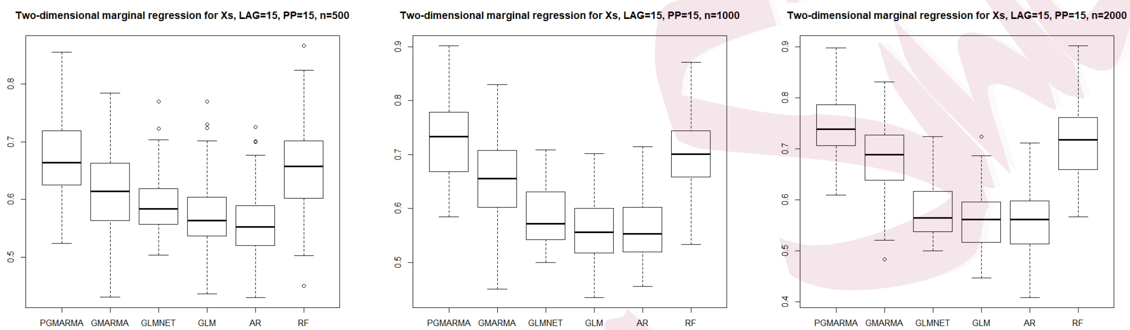

Figure 1: Boxplots of testing AUC values (with test sample size $n_\tau = 50$) of 100 replications for one-step ahead classification predictions using 15 covariates and 15 autoregressive lag terms for DGP 1 (true containment of 6 covariates and 3 autoregressive terms), by different methods (PGMAFMA, GMAFMA, GLMNET, GLM, AR, Random Forest (RF)) modeled with training sample size $n = 500$ (left), $n = 1000$ (middle), $n = 2000$ (right), respectively.

diction. Their performances are evaluated in terms of mean absolute prediction error for 100 Monte-Carlo replications in boxplot in Figure 2, with the smaller the error the better the method. See Section A.4.2 of the online Supplementary Materials for more details.

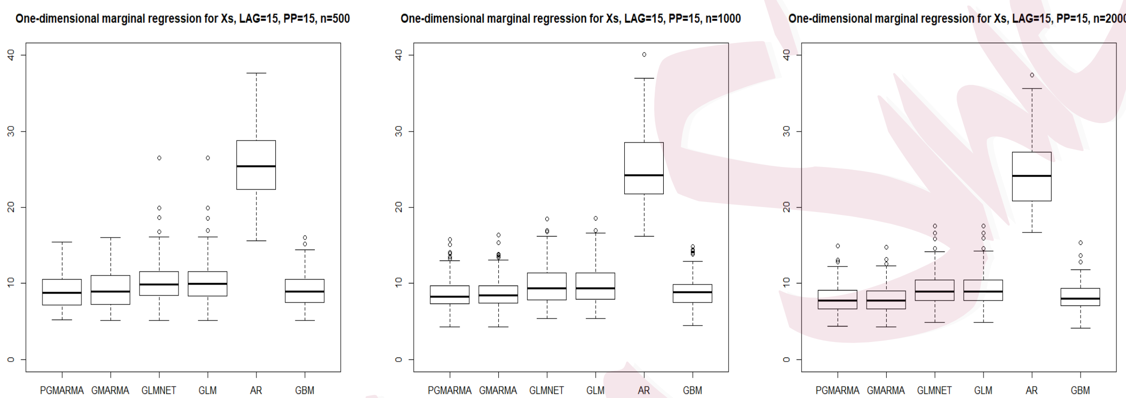

Figure 2: Boxplots of testing Mean Absolute Errors (test sample size $n_\tau = 50$) of 100 replications for one-step ahead predictions of Y_t using 15 covariates and 15 autoregressive lags for DGP 2 (containing 3 covariates and 3 autoregressive lags), by different methods (PGMAFMA, GMAFMA, GLMNET, GLM, AR, Gradient Boosting Machine (GBM)) modeled with training sample size $n = 500$ (left), $n = 1000$ (middle), $n = 2000$ (right), respectively.

We summarise the numerical findings as follows, with readers referred to Sections A.4.1 and A.4.2 of online Supplementary Materials for further discussions:

(i) Our proposed PGMAFMA and GMAFMA models are the most competitive models and are much better than the linear-type models of GLM, LASSO-based GLMNET, and linear AR among all cases (especially for DGP 1 of a non-GAM form). This shows that PGMAFMA and GMAFMA can well capture the useful nonlinear features present in data.

(ii) The predictive power increases overall with the larger training sample size for better model estimation. However, the linear-type models do not seem to improve even with a large sample size as the linear assumptions are indeed violated in this case. Interestingly, even with a sample size of $n = 500$, our PGMAFMA has worked quite well.

