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*

Supplementary Material

In this supplementary document, we will provide the technical details for the theorems present in the main text with the explicit assumptions and sketches of proof given in Section S1. An algorithm for the penalized GMAFMA procedure is presented in Section S2. Additional details on the simulation and real-world application are then provided in Section S3 and S4, respectively. Sections S5 and S6 are added, as suggested by a referee, to examine the Mallows-type model averaging in our context of nonparametric marginal regression models and the tuning parameter selection by a forward-CV compared with our applied *cv.glmnet* (from R package **glmnet**) for PGMAFMA, respectively. We remark here all equations in this supplementary are numbered with A to distinguish them from those in the main text.

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

S1 Technical details

S1.1 Proof of Proposition 1

Suppose model (1) holds for I_{t-1} representing $X_t = (X_{1t}, \cdots, X_{dt})^T$ and its component X_{jt}, $j = 1, \cdots, d$, respectively. Then the conditional distributions (either probability density, or probability, functions) of Y_t given $X_t = (X_{1t}, \cdots, X_{dt})^T$ and its component X_{jt}, $j = 1, \cdots, d$, are given, respectively, by $f_{Y_t|X_t}(y) = \exp\{y\theta_t - \psi(\theta_t) + \phi(y, \Theta_t)\}$, and $f_{Y_t|X_{jt}}(y) = \exp\{y\theta_{jt} - \psi(\theta_{jt}) + \phi(y, \Theta_{jt})\}$, where $\theta_t = \eta(\mu_t) \equiv \theta(X_t)$ and $\theta_{jt} = \eta(\mu_{jt}) \equiv \theta_j(X_{jt})$, with $\eta(\cdot) = (\psi')^{-1}(\cdot)$ known, and $\mu_t = E(Y_t|X_t)$ and $\mu_{jt} = E(Y_t|X_{jt})$. Thus for $x = (x_1, \cdots, x_d)'$, $f_{Y_t|X_t=x}(y) = \exp\{y\theta(x) - \psi(\theta(x)) + \phi(y, \Theta(x))\}$, $f_{Y_t|X_{jt}=x_j}(y) = \exp\{y\theta_j(x_j) - \psi(\theta_j(x_j)) + \phi(y, \Theta_j(x_j))\}$, where $\Theta_t = \Theta(X_t)$ and $\Theta_{jt} = \Theta_j(X_{jt})$.

Now given that $X_{1t}, \cdots, X_{dt}$ are conditionally independent given Y_t, we have the conditional distribution of $X_t = (X_{1t}, \cdots, X_{dt})^T$ given $Y_t = y$ is

$$f_{X_t|Y_t}(x|y) = \prod_{j=1}^{d} f_{X_{jt}|Y_t}(x_j|y), \tag{S1.1}$$

where $f_{X_{jt}|Y_t}(x_j|y)$ is conditional distribution of X_{jt} given $Y_t = y$, for $j = 1, \cdots, d$. Note that $f_{X_t|Y_t}(x|y) = f_{Y_t|X_t=x}(y) f_X(x)/f_Y(y)$ and $f_{X_{jt}|Y_t}(x_j|y) = f_{Y_t|X_{jt}=x_j}(y) f_{X_j}(x_j)/f_Y(y)$, where $f_X(x)$, $f_{X_j}(x_j)$ and $f_Y(y)$ are the probability density, or probability, functions for X_t, X_{jt} and Y_t, respectively. Taking a logarithm on both sides of (S1.1), we easily get

$$y\theta(x) - \psi(\theta(x)) + \phi(y, \Theta(x)) + \log f_X(x) - \log f_Y(y)$$

$$= \sum_{j=1}^{d} [y\theta_j(x_j) - \psi(\theta_j(x_j)) + \phi(y, \Theta_j(x_j)) + \log f_{X_j}(x_j) - \log f_Y(y)], \tag{S1.2}$$

for y and x appropriately given. Thus

$$\begin{aligned} & y[\theta(x) - \sum_{j=1}^{d} \theta_j(x_j)] + \phi(y, \Theta(x)) - \sum_{j=1}^{d} \phi(y, \Theta_j(x_j)) + (d-1)\log f_Y(y) \\ & = \psi(\theta(x)) - \sum_{j=1}^{d} \psi(\theta_j(x_j)) + \sum_{j=1}^{d} \log f_{X_j}(x_j) - \log f_X(x). \end{aligned} \tag{S1.3}$$

Letting $y = 0$ in (S1.3), we get $\phi(0, \Theta(x)) - \sum_{j=1}^{d} \phi(0, \Theta_j(x_j)) + (d-1)\log f_Y(0)$

$$= \psi(\theta(x)) - \sum_{j=1}^{d} \psi(\theta_j(x_j)) + \sum_{j=1}^{d} \log f_{X_j}(x_j) - \log f_X(x).$$

Then it follows from (S1.3) that

$$\begin{aligned} & y[\theta(x) - \sum_{j=1}^{d} \theta_j(x_j)] + \phi(y, \Theta(x)) - \sum_{j=1}^{d} \phi(y, \Theta_j(x_j)) + (d-1)\log f_Y(y) \\ & = \phi(0, \Theta(x)) - \sum_{j=1}^{d} \phi(0, \Theta_j(x_j)) + (d-1)\log f_Y(0). \end{aligned} \tag{S1.4}$$

Under the assumptions that $\phi(y, \Theta) \equiv \phi(y)$ is independent of a nuisance parameter Θ, we simplify (S1.4) as follows

$$y[\theta(x) - \sum_{j=1}^{d} \theta_j(x_j)] = (d-1)[\phi(y) - \phi(0)] + (d-1)[\log f_Y(0) - \log(f_Y(y))]. \tag{S1.5}$$

We hence have $\theta(x) - \sum_{j=1}^{d} \theta_j(x_j) = c$, where c is a constant independent of $x = (x_1, \cdots, x_d)^T$ and y. Thus $\theta(x) = \sum_{j=1}^{d} \theta_j(x_j) + c = \alpha_0 + \sum_{j=1}^{d} \alpha_j \theta_j(x_j)$, with $\alpha_0 = c$ and $\alpha_j = 1$ for $j = 1, \cdots, d$.

S1.2 Regularity assumptions

We first introduce some necessary assumptions. Recall we suppose (Y_t, X_t^T) is β-mixing, for which we give the following definition:

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

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

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

Here are some assumptions introduced.

Assumption 1. A1. (i) $Z_t = (Y_t, X_t)$ is β-mixing with the mixing coefficient $\beta(t) = O(t^{-b})$, as $t \to \infty$, for some $b > max(2(\rho r + 1)/(\rho r - 2),\ (r + a)/(1 - 2/\rho))$ with $a \geq (r\rho - 2)r/(2 + r\rho - 4r)$, and Y_t, given X_t, has a conditional distribution in the exponential family as specified in (1) with $\psi(\cdot)$ having continuous first and second order derivatives, and so does Y_t, given the k-th component X_{kt}, for $k = 1, \cdots, d$; (ii) the joint probability density function $g_{X_{t_0}, X_{t_1}, \cdots, X_{t_s}}(x_0, x_1, \cdots, x_s)$ is bounded uniformly for any $t_0 < t_1 < \cdots < t_s$ and $0 \leq s \leq 2(r-1)$; (iii) $E|\omega(Y_t, f(X_t))|^{\rho r} < \infty$, $E|Y_t|^{\rho r} < \infty$ for some real number $\rho > 4 - 2/r$, where $\omega(Y_t, f(X_t))$ is an appropriate defined function denoting the distance between Y_t and $f(X_t)$, and $r \geq 1$ is some positive integer.

A2. (i) The bandwidth $h = h_n$ satisfies the conditions $\lim_{n\to\infty} h = 0$ and $\liminf_{n\to\infty} nh^{\frac{2(r-1)a+(\rho r-2)}{(a+1)\rho}} > 0$ for some integer $r \geq 3$; (ii) There exists a sequence of positive integers $s_n \to \infty$ such that $s_n = o((nh)^{1/2}), ns_n^{-b} \to 0$ and $s_n h^{\frac{2(\rho r-2)}{[2+b(\rho r-2)]}} > 1$ as $n \to \infty$.

This assumption A2 is easily satisfied. For example, if we take $s_n = [(nh)^{\widetilde{s}/2}]$ (with $[\cdot]$ standing for the integer part here) for some $0 < \widetilde{s} < 1$, then $s_n = o((nh)^{1/2})$ naturally holds, and $ns_n^{-b} = n(nh)^{-b\widetilde{s}/2} = (n^{1-2/(b\widetilde{s})}h)^{-b\widetilde{s}/2} \to 0$ and $s_n h^{\frac{2(\rho r-2)}{[2+b(\rho r-2)]}} = \left\{nh^{1+\frac{4(\rho r-2)}{[2+b(\rho r-2)]\widetilde{s}}}\right\}^{\widetilde{s}/2} > 1$ also easily hold if $n^{1-2/(b\widetilde{s})}h \to \infty$, which is easily satisfied by letting b be sufficiently large.

A3. The weight function $W(X_t) = \prod_{k=1}^{d} I_{(c_{k0} \leq X_{kt} \leq c_{k1})}$ with $c_{k0} < c_{k1}$ appropriately chosen to be sufficiently large, where $I_{(\cdot)}$ is an indicator function.

A4. The kernel $K(\cdot)$ is a bounded and symmetric density function on R^1 with bounded support S_K. Furthermore, $|K(z) - K(x)| \leq C|z - x|$ for $z, x \in S_K$ and some $0 < C < \infty$.

A5. Let $\mathbf{f}^0(x) = (f_1^0(x_1), ..., f_d^0(x_d))^T$ be the vector of the true conditional regression functions for $x = (x_1, \cdots, x_d)^T \in R^d$, with $f_k^0(\cdot)$'s the true functions of $f_k(\cdot)$'s given in Equation (3), more precisely defined right after Equation (11). For an $\mathbf{f}(\cdot)$, define its Lipschitz norm: For some $\varsigma > 0$, let $[\varsigma]$ be the largest integer not greater than ς, and define (if it exists)

$$\|\mathbf{f}\|_{\infty,\varsigma} = \max_{0\le\kappa\le[\varsigma]} \sup_{x\in A} \|\mathbf{f}^{(\kappa)}(x)\| + \sup_{x\neq x';x,x'\in A} \frac{\|\mathbf{f}^{([\varsigma])}(x) - \mathbf{f}^{([\varsigma])}(x')\|}{\|x-x'\|^{\varsigma-[\varsigma]}}, \tag{S1.6}$$

where $\|\cdot\|$ stands for the Euclidean norm, $\mathbf{f}^{(\kappa)}(x)$ is the component-wise κ-th derivative of $\mathbf{f}(x)$, and $A = \prod_{k=1}^{d}[c_{k0}, c_{k1}]$ with some real values of c_{k0} and c_{k1} satisfying $c_{k0} < c_{k1}$ given in A2 of Assumption 1. We suppose $\mathbf{f}^0(\cdot)$ with f_k^0's belongs to the functional space $\mathbf{F}$ with $\varsigma \geq 2$:

$$\mathbf{F} := \{\mathbf{f} : \text{continuous from } A \text{ to } R^d \text{ with } \|\mathbf{f}\|_{\infty,\varsigma} \leq c\}, \tag{S1.7}$$

where c is a positive constant. This functional space $\mathbf{F}$ (containing functions $\mathbf{f}$ whose Lipschitz norm is bounded) is often denoted by $C_c^{\varsigma}(A)$.

A6. For the local likelihood function (9), define $\Phi(Y_t, z_k) = Y_t - \psi'(z_k)$, and

$$m(x_k, z_k) = E[\Phi(Y_t, z_k)|X_{kt} = x_k], \tag{S1.8}$$

satisfying $(x_k, z_k) \to m(x_k, z_k) \cdot g_k(x_k)$ is three times continuously differentiable as a function from R^2 to R^1, where $g_k(x_k)$ is the marginal density of X_{kt}, which is strictly positive and continuous over $A_k = [c_{k0}, c_{k1}]$. We denote the derivative of m with respect to x_k by m'_1, and the derivative with respect z_k by m'_2, etc.

Remark 1. Note that A1 of Assumption 1 gives the weak dependency of time series, which is β-mixing (Fan and Yao, 2003; Lu, Tjøstheim and Yao, 2007) with the condition on the mixing coefficient borrowed from Peng and Lu (2023). A2 of Assumption 1 is standard in time series topics (Fan, Yao and Cai, 2003; Lu, Tjøstheim and Yao, 2007). The edge effect is controlled by A3 of Assumption 1, which removes the extreme estimates around the boundaries of X_t, to improve the practical performance of the estimation (c.f. Fan, Härdle and Mammen (1998); Fan, Yao and Cai (2003) and Lu, Tjøstheim and Yao (2007)). The kernel is guaranteed to be bounded by A4 of Assumption 1 which is commonly seen in this type of problem (Hardle et al., 1993; Xia and Li, 1999). A5 and A6 of Assumption 1 give smoothness conditions on the conditional regression and marginal density functions. The Lipschitz norm conditions (A5 of Assumption 1) are introduced to give a tighter bound than uniform norm (Nielsen, 2005). For more information on Lipschitz norm, the reader is referred to Van Der Vaart and Wellner (1996). For more details on the assumptions in Assumption 1, the reader is referred to Peng and Lu (2023).

For PGMAFMA, we further introduce some mild conditions to establish the asymptotic results, given in Assumption 2 in addition to Assumption 1.

Assumption 2. B1. We assume under the true parameter $\boldsymbol{\alpha}^*$, $E[\frac{\partial L(\boldsymbol{\alpha}^*;\mathbf{f}^0)}{\partial \boldsymbol{\alpha}}] = 0$, where $L(\boldsymbol{\alpha};\mathbf{f})$ is defined in (13). The matrix $\boldsymbol{U} = E[\phi''(\mathbf{f}^0,\boldsymbol{\alpha}^*)]\widetilde{\chi}_t(\mathbf{f}^0)\widetilde{\chi}_t(\mathbf{f}^0)^T W(X_t)$ under true model is finite and positive definite.

B2. There is an sufficiently large enough open subset Ω that contains $\boldsymbol{\alpha}^*$(true parameter), such that $\forall \boldsymbol{\alpha} \in \Omega$, there exists a finite function $\psi_{jks} = E_{\boldsymbol{\alpha}^*}[\Psi_{jks}(X)] < \infty$: $|\frac{\partial^3 L(X,\alpha)}{\partial\alpha_k\partial\alpha_k\partial\alpha_s}| \leq \Psi_{jks}(X)$, where this Ψ is the upper bound uniformly with respect to α.

Remark 2. Assumption B1 and Assumption B2 are often adopted in conventional models to guarantee asymptotic normality of the maximum likelihood estimates (Fan and Li, 2001).

S1.3 Sketch of Proof of Theorem 1

Proof. (i) *Proof of consistency*:

First of all, we show the consistency of $\widehat{\boldsymbol{\alpha}}^{*(n)}$. That is, we would like to show:

$$\forall \delta > 0, P(\|\widehat{\boldsymbol{\alpha}}^{*(n)} - \boldsymbol{\alpha}^*\| > \delta) \to 0, \text{as } n \to \infty.$$

Here to show this consistency, we will need Proposition 1 (Consistency Lemma) stated below, which is adapted from Lemma 4.1 in Lu, Tjøstheim and Yao (2007). The consistency of $\widehat{\boldsymbol{\alpha}}^{*(n)}$ can then be established by checking the conditions specified in Proposition 1.