(iii) Our PGMAFMA with adaptive LASSO penalization technique has clearly enhanced the prediction of our GMAFMA by reducing the prediction errors. In general, the predictive powers of penalized models are better than those of non-penalized ones.

(iv) Our PGMAFMA even beats the popular machine learning models like the Random Forest (RF) and the Gradient Boosting Machine (GBM), while the GMAFMA performs very closely to them. We argue that the machine learning models are more complex and much harder to interpret, while our PGMAFMA method provides more explainable results.

6.2 Real data examples

Three non-Gaussian real datasets in labor, finance, and health economics are examined.

6.2.1 Number of strikes in the US manufacturing

When the world's economy decreases, strikes are widely observed. Strikes play a significant role and their relationship with economic conditions is important to study in labor economics. For instance, the number of strikes in US manufacturing (available in R package **Ecdat**) was studied in Kennan (1985) and Cameron and Trivedi (1990) based on regression. However, none of them considers the temporal lagged information of historical impact and it may be more sensible to extract the relationship from a prediction perspective on strikes since the out-of-sample prediction would help plan on managing future strikes. In this section, we demonstrate the performance of our model in predicting the number of strikes by a conditional Poisson distribution with lagged information included to help unveil the relationships between the number of strikes and economic conditions.

Denote by Y_t the number of strikes (number of contract strikes in U.S. manufacturing beginning each month) and Z_t the level of economic activity (measured as the cyclical departure of aggregate production from its trend level), for $t = 1, ..., 108$, being the monthly observations from Jan 1968 to Dec 1976, in the dataset 'StrikeNb' of R package **Ecdat**. We use the first 84 observations (from Jan 1968 to Dec 1974) for the training and the remaining 24 observations (Jan 1975 to Dec 1976) for prediction.

We benchmark the performance using an AR model for Y_t of lag order 2, optimally decided via the AIC criterion. We hence use order 2 lagged information in other working models as our dataset is short. Our GMAFMA is expressed as $Y_t|I_{t-1} \sim Poisson(\mu_t)$

with

$$\log(\mu_t) \approx \alpha_0 + \alpha_1 f_1(Y_{t-1}) + \alpha_2 f_2(Y_{t-2}) + \alpha_3 f_3(Z_t) + \alpha_4 f_4(Z_{t-1}) + \alpha_5 f_5(Z_{t-2}), \tag{6.18}$$

where $f_j(\cdot)$'s are the marginal forecasts of $\log(\mu_{jt})$'s by the involved marginal information.

For comparisons, we also considered two Poisson (linear) regression models by GLM, one including lagged information by replacing each $\alpha_j f_j(\cdot)$ with its linear form in (6.18), and one with no lags like Cameron and Trivedi (1986) considering only Z_t in regression.

Table 1: Prediction performances of candidate models for strike data

Model	MAE of prediction
AR(2)	1.98
GLM with lagged information	1.91
GLM without lagged information	2.55
GMAFMA	1.80
PGMAFMA	**1.67**

Note: The performance with a smaller MAE (mean absolute error) for prediction is preferred. The AR model is estimated by the **ar** in R; the GMAFMA model is made with a tentative bandwidth of 0.3 used; the penalized GMAFMA model is estimated using Algorithm 1, detailed in Section A.3 of online Supplementary Material, which applies the adaptive LASSO to the GMAFMA model (6.18).

The out-of-sample performance of one-step prediction in terms of mean absolute error (MAE), reported in Table 1, confirms that adding lagged information indeed helps in prediction. In particular, the true relationship between the number of strikes and the level of economic activities appears nonlinear and it is well captured by the GMAFMA model via the nonparametric marginal estimation. More interestingly, after applying the penalty, the PGMAFMA model contains only the lagged information: Y_{t-1}, Y_{t-2}, and Z_{t-2}; viz:

$$\log(\mu_t) = -2.1834 + 0.7850 f_1(Y_{t-1}) + 0.5172 f_2(Y_{t-2}) + 0.9384 f_5(Z_{t-2}). \tag{6.19}$$

This indicates a clear difference of the economic activity Z_t contributing to strikes in a *lagged manner* from that in Cameron and Trivedi (1986) who did not consider lagged impact. The improvement in predictability by our best PGMAFMA model (6.19) over the non-lagged GLM model (c.f., Cameron and Trivedi (1986)) is profound with the prediction error in MAE reduced as high as $(2.55 - 1.67)/2.55 = 34.5\%$, and PGMAFMA outperforming other models, following from Table 1. This shows that the number of strikes is *non-linearly* impacted by the *past* level of economic activity (Z_{t-2}), in addition to the *past* strikes.

Moreover, as suggested by a referee, we have additionally implemented a forward cross-validation for tuning λ on the US Strike dataset, which leads to the same result as above. See Appendix A.7 for more details and discussions.

6.2.2 Prediction of FTSE100 index market moving direction

Our second dataset to be examined is in finance. Forecasting the financial market is well known to be challenging as explained by the well-known market efficient hypothesis theory. In this section, we demonstrate the strength of our penalized GMAFMA (PGMAFMA) procedure in forecasting the FTSE100 index market moving direction by a conditional binomial distribution via variable selection with more complex lag interactions than accounted for in Peng and Lu (2024). As in the mentioned paper, we consider the daily time series, from 1 May 2013 to 1 May 2018, for 4 variables, namely market direction indicator (Y_t), volatility (v_t), volume (V_t), and geometric return (G_t). We use the first 1200 observations for training and the remaining 62 observations for prediction evaluation.

Differently from Peng and Lu (2024) who considered one-dimensional marginal forecasts only, we are examining if the forecasts with additional two-dimensional marginals of paired lags of Y_t, v_t, V_t and G_t and further penalization, taking account of more historical information, can help improve the prediction of moving direction Y_t. We consider one-step-ahead prediction of Y_t based on the past information of a week lag order equal to 7, that is, $X_t = (Y_{t-1}, \ldots, Y_{t-7}, v_{t-1}, \ldots, v_{t-7}, V_{t-1}, \ldots, V_{t-7}, G_{t-1}, \ldots, G_{t-7})$ with their one- and two-dimensional marginal forecasts in GMAFMA and PMAFMA to predict Y_t, viz:

$$\begin{aligned}\text{logit}(P(Y_t = 1|I_{t-1})) &= \alpha_0 + \alpha_1 f_1(Y_{t-1}) + \cdots + \alpha_7 f_7(Y_{t-7}) + \alpha_8 f_8(v_{t-1}) + \cdots + \alpha_{14} f_{14}(v_{t-7}) \\ &\quad + \alpha_{15} f_{15}(V_{t-1}) + \cdots + \alpha_{21} f_{21}(V_{t-7}) + \alpha_{22} f_{22}(G_{t-1}) + \cdots + \alpha_{28} f_{28}(G_{t-7}) \\ &\quad + \alpha_{29} f_{29}(Y_{t-1}, Y_{t-2}) + \cdots + \alpha_{406} f_{406}(G_{t-6}, G_{t-7}). \qquad (6.20)\end{aligned}$$

This has made the total number of marginal (one- and two-dimensional) forecasts as high as 406 in GMAFMA, extremely larger and more challenging than that of only 28 one-dimensional marginal forecasts (as done in Peng and Lu (2024)), thus motivating us to apply our penalized method, PGMAFMA; see Table 2. The GMAFMA and PGMAFMA with one-dimensional marginals only are also reported in Table 2 (indicated in 1-dim).

As benchmark comparisons, we have also examined the linear logistic regression with 406 variables including interactions (i.e., the functions being linear in (6.20)), denoted by GLM, and its LASSO penalized version (denoted GLMNET) for Y_t in Table 2. A popular learning method of random forecast classification for Y_t and the simplest autoregressive (AR) model of G_t based classification (i.e., by $Y_t = I(G_t > 0)$) are further considered. See Table 2 for evaluation of different methods with the test data in terms of the AUC (area under curve) values, and Section A.5 of online Supplementary Materials for more details.

The results in Table 2 interestingly suggest that our proposed GMAFMA and PGMAFMA perform outstandingly well among all considered models. In particular, our PGMAFMA including two-dimensional marginals performs the best with the largest AUC value of 0.6953 among all candidate methods. It outperforms the GLM and the GLMNET with interactions by $(0.6953 - 0.5116)/0.5116 = 35.9\%$ and $(0.6953 - 0.6208)/0.6208 = 12.0\%$, respectively, and improves the one-dimensional marginals based GMAFMA: 1-dim and PGMAFMA: 1-dim by $(0.6953 - 0.6071)/0.6071 = 14.5\%$ and $(0.6953-0.6639)/0.6639 = 4.73\%$, respectively. The PGMAFMA has also improved

Table 2: The performances in AUC of one step ahead prediction for candidate models

Model	Prediction AUC	Number of non-zero predictors
GMAFMA	0.6176	406
PGMAFMA	**0.6953**	206
GMAFMA: 1-dim	0.6071	28
PGMAFMA: 1-dim	0.6639	11
GLM	0.5116	406
GLMNET	0.6208	32
AR$^{(1)}$	0.5557	7
Random Forest$^{(2)}$	0.5058	500$^{(3)}$