Proposition 1. (Consistency) Suppose $\boldsymbol{\alpha}^* \in \mathfrak{A}$ satisfies $L(\boldsymbol{\alpha}^*, \mathbf{f}^0(\cdot)) = \max_{\boldsymbol{\alpha}\in\mathfrak{A}} L(\boldsymbol{\alpha}, \mathbf{f}^0(\cdot))$, where $L(\boldsymbol{\alpha}, \mathbf{f}^0(\cdot))$ is the $L(\boldsymbol{\alpha})$ defined in (11), with $\mathbf{f}^0(\cdot)$ the true function vector in Assumption A5, $\mathfrak{A}$ is a closed set in $\mathbb{R}^{d+1}$ with $\boldsymbol{\alpha}^*$ an interior point of $\mathfrak{A}$, and that

i. $L_n(\widehat{\boldsymbol{\alpha}}^{*(n)}, \widehat{\mathbf{f}}(\cdot)) \leq \max_{\boldsymbol{\alpha}\in\mathfrak{A}} L_n(\boldsymbol{\alpha}, \widehat{\mathbf{f}}(\cdot)) + o_P(1)$, where $L_n(\boldsymbol{\alpha}, \widehat{\mathbf{f}}(\cdot))$ is defined in (10).

ii. For all $\delta > 0$, there exists $\epsilon(\delta) > 0$ such that

$$\inf_{\|\boldsymbol{\alpha}-\boldsymbol{\alpha}^*\|>\delta} |L(\boldsymbol{\alpha}, \mathbf{f}^0(\cdot)) - L(\boldsymbol{\alpha}^*, \mathbf{f}^0(\cdot)| \geq \epsilon(\delta).$$

iii. Uniformly for all $\boldsymbol{\alpha} \in \mathfrak{A}$, $L(\boldsymbol{\alpha}, \mathbf{f}(\cdot))$ is continuous with respect to the metric $\|\cdot\|_{\mathbf{F}}$ in $\mathbf{f}(\cdot)$ at $\mathbf{f}^0(\cdot)$, where $\|\mathbf{f}(\cdot)\|_{\mathbf{F}} = \sup_{x\in A} \|\mathbf{f}(x)\|$ with $\|\cdot\|$ being the Euclidean norm of R^d, and $A = \prod_{k=1}^d [c_{k0}, c_{k1}]$ is the support of the weight function $W(x)$ defined in Assumption 1 A3.

iv. $\|\widehat{\mathbf{f}}(\cdot) - \mathbf{f}^0(\cdot)\|_{\mathbf{F}} = o_P(1)$.

v. For all δ_n with $\delta_n = o(1)$, $\sup_{\boldsymbol{\alpha}\in\mathfrak{A}} \sup_{\|\mathbf{f}-\mathbf{f}^0\|_{\mathbf{F}}\le\delta_n} |L_n(\boldsymbol{\alpha}, \mathbf{f}(\cdot)) - L(\boldsymbol{\alpha}, \mathbf{f}(\cdot))| = o_P(1)$.

Then $\widehat{\boldsymbol{\alpha}}^{*(n)} - \boldsymbol{\alpha}^* = o_P(1)$

The proof of Proposition 1 is omitted, which follows from Lemma 4.1 in Lu, Tjøstheim and Yao (2007). We now check the conditions of Proposition 1 one by one. As $\widehat{\boldsymbol{\alpha}}^{*(n)}$ and $\boldsymbol{\alpha}^*$ are the maximizers of $L_n(\boldsymbol{\alpha}, \widehat{\mathbf{f}}(\cdot))$ and $L(\boldsymbol{\alpha}, \mathbf{f}^0(\cdot))$, respectively, (i) and (ii) hold obviously, where (ii) holds by noticing that $L(\boldsymbol{\alpha}, \mathbf{f}^0(\cdot))$ is a continuous function of $\boldsymbol{\alpha}$, as defined in (11). (iii) also holds clearly by the following fact:

$$L(\boldsymbol{\alpha}, \mathbf{f}) = E[\{Y_t(\alpha_0 + \sum_{k=1}^d \alpha_k f_k(X_{kt}))\} - \psi(\alpha_0 + \sum_{k=1}^d \alpha_k f_k(X_{kt})) + \phi(Y_t, \Theta)]W(X_t) \tag{S1.9}$$

$$\equiv E[(Y_t \widetilde{\chi}_t(\mathbf{f})^T \boldsymbol{\alpha}) - \psi(\boldsymbol{\alpha}, \mathbf{f}) + \phi(Y_t, \Theta)]W(X_t),$$

where $\widetilde{\chi}_t(\mathbf{f}) = (1, f_1(x_{1t})..., f_d(x_{dt}))^T$ with f_k's being marginal functions that are generally different from those in $\mathbf{f}^0$ given in A5 of Assumption 1 at a cost of slight notation confusion, and $\psi(\boldsymbol{\alpha}, \mathbf{f}) = \psi(\widetilde{\chi}_t(\mathbf{f})^T \boldsymbol{\alpha})$. Then

$$\sup_{\boldsymbol{\alpha}\in\mathfrak{A}} |L(\boldsymbol{\alpha}, \mathbf{f}(\cdot)) - L(\boldsymbol{\alpha}, \mathbf{f}^0(\cdot))| \le \sup_{\boldsymbol{\alpha}\in\mathfrak{A}} E[Y_t \|\widetilde{\chi}_t(\mathbf{f}) - \widetilde{\chi}_t(\mathbf{f}^0)\| \boldsymbol{\alpha} + |\psi(\boldsymbol{\alpha}, \mathbf{f}) - \psi(\boldsymbol{\alpha}, \mathbf{f}^0)|]W(X_t)$$

$$\leq \sup_{\boldsymbol{\alpha}\in\mathfrak{A}} E[|Y_t| + |\psi'(\boldsymbol{\alpha},\mathbf{f})|]\|\widetilde{\chi}_t(\mathbf{f}) - \widetilde{\chi}_t(\mathbf{f}^0)\|\|\boldsymbol{\alpha}\|W(X_t) \leq C\|\mathbf{f}-\mathbf{f}^0\|_{\mathbf{F}}, \tag{S1.10}$$

where C is a generic constant. Here the second inequality follows by taking the Taylor expansion of the second term on the RHS and noting that $E(Y_t) = \mu_t = \psi'(\boldsymbol{\alpha},\mathbf{f})$. The last inequality follows from the fact that $\|\mathbf{f}-\mathbf{f}^0\|_{\mathbf{F}} = \|\mathbf{f}-\mathbf{f}^0\|_{\infty} = \max_{i=1,\dots,d}\sup_{x_i}|f_i(x_i) - f_{i0}(x_i)|W(x)$ with $x = (x_1,\cdots,x_d)^T$.

Now, to prove (iv), we need to show that the estimator $\widehat{f}_k(.)$ replacing the $f_k(.)$ function in the model averaging step is uniformly consistent under A1, A2, A4, A5 and A6 of Assumption 1. This uniform consistency of the local fitting technique follows from Theorem 3.2 of Peng and Lu (2023).

To check (v), let $\delta_n = o(1)$ and $\|\mathbf{f}-\mathbf{f}^0\|_{\mathbf{F}} \leq \delta_n$. Then we have:

$$\begin{aligned} L_n(\boldsymbol{\alpha},\mathbf{f}(\cdot)) - L(\boldsymbol{\alpha},\mathbf{f}(\cdot)) &= \{L_n(\boldsymbol{\alpha},\mathbf{f}(\cdot)) - L_n(\boldsymbol{\alpha},\mathbf{f}^0(\cdot))\} \\ &\quad + \{L_n(\boldsymbol{\alpha},\mathbf{f}^0(\cdot)) - L(\boldsymbol{\alpha},\mathbf{f}^0(\cdot))\} \qquad \text{(S1.11)} \\ &\quad + \{L(\boldsymbol{\alpha},\mathbf{f}^0(\cdot)) - L(\boldsymbol{\alpha},\mathbf{f}(\cdot))\} \\ &=: I + II + III. \qquad \text{(S1.12)} \end{aligned}$$

We need to show that uniformly, for $\boldsymbol{\alpha}\in\mathfrak{A}$ and $\mathbf{f}$ satisfying $\|\mathbf{f}-\mathbf{f}^0\|_{\mathbf{F}} \leq \delta_n$, terms I, II and III tend to zero in probability as $n\to\infty$. First, it is easy to show that term III tending to zero follows from equation (S1.10), that is, $|L(\boldsymbol{\alpha},\mathbf{f}^0(\cdot)) - L(\boldsymbol{\alpha},\mathbf{f}(\cdot))| \leq C\|\mathbf{f}-\mathbf{f}^0\|_{\mathbf{F}} \leq C\delta_n \to 0, \text{ as } n\to\infty$. Next, by the fact that

$$L_n(\boldsymbol{\alpha},\mathbf{f}^0(\cdot)) = \frac{1}{n}\sum_{t=1}^{n}\left[(Y_t\widetilde{\chi}_t(\mathbf{f}^0)^T\boldsymbol{\alpha}) - \psi(\boldsymbol{\alpha},\mathbf{f}^0) + \phi(Y_t,\Theta)\right]W(X_t), \tag{S1.13}$$

term II tending to zero in probability uniformly with respect to $\boldsymbol{\alpha} \in \mathfrak{A}$ can be easily proved by the law of large number together with $\mathfrak{A}$ being a compact set. Thirdly, note that III is the expected value of I. That term I tends to zero in probability uniformly with respect to $\boldsymbol{\alpha} \in \mathfrak{A}$ and $\mathbf{f}$ satisfying $\|\mathbf{f} - \mathbf{f}^0\|_{\mathbf{F}} \leq \delta_n$ can be proved similarly as that for III. Hence we know that $I + II + III$ tends to zero in probability uniformly with respect to $\boldsymbol{\alpha} \in \mathfrak{A}$ and $\mathbf{f}$ satisfying $\|\mathbf{f} - \mathbf{f}^0\|_{\mathbf{F}} \leq \delta_n$.

By completing the checking of the conditions of Proposition 1 (Consistency), the proof of the consistency of $\widehat{\boldsymbol{\alpha}}^{*(n)}$ converging to α^* in probability is completed.

(ii) *Proof of asymptotic normality*:

Now we turn to the proof of asymptotic normality of $\widehat{\boldsymbol{\alpha}}^{*(n)}$. By definition of $\widehat{\boldsymbol{\alpha}}^{*(n)}$ with $L_n(\boldsymbol{\alpha})$ defined in (10), we have $\frac{\partial L_n(\widehat{\boldsymbol{\alpha}}^{*(n)})}{\partial \alpha} = 0$, that is

$$\boldsymbol{V}_n(\widehat{\boldsymbol{\alpha}}^{*(n)}, \widehat{\mathbf{f}}) =: n^{-1} \sum_{t=1}^{n} [Y_t - \psi'(\widehat{\alpha}_0 + \sum_{k=1}^{d} \widehat{\alpha}_k \widehat{f}_k(X_{kt}))] \widetilde{\chi}_t(\widehat{\mathbf{f}}) W(X_t) = 0, \tag{S1.14}$$

where $\widetilde{\chi}_t(\widehat{\mathbf{f}}) = (1, \widehat{f}_1(x_{1t})..., \widehat{f}_d(x_{dt}))^T$. Similarly, we denote $\widetilde{\chi}_t(\mathbf{f}^0) = (1, f_1(x_{1t})..., f_d(x_{dt}))^T$. By Taylor's expansion,

$$\widehat{\boldsymbol{\alpha}}^{*(n)} - \boldsymbol{\alpha}^* = \boldsymbol{U}_n^{-1}(\widetilde{\alpha}, \widehat{\mathbf{f}}) \boldsymbol{V}_n(\alpha^*, \widehat{\mathbf{f}}), \tag{S1.15}$$

where $\boldsymbol{V}_n(\alpha^*, \widehat{\mathbf{f}}) = n^{-1} \sum_{t=1}^{n} [Y_t - \psi'(\alpha_0^* + \sum_{k=1}^{d} \alpha_k^* \widehat{f}_k(X_{kt}))] \widetilde{\chi}_t(\widehat{\mathbf{f}}) W(X_t)$, and

$$\boldsymbol{U}_n(\widetilde{\alpha}, \widehat{\mathbf{f}}) = n^{-1} \sum_{t=1}^{n} [\psi''(\widetilde{\alpha}_0 + \sum_{k=1}^{d} \widetilde{\alpha}_k \widehat{f}_k(X_{kt}))] \widetilde{\chi}_t(\widehat{\mathbf{f}}) \widetilde{\chi}_t(\widehat{\mathbf{f}})^T W(X_t),$$

with $\widetilde{\alpha} = \alpha^* + a(\widehat{\alpha}^{*(n)} - \alpha^*)$, for some $a \in [0, 1]$.