Note: The performance with a larger AUC (area under curve) is preferred. A bandwidth $h = 0.6$ was used in estimation of all marginals f_k's in (6.20); the PGMAFMA method applies the adaptive LASSO to the GMAFMA in (6.20) with a penalty tuning parameter $\lambda = 0.001179$ for two-dimensional case and $\lambda = 0.024322$ for benchmark; the GLMNET model applies the LASSO penalty by the **glmnet** package in R; the Random Forest model is estimated by the **RandomForest** in R; the AR model is estimated by the **ar** in R. Here the superscripts in the table stand for: $^{(1)}$ the AR model uses only the linear form of past market return G_t; $^{(2)}$ the Random Forest model should, in theory, be able to detect all the interactions automatically, $^{(3)}$ using 500 trees with 5 variables randomly tried per split, the final model of which is thus much more complicated than all other models reported here.

its non-penalized version of GMAFMA by $(0.6953 - 0.6176)/0.6176 = 12.6\%$, and the AR and the Random Forest (RF) by $(0.6953 - 0.5557)/0.5557 = 25.1\%$ and $(0.6953 - 0.5058)/0.5058 = 37.5\%$, respectively. As indicated, the financial market is somehow predictable (c.f., Murray, Xia and Xiao (2024)), which violates the well-known efficient market hypothesis that the market moving direction is hard to predict with an AUC value for prediction approximately about 0.5. The AUC being equal to 0.5058 for the popular machine learning of RF confirms this, unexpectedly performing the worst, partially due to the bootstrap in "randomForest" R package being unable to capture time series dependence. All these further highlight the advantages of our GMAFMA and PGMAFMA.

6.2.3 Multi-step ahead forecasting in causal analysis of UK road casualty

We now illustrate the multi-step ahead forecasting by our methods in this last empirical study for analysis of road casualties in health economics (García-Ferrer et al., 2007). We are concerned with estimation of the causal effect owing to a seatbelt law, in the UK, in reducing the monthly number of killed van drivers, denoted by Y_t for the t-th month, with the data from Jan. 1969 to Dec. 1984, available in the dataset 'Seatbelts' in R package *datasets*. The time series plot of the data Y_t is displayed in Figure 3, and it has been studied by Harvey and Durbin (1986) for the effect of seatbelt legislation introduced on 31 Jan 1983. A recent study by Liboschik, Fokianos and Fried (2017) suggests that the monthly number of killed drivers is much smaller than that analyzed by Harvey and Durbin (1986). We hence apply our methods to the count data Y_t modelled by conditional Poisson distribution for estimating causal effect requiring multi-step-ahead forecasting.

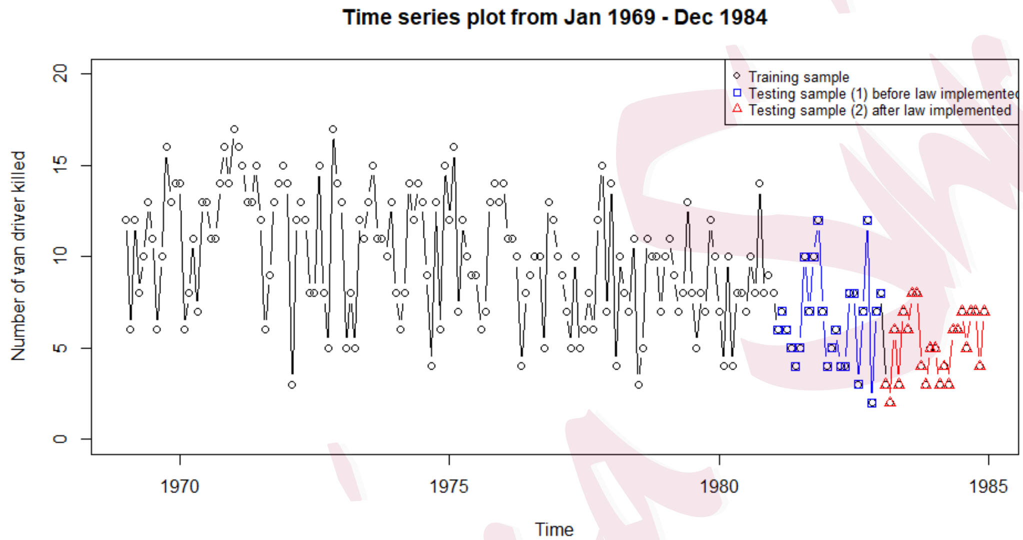

Figure 3: UK road causality of van drivers

To assess the effect of seatbelt legislation, we first segment the data into two parts according to the date before/after the law implementation, as depicted in Figure 3: (1) *Sample One* for the first 169 observations from Jan. 1969 to Jan. 1983 (including the black and blue colored parts) with sample size $T_0 = 169$; (2) *Sample Two* for the remaining 23 observations from Feb. 1983 to Dec. 1984 (the red colored part). We then need to predict the *counterfactual* monthly number of killed van drivers after the law implementation on 31 Jan. 1983, denoted by Y_t^0 for $t > T_0$ (see Rubin (1974)), based on the information of Sample One, by examining different methods. Thus we can evaluate the seatbelt law effect using Sample Two, by $\tau_t = Y_t - Y_t^0$ for the t-th month during the period from Feb. 1983 to Dec. 1984, where Y_t is observed in Sample Two, but Y_t^0 need be predicted by *multi-step-ahead* forecasting using our methods fully based on Sample One.

We need to evaluate which method among different methods (see Table 3) can do better prediction, estimated using Sample One. We can then rationally expect that the optimal method would make a better prediction of $Y_{T_0+i}^0$ with $T_0 = 169$ by i-*step-ahead* forecasting, for $i = 1, \cdots, 23$, using the whole Sample One of size T_0, needed to assess the law effect with Sample Two. To do this, we have further partitioned Sample One into two parts, the training sample from the 1st observation to the 145th observation (the black colored part in Figure 3), and the validating sample containing 24 observations from the 146th observation to the 169th observation (the blue colored part in Figure 3) for evaluation of the predictions by the different methods in Table 3. Here we do the multi-step ahead prediction of Y_{145+i}, for $i = 1, \ldots, 24$, based on the data $Y_1, \cdots, Y_{145}$, i.e., the training data of Sample One.

In Liboschik, Fokianos and Fried (2017), following Harvey and Durbin (1986), they found the lagged terms, Y_{t-1} and Y_{t-12} (indicating the short-term dependency and yearly seasonality, respectively), the real price of petrol P_t and a linear trend term T_t helpful to predict Y_t. Their fitted model ($Y_t|I_{t-1} \sim Poisson(\mu_t)$ in R package 'tsglm') is

$$\log(\mu_t) = 1.83 + 0.09Y_{t-1} + 0.15Y_{t-12} + 0.78P_t - 0.03T_t, \ \textit{with } T_t = t/12. \qquad (6.21)$$

It is worth pointing out that Liboschik, Fokianos and Fried (2017) have also reported that the petrol price P_t does not seem to influence the number of killed van drivers Y_t, as its coefficient has a much larger standard error compared to its estimated value. They also concluded that there is little short-term dependency indicated by the small coefficient of Y_{t-1}.

Inspired by those references, we are first examining if including more lagged information on the number of killed van drivers Y_t can help improve the prediction. For this aim, we have used the lagged information of the monthly number of killed van drivers Y_t, up to 24 lags (from lag 1 to lag 24), i.e., using $X_t = (Y_{t-1}, \ldots, Y_{t-24})$ to predict Y_{t-1+i} for $i = 1, 2, \cdots, 24$. For the i-step ahead forecasting, we formulate our GMAFMA model as follows:

$$Y_{t-1+i}|I_{t-1} \sim Poisson(\mu_t), \quad \log(\mu_t) = \alpha_0 + \alpha_1 f_1(Y_{t-1}) + \cdots + \alpha_{24} f_{24}(Y_{t-24}), \quad (6.22)$$

where, for simplicity, f_k's are one-dimensional nonlinear marginal forecasts of Y_{t-1+i} given Y_{t-k}, which are pre-estimated, and we should note they depend on i. Then α_k's are estimated for GMAFMA, as detailed in Section 3, and for PGMAFMA, in Section 4.

As it can be seen from the 2nd column of Table 3, evaluated by the validating part of Sample One, our PGMAFMA achieves the smallest MAE of 2.25, and the 2nd smallest MAE of 2.42 was achieved by GMAFMA and tsglm. In this sense, our PGMAFMA outperforms all other methods, and, therefore, estimating the law effect by it seems expected to be the most rationally reliable. The total reduced numbers of van drivers killed, calculated by $\sum_{i=1}^{23}(\widehat{Y}^0_{T_0+i} - Y_{T_0+i})$, after the law implementation are reported in

Table 3: Prediction performances of candidate models for road casualty data

Model	MAE of prediction evaluation before law implementation	Reduced van-drivers-killed number due to law implementation$^{(*)}$
GMAFMA	2.42	34.784
PGMAFMA	**2.25**	**32.142**
GLM	2.79	20.549
GLMNET	2.63	19.746
GBM	2.71	38.174
ar(24)	2.94	24.851
arima(12,1,0)	2.50	43.053
tsglm	2.42	24.851

Note: The PGMAFMA model is estimated using Algorithm 1, detailed in Section A.3 of Online Materials, which applies the adaptive LASSO on the GMAFMA model (6.22) with a tentatively chosen bandwidth $h = 0.3$; the GLMNET model applies the LASSO penalty to Poisson GLM model, which can be estimated by the **glmnet** package in R; The AR model is estimated by the **ar** in R; The ARIMA(12,1,0) is the model selected as the best using **arima** in R; the Gradient Boosting Machine model is estimated by the **gbm** in R; and the tsglm model is given in (6.21), which is also available via **tscount** in R. (*) The predicted reduced number of van drivers killed is over the period from Feb 1983 to Dec 1984: $\sum_t (predicted_t - real_t)$.