To show the asymptotic normality on $\widehat{\alpha}^{*(n)}$, we will need an asymptotic expression on $\widehat{f}_k(x_k) - f_k^0(x_k)$ uniformly over $x_k \in [c_{k0}, c_{k1}]$ for $k = 1, \cdots, d$. Note that differently from the least squares MAFMA procedure with continuous-valued Y_t in Li, Linton and Lu (2015), in which the least squares local linear marginal regressions have analytical expressions, our maximum likelihood based local linear marginal regressions $\widehat{f}_k(x_k)$'s, defined in (9), however, do not own analytical solutions. So we need to derive their asymptotic expressions to facilitate our needs below. We introduce some necessary notations for the k-th marginal regression as similarly given in Peng and Lu (2023). Let $\boldsymbol{\beta} = (f_k(x_k), f_k'(x_k))^T$ and $\boldsymbol{\beta}_0 = (f_k^0(x_k), (f_k^0)'(x_k))^T$. Denote $\omega(Y_t, z) = Y_t - \psi'(z)$, $\Phi(x_k, z) = E[\omega(Y_t, z)|X_{kt} = x_k] = E(Y_t|X_{kt} = x_k) - \psi'(z) = \psi'(f_k^0(x_k)) - \psi'(z)$. Then the derivative of $\Phi(x_k, z)$ with respect to z is $\dot{\Phi}_z(x_k, z) = -\psi''(z)$, and $\dot{\Phi}_{x_k}(x_k, z) = \psi''(f_k^0(x_k))(f_k^0(x_k))'$. Clearly $\Phi(x_k, f_k^0(x_k)) = 0$, where $f_k^0(x)$ is the true function for equation (4), also denoted as β_{10} for the first component of $\boldsymbol{\beta}_0$. Then we define $\Omega_{k,0}(\boldsymbol{\beta}, x_k) = (\Omega_{k,0}^{(1)}(\boldsymbol{\beta}, x_k), \Omega_{k,0}^{(2)}(\boldsymbol{\beta}, x_k))^T$, with $\Omega_{k,0}^{(1)}(\boldsymbol{\beta}, x_k) = \Phi(x_k, \beta_1)g_k(x_k)$ and

$$\Omega_{k,0}^{(2)}(\boldsymbol{\beta}, x_k) = (\beta_2\dot{\Phi}_z(x_k, \beta_1) + \dot{\Phi}_{x_k}(x_k, \beta_1))g_k(x_k) + \Phi(x_k, \beta_1)g_k'(x_k),$$

with β_ℓ for the ℓ-th component of $\boldsymbol{\beta}$ for $\ell = 1, 2$, and our estimator $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_n(x_k) = (\widehat{f}_k(x_k), \widehat{f}_k'(x_k))^T$ defined in (9) is the solution to:

$$\Omega_{k,n}(\boldsymbol{\beta}, x_k, h) = \begin{pmatrix} \Omega_{k,n}^{(1)}(\boldsymbol{\beta}, x_k, h) \\ \Omega_{k,n}^{(2)}(\boldsymbol{\beta}, x_k, h) \end{pmatrix} = 0, \tag{S1.16}$$

with $\Omega_{k,n}^{(1)}(\boldsymbol{\beta}, x_k, h) = \frac{1}{n}\sum_{t=1}^n\{\omega(Y_t; \beta_1 + h\beta_2((X_{kt} - x_k)/h))K_h(X_{kt} - x_k)\}$,

$$\Omega_{k,n}^{(2)}(\boldsymbol{\beta}, x_k, h) = \frac{1}{n}\sum_{t=1}^n\{\omega(Y_t; \beta_1 + h\beta_2((X_{kt} - x_k)/h))[(X_{kt} - x_k)/h]K_h(X_{kt} - x_k)\}.$$

Based on $\Omega_{k,n}(\widehat{\boldsymbol{\beta}}, x_k, h) = 0$, by Taylor's expansion and the uniform consistency of $\widehat{\boldsymbol{\beta}}(x_k)$ to $\boldsymbol{\beta}_0(x_k)$ over $x_k \in A_k = [c_{k0}, c_{j1}]$ (from Peng and Lu (2023)), we easily have

$$0 = \Omega_{k,n}(\widehat{\boldsymbol{\beta}}, x_k, h) = \Omega_{k,n}(\boldsymbol{\beta}_0, x_k, h) + (1 + o_P(1))\dot{\Omega}_{k,\boldsymbol{\beta}_0}(x_k)(\widehat{\boldsymbol{\beta}}(x_k) - \boldsymbol{\beta}_0(x_k)),$$

and hence

$$\widehat{f}_k(x_k) - f_k^0(x_k) = (1,0)[\widehat{\boldsymbol{\beta}}(x_k) - \boldsymbol{\beta}_0(x_k)] = -(1 + o_P(1))(1,0)(\dot{\Omega}_{k,\boldsymbol{\beta}_0}(x_k))^{-1}\Omega_{k,n}(\boldsymbol{\beta}_0, x_k, h), \tag{S1.17}$$

where $o_P(1)$ is uniform with respect to $x_k \in A_k = [c_{k0}, c_{j1}]$, and $\dot{\Omega}_{k,\boldsymbol{\beta}_0}(x_k)$ denotes the derivative of $\Omega_{k,0}(\boldsymbol{\beta}, x_k)$ with respect to $\boldsymbol{\beta}$ at $\boldsymbol{\beta} = \boldsymbol{\beta}_0(x_k)$, and hence $\dot{\Omega}_{k,\boldsymbol{\beta}_0}(x_k)$ is a 2×2 matrix whose $(1,1)$th element equal to $\dot{\Phi}_z(x_k, \beta_{10})g_k(x_k)$, the $(1,2)$th elements 0's, the $(2,1)$th element equal to $\mu_{2K}\{[\boldsymbol{\beta}_{20}\ddot{\Phi}_{zz}(x_k, \beta_{10}) + \ddot{\Phi}_{xz}(x_k, \beta_{10})]g_k(x_k) + \dot{\Phi}_z(x_k, \beta_{10})g_k'(x_k)\}$, and the $(2,2)$th element $\mu_{2K}\dot{\Phi}_z(x_k, \beta_{10})g_k(x_k)$. Thus it follows from (S1.17) that

$$\widehat{\mathbf{f}}(x) - \mathbf{f}^0(x) = -(1 + o_P(1))\mathcal{C}_n(x),$$

where $\mathcal{C}_n(x)$ is a $d\times 1$ vector whose k-th element is $\mathcal{C}_{k,n}(x_k) = (1,0)(\dot{\Omega}_{k,\boldsymbol{\beta}_0}(x_k))^{-1}\Omega_{k,n}(\boldsymbol{\beta}_0, x_k, h)$, expressed by (S1.16) as follows

$$\Omega_{k,n}(\boldsymbol{\beta}_0, x_k, h) = \frac{1}{n}\sum_{t=1}^{n}\omega_{t,k,n}(\boldsymbol{\beta}_0, x_k, h), \quad \textit{and} \quad \mathcal{C}_{k,n}(x_k) = \frac{1}{n}\sum_{t=1}^{n}\mathcal{C}_{t,k,n}(x_k),$$

with $\omega_{t,k,n}(\boldsymbol{\beta}_0, x_k, h) = \{\omega(Y_t; f_k^0(x_k) + h\dot{f}_k^0(x_k)((X_{kt} - x_k)/h))[1, (X_{kt} - x_k)/h]^T K_h(X_{kt} - x_k)\}$, and $\mathcal{C}_{t,k,n}(x_k) = (1,0)(\dot{\Omega}_{k,\boldsymbol{\beta}_0}(x_k))^{-1}\omega_{t,k,n}(\boldsymbol{\beta}_0, x_k, h)$. Note that

$$E\omega_{t,k,n}(\boldsymbol{\beta}_0, x_k, h) = E[\psi(f_k(X_{kt})) - \psi(f_k^0(x_k) + \dot{f}_k^0(x_k)(X_{kt} - x_k))][1, \frac{X_{kt} - x_k}{h}]^T K_h(X_{kt} - x_k)$$

$$
\begin{aligned}
&= \frac{1}{2}h^2\psi'(f_k^0(x_k))\ddot{f}_k^0(x_k)\{\int_{R^1} u^2K(u)du, 0\}^T(1+o(1)),\\
E\mathcal{C}_{t,k,n}(x_k) &= \frac{1}{2}h^2(1,0)(\dot{\Omega}_{k,\boldsymbol{\beta}_0}(x_k))^{-1}\psi'(f_k^0(x_k))\ddot{f}_k^0(x_k)\{\int_{R^1} u^2K(u)du, 0\}^T(1+o(1))\\
&= \frac{1}{2}h^2B_k(x_k)(1+o(1)), \qquad \text{(S1.18)}
\end{aligned}
$$

where $B_k(x_k) = (1,0)(\dot{\Omega}_{k,\boldsymbol{\beta}_0}(x_k))^{-1}\psi'(f_k^0(x_k))\ddot{f}_k^0(x_k)\{\int_{R^1} u^2K(u)du, 0\}^T$, and $o(1)$ is uniform with respect to $x_k \in A_k = [c_{k0}, c_{k1}]$. Thus

$$
\widehat{\mathbf{f}}(x) - \mathbf{f}^0(x) = -(1+o_P(1))\frac{1}{n}\sum_{t=1}^n[\mathcal{C}_{t,n}(x) - E\mathcal{C}_{t,n}(x)] - \frac{1}{2}h^2\widetilde{B}(x)(1+o_P(1)), \quad \text{(S1.19)}
$$

where $\mathcal{C}_{t,n}(x)$ is a $d\times 1$ vector whose k-th element is $\mathcal{C}_{t,k,n}(x_k)$, and $\widetilde{B}(x)$ is similarly a $d \times 1$ vector whose k-th element is $B_k(x_k)$. Here $o_P(1)$ is uniform with respect to $x \in A = \prod_{k=1}^d [c_{k0}, c_{k1}]$.

Now note from (S1.15) we can rewrite

$$
\begin{aligned}
\boldsymbol{V}_n(\alpha^*, \widehat{\mathbf{f}}) &= \boldsymbol{V}_n(\alpha^*, \mathbf{f}^0) + [\boldsymbol{V}_n(\alpha^*, \widehat{\mathbf{f}}) - \boldsymbol{V}_n(\alpha^*, \mathbf{f}^0)]\\
&=: \boldsymbol{V}_n^1 + \boldsymbol{V}_n^2, \qquad \text{(S1.20)}
\end{aligned}
$$

where we need to consider, for $\widehat{\mathbf{f}}$ satisfying $\|\widehat{\mathbf{f}} - \mathbf{f}^0\|_\infty \le \delta_n$ with $\delta_n = (nh)^{-1/2} + h^2 \to 0$,

$$
\begin{aligned}
\boldsymbol{V}_n^1 &= \boldsymbol{V}_n(\alpha^*, \mathbf{f}^0) = n^{-1}\sum_{t=1}^n [Y_t - \psi'(\alpha_0^* + \sum_{k=1}^d \alpha_k^* f_k^0(X_{kt}))]\widetilde{\chi}_t(\mathbf{f}^0)W(X_t)\\
&= n^{-1}\sum_{t=1}^n m^*(Z_t, \boldsymbol{\alpha}^*, \mathbf{f}^0), \qquad \text{(S1.21)}\\
\boldsymbol{V}_n^2 &= \boldsymbol{V}_n(\alpha^*, \widehat{\mathbf{f}}) - \boldsymbol{V}_n(\alpha^*, \mathbf{f}^0) = n^{-1/2}(\nu_n(\widehat{\mathbf{f}}) - \nu_n(\mathbf{f}^0)) + (M^*(\widehat{\mathbf{f}}) - M^*(\mathbf{f}^0))\\
&=: \boldsymbol{V}_n^{21} + \boldsymbol{V}_n^{22}, \qquad \text{(S1.22)}
\end{aligned}
$$

where we define an empirical process $\nu_n(\mathbf{f}) = \frac{1}{\sqrt{n}}\sum_{t=1}^n(m^*(Z_t, \boldsymbol{\alpha}^*, \mathbf{f}) - Em^*(Z_t, \boldsymbol{\alpha}^*, \mathbf{f}))$ with $m^*(Z_t, \boldsymbol{\alpha}, \mathbf{f}) = [Y_t - \psi'(\alpha_0 + \sum_{k=1}^d \alpha_k f_k(X_{kt}))]\widetilde{\chi}_t(\mathbf{f})W(X_t)$, and denote by $M^*(\mathbf{f}) = Em^*(Z_t, \boldsymbol{\alpha}^*, \mathbf{f})$.

We first consider $\boldsymbol{V}_n^{22}$. For notational convenience, recall $Z_t = (Y_t, X_t)$, and we introduce a notation $\boldsymbol{\alpha}^*(\mathbf{f})$ that maximizes $L(\boldsymbol{\alpha}, \mathbf{f})$, defined in (11), with respect to $\boldsymbol{\alpha} \in \mathfrak{A}$ for a given generic function vector $\mathbf{f}$. We then have

$$\boldsymbol{V}(\alpha^*(\mathbf{f}), \mathbf{f}) := \frac{\partial L(\boldsymbol{\alpha}^*(\mathbf{f}), \mathbf{f})}{\partial \boldsymbol{\alpha}} = E[m^*(Z_t, \boldsymbol{\alpha}^*(\mathbf{f}), \mathbf{f})] = 0. \tag{S1.23}$$

Then $\boldsymbol{\alpha}^* = \boldsymbol{\alpha}^*(\mathbf{f}^0)$, and $\boldsymbol{V}(\alpha^*, \mathbf{f}^0) = E[m^*(Z_t, \boldsymbol{\alpha}^*, \mathbf{f}^0)] = 0$. Then note that $M^*(\mathbf{f}^0) = \boldsymbol{V}(\alpha^*, \mathbf{f}^0) = E[m^*(Z_t, \boldsymbol{\alpha}^*, \mathbf{f}^0)] = 0$, and $M^*(\mathbf{f}) = M^*(\mathbf{f}) - M^*(\mathbf{f}^0) = E[\dot{m}_{\mathbf{f}}^*(Z_t, \boldsymbol{\alpha}^*, \mathbf{f}^0 + a(\mathbf{f} - \mathbf{f}^0))(\mathbf{f} - \mathbf{f}^0))]$ for some $a \in [0, 1]$, where $\dot{m}_{\mathbf{f}}^* =: \dot{m}_{\mathbf{f}}^*(Z_t, \boldsymbol{\alpha}^*, \mathbf{f}^0)$ is the directional derivative of $m^*(Z_t, \boldsymbol{\alpha}^*, \mathbf{f}) = [Y_t - \psi'(\alpha_0^* + \sum_{k=1}^d \alpha_k^* f_k(X_{kt}))]\widetilde{\chi}_t(\mathbf{f})W(X_t)$ with respect to $\mathbf{f}$ at $\mathbf{f}^0$ along $(\mathbf{f} - \mathbf{f}^0)$. Then the k-th column of $\dot{m}_{\mathbf{f}}^*$ is $\dot{m}_{\mathbf{f},k}^*(Z_t, \boldsymbol{\alpha}^*, \mathbf{f}^0) = -\psi''(\alpha_0^* + \sum_{k=1}^d \alpha_k^* f_k^0(X_{kt}))\alpha_k^* \widetilde{\chi}_t(\mathbf{f}^0)W(X_t) + [Y_t - \psi'(\alpha_0^* + \sum_{k=1}^d \alpha_k^* f_k^0(X_{kt}))]\boldsymbol{\gamma}_k W(X_t)$, where $\boldsymbol{\gamma}_k$ is a vector of dimension $(d+1)$ whose elements being zeros except $(k+1)$-th element being 1. Thus, in view of the uniform consistency of $\|\widehat{\mathbf{f}} - \mathbf{f}^0\|_\infty = \max_{1\le k\le d}\sup_{x_k\in[c_{k0}, c_{k1}]}|\widehat{f}_k(x_k) - f_k^0(x_k)| = o_P(1)$ together with (S1.19),

$$\begin{aligned}
\boldsymbol{V}_n^{22} &= M^*(\widehat{\mathbf{f}}) = M^*(\widehat{\mathbf{f}}) - M^*(\mathbf{f}^0) = (1 + o_P(1))E[\dot{m}_{\mathbf{f}}^*(Z_t, \boldsymbol{\alpha}^*, \mathbf{f}^0)(\widehat{\mathbf{f}} - \mathbf{f}^0)] \\
&= (1 + o_P(1))\int_{R^{1+d}} \dot{m}_{\mathbf{f}}^*((y, x), \boldsymbol{\alpha}^*, \mathbf{f}^0)(\widehat{\mathbf{f}}(x) - \mathbf{f}^0(x))g_{Y,X}(y, x)dydx \\
&= -(1 + o_P(1))\frac{1}{n}\sum_{t=1}^n \int_{R^{1+d}} \dot{m}_{\mathbf{f}}^*((y, x), \boldsymbol{\alpha}^*, \mathbf{f}^0)[\mathcal{C}_{t,n}(x) - E\mathcal{C}_{t,n}(x)]g_{Y,X}(y, x)dydx
\end{aligned}$$

$$
\begin{aligned}
&-\frac{1}{2}h^2\int_{R^{1+d}}\dot{m}^*_{\mathbf{f}}((y,x),\boldsymbol{\alpha}^*,\mathbf{f}^0)\widetilde{B}(x)g_{Y,X}(y,x)dydx(1+o_P(1))\\
&=(1+o_P(1))\frac{1}{n}\sum_{t=1}^{n}[\mathcal{D}_{t,n}-E\mathcal{D}_{t,n}]+O_P(h^2), \qquad \text{(S1.24)}
\end{aligned}
$$

where $g_{Y,X}(y,x)$ stands for the joint probability density function of (Y_t,X_t) if all components of (Y_t,X_t) are continuous-valued random variables, or the probability function for the discrete-valued ones in (Y_t,X_t) with an integration seen as a summation over the support of a discrete-valued random variable, and

$$
\mathcal{D}_{t,n}=:-\int\dot{m}^*_{\mathbf{f}}((y,x),\boldsymbol{\alpha}^*,\mathbf{f}^0)\mathcal{C}_{t,n}(x)g_{Y,X}(y,x)dydx.
$$