the 3rd column of Table 3, where $\widehat{Y}^0_{T_0+i}$'s are the predicted values by (6.22) estimated with the whole Sample One while Y_{T_0+i}'s are the observations in Sample Two. It follows from this 3rd column that after the law implementation, the total reduced number of van drivers killed, estimated by our PGMAFMA, appears reasonable, which is about 32. It is larger than the estimated effects of about 25 by the tsglm, about 25 by the AR of order 24, about 20 by the GLMNET, and about 21 by the GLM, which seem to underestimate the law effects, while it is smaller than the estimated effects of about 35 by the GMAFMA, about 38 by the GBM, and about 43 by the ARIMA(12,1,0), which seem to overestimate the law effects. The estimated effect of 32 by PGMAFMA is similar to that of 35 by our GMAFMA, appearing most reasonable.

7. Conclusion

In this paper, we have proposed a novel semiparametric procedure GMAFMA and its penalized version of PGMAFMA with adaptive LASSO, such that it is enabled to deal with the challenges due to large-scale lagged information of non-Gaussian time series data of unknown forms in the conditional exponential family. They particularly adapt to not only high dimensional but also non-Gaussian discrete- and continuous-valued time series data. The computation of both GMAFMA and PGMAFMA is cheap and easy to implement in practice. Theoretical results, including the consistency and asymptotic normality of the proposed procedures, are established. Numerical examples by Monte-Carlo simulations have demonstrated the significant performances of our proposed methods for non-Gaussian time series under conditional binomial and Poisson distributions, interestingly outperforming traditional GLM methods and popular machine learning methods like Random Forest and Gradient Boosting Machine. Three applications to the US strike data in labor economics, the FTSE 100 Index data in finance economics, and the causal analysis of UK road casualty data in health economics have validated the power of our proposed methods in real data analysis. We believe our methods provide a unified way of dealing with time-dependent and possibly nonlinear semiparametric structures of data that can overcome the "curse of dimensionality" for non-Gaussian time series modeling, and provide practitioners with a more flexible framework than the traditional generalized linear and additive models to deal with real data of unknown form, in particular where variables are not well understood, in prediction.

The methods of this paper can be further extended to a spatio-temporal domain such that not only the relations of time dependency but also location dependency can be included (Al-Sulami et al., 2019). This is left for future work.

Supplementary Materials

The supplementary material contains technical details and proofs for the results in the main paper with additional numerical results.

Acknowledgement

The authors would like to express sincere gratitude to the editor Prof Huixia Wang, the Associate Editor and both referees for their careful reading and constructive comments, which have greatly improved the early version of this paper. The research of Lu was partially supported by the Startup Fund (No.7200813) of City University of Hong Kong, which is also acknowledged.

References

Al-Sulami, D., Jiang, Z., Lu, Z. and Zhu, J. (2019). On a Semiparametric Data-Driven Nonlinear Model with Penalized Spatio-Temporal Lag Interactions. *Journal of Time Series Analysis* **40** (3), 327–342.

An, H. and Huang, F. (1996). The Geometrical Ergodicity of Nonlinear Autoregressive Models. *Statistica Sinica* **6** (4), 943–956.

Box, G. E. P., Jenkins, G. M. and Reinsel, G. C. (1994). *Time Series Analysis: Forecasting and Control*. 3rd Edition. Prentice Hall, Englewood Cliffs, New Jersey.

Cameron, A. and Trivedi, P. (1998). *Regression Analysis of Count Data*. Cambridge University Press.

Cameron, A. C. and Trivedi, P. K. (1986). Econometric Models Based on Count Data.

Comparisons and Applications of Some Estimators and Tests. *Journal of Applied Econometrics* **1** (1), 29–53.

Cameron, A. C. and Trivedi, P. K. (1990). Regression-Based Tests for Overdispersion in the Poisson Model. *Journal of Econometrics* **46** (3), 347–364.

Cameron, A. C. and Trivedi, P. K. (1996). 12 Count Data Models for Financial Data. *Handbook of Statistics* **14**, 363–391.

Cameron, A. C. and Trivedi, P. K. (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press.

Chen, J., Li, D. and Linton, O. (2019). A New Semiparametric Estimation Approach for Large Dynamic Covariance Matrices with Multiple Conditioning Variables. *Journal of Econometrics* **212** (1), 155–176.

Chen, J., Li, D., Linton, O. and Lu, Z. (2016). Semiparametric Dynamic Portfolio Choice with Multiple Conditioning Variables. *Journal of Econometrics* **194** (2), 309–318.

Chen, J., Li, D., Linton, O. and Lu, Z. (2018). Semiparametric Ultra-High Dimensional Model Averaging of Nonlinear Dynamic Time Series. *Journal of the American Statistical Association* **113** (522), 919–932.

Dong, C., Gao, J. and Linton, O. (2023). High Dimensional Semiparametric Moment Restriction Models. *Journal of Econometrics* **232** (2), 320–345.

Doukhan, P., Massart, P. and Rio, E. (1995). Invariance Principles for Absolutely Regular Empirical Processes. *Ann. Inst. H. Poincaré Probab. Statist.* **31**, 393–427.

Drost, F. C., Van den Akker, R. and Werker, B. J. (2009). Efficient Estimation of Auto-Regression Parameters and Innovation Distributions for Semiparametric Integer-Valued AR (p) Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **71** (2), 467–485.

Fan, J. (2018). *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability 66*. Routledge.

Fan, J., Farmen, M. and Gijbels, I. (1998). Local Maximum Likelihood Estimation and Inference. *Journal of the Royal Statistical Society: Series B Statistical Methodology* **60** (3), 591–608.

Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association* **96** (456), 1348–1360.

Fan, J. and Yao, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer Science & Business Media.

Fang, F., Li, J. and Xia, X. (2022). Semiparametric Model Averaging Prediction for Dichotomous Response. *Journal of Econometrics* **229** (2), 219–245.

Fokianos, K., Fried, R., Kharin, Y. and Voloshko, V. (2021). Statistical Analysis of Multivariate Discrete-Valued Time Series. *Journal of Multivariate Analysis* **188**, 104805.

Fokianos, K., Rahbek, A. and Tjøstheim, D. (2009). Poisson Autoregression. *Journal of the American Statistical Association* **104** (488), 1430–1439.

Freund, Y. and Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences* **55** (1), 119–139.

Gao, J. (2007). *Nonlinear Time Series: Semiparametric and Nonparametric Methods*. Chapman and Hall/CRC.

Gao, J., Kanaya, S., Li, D. and Tjøstheim, D. (2015). Uniform Consistency for Nonparametric Estimators in Null Recurrent Time Series. *Econometric Theory* **31** (5), 911–952.

Gao, J., Wang, Q. and Yin, J. (2013). Long-Range Dependent Time Series Specification. *Bernoulli* **19** (5A), 1714–1749.

García-Ferrer, A., De Juan, A. and Poncela, P. (2007). The Relationship Between Road Traffic Accidents and Real Economic Activity in Spain: Common Cycles and Health Issues. *Health Economics* **16** (6), 603–626.

Hansen, B. E. (2007). Least Squares Model Averaging. *Econometrica* **75** (4), 1175–1189.

Harvey, A. C. and Durbin, J. (1986). The Effects of Seat Belt Legislation on British Road Casualties: A Case Study in Structural Time Series Modelling. *Journal of the Royal Statistical Society: Series A Statistics in Society* **149** (3), 187–210.

Kennan, J. (1985). The Duration of Contract Strikes in US Manufacturing. *Journal of Econometrics* **28** (1), 5–28.

Li, C., Li, Q., Racine, J. S. and Zhang, D. (2018). Optimal Model Averaging of Varying Coefficient Models. *Statistica Sinica* **28** (4), 2795–2809.

Li, D., Linton, O. and Lu, Z. (2015). A Flexible Semiparametric Forecasting Model for Time Series. *Journal of Econometrics* **187** (1), 345–357.

Li, J., Lv, J., Wan, A. T. and Liao, J. (2022). Adaboost Semiparametric Model Averaging Prediction for Multiple Categories. *Journal of the American Statistical Association* **117** (537), 495–509.

Li, Q. and Racine, J. (2003). Nonparametric Estimation of Distributions with Categorical and Continuous Data. *Journal of Multivariate Analysis* **86** (2), 266–292.

Liao, J., Zou, G., Gao, Y. and Zhang, X. (2021). Model Averaging Prediction for Time Series Models with a Diverging Number of Parameters. *Journal of Econometrics* **223** (1), 190–221.

Liboschik, T., Fokianos, K. and Fried, R. (2017). tscount: An R Package for Analysis of Count Time Series Following Generalized Linear Models. *Journal of Statistical Software* **82**, 1–51.

Lu, Z. (1998). On the Geometric Ergodicity of a Non-Linear Autoregressive Model with an Autoregressive Conditional Heteroscedastic Term. *Statistica Sinica* **8** (4), 1205–1217.

Lu, Z., Steinskog, D. J., Tjøstheim, D. and Yao, Q. (2009). Adaptively Varying-Coefficient Spatiotemporal Models. *Journal of the Royal Statistical Society: Series B Statistical Methodology* **71** (4), 859–880.