Here, recall $\dot{m}^*_{\mathbf{f},k}$ is the k-th component of $\dot{m}^*_{\mathbf{f}}$, and $\mathcal{D}_{t,n}$ is a $d\times 1$ vector whose k-th element is

$$
\begin{aligned}
&\mathcal{D}_{t,k,n}=-\int_{R^{1+d}}\dot{m}^*_{\mathbf{f},k}((y,x),\boldsymbol{\alpha}^*,\mathbf{f}^0)\mathcal{C}_{t,k,n}(x)g_{Y,X}(y,x)dydx\\
&=-\int_{R^{1+d}}\dot{m}^*_{\mathbf{f},k}((y,x),\boldsymbol{\alpha}^*,\mathbf{f}^0)(1,0)(\dot{\Omega}_{k,\boldsymbol{\beta}_0}(x_k))^{-1}\omega(Y_t;f^0_k(x_k)+h\dot{f}^0_k(x_k)((X_{kt}-x_k)/h))\\
&\qquad[1,(X_{kt}-x_k)/h]^TK_h(X_{kt}-x_k)g_{Y,X}(y,x)dydx\\
&=-\int_{R^{1+d}}[\dot{m}^*_{\mathbf{f},k}((y,x_{-k},X_{kt}-hu),\boldsymbol{\alpha}^*,\mathbf{f}^0)(1,0)(\dot{\Omega}_{k,\boldsymbol{\beta}_0}(X_{kt}-hu))^{-1}\\
&\qquad\omega(Y_t;f^0_k(X_{kt}-hu)+h\dot{f}^0_k(X_{kt}-hu)u)(1,u)^TK(u)\\
&\qquad g_{Y,X_{-k},X_k}(y,x_{-k},X_{kt}-hu)]dydx_{-k}du=(1+o_P(1))\mathcal{D}_{t,k}, \qquad \text{(S1.25)}
\end{aligned}
$$

where $X_{-k,t}$ and x_{-k} stand for the X_t and x with k-th component X_{kt} and x_k removed, respectively, and, in view of $\dot{\Omega}_{k,\boldsymbol{\beta}_0}(X_{kt})$ defined in (S1.17) being a diagonal matrix with its (1,1)-th element equal to $-\psi''(f_k^0(X_{kt}))g_k(X_{kt})$,

$$\begin{aligned}\mathcal{D}_{t,k} = & -\int_{R^d}[\dot{m}^*_{\mathbf{f},k}((y,x_{-k},X_{kt}),\boldsymbol{\alpha}^*,\mathbf{f}^0)(1,0)(\dot{\Omega}_{k,\boldsymbol{\beta}_0}(X_{kt}))^{-1}\omega(Y_t;f_k^0(X_{kt}))(1,0)^T \\ & \qquad g_{Y,X_{-k},X_k}(y,x_{-k},X_{kt})]dydx_{-k} \\ = & (\psi''(f_k^0(X_{kt}))g_k(X_{kt}))^{-1}\omega(Y_t;f_k^0(X_{kt}))D_k(X_{kt}),\end{aligned} \tag{S1.26}$$

where $D_k(x_k) = \int_{R^d}[\dot{m}^*_{\mathbf{f},k}(y,x_{-k},x_k)g_{Y,X_{-k},X_k}(y,x_{-k},x_k)]dydx_{-k}$ with

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

with $\boldsymbol{\gamma}_k$ a $(d+1)\times 1$ vector whose $(k+1)$-th element equal to 1 and zeros otherwise. Here, note that if some component X_{kt} is discrete-valued, we should understand the kernel smoothing for continuous variable replaced by the discrete-valued case as explained in Section 4.1, so we only treat the continuous case for simplicity in this proof. Then, by some tedious calculations, it is easy to follow from (S1.25) under A4 of Assumption 1 that $E\mathcal{D}_{t,k,n} = O(h^2)$ and $E(\mathcal{D}_{t,k}) = 0$. It now follows from (S1.24) that

$$\boldsymbol{V}_n^{22} = M^*(\widehat{\mathbf{f}}) = M^*(\widehat{\mathbf{f}}) - M^*(\mathbf{f}^0) = (1+o_P(1))\frac{1}{n}\sum_{t=1}^n\mathcal{D}_t + O_P(h^2), \tag{S1.27}$$

where $\mathcal{D}_t$ is a $d\times 1$ vector whose k-th component is $\mathcal{D}_{t,k}$, defined in (S1.26).

We next show $|\boldsymbol{V}_n^{21}| \le n^{-1/2}\sup_{\|\mathbf{f}-\mathbf{f}^0\|\le\delta_n}|\nu_n(\mathbf{f})-\nu_n(\mathbf{f}^0)| = o_P(n^{-1/2})$. It suffices to prove the stochastic equicontinuity of the empirical process $\{\nu_n(\mathbf{f}),\mathbf{f}\in\mathbf{F}\}$, where $\mathbf{F}$ is defined in (S1.7), and suffices for our proof as $\delta_n < 1$ for n large enough by $\delta_n \to 0$. This stochastic equicontinuity follows from Doukhan, Massart and Rio (1995) (p.405) by checking the conditions:

(a) $\{Z_t : t \geq 1\}$ is a stationary absolutely regular sequence with mixing coefficient $\beta(s) \leq Cs^{-b}$ for some $b > r/(r-1)$ and some $r > 1$; (b) $E[\widetilde{m}^{2r}(Z_t)] < \infty$ for r as in (a), where $\widetilde{m}(\cdot)$ is the envelope of $\mathcal{M} = \{m^*(\cdot, \alpha^*, \mathbf{f}) : \mathbf{f} \in \mathbf{F}\}$, that is $|m^*(\cdot, \alpha^*, \mathbf{f})| \leq |\widetilde{m}(\cdot)|$ for any $\mathbf{f} \in \mathbf{F}$; (c) For any $\varepsilon > 0$, $\log N_2(\varepsilon, \mathcal{M}) \leq C\varepsilon^{-2\eta}$ for some $\eta > 0$, with $b(1-\eta) > r/(r-1)$ for r as in (a), where $N_2(\varepsilon, \mathcal{M})$ is the L_2-bracketing cover number of $\mathcal{M}$ in (b).

We check those conditions as follows. Here, (a) holds by the condition A1 of Assumption 1. To show (b), notice that for $\mathbf{f} \in \mathbf{F}$ we have $\mathbf{f}(X_t) = (f_1(X_{1t}), \cdots, f_d(X_{dt}))^T$ is bounded over $W(X_t) = \prod_{k=1}^d I_{[c_{k0}, c_{k1}]}(X_{kt}) > 0$, and hence for $m^* \in \mathcal{M}$, $|m^*(Z_t, \alpha^*, \mathbf{f})| \leq (|Y_t| + C)w_0$, where $w_0 = \sup_{x \in A = \prod_{k=1}^d [c_{k0}, c_{k1}]} W(x) \leq 1$, and C is a generic positive constant. So we can take $\widetilde{m}(Z_t) = (|Y_t| + C)w_0$, and hence (b) holds by condition A5 of Assumption 1. Finally for (c), as $\mathbf{F} \subset C_c^\zeta$, it is well known (c.f., Van Der Vaart and Wellner (1996), Theorem 2.7.1) that we can cover C_c^ζ by finite number $N_1 = N(\varepsilon, C_c^\zeta, \|\cdot\|_\infty)$, of balls of functions centered at, say, $\mathbf{f}^j$, $j = 1, \cdots, N_1$, in C_c^ζ, such that $\log N(\varepsilon, C_c^\zeta, \|\cdot\|_\infty) \leq const. \varepsilon^{-1/\zeta}$, and for any $\mathbf{f} \in C_c^\zeta$, there is a $\mathbf{f}^j$ such that $\|\mathbf{f} - \mathbf{f}^j\|_\infty = \max_{1 \leq k \leq d} \sup_{x_k \in [c_{k0}, c_{k1}]} |f_k(x_k) - f_k^j(x_k)| < \varepsilon$. Notice, in view of the continuity of $\psi'(\cdot)$ and $\psi''(\cdot)$ in A1 of Assumption 1 and the bounded support of $W(\cdot)$ in A3 of Assumption 1, that

$$
\begin{aligned}
&(E|m^*(Z_t, \boldsymbol{\alpha}^*, \mathbf{f}) - m^*(Z_t, \boldsymbol{\alpha}^*, \mathbf{f}^j)|^2)^{1/2} \\
&= (E|[Y_t - \psi'(\alpha_0^* + \sum_{k=1}^d \alpha_k^* f_k(X_{kt}))]\widetilde{\chi}_t(\mathbf{f})W(X_t)
\end{aligned}
$$

$$- [Y_t - \psi'(\alpha_0^* + \sum_{k=1}^{d} \alpha_k^* f_k^j(X_{kt}))]\widetilde{\chi}_t(\mathbf{f}^j)W(X_t)|^2)^{1/2}$$

$$\leq C(E(|Y_t| + C)^2)^{1/2}\|\mathbf{f} - \mathbf{f}^j\|_\infty \leq C\varepsilon, \tag{S1.28}$$

by A1(iii) of Assumption 1, with C standing for a generic constant. Therefore we can cover $\mathcal{M}$ by a finite number of balls of functions centered at $m_k = m^*(Z_t, \boldsymbol{\alpha}^*, \mathbf{f}^j)$, $j = 1, \cdots, N_1$, in $\mathcal{M}$, such that $N_2 = N(C\varepsilon, \mathcal{M}, \|\cdot\|_{L_2}) \leq N_1$, and hence $\log N(C\varepsilon, \mathcal{M}, \|\cdot\|_{L_2}) \leq \log N(\varepsilon, C_c^\zeta, \|\cdot\|_\infty) \leq const.\varepsilon^{-1/\zeta} = const.\varepsilon^{-2\eta}$ with $\eta = 1/(2\zeta)$. Thus under $b > r/[(r-1)(1-\eta)] = r/[(r-1)(1-1/(2\zeta))]$ with the beta-mixing coefficient $\beta(s) \leq Cs^{-b}$, the conditions (a), (b) and (c) hold, and hence the stochastic equicontinuity of the empirical process $\{\nu_n(\mathbf{f}), \mathbf{f} \in \mathbf{F}\}$ holds true, and it follows that $\boldsymbol{V}_n^{21} = o_P(n^{-1/2})$.

Now if follows from (S1.20), (S1.21), (S1.22), (S1.27) and $\boldsymbol{V}_n^{21} = o_P(n^{-1/2})$ that

$$\sqrt{n}\boldsymbol{V}_n(\alpha^*, \widehat{\mathbf{f}}) = (1 + o_P(1))\frac{1}{\sqrt{n}}\sum_{t=1}^{n}[m^*(Z_t, \boldsymbol{\alpha}^*, \mathbf{f}^0) + \mathcal{D}_t] + o_P(1) + \sqrt{n}O_P(h^2). \tag{S1.29}$$

Hence it follows from (S1.15) with $nh^4 = o(1)$ that $\widehat{\boldsymbol{\alpha}}^{*(n)} - \boldsymbol{\alpha}^* = O_P(n^{-1/2})$ and

$$\sqrt{n}(\widehat{\boldsymbol{\alpha}}^{*(n)} - \boldsymbol{\alpha}^*) = (1 + o_P(1))\mathbf{U}^{-1}\sqrt{n}\boldsymbol{V}_n(\alpha^*, \widehat{\mathbf{f}}) \xrightarrow{L} N(0, \mathbf{U}^{-1}\mathbf{V}\mathbf{U}^{-1}), \tag{S1.30}$$

where $\mathbf{V} = \sum_{k=-\infty}^{\infty} Cov(\mathcal{V}_t, \mathcal{V}_{t-k})$, with $\mathcal{V}_t = m^*(Z_t, \boldsymbol{\alpha}^*, \mathbf{f}^0) + \mathcal{D}_t$, and $\boldsymbol{U} = E[\psi''(\alpha_0^* + \sum_{k=1}^{d} \alpha_k^* f_k^0(X_{kt}))]\widetilde{\chi}_t(\mathbf{f}^0)\widetilde{\chi}_t(\mathbf{f}^0)^T W(X_t)$. □

S1.4 Sketch of Proof of Theorem 2

Proof. Let $c_n = \frac{1}{\sqrt{n}}$, and recall $\gamma_k = |\widehat{\alpha}_k^*|^{-\iota}$ with $\widehat{\alpha}_k^*$ the root-n consistent estimator of α_k^*, the k-th component of the maximizer of $L(\cdot;\cdot)$ defined in (11).

Let $\mathfrak{A} = \{\alpha^* + c_n\boldsymbol{\delta} : \|\boldsymbol{\delta}\| \le \xi\}$ denote a ball centred around α^*, the minimizer of true function $R_n(\cdot)$, with radius $c_n\boldsymbol{\delta}$. Here $\|.\|$ is the $\mathfrak{L}_2$ Euclidean norm, and ξ is a large constant. Then if the global minimum of $\widehat{R}(\cdot)$, $\widehat{\boldsymbol{\alpha}}$, satisfies $\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\| = O_P(\frac{1}{\sqrt{n}})$, it equivalently belongs, with probability tending to 1, in the ball $\mathfrak{A}$.

It follows from (14) that

$$\begin{aligned}
&\widehat{R}_n(\boldsymbol{\alpha}^* + c_n\boldsymbol{\delta}) - \widehat{R}_n(\boldsymbol{\alpha}^*)\\
&\ge -(L(\boldsymbol{\alpha}^* + c_n\boldsymbol{\delta}; \widehat{\mathbf{f}}) - L(\boldsymbol{\alpha}^*; \widehat{\mathbf{f}})) + \lambda_n \sum_{k=1}^{d} \gamma_k(|\boldsymbol{\alpha}^* + c_n\boldsymbol{\delta}| - |\boldsymbol{\alpha}^*|)\\
&\ge -(L(\boldsymbol{\alpha}^* + c_n\boldsymbol{\delta}; \widehat{\mathbf{f}}) - L(\boldsymbol{\alpha}^*; \widehat{\mathbf{f}})) - \lambda_n c_n \sum_{k=1}^{d} \gamma_k|\boldsymbol{\delta}|. \qquad \text{(S1.31)}
\end{aligned}$$

Then, by a Taylor's expansion

$$\begin{aligned}
&L(\boldsymbol{\alpha}^* + c_n\boldsymbol{\delta}; \widehat{\mathbf{f}}) - L(\boldsymbol{\alpha}^*; \widehat{\mathbf{f}})\\
&= c_n\boldsymbol{\delta}'[\frac{\partial L(\boldsymbol{\alpha}^*; \widehat{\mathbf{f}})}{\partial \boldsymbol{\alpha}}] + (1/2)c_n^2\boldsymbol{\delta}'\frac{\partial^2 L(\boldsymbol{\alpha}^*; \widehat{\mathbf{f}})}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}^T}\boldsymbol{\delta}\{1 + o_P(1)\}\\
&= \sum_{t=1}^{n}[(Y_t - \psi'(\widehat{\mathbf{f}}, \boldsymbol{\alpha}^*))]c_n\widetilde{\chi}_t(\widehat{\mathbf{f}})^T\boldsymbol{\delta}W(X_t)\\
&- (1/2)\sum_{t=1}^{n}[\psi''(\widehat{\mathbf{f}}, \boldsymbol{\alpha}^*)\boldsymbol{\delta}^T\widetilde{\chi}_t(\widehat{\mathbf{f}})\widetilde{\chi}_t(\widehat{\mathbf{f}})^T\boldsymbol{\delta}]c_n^2W(X_t)\{1 + o_P(1)\}. \qquad \text{(S1.32)}
\end{aligned}$$

Thus

$$\begin{aligned}
&\widehat{R}_n(\boldsymbol{\alpha}^* + c_n\boldsymbol{\delta}) - \widehat{R}_n(\boldsymbol{\alpha}^*)\\
&\ge -\sum_{t=1}^{n}[(Y_t - \psi'(\widehat{\mathbf{f}}, \boldsymbol{\alpha}^*))]c_n\widetilde{\chi}_t(\widehat{\mathbf{f}})^T\boldsymbol{\delta}W(X_t) \qquad \text{(S1.33)}\\
&+ (1/2)\sum_{t=1}^{n}[\psi''(\widehat{\mathbf{f}}, \boldsymbol{\alpha}^*)\boldsymbol{\delta}^T\widetilde{\chi}_t(\widehat{\mathbf{f}})\widetilde{\chi}_t(\widehat{\mathbf{f}})^T\boldsymbol{\delta}]c_n^2W(X_t)\{1 + o_P(1)\} \qquad \text{(S1.34)}
\end{aligned}$$

$$- \lambda_n c_n \sum_{k=1}^{d} \gamma_k |\boldsymbol{\delta}| \tag{S1.35}$$

$$=: A_{n1} + A_{n2} + A_{n3}. \tag{S1.36}$$

By using the uniform consistency of local linear fitting under β-mixing in Peng and Lu (2023) and following the argument of proof of Theorem 1 above, we have

$$(1/n) \sum_{t=1}^{n} [(Y_t - \psi'(\widehat{\mathbf{f}}, \boldsymbol{\alpha}^*))] \widetilde{\chi}_t(\widehat{\mathbf{f}})^T W(X_t) = O_P(n^{-1/2}). \tag{S1.37}$$

We then have

$$A_{n1} = O_P(n c_n^2 \xi), \tag{S1.38}$$

For A_{n2},

$$A_{n2} = O_P(n c_n^2 \xi^2), \tag{S1.39}$$

as $1/n \frac{\partial^2 L(\boldsymbol{\alpha}^*; \widehat{\mathbf{f}})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^T} \xrightarrow{\mathrm{p}} E[\phi''(\mathbf{f}^0, \boldsymbol{\alpha}^*)] \widetilde{\chi}_t(\mathbf{f}^0) \widetilde{\chi}_t(\mathbf{f}^0)^T W(X_t)$, which is finite and positive according to Assumption B1.

For A_{n3}, if $\alpha_k^* \neq 0$ for $k \in \mathcal{A}$, then $\gamma_k = \frac{1}{|\widehat{\alpha}_k^*|^\iota} \to \frac{1}{|\alpha_k^*|^\iota}$ and $c_n^{-1}(|\alpha_k^* + c_n \delta_k| - |\alpha_k^*|) \to \delta_k \mathrm{sgn}(\alpha_k^*)$. Thus

$$\lambda_n c_n \sum_{k \in \mathcal{A}} \gamma_k c_n^{-1} (|\alpha_k^* + c_n \delta_k| - |\alpha_k^*|) = \lambda_n c_n \sum_{k \in \mathcal{A}} \delta_k \mathrm{sgn}(\alpha_k^*) O_P(1) = O_P(\lambda_n c_n \xi) \xrightarrow{\mathrm{P}} 0$$

for any $\|\delta\| \leq \xi$ because $\lambda_n c_n = O(\lambda_n / \sqrt{n}) \to 0$ by the assumption of this theorem.

If $\alpha_k^* = 0$ for $k \in \mathcal{A}^c$, then $c_n^{-1}(|\alpha_k^* + c_n \delta_k| - |\alpha_k^*|) = |\delta_k|$ and $|c_n^{-1} \widehat{\alpha}_k^*| = O_P(1)$ by the root-n consistency of the pre-estimator $\widehat{\alpha}_k^*$. Hence

$$\lambda_n c_n \gamma_k c_n^{-1} (|\alpha_k^* + c_n \delta_k| - |\alpha_k^*|) = \lambda_n c_n \gamma_k |\delta_k| = \lambda_n c_n c_n^{-\iota} (|c_n^{-1} \widehat{\alpha}_k^*|)^{-\iota} |\delta_k| = \lambda_n c_n c_n^{-\iota} |\delta_k| O_P(1),$$

which tends to $+\infty$ in probability if $\delta_k \neq 0$ for $k \in \mathcal{A}^c$, and is equal to zero otherwise, because $\lambda_n c_n c_n^{-\iota} = O(\lambda_n n^{(\iota-1)/2}) \to \infty$ by the assumption of this theorem. Hence by choosing a sufficiently large ξ, we have:

$$A_{n3} = \lambda_n c_n \left(\sum_{k \in \mathcal{A}} \gamma_k c_n^{-1}(|\alpha_k^* + c_n\delta_k| - |\alpha_k^*|) + \sum_{k \in \mathcal{A}^c} \gamma_k c_n^{-1}(|\alpha_k^* + c_n\delta_k| - |\alpha_k^*|) \right), \quad \text{(S1.40)}$$

which tends to $+\infty$ in probability if $\delta_k \neq 0$ for $k \in \mathcal{A}^c$, and zero otherwise.

Note that $\widehat{\boldsymbol{\alpha}}$ minimizes $\widehat{R}(\boldsymbol{\alpha}) = R_n(\boldsymbol{\alpha})(1 + o_P(1))$. If $\widehat{\boldsymbol{\alpha}}$ is not within $\{\boldsymbol{\alpha}^* + c_n\boldsymbol{\delta} : \|\boldsymbol{\delta}\| \leq \xi\}$, that is $\widehat{\boldsymbol{\alpha}}$ is in $\{\boldsymbol{\alpha}^* + c_n\boldsymbol{\delta} : \|\boldsymbol{\delta}\| \geq \xi\}$, then, owing to convexity of $R_n(\boldsymbol{\alpha}), \widehat{\boldsymbol{\alpha}}$ must be on $\{\boldsymbol{\alpha}^* + c_n\boldsymbol{\delta} : \|\boldsymbol{\delta}\| = \xi\}$, with probability tending to one.

When $\|\boldsymbol{\delta}\| = \xi$ holds, by noticing that $A_2 \geq 0$ is the largest term, we have $\inf_{\|\boldsymbol{\delta}\|=\xi} R_n(\boldsymbol{\alpha}^* + c_n\boldsymbol{\delta}) \geq R_n(\boldsymbol{\alpha}^*)$. Thus, $P(\sqrt{n}|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*| > \xi) = P(\widehat{\boldsymbol{\alpha}} \notin \mathfrak{A}) \leq P(\inf_{\|\boldsymbol{\delta}\|=\xi} R_n(\boldsymbol{\alpha}^* + c_n\boldsymbol{\delta}) \leq R_n(\boldsymbol{\alpha}^*)) \to 0$, and this completes the proof.

□

S1.5 Sketch of Proof of Theorem 3

Proof. Recall that $\mathcal{A} = \{k : \alpha_k^* \neq 0\}$. Now we define $\widehat{\mathcal{A}} = \{k : \widehat{\alpha}_k \neq 0\}$. If we have $\forall j \in \widehat{\mathcal{A}}, P(j \in \mathcal{A}) \to 1$, then it suffices to show that $\forall j' \in \widehat{\mathcal{A}}^c, P(j' \in \mathcal{A}^c) \to 1$.

Denote by w^2 the part of $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_d)^T$ corresponding to γ_j's with $j' \in \mathcal{A}^c$, where $\mathcal{A}^c$ stands for the complement of $\mathcal{A}$, and componentwise operations apply where easily seen. Consider $j \in \mathcal{A}^c$. By taking the derivative of $\widehat{R}(\boldsymbol{\alpha})$ with respect to $\boldsymbol{\alpha}^2$, with $\boldsymbol{\alpha}^2$ denoting all the corresponding α_j's, it holds that

$$\begin{aligned}\frac{\partial \widehat{R}(\widehat{\boldsymbol{\alpha}})}{\partial \boldsymbol{\alpha}^2} &= -\frac{\partial L(\widehat{\boldsymbol{\alpha}}; \widehat{\mathbf{f}})}{\partial \boldsymbol{\alpha}^2} + \lambda_n w^2 \mathrm{sgn}(\boldsymbol{\alpha}^{*2}) \\ &= -\frac{\partial L(\boldsymbol{\alpha}^*; \widehat{\mathbf{f}})}{\partial \boldsymbol{\alpha}^2} - \frac{\partial^2 L(\boldsymbol{\alpha}^*; \widehat{\mathbf{f}})}{\partial \boldsymbol{\alpha}^2 \partial \boldsymbol{\alpha}^{2T}} (\widehat{\boldsymbol{\alpha}}^2 - \boldsymbol{\alpha}^{*2})\{1 + o_P(1)\} + \lambda_n w^2 \mathrm{sgn}(\boldsymbol{\alpha}^{*2}) \\ &=: B_1 + B_2 + B_3, \end{aligned} \tag{S1.41}$$

where the second equation applies the Taylor expansion.

From (A.21) in the proof of Theorem 1 above, it suffices to show that $B_1/\sqrt{n} \xrightarrow{\mathrm{d}} N(0, \boldsymbol{V}_2)$, where $\boldsymbol{V}_2$ corresponds to j'th row of $\boldsymbol{V}$.

Similarly, $B_2/\sqrt{n} = 1/n \frac{\partial^2 L(\boldsymbol{\alpha}^*; \widehat{\mathbf{f}})}{\partial \boldsymbol{\alpha}^2 \partial \boldsymbol{\alpha}^{2T}} \sqrt{n}(\widehat{\boldsymbol{\alpha}}^2 - \boldsymbol{\alpha}^{*2})$, where the first term converges in probability to $E[\phi''(\mathbf{f}^0_{\mathcal{A}^c}, \boldsymbol{\alpha}^{*2})] \widetilde{\chi}_{t,\mathcal{A}^c}(\mathbf{f}^0) \widetilde{\chi}_{t,\mathcal{A}^c}(\mathbf{f}^0)^T W(X_t)$, where $\mathbf{f}^0_{\mathcal{A}^c}$ and $\widetilde{\chi}_{t,\mathcal{A}^c}(\mathbf{f}^0)$ stand for the component of $\mathbf{f}^0$) and $\widetilde{\chi}_t(\mathbf{f}^0)$ corresponding to index $j' \in \mathcal{A}^c$, that is finite, and the second term converges to zero with probability tending to 1 as shown in Theorem 2.

$B_3/\sqrt{n}$ is of order $O_P(\lambda_n w^2 \mathrm{sgn}(\boldsymbol{\alpha}^{*2})/\sqrt{n}) = O_P(\lambda_n n^{(\iota-1)/2} \mathrm{sgn}(\boldsymbol{\alpha}^{*2}))$, under adaptive weights of w^2, which tends to zero or ∞ component-wisely, depending on $\boldsymbol{\alpha}^{*2} = 0$ or not component-wisely, by the assumption of this theorem. Hence $\frac{1}{\sqrt{n}} \frac{\partial \widehat{R}(\widehat{\boldsymbol{\alpha}})}{\partial \boldsymbol{\alpha}^2} = \frac{1}{\sqrt{n}}(B_1 + B_2 + B_3)$, for which $\frac{1}{\sqrt{n}} \frac{\partial \widehat{R}(\widehat{\boldsymbol{\alpha}})}{\partial \boldsymbol{\alpha}^2}$ tends to zero in probability (as $\widehat{\boldsymbol{\alpha}}$ is the minimizer of $\widehat{R}(\boldsymbol{\alpha})$), is determined by the sign of $\boldsymbol{\alpha}^{*2}$. So we have $P(\widehat{\boldsymbol{\alpha}}^2 = \boldsymbol{\alpha}^{*2} = 0) \to 1$ as $n \to \infty$.

□

S1.6 Sketch of Proof of Theorem 4

Proof. Recall that $\mathcal{A} = \{k : \alpha_k^* \neq 0\}$. From Theorem 3, there exists a global minimizer of the objective function $\widehat{R}(\boldsymbol{\alpha})$, which is as same as the minimizer of $\widehat{R}(\boldsymbol{\alpha}^1)$ with $\boldsymbol{\alpha}^1$ denoting all the α_j's, $\forall j \in \mathcal{A}$.

By Taylor's expansion,

$$0 = \frac{\partial \widehat{R}(\widehat{\boldsymbol{\alpha}})}{\partial \boldsymbol{\alpha}^1} = C_{1n} + C_{2n} + C_{3n}, \tag{S1.42}$$

with

$$C_{1n} := -\sum_{t=1}^{n} [(Y_t - \psi'(\widehat{\mathbf{f}}_{\mathcal{A}}, \boldsymbol{\alpha}^{*1})] \widetilde{\chi}_{t,A}(\widehat{\mathbf{f}}) W(X_t), \tag{S1.43}$$

where $\widehat{\mathbf{f}}_{\mathcal{A}}$ and $\widetilde{\chi}_{t,\mathcal{A}}(\widehat{\mathbf{f}})$ stand for the components of $\widehat{\mathbf{f}}$ and $\widetilde{\chi}_t(\widehat{\mathbf{f}})$ corresponding to index $j \in \mathcal{A}$,

$$C_{2n} := \sum_{t=1}^{n} [\psi''(\widehat{\mathbf{f}}_{\mathcal{A}}, \boldsymbol{\alpha}^{*1}) \widetilde{\chi}_{t,A}(\widehat{\mathbf{f}}) \widetilde{\chi}_{t,A}(\widehat{\mathbf{f}})^T] W(X_t) (\boldsymbol{\alpha}^{*1} - \widehat{\boldsymbol{\alpha}}^1) \{1 + o_P(1)\}$$

$$, \tag{S1.44}$$

$$C_{3n} =: \lambda_n w^1 \mathrm{sgn}(\widehat{\boldsymbol{\alpha}}^1), \tag{S1.45}$$

where w^1 denotes the part of $\boldsymbol{\gamma}$ corresponding to γ_k's with $k \in \mathcal{A}$, and componentwise operations apply where easily seen. $\mathrm{sgn}(\widehat{\boldsymbol{\alpha}}^1)$ is the point-wise sign function of vector $\widehat{\boldsymbol{\alpha}}^1$.