Lu, Z., Tjøstheim, D. and Yao, Q. (2007). Adaptive Varying-Coefficient Linear Models for Stochastic Processes: Asymptotic Theory. *Statistica Sinica* **17** (1), 177–198.

Masry, E. and Tjøstheim, D. (1995). Nonparametric Estimation and Identification of Nonlinear ARCH Time Series Strong Convergence and Asymptotic Normality: Strong Convergence and Asymptotic Normality. *Econometric Theory* **11** (2), 258–289.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Volume 37. CRC Press.

Murray, S., Xia, Y. and Xiao, H. (2024). Charting by Machines. *Journal of Financial Economics* **153**, 103791.

Peng, R. and Lu, Z. (2023). Uniform Consistency for Local Fitting of Time Series Non-Parametric Regression Allowing for Discrete-Valued Response. *Statistics and Its Interface* **16**, 305–318.

Peng, R. and Lu, Z. (2024). Semiparametric Averaging of Nonlinear Marginal Logistic Regressions and Forecasting for Time Series Classification. *Econometrics and Statistics* **31**, 19–37.

Qaqish, B. F. (2003). A Family of Multivariate Binary Distributions for Simulating Correlated Binary Variables with Specified Marginal Means and Correlations. *Biometrika* **90** (2), 455–463.

Racine, J. and Li, Q. (2004). Nonparametric Estimation of Regression Functions with Both Categorical and Continuous Data. *Journal of Econometrics* **119** (1), 99–130.

Racine, J. S., Li, Q., Yu, D. and Zheng, L. (2023). Optimal Model Averaging of Mixed-Data Kernel-Weighted Spline Regressions. *Journal of Business & Economic Statistics* **41** (4), 1251–1261.

Rubin, D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology* **66** (5), 688.

Steel, M. F. (2020). Model Averaging and Its Use in Economics. *Journal of Economic Literature* **58** (3), 644–719.

Stock, J. H. and Watson, M. W. (2006). Forecasting with Many Predictors. In *Handbook of Economic Forecasting* (Edited by G. Elliott, C. W. J. Granger and A. Timmermann), Volume 1, Elsevier.

Terasvirta, T., Tjøstheim, D. and Granger, C. W. (2010). *Modelling Nonlinear Economic Time Series*. Oxford University Press.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society: Series B Statistical Methodology* **58** (1), 267–288.

Tjøstheim, D. (1990). Non-Linear Time Series and Markov Chains. *Advances in Applied Probability* **22** (3), 587–611.

Weiß, C. H. (2018). *An Introduction to Discrete-Valued Time Series*. John Wiley & Sons.

Weiß, C. H. and Schnurr, A. (2023). Generalized Ordinal Patterns in Discrete-Valued Time Series: Nonparametric Testing for Serial Dependence. *Journal of Nonparametric Statistics*, 1–27.

Winkelmann, R. and Zimmermann, K. F. (1994). Count Data Models for Demographic Data. *Mathematical Population Studies* **4** (3), 205–221.

Yu, D., Tang, N.-S. and Shi, Y. (2025). Adaptively Aggregated Forecast for Exponential Family Panel Model. *International Journal of Forecasting* **41** (2), 733–747.

Yu, D., Zhang, X. and Yau, K. K. (2018). Asymptotic Properties and Information Criteria for Misspecified Generalized Linear Mixed Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **80** (4), 817–836.

Zhang, X., Lu, Z. and Zou, G. (2013). Adaptively Combined Forecasting for Discrete Response Time Series. *Journal of Econometrics* **176** (1), 80–91.

Zhang, X., Yu, D., Zou, G. and Liang, H. (2016). Optimal Model Averaging Estimation for Generalized Linear Models and Generalized Linear Mixed-Effects Models. *Journal of the American Statistical Association* **111** (516), 1775–1790.

Zhang, X. and Zhang, X. (2023). Optimal Model Averaging Based on Forward-Validation. *Journal of Econometrics* **237** (2), 105295.

Zhao, P. and Yu, B. (2006). On Model Selection Consistency of Lasso. *Journal of Machine Learning Research* **7** (Nov), 2541–2563.

Zhou, W., Gao, J., Harris, D. and Kew, H. (2024). Semi-Parametric Single-Index Predictive Regression Models with Cointegrated Regressors. *Journal of Econometrics* **238** (1), 105577.

Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association* **101** (476), 1418–1429.

School of Mathematics, Hunan University, Hunan, China.

E-mail: (rongpeng@hnu.edu.cn)

Department of Biostatistics, City University of Hong Kong, Hong Kong SAR, China.

E-mail: (zudilu@cityu.edu.hk)

University of Southampton Business School, University of Southampton, Southampton, UK.

E-mail: (f.ge@soton.ac.uk)