From the argument of proof of Theorem 1,

$$\frac{1}{\sqrt{n}}C_{1n} = (1+op(1))\frac{1}{\sqrt{n}}\sum_{t=1}^{n}[m^*(Z_t,\boldsymbol{\alpha}^{*1},\mathbf{f}_A^0)+\mathcal{D}_{t,A}]+o_P(1)+\sqrt{n}O_P(h^2) =: \boldsymbol{V}_{n1}, \tag{S1.46}$$

where $\boldsymbol{\alpha}^{*1}$ is component of $\boldsymbol{\alpha}^*$, $\mathbf{f}_A^0$ denotes component of $\mathbf{f}^0$ and $\mathcal{D}_{t,A}$ stands for component of $\mathcal{D}_t$, corresponding to $j \in \mathcal{A}$.

Similarly, it suffices to show that

$$1/n\sum_{t=1}^{n}[\psi''(\widehat{\mathbf{f}}_{\mathcal{A}},\boldsymbol{\alpha}^{*1})\widetilde{\chi}_{t,A}(\widehat{\mathbf{f}})\widetilde{\chi}_{t,A}(\widehat{\mathbf{f}})^T]W(X_t) \xrightarrow{\text{P}} E[\phi''(\mathbf{f}_A^0,\boldsymbol{\alpha}^{*1})]\widetilde{\chi}_{t,A}(\mathbf{f}^0)\widetilde{\chi}_{t,A}(\mathbf{f}^0)^TW(X_t) =: \boldsymbol{U}_1. \tag{S1.47}$$

and thus

$$\frac{1}{\sqrt{n}}C_{2n} = (1+op(1))\boldsymbol{U}_1\sqrt{n}(\boldsymbol{\alpha}^{*1}-\widehat{\boldsymbol{\alpha}}^1). \tag{S1.48}$$

For C_{3n}, when $n \to \infty$, $\lambda_n\boldsymbol{w}^1\text{sgn}(\widehat{\boldsymbol{\alpha}}^1) = \lambda_n\boldsymbol{w}^1\text{sgn}(\boldsymbol{\alpha}^{*1})\{1+o_P(1)\}$. Also we know that $\frac{1}{\sqrt{n}}C_{3n} = O_P(\lambda_n/\sqrt{n}) \to 0$ by the assumption of this theorem. Therefore,

$$\sqrt{n}(\boldsymbol{\alpha}^{*1}-\widehat{\boldsymbol{\alpha}}^1) = (1+o_P(1))\boldsymbol{U}_1^{-1}\sqrt{n}\boldsymbol{V}_{n1} \xrightarrow{\text{L}} N(0,\boldsymbol{U}_1^{-1}\boldsymbol{V}_1\boldsymbol{U}_1^{-1}), \tag{S1.49}$$

where $\boldsymbol{V}_1$ and $\boldsymbol{U}_1$ are the submatrices, corresponding to those components in A, of the $\boldsymbol{V}$ and $\boldsymbol{U}$ defined in Theorem 1, and the asymptotic normality is proven.

□

S2 Algorithm for PGMAFMA

In this section, we present the algorithm for estimating our PGMAFMA procedure. It is, however, noted that there are many alternative methods available, and the computation is not expensive.

S3 More details on simulation

In this Section, we consider two DGPs as presented in the main text: DGP 1 - the binary classification forecasting problem (Binomial distribution) with two-dimensional marginals, and DGP 2 - the count data prediction problem (Poisson distribution), with discrete-valued lagged and other exogenous information accounted for.

We will first examine binomial data for time series classification based on two-dimensional functions, with a true model of a non-GAM form detailed in Section S3.1. In such cases of binary classifications, we compare the predictive power of candidate models with the aid of a widely used measure of the area under the curve (AUC) of receiver operating characteristic (ROC), which gives the plot of true positive rate against false positive rate. In general, the higher the AUC value, the better the prediction is.

A Poison example is then examined in Subsection S3.2, where the expected mean of a Poison distributed random variable Y_i given the past information I_{i-1} up to time $(i-1)$, i.e, $\lambda_i = E(Y_i|I_{i-1})$, is generated again based on exogenous covariates and past value of observations. Here we will only look at the case that a true model is of a GAM

(The algorithm for the GMAFMA model with adaptive LASSO)

1. Solve the GMAFMA model to get the initial estimator $\widehat{\boldsymbol{\alpha}}^{*(n)}$;

2. Compute the weight of adaptive LASSO: $\widehat{\gamma}_k = \frac{1}{|\widehat{\boldsymbol{\alpha}}_k^{*(n)}|^{\iota}}$, where $\iota > 0$ can be chosen as 1 for simplicity;

3. Define $\widetilde{f}_k(X_{kt}) = \widehat{f}_k(X_{kt})/\widehat{\gamma}_k$, for $k = 1, ..., d$;

4. Solve the LASSO model for all λ_n's that are considered by tackling the following minimization problem:

$$\widehat{\boldsymbol{\alpha}}^{**}(\lambda_n) = \arg\min_{\boldsymbol{\alpha}} \sum_{t=1}^{n} \{[-Y_t(\alpha_0 + \sum_{k=1}^{d} \alpha_k \widetilde{f}_k(X_{kt}))] \\ + \psi(\alpha_0 + \sum_{k=1}^{d} \alpha_k \widetilde{f}_k(X_{kt})) - \phi(Y_t, \Theta_t)\} W(X_t) + \lambda_n \sum_{k=1}^{d} |\alpha_k|, \quad \text{(S2.50)}$$

   where we use the **lbfgs** package in R, which would handle the adaptive LASSO problem by treating it as an optimization problem of the log-likelihood function plus the $\mathfrak{L}_1$ norm penalization;

5. Compute the adaptive LASSO estimate: $\widehat{\boldsymbol{\alpha}}^{*}(\lambda_n) = \widehat{\boldsymbol{\alpha}}^{**}(\lambda_n)/\widehat{\gamma}_k$, for $k = 1, ..., d$;

6. Define the best estimation $\widehat{\boldsymbol{\alpha}}^{*}(\lambda_n^*)$ and choose the best penalisation coefficient λ_n^* by, e.g., finding the minimum BIC value:

$$BIC(\lambda_n) = \sum_{t=1}^{n} \{[y_t(\widehat{\alpha}_0^*(\lambda_n) + \sum_{k=1}^{d} \widehat{\alpha}_k^*(\lambda_n) \widehat{f}_k(X_{kt}))] \\ - \psi(\widehat{\alpha}_0^*(\lambda_n) + \sum_{k=1}^{d} \widehat{\alpha}_k^*(\lambda_n) \widehat{f}_k(X_{kt}))\} W(X_t) + k \log(n), \quad \text{(S2.51)}$$

   where $k \neq 0$ is the number of non-zero parameters estimated by the model with λ_n;

form for simplicity. To compare the performance of predictions denoted by $\widehat{\lambda}_i$, we will apply the measure of Mean Absolute Error (MAE) based on a test data set as follows:

$$MAE = \frac{\sum_{i=1}^{n_\tau} |Y_i - \widehat{\lambda}_i|}{n_\tau}, \tag{S3.52}$$

where n_τ is the test data sample size. In general, the lower the MAE value, the better the prediction.

All the simulations consist of the data generated with the sample size equal to $n = 500$, $n = 1000$, and $n = 2000$, respectively, and a testing sample of size of $n_\tau = 50$ for prediction evaluation. Note that when generating the time series data, in view of a necessary warming-up step, we deleted the first 100 observations from the $(100+n+n_\tau)$ observations generated through the iterations for Y_t based on a simulating model with initial values taken to be zero. In addition, all the simulations are repeated 100 times for each setting. A simple cross-validation using h.select in R package *sm* is applied for bandwidth h in nonparametric local linear estimation, which is essentially based on a direct estimation of $\mu_{kt} = E(Y_t|X_{kt})$ and appears to work well.

S3.1 DGP 1: Marginally interacted logistic regression

Consider a Bernoulli Y_t series jointly impacted by exogenous covariates x_{t1}, x_{t2}, x_{t3}, x_{t4}, x_{t5} and x_{t6} with interactions, different from those considered in Peng and Lu (2024), viz:

$$Y_t = I(y_t > 0), \quad y_t = \sum_{k=1}^{3} g_{0k}(y_{t-k}) + \cos(2x_{t1}x_{t2}) + \log(1 + (x_{t3}x_{t4})^2) + x_{t5}x_{t6} + \epsilon_t,$$

$$with\ g_{0k}(y_{t-k}) = -\sin(2y_{t-k}),\ k = 1, 2, 3, \tag{S3.53}$$

where the ϵ_t's are *i.i.d.* following a logistic distribution. We consider the exogenous covariates, $x_{t1}, \ldots, x_{tp}$, with $p = 15$ below, are independently generated from a normal distribution $N(0, 1)$. Note that in (S3.53), only $p = 6$ of the exogenous covariates are truly relevant and the true lag order of the autoregressive terms is $L = 3$.

The data generated by (S3.53) is present with periodic, logarithmic, and linear structural interactions. We are examining the case of $(p, L) = (15, 15)$ in the working model (S3.54) given below. For simplicity, we are considering adding the two-dimensional marginal estimates of $f_k(x_{tj}, x_{t,j+1})(= logit(P(Y_t = 1|x_{tj}, x_{t,j+1}))$ for all the exogenous covariates x's, with $x_{t,p+1} = x_{t1}$ and the one-dimensional marginal estimates of $f_{p+k}(y_{t-k})(= logit(P(Y_t = 1|y_{t-k}))$ for all the lagged terms. The GMAFMA model is thus formulated as:

$$\begin{aligned} logit(P(Y_t = 1|I_{t-1})) &\approx \alpha_0 + \alpha_1 f_1(x_{t1}, x_{t2}) + ... + \alpha_p f_p(x_{tp}, x_{t,p+1}) \\ &+ \alpha_{p+1} f_{p+1}(y_{t-1}) + ... + \alpha_{p+L+1} f_{p+L+1}(y_{t-L}) \equiv f_t^{MA}, \end{aligned} \tag{S3.54}$$

where the unknown coefficients α_j and α_{p+k}, for $j = 0, 1, ..., p(= 15)$ and $k = 1, ..., L(= 15)$, with the GMAFMA and PGMAFMA can be estimated by combining the ideas in Section 5.2 in the main text.

The performance of the PGMAFMA method is compared with the GMAFMA, GLM, GLMNET (GLM with LASSO), AR, and Random Forest (RF) methods. The boxplots of the area under the curve (AUC) with 100 replications of one-step ahead classification predictions for these methods, with $n_\tau = 50$ observations for testing, are depicted in Figure 1, where the used sample sizes are $n = 500$, $n = 1000$ and $n = 2000$, respectively, with the panels from the left to the right.

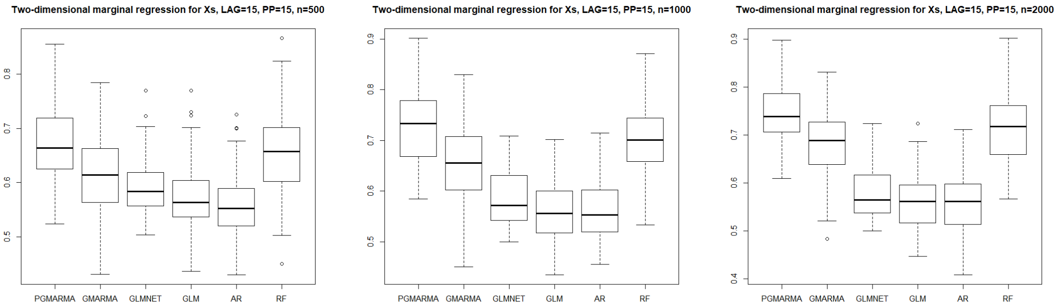

Figure 1: Boxplots of the area under curve (AUC) with 100 repetitions for one-step ahead classification predictions, with $n_\tau = 50$ for testing, for different methods (Penalized GMAFMA, GMAFMA, GLMNET, GLM, AR, Random Forest) with two-dimensional interactions based on $n = 500$ (left), $n = 1000$ (middle), $n = 2000$ (right) for training

Some additional insights into the case of two-dimensional interactions are summarized.

(i) The DGP is highly nonlinear in this subsection. However, we can see from the panels of Figure 1 that our proposed PGMAFMA outperforms all other models in all cases. We also find that the linear models like GLM, AR, and even GLMNET, perform much worse, especially with poor performance of the GLM and AR methods in this example. This is heuristically understandable as the nonlinear interactions with the covariates $x_{t1}, \ldots, x_{t6}$ are involved in the true model.

(ii) Overall, the performance of the Random Forest (RF) model is competitive, but it is worse than our PGMAFMA model in terms of the AUC for classification prediction. From the figure panels, the RF model has well demonstrated its ability to capture the interactions of covariates in tree structures. However, compared to our PGMAFMA, the RF model is more complex with a much larger model dimension (c.f. Jo et al. (2023)). It is worth noting that the RF is a black-box model that is hard to understand, while our PGMAFMA method can provide explainable results and a cheaper computational cost.

S3.2 DGP 2: Marginal Poisson regression

Now we consider a DGP for Y_t following a conditional $Poisson(\lambda_t)$ distribution, viz:

$$\log \lambda_t = \frac{1}{4}\sum_{k=1}^{3} g_{0k}(Y_{t-k}) + 3\cos(x_{t1}) + 2e^{2x_{t2}} + 6x_{t3}^2,\ \ with\ g_{0k}(Y_{t-k}) = -\sin(Y_{t-k}), \tag{S3.55}$$

where $\lambda_t = E[Y_t|I_{t-1}]$, and we consider the true Poisson (auto)regression functions involving $p = 3$ exogenous covariates, x_{t1}, x_{t2}, x_{t3}, and $L = 3$ lag variables of Y_t, which we note are discrete-valued. The exogenous covariates, $x_{t1}, \ldots, x_{tp}$, are independently generated from a uniform distribution $U(0, 1)$; we use the *rpois* function based in R to generate the observations Y_t according to the expected mean λ_t.

Considering $(p, L) = (15, 15)$ for working models, as indicated in (S3.56), which involve 30 marginal forecasts even for one-dimensional marginals only. We are comparing the performances of the Penalized GMAFMA (PGMAFMA) with those of GMAFMA, GLM, GLMNET (GLM with LASSO), AR, and Gradient Boosting Machines (GBM, using gbm function in R package *gbm*) , respectively. We consider the one-dimensional marginal Poisson regression estimation of $f_k(\cdot)(=\log(E(Y_t|\cdot)))$ for each of the p exogenous covariates in $\mathbb{X}_t$ and the autoregressive terms Y_{t-k}'s up to lag order L. Then the GMAFMA model is specified as follows:

$$\begin{aligned}\log(\lambda_t) &\approx \alpha_0 + \alpha_1 f_1(x_{t1}) + ... + \alpha_p f_p(x_{tp}) \\ &\quad + \alpha_{p+1} f_{p+1}(Y_{t-1}) + ... + \alpha_{p+L} f_{p+L}(Y_{t-L}) \equiv f_t^{MA}. \end{aligned} \tag{S3.56}$$

Then $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, ..., \alpha_{p+l})$ is the vector of unknown coefficients to be estimated, as detailed in Section 2 for GMAFMA and in Section 3 for PGMAFMA. The boxplots of the mean absolute errors (MAE) for 100 repetitions of one-step-ahead predictions, with $n_\tau = 50$ observations for testing, by different methods are depicted in Figure 2, with the average of MAE values for 100 repetitions reported in Table 1. Note that for the MAE value used as a criterion, the smaller value is preferred.

We summarize our findings on the results for count data prediction from Figure 2 and Table 1.

(i) Compared with other popular methods in machine learning, our PGMAFMA and GMAFMA models are again the most competitive candidates for all sample sizes, even for $n = 500$. The AR model for prediction of Y_t performs the worst, as the true model of the data is nonlinear, far away from the AR model, with the count data seen

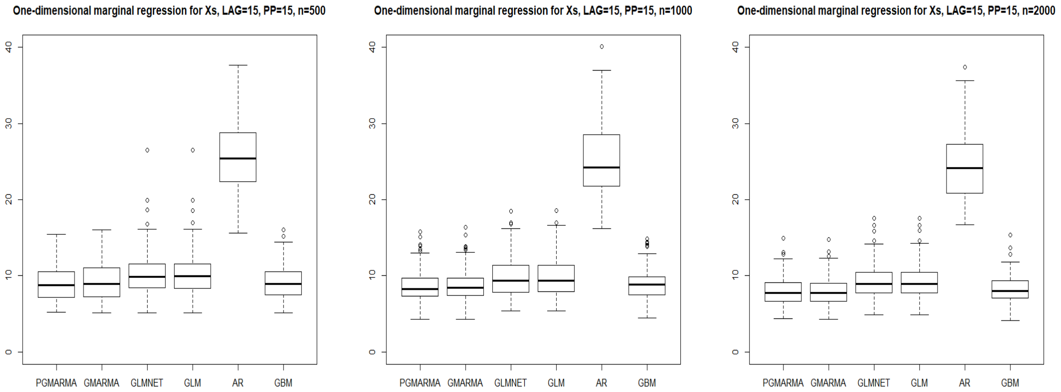

Figure 2: Boxplots of the Mean Absolute Error values with 100 repetitions for one-step-ahead predictions, with $n_\tau = 50$ observations for testing, for different methods (PGMAFMA, GMAFMA, GLMNET, GLM, AR, GBM) based on $n = 500$,$n = 1000$,$n = 2000$ observations for training.

Table 1: Average of mean absolute error values with 100 repetitions for one-step-ahead predictions

Model	PGMAFMA	GMAFMA	GLMNET	GLM	AR	GBM
N=500	9.037898	9.368950	10.337029	10.361082	25.485584	9.189626
N=1000	8.721731	8.841820	9.845152	9.843991	25.327154	9.027817
N=2000	7.924985	7.943700	9.270446	9.275893	24.291927	8.292361

as continuous-valued.

(ii) It is noticed that the GBM model is in general working well, better than the GLMNET, GLM, and AR models. But our PGMAFMA beats the GBM clearly both from Figure 2 and Table 1.

(iii) With the increase in sample size, the improvement of the predictive power of our PGMAFMA model is significant compared to the performance of the GLMNET (GLM with LASSO) model. It seems that the sample size $n = 500$ can work sufficiently well. This gives the credit to our GMAFMA model, as we have well estimated the marginal information with nonlinear structures.

S4 More details on prediction of FTSE100 index market moving direction

As an extension to Peng and Lu (2024), we again consider the forecasting of FTSE100 index market moving direction to demonstrate the strength of our penalized GMAFMA procedure. The data include the open price op_t, close price cp_t, daily maximum price $maxp_t$ and minimum price $minp_t$ and the trading volume Vlm_t, each having 1263 observations from 1 May 2013 to 1 May 2018. We are concerned with whether the market price going up ($Y_t = 1$) or not ($Y_t = 0$) is related to the past volatility, volume, and geometric return, with the relevant quantities defined in the same manner as in Li, Linton and Lu (2015) and Peng and Lu (2024), which are depicted in Figure 3. We use the first 1200 observations for training and the remaining 62 observations are used for testing.

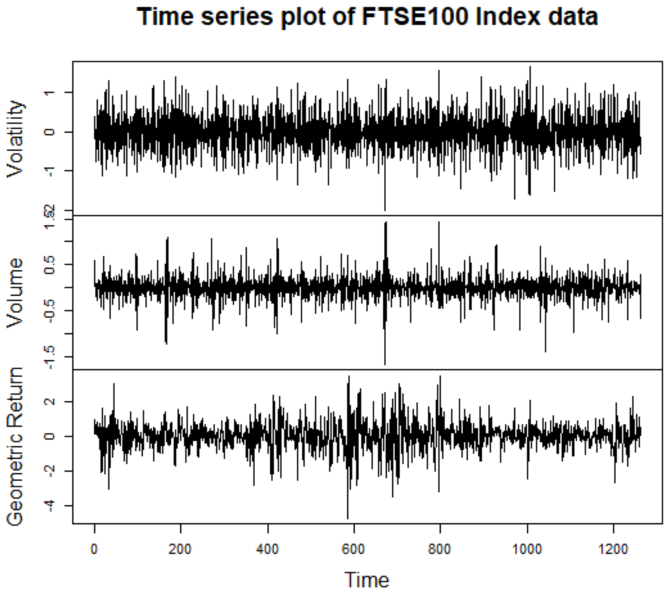

Figure 3: The time series plot of volatility v_t, logarithm of volume V_t and geometric return G_t

Differently from Peng and Lu (2024) who only considered the one-dimensional marginal regression estimates for the past information of the predictors with manual variable selection in their work, we are examining if the estimates of more additional two-dimensional marginal interactions of every two predictor variables, and hence with more marginal historical information taken into account, can help to improve the prediction of price moving direction. However, this makes the number of marginal (one- and two-dimensional) estimates, i.e., model dimension in GMAFMA, much larger, and hence more challenging, than that in Peng and Lu (2024), thus motivating us to apply our penalized method, PGMAFMA, to the proposed setting.

In particular, we consider the one-step-ahead prediction of the market price moving direction Y_t based on the past information of a weekly lag order equal to 7 (i.e., from lag 1 to lag 7) of the market direction, volatility, volume, and geometric return, with a total number of predictors equal to 28, to check if they help improve the explanation or prediction of market price moving. That is, we are using $X_t = (Y_{t-1}, \ldots, Y_{t-7}, v_{t-1}, \ldots, v_{t-7}, V_{t-1}, \ldots, V_{t-7},$
$G_{t-1}, \ldots, G_{t-7})$ and considering the estimates of their one-dimensional marginal regressions and two-dimensional marginal interactions in GMAFMA and PMAFMA to predict Y_t. For simplicity, three scenarios below are compared, viz:

- Scenario 1: no past market moving direction information (i.e., no $Y_{t-1}, \ldots, Y_{t-7}$; c.f., (S4.57));
- Scenario 2: a linear form of the past market moving direction (i.e., $Y_{t-1}, \ldots, Y_{t-7}$; c.f., (S4.58));
- Scenario 3: additional interactions of past market moving direction with other covariates (i.e., $Y_{t-1}, \ldots, Y_{t-7}, f(Y_{t-1}, Y_{t-2}), \ldots, f(G_{t-6}, G_{t-7})$; c.f., (S4.59)).

We remark here that as the market moving direction Y_t is binary and the other variables are continuous, the marginal estimation of $f(.)$ for mixed type of data can be done via **np** package available in R (Li and Racine, 2003; Racine and Li, 2004). The optimal selection of lags in such a high-dimensional case is interesting but challenging.

Following the simulation, we have considered the GMAFMA, PGMAFMA, GLM, GLMNET (GLM with LASSO), AR, and Random Forest models to predict the market moving direction Y_t. We can then formulate the GMAFMA models similarly to that given in (S3.54) under the three scenarios as follows:

GMAFMA model under Scenario 1 (without past market direction):

$$logit(P(Y_t = 1|I_{t-1})) = \alpha_0 + \alpha_1 f_1(v_{t-1}) + \cdots + \alpha_7 f_7(v_{t-7}) + \alpha_8 f_8(V_{t-1}) + \cdots + \alpha_{14} f_{14}(V_{t-7})$$
$$+ \alpha_{15} f_{15}(G_{t-1}) + \cdots + \alpha_{21} f_{21}(G_{t-7}) + \alpha_{22} f_{22}(v_{t-1}, v_{t-2}) + \cdots + \alpha_{231} f_{231}(G_{t-6}, G_{t-7}), \tag{S4.57}$$

with a total dimension of 231 past marginals.

GMAFMA model under Scenario 2 (linear past market direction):

$$logit(P(Y_t = 1|I_{t-1})) = \alpha_0 + \alpha_1 Y_{t-1} + \cdots + \alpha_7 Y_{t-7} + \alpha_1 f_1(v_{t-1}) + \cdots + \alpha_{30} f_{30}(v_{t-30})$$
$$+ \alpha_8 f_8(V_{t-1}) + \cdots + \alpha_{14} f_{14}(V_{t-7}) + \alpha_{15} f_{15}(G_{t-1}) + \cdots + \alpha_{21} f_{21}(G_{t-7})$$
$$+ \alpha_{22} f_{22}(v_{t-1}, v_{t-2}) + \cdots + \alpha_{238} f_{238}(G_{t-6}, G_{t-7}), \tag{S4.58}$$

with a total dimension of 238 past marginals.

GMAFMA model under Scenario 3 (also interactions of past market direction):

$$logit(P(Y_t = 1|I_{t-1})) = \alpha_0 + \alpha_1 f_1(Y_{t-1}) + \cdots + \alpha_7 f_7(Y_{t-7}) + \alpha_8 f_8(v_{t-1}) + \cdots + \alpha_{14} f_{14}(v_{t-7})$$
$$+ \alpha_{15} f_{15}(V_{t-1}) + \cdots + \alpha_{21} f_{21}(V_{t-7}) + \alpha_{22} f_{22}(G_{t-1}) + \cdots + \alpha_{28} f_{28}(G_{t-7})$$
$$+ \alpha_{29} f_{29}(Y_{t-1}, Y_{t-2}) + \cdots + \alpha_{406} f_{406}(G_{t-6}, G_{t-7}), \tag{S4.59}$$

with a total dimension of 406 past marginals.

All the f_k's in (S4.57)-(S4.59) are low-dimensional (one or two) nonlinear marginal regressions that are pre-estimated, which can be viewed as weak learners in the sense of machine learning. Then α_k's are estimated by the procedures for GMAFMA and PGMAFMA, respectively, in Sections 2 and 3, and the ideas in Sections 4.1 and 4.2 when discrete covariates and interactions are involved.

We can specify the GLM models under Scenarios 1-3 correspondingly to (S4.57)-(S4.59) with the f's being identity functions. All the α_k's are estimated, respectively, by the **glm** and **glmnet** in R.

The fitted models are summarized in Tables 2 and 3, due to the limited space by noting that there are in total 406 (max) regressors (28 predictors and 378 two-way interactions of them) in the full model (S4.59). In such a high-dimensional case involving so much lagged information, it is interesting to see the number of variables selected by the penalized models to avoid over-fitting. We have therefore reported the number of non-zero predictors kept in the corresponding models that are used for the prediction in Table 2. For simplicity, we tentatively used a global bandwidth $h \in (0.2, 0.9)$ for the estimation of all the low dimensional nonlinear marginal regressions f_k's in this numerical example. The optimal bandwidth h and tuning parameter λ are reported in Table 2.

We remark that the selected bandwidth h listed in Table 2 works well for each scenario with GMAFMA and PGMAFMA in this numerical example, and the PG-

Table 2: The AUC values for one step ahead prediction by GMAFMA and PGMAFMA with the best-found bandwidth h, penalty tuning parameter λ and number of non-zero parameters selected by PGMAFMA

Scenario	Bandwidth h	AUC(GMAFMA)	AUC(PGMAFMA)	Penalty tuning λ	Number of non-zero parameters
Scenario 1	0.6	0.6050	0.6271	0.002649	98 (in total 231)
Scenario 2	0.6	0.6113	0.6187	0.000738	165 (in total 238)
Scenario 3	0.6	0.6176	0.6954	0.001179	206 (in total 406)

MAFMA has improved the prediction for all settings with a considerable reduction of the model dimension compared to GMAFMA. Further, Table 2 being combined with the results in Table 3, it also suggests that only including the linear form of past market moving directions in the GMAFMA and PGMAFMA would not help improve the prediction that much (by comparing the results for Scenario 2 with those of Scenario 1), but the interactions of past market moving directions with other covariates have largely enhanced the performances for both GMAFMA and PGMAFMA models (where the results for Scenario 3 are compared with those of Scenarios 1 and 2). It appears that such interactions are nonlinear, playing an important role in enhancing the prediction, which however could not be recognized by linear models, except for the case of GLMNET model. Our PGMAFMA under Scenario 3 achieves the highest AUC value of 0.6954, much higher than other AUC values by any other methods including the random forest.

As it is well known, predicting financial markets is hard as it is somehow explained

Table 3: The performances of the AUC values in one step ahead prediction for all candidate models

Model	Prediction AUC	Number of non-zero predictors
GMAFMA (Scenario 1)	0.6050	231
GMAFMA (Scenario 2)	0.6113	238
GMAFMA (Scenario 3)	0.6176	406
PGMAFMA (Scenario 1)	0.6271	98
PGMAFMA (Scenario 2)	0.6187	165
PGMAFMA (Scenario 3)	0.6953	206
GLM (Scenario 1)	0.5452	231
GLM (Scenario 2)	0.5399	238
GLM (Scenario 3)	0.5116	406
GLMNET (Scenario 1)	0.5546	189
GLMNET (Scenario 2)	0.5746	32
GLMNET (Scenario 3)	0.6208	32
AR(1)	0.5872	7
Random Forest(2)	0.5116	500(3)

Note: The penalized GMAFMA model is estimated using Algorithm 1, detailed in Section 3, which applies the adaptive LASSO on the GMAFMA model (S4.57); The GLMNET model applies the LASSO penalty, which can be estimated by the **glmnet** package in R; The AR model is estimated by the **ar** in R; And the Random Forest model is estimated by the **RandomForest** in R. (1) The AR model uses only the linear form of past market direction Y_t in any scenario; (2) The Random Forest model should, in theory, be able to detect all the interactions automatically, we report here the case where past market direction Y_t is fed to the model; (3) The Random Forest model uses 500 trees and 5 variables tried per split, the final model is thus much more complicated than all other models reported here.

by the market efficient hypothesis theory, involving high irregularity, fluctuation, and noises. This is evidenced by the AUC values for, e.g., the AR model, and the mediocre performance of the Random Forest (RF) model, which may result from its failure to capture the dependent structure of data as the RF assumes i.i.d data, again pointing to the difficulty of such prediction. In this example, we have shown in Tables 2 and 3 that our PGMAFMA model can significantly boost the predictive power compared to the GMAFMA model, as it reduces the prediction error by penalizing/removing uncorrelated covariates. The performances of GLMNET compared with GLM also confirm this benefit.

S5 Poisson model averaging with Mallows criterion

Alternative to our GMAFMA framework, one can follow Hansen (2007) and estimate the optimal weights of each marginal regression via minimizing the Mallows criterion. For specificity, we examine the Mallows criterion under conditional Poisson distribution of Y_t given the past information I_{t-1}. Adapting the original criterion of Hansen (2007), which is designed based on squared error, we have the following maximum likelihood version of the Mallows criterion for Poisson distribution:

$$C_n(W) = D(W) + 2\sum_{k=1}^{d} w_k K_k^*, \tag{S5.60}$$

where $W = (w_1, w_2, ..., w_d)$ denotes the weights assigned for each of the d model forecasts, $\widehat{y}_{ik}$'s, of Y_i such that $\widehat{y}_i^W = \sum_{k=1}^d w_k \widehat{y}_{ik}$ (subjected to $\sum_{k=1}^d w_k = 1$) is the model-averaged forecast. Here, more complex than that for parametric model average, K_k^*, namely the model size, refers to the effective degree of freedom for nonparamatric prediction of the k-th model, i.e., K_k^* is the trace of the hat matrix H_k, where $\widehat{Y}^{(k)} := (\widehat{y}_{1k}, \cdots, \widehat{y}_{nk})^T = H_k Y$ with $Y = (Y_1, \cdots, Y_n)^T$, under conditional mean regression (cf, Fan and Yao (2003)). However, for non-Gaussian distribution under exponential family, it becomes much more complex. From the local linear maximum likelihood fitting in Eq. (9) in the main text, by taking the first-order derivative of the local log-likelihood (i.e., E.q. (9) in the main text) with respect to β_1 and letting it equal to zero, we have:

$$\frac{\partial \ell_{h,x_{k0}}}{\partial \beta} = \sum_{t=1}^n [y_t - \psi'(\theta_{kt})] K_h(x_{kt} - x_{k0}) = 0, \tag{S5.61}$$

$$\approx \sum_{t=1}^n [y_t K_h(x_{kt} - x_{k0}) - \psi'(\beta_1) K_h(x_{kt} - x_{k0})] = 0, \tag{S5.62}$$

owing to the fact that $\widehat{\theta}_{kt} = \widehat{f}_k(x_{kt}) \approx \widehat{\beta}_1 = \widehat{f}_k(x_{k0})$ when x_{kt} is close to x_{k0}. We thus have:

$$\psi'(\beta_1) = \sum_{t=1}^n y_t K_h(x_{kt} - x_{k0}) / \sum_{t=1}^n K_h(x_{kt} - x_{k0}) \triangleq A_{kh}(x_{k0}), \tag{S5.63}$$

which leads to $\widehat{\beta}_1 = \widehat{f}_k(x_{k0}) = \psi'^{-1}(A_{kh}(x_{k0}))$, and hence $\widehat{y}_i = \psi'(\widehat{f}_k(x_{ki})) = A_{kh}(x_{ki})$. Therefore, the effective degrees of freedom for k-th component marginal regression can be calculated as follows:

$$K_k^* = tr(H_k) = \sum_{i=1}^n \frac{\partial \widehat{y}_i}{\partial y_i} = \sum_{i=1}^n \frac{\partial A_{kh}(x_{ki})}{\partial y_i} = \sum_{i=1}^n K_h(0) / \sum_{t=1}^n K_h(x_{kt} - x_{ki}) \tag{S5.64}$$

$$= \frac{1}{n} K_h(0) \sum_{i=1}^{n} [\frac{1}{n} \sum_{t=1}^{n} K_h(x_{kt} - x_{ki})]^{-1}. \tag{S5.65}$$

Analogue to Hansen (2007) of the quadratic form of the weighted error, $D(W)$ denotes the total deviance of the averaged predictions:

$$D(W) = 2 \sum_{i=1}^{n} [y_i \log(\frac{y_i}{\widehat{y}_i^W}) - (y_i - \widehat{y}_i^W)]. \tag{S5.66}$$

We thus have:

$$C_n(W) = 2 \sum_{i=1}^{n} [y_i \log(\frac{y_i}{\widehat{y}_i^W}) - (y_i - \widehat{y}_i^W)] + 2 \sum_{k=1}^{d} w_k \frac{1}{n} K_h(0) \sum_{i=1}^{n} [\frac{1}{n} \sum_{t=1}^{n} K_h(x_{kt} - x_{ki})]^{-1}. \tag{S5.67}$$

Optimal weights W^* are thus obtained by minimizing $C_n(W)$, which can be done via Constrained Optimization BY Linear Approximation (COBLA) in R package **nloptr**.

To test its performance, we apply the above approach to the US Strike dataset examined in the main text (Subsection 6.2.1). For this dataset, we have $X_t = (Y_{t-1}, Y_{t-2}, Z_t, Z_{t-1}, Z_{t-2})$ with $d = 5$, and $f_k(X_{kt})$, with X_{kt} the k-th component of X_t, being our one-dimensional marginal regression:

$$\mu_t = \sum_{k=1}^{5} w_k \widehat{y}_{kt}, \tag{S5.68}$$

where $\widehat{y}_{kt}$ is obtained from the k-th marginal model:

$$\log(\mu_{kt}) = f_k(X_{kt}). \tag{S5.69}$$

We report the results in Table 4 and summarize them as follows. With the default setting of the bandwidth for each marginal regression (i.e., $h = 0.3$), the optimal

Table 4: Prediction performances of candidate models for strike data with Mallows criterion under Poisson

Method	MAE
Mallows (h=0.3)	3.75
Mallows (h via ThumbBw)	2.02
AR(2)	1.98
GLM (with lagged info)	1.91
GLM (without lagged info)	2.55
GMAFMA	1.80
PGMAFMA	**1.67**

Note: Forward CV with bandwidth $h = 0.3$ selects optimal weight $w_1 = 1$. Forward CV with bandwidth optimized via ThumbBw uses weights $w_2 = 2/3$ and $w_3 = 1/3$. The choice of bandwidth directly determines the effective degrees of freedom in our marginal nonparametric estimator (recall K_h in (S5.67)).

solution is to place all weight on the $\widehat{y}_1$, which reduces the model back to an AR(1) process. The mean absolute error for prediction is, therefore, 3.75, and much higher than our proposed GMAFMA and PGMAFMA framework. After tuning the bandwidth for each marginal regression using *ThumbBw* function provided in R package **locpol**, the optimal solution is to place approximately 2/3 weights on $\widehat{y}_2$ and 1/3 on $\widehat{y}_3$. This leads to an improved mean absolute error for prediction that is 2.02, which is still much worse than the MAE of 1.67 of our proposed PGMAFMA (c.f., Table 4 with results for other models from Table 1 in Subsection 6.2.1). One possible reason for this gap may stem from the difficulty of reliably estimating the *effective degrees of freedom* for complex nonparametric components with Mallows criterion in our setting.

In short, our proposed PGMAFMA appears to be much easier to be implemented than the Mallows–type penalisation and performs well in our context. Coupled with the fact that Mallows–type model averaging is most commonly developed for OLS and other parametric settings (c.f., Zhang et al. (2016)), our context of nonparametric marginal regressions model combination by Mallows criterion is inherently more complex than that of the parametric models and likely warrants a dedicated investigation, which is beyond the scope of the present study.

S6 Forward cross-validation for tuning parameter

In this section, we further examine a method of forward cross-validation, like in Sun et al. (2023) and Zhang and Zhang (2023), for tuning parameter selection in our PG-MAFMA, as a referee suggested. We consider two rolling-window schemes for selection of λ for the dataset of US strikes in Subsection 6.2.1: a fixed-size (sliding) window and an expanding window that begins at the same initial size and grows as new observations arrive. The ideas have been illustrated below in Figure 4.

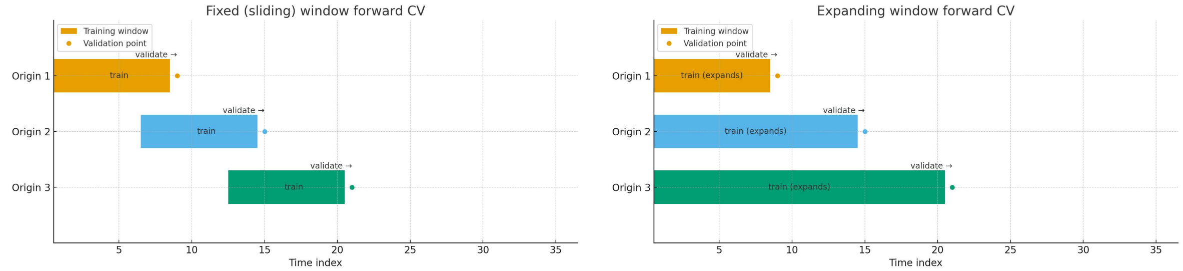

Figure 4: Forward cross-validation window schemes: fixed (sliding) vs. expanding (one-step ahead).

For the US Strike dataset analyzed in Subsection 6.2.1 with $X_t = (Y_{t-1}, Y_{t-2}, Z_t, Z_{t-1}, Z_{t-2})$ and $d = 5$ for lag order 2, we noticed that the forward cross-validation either by the fixed or expanding scheme (with several window sizes, such as 50 and 65) for our PGMAFMA (implemented via the *glmnet*) selected the same penalty parameter λ (≈ 0.001) as that by the *cv.glmnet* from the R package **glmnet** in Subsection 6.2.1 with Table 1.

To facilitate a more insight into forward-CV and *cv.glmnet* for tuning parameter selection, we further considered an enlarged lag order from 2 to 5, yielding a total of 11 predictors: 10 lagged terms (5 for X_t and 5 for Z_t) plus the contemporaneous Z_t. The sample comprises 103 observations, of which the first 93 are used for training and the remaining 10 for out-of-sample evaluation with the two schemes of forward CV and the cv.glmnet. As a benchmark, we also built an AR(5) model and other models. Summary results for the coefficient estimates and the MAEs of one-step-ahead predictions for the out-of-sample evaluation for those different models considered are reported in Table 5.

As shown by Table 5 with our experiments, the forward cross-validation (forward-CV) with each of the two window schemes achieves the mean absolute errors (MAEs) for prediction comparable to (nearly the same as) those from *cv.glmnet*, but at a higher computational cost. Unlike the *cv.glmnet* that is easily implemented, the accuracy of forward-CV appears to be highly sensitive to the rolling-window design: the best MAE (1.994) occurs with both windows of size 65, whereas the worst MAE (2.096) arises with a fixed window (FW) of size 35. Consistent with Zhang and Zhang (2023), these results indicate that the rolling-window length influences the selected penalty level λ_n and, consequently, predictive performance. It thus introduces an additional hyperparameter for tuning. From a computational perspective, forward-CV, especially with an expanding window, is considerably more time-consuming than *cv.glmnet* because it evaluates many more folds. Nevertheless, by Table 5, the penalized models selected by forward CV and by *cv.glmnet* perform nearly the same (no essentially big difference) in MAE for prediction, both of which are however worse than the PGMAFMA (with

Table 5: Estimated coefficients and prediction errors for US Strike data for lag order 5.

						Forward CV						*cv.glmnet*	
	AR(5)	AR(2)	GLM (lagged)	GLM (no lag)	GMAFMA	FW			EW			$\lambda_{\min}$	λ_{1se}
						35	50	65	35	50	65		
Penalty λ						0.002	0.008	0.039	0.008	0.017	0.039	0.008	0.140
Intercept			1.041	1.641	0.080	0.008	0.253	0.776	0.253	0.551	0.776	0.250	1.209
Z_t			-2.957	3.194	-0.248								
Y_{t-1}	0.388	0.415	0.059		0.081	1.511	1.648	1.946	1.648	1.835	1.946	1.646	1.398
Y_{t-2}	0.235	0.254	0.037		0.061	1.509	1.517	1.100	1.517	1.504	1.100	1.517	
Y_{t-3}	0.163		0.027		0.067	1.845	1.518		1.518	0.963		1.522	
Y_{t-4}	-0.037		-0.008		-0.029								
Y_{t-5}	-0.122		-0.018		0.066	0.508							
Z_{t-1}			3.671		0.511								
Z_{t-2}			3.951		-0.095								
Z_{t-3}			-20.209		-0.744								
Z_{t-4}			28.369		0.860	4.282	2.419		2.419			2.440	
Z_{t-5}			-11.076		-0.244								
MAE	2.179	2.128	6.898	2.097	4.634	2.096	2.046	1.994	2.046	2.008	1.994	2.047	2.043

Notes: FW denotes a fixed rolling window and EW an expanding window. All values are rounded to three decimals. Results are reported with bandwidth $h = 0.3$ because the data-driven selector *ThumbBw* yields optimal bandwidths that, in this application, are clustered near 0.3 and do not affect cross-validation performance.

MAE of 1.67) reported in Table 1 of Subsection 6.2.1.

In short, the chosen penalty level λ_n looks to be robust with the easy use of *cv.glmnet* in Subsection 6.2.1, getting similar results of prediction to those by using forward CV that requires additional tunings of window sizes at a cost of more computations. A comprehensive investigation into forward CV with optimal window design is however beyond the scope of this study, which is left for a potential future work.

References

Doukhan, P., Massart, P. and Rio, E. (1995). Invariance principles for absolutely regular empirical processes. *Ann. Inst. H. Poincaré Probab. Statist.* **31**, 393–427.

Fan, J., Härdle, W. and Mammen, E. (1998). Direct estimation of low-dimensional components in additive models. *The Annals of Statistics* **26** (3), 943–971.

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.

Fan, J., Yao, Q. and Cai, Z. (2003). Adaptive varying-coefficient linear models. *Journal of the Royal Statistical Society: Series B statistical methodology* **65** (1), 57–80.

Hansen, B. E. (2007). Least squares model averaging. *Econometrica* **75** (4), 1175–1189.

Hardle, W., Hall, P., Ichimura, H. et al. (1993). Optimal smoothing in single-index models. *The Annals of Statistics* **21** (1), 157–178.

Jo, N., Aghaei, S., Benson, J., Gomez, A. and Vayanos, P. (2023). Learning optimal fair decision trees: Trade-offs between interpretability, fairness, and accuracy. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 181–192.

Li, D., Linton, O. and Lu, Z. (2015). A flexible semiparametric forecasting model for time series. *Journal of Econometrics* **187** (1), 345–357.

Li, Q. and Racine, J. (2003). Nonparametric estimation of distributions with categorical and continuous data. *Journal of Multivariate Analysis* **86** (2), 266–292.

Lu, Z., Tjøstheim, D. and Yao, Q. (2007). Adaptive varying-coefficient linear models for stochastic processes: asymptotic theory. *Statistica Sinica* **17** (1), 177–198.

Nielsen, S. F. (2005). Local linear estimating equations: Uniform consistency and rate of convergence. *Nonparametric Statistics* **17** (4), 493–511.

Peng, R. and Lu, Z. (2023). Uniform consistency for local fitting of time series nonparametric 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.

Racine, J. and Li, Q. (2004). Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics* **119** (1), 99–130.

Sun, Y., Hong, Y., Wang, S. and Zhang, X. (2023). Penalized time-varying model averaging. *Journal of Econometrics* **235** (2), 1355–1377.

Van Der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer.

Xia, Y. and Li, W. (1999). On single-index coefficient regression models. *Journal of the American Statistical Association* **94** (448), 1275–1285.

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.