Statistica Sinica Preprint No: SS-2020-0340

Title	Frequentist Model Averaging for the Nonparametric Additive Model
Manuscript ID	SS-2020-0340
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202020.0340
Complete List of Authors	Jun Liao, Alan Wan, Shuyuan He and Guohua Zou
Corresponding Author	Guohua Zou
E-mail	ghzou@amss.ac.cn
Notice: Accepted version subject to English editing.	

Statistica Sinica

Frequentist Model Averaging for the Nonparametric Additive Model

Jun Liao[a,c], Alan T.K. Wan[b], Shuyuan He[c], and Guohua Zou[c,*]

[a] *Center for Applied Statistics and School of Statistics, Renmin University of China*
[b] *Department of Management Sciences, City University of Hong Kong*
[c] *School of Mathematical Sciences, Capital Normal University*

Abstract: This paper develops an optimal frequentist model averaging approach for estimating the unknown conditional mean function in the nonparametric additive model when the covariates and the degree of smoothing are subject to uncertainty. Our weight choice criterion selects model weights by minimising a plug-in estimator of the risk of the model average estimator under a squared error loss function. We derive the convergence rate of the model weights obtained from our proposed method to the infeasible optimal weights, and prove that the resultant model average estimators are asymptotically optimal. An extension to the additive autoregressive model for time series data is also considered. Our simulation analysis shows that the proposed model average estimators can significantly outperform several commonly used model selection estimators and their model averaging counterparts in terms of mean squared error in a large part of the parameter space. We further illustrate our methods in two real data studies.

Key words and phrases: Additive model, Asymptotic optimality, Autoregressive model, Consistency, Model averaging.

[*]Corresponding author: Guohua Zou, Email: *ghzou@amss.ac.cn*.

1. Introduction

The nonparametric additive model (AM) (Stone, 1985; Hastie and Tibshirani, 1990) is a well-known statistical modeling approach. The AM is a class of regression models where the usual linear relationship between the response and covariates is replaced by a sum of univariate smooth functions. AMs thus avoid much of the curse of dimensionality that afflicts fully nonparametric regression and afford more flexibility than traditional linear models with respect to the covariate effects. The smooth functions in AMs are commonly estimated by backfitting (Buja et al., 1989; Mammen et al., 1999; Opsomer, 2000; Nielsen and Sperlich, 2005; Ravikumar et al., 2009), smoothing splines (Stone, 1985; Doksum and Koo, 2000; Huang and Yang, 2004; Chen et al., 2018b), or marginal integration methods (Tjostheim and Auestad, 1994; Linton and Nielsen, 1995; Fan et al., 1998). AMs have been widely applied in many disciplines. These applications are numerous in such areas as ecology, economics, environmental research and medicine. Recent examples include Eyto and Irvine (2007), Bontemps et al. (2008), to name a few.

Model selection is a vital aspect of any statistical analysis. Within the framework of the AM, model selection typically involves choosing covariates and their degrees of smoothing. Huang and Yang (2004) proposed a spline-based Bayesian Information Criterion (BIC) for selecting the lag order in a non-linear additive autoregressive model. They proved the consistency of the proposed method, and examined the analogous spline splitting methods based on the Akaike Information Criterion (AIC) and generalised cross validation (GCV) in a simulation study. Xue (2009) introduced a penalised polynomial spline method for simultaneous model estimation and variable selection in AM. Huang et al. (2010) considered the group lasso. Belitz and Lang (2008) and Cantoni et al. (2011) developed algorithms for component selection in AMs. Other well-known works on model selection in AMs include Härdle and Korostelev (1996), Chen et al. (2011), Fan et al. (2011), among others.

In recent years, model averaging has emerged as a viable alternative to model selection. Unlike model selection that chooses one model for the data at hand, model averaging adopts a weighted ensemble approach that allows multiple models to contribute to the analysis in proportion to their estimated performance, and in doing so, it can capture all the useful information of the models and avoid the risk of "putting all eggs in one basket". Model averaging usually produces more stable estimates and more precise forecasts than those obtained from a single model. As well, as model averaging properly accommodates uncertainty over models in situations where there is no predominant model to call on, it benefits statistical inference (Hjort and Claeskens, 2003; Liu, 2015; Zhang and Liu, 2019). Model averaging may be viewed as a smoothed extension of model selection from the point of view of estimation and prediction.

A major part of the model averaging literature is about choosing model weights. When approached from a Bayesian perspective, the model weights are usually determined by the individual models' posterior probabilities. In this paper, we focus on the frequentist approach to model averaging, which has been making inroads into the realm of statistics and data analysis in recent years. Frequentist model averaging (FMA) precludes the need to specify any prior distribution, although determining an optimal weight choice by a data-driven method is arguably more challenging for the frequentist formulation than its Bayesian counterpart. The FMA strategies that have been developed include weighting by information criterion scores (Buckland et al., 1997), adaptive regression by mixing (Yang, 2001), Mallows model averaging (Hansen, 2007), Jackknife model averaging (Hansen and Racine, 2012; Ando and Li, 2014), optimal MSE averaging (Liang et al., 2011), and averaging by Kullback-Leibler-type measures (Zhang et al., 2016). Claeskens (2016) provided a survey of this rapidly expanding literature.

While the optimal choice of weights in FMA has been extensively researched, the vast majority of the literature focuses on parametric models. The optimal FMA methods are relatively less well developed for nonparametric and semiparametric models. Hansen (2014) suggested the Jackknife criterion for the weight choice of nonparametric sieve regression averaging estimator. For the partially linear model (PLM), Zhang and Wang (2019) developed a Mallows-type weight choice criterion and studied the asymptotic optimality of the corresponding model averaging estimator. Zhu et al. (2019) proposed a weight choice criterion in a PLM with varying coefficients. Other studies of FMA on the weight choice in nonparametric and semiparametric models include Gao (2015), Chen et al. (2018a), Li et al. (2018), among others.

The principal scientific contribution of this article is to develop the optimal FMA approaches for the nonparametric additive models and additive autoregressive models, which have not been thoroughly studied in the existing literature. Our weight choice criterion is based on minimising a plug-in estimator of the squared error risk of the FMA estimator. We consider two plug-in estimators that have similar forms but different penalties. We derive the convergence rate of our weights to the infeasible optimal weights, and prove that the model average estimators obtained from the weight choice criterion are asymptotically optimal in the sense of achieving the smallest possible squared error. It should be pointed out that the development of asymptotic theory in the present context is nontrivial, because in addition to the uncertainty in the covariate set, we also allow for uncertainty in the degree of smoothing and a dependent data structure.

The remainder of this paper is organised as follows. Section 2 describes the model setup, discusses the FMA scheme and presents the main results on the asymptotic properties of the proposed model averaging method. In Section 3, we extend our analysis to the additive autoregressive models. Section 4 reports the results of a simulation study that examines the finite sample performance of the proposed model averaging estimators. Section 5 applies the proposed method to two real data sets related to medical and environmental research. Section 6 concludes. Proofs of results are contained in the appendix.

2. Model setup and the weight choice criterion

Consider the nonparametric additive model

$$Y = \sum_{j=1}^{d_0} g_j(X^{(j)}) + e \equiv \mu + e, \tag{2.1}$$

where $Y = (y_1, \cdots, y_n)'$ is a random vector, $X^{(j)} = (x_{1j}, ..., x_{nj})'$ is the j^{th} covariate, g_j's are one-dimensional nonparametric functions with $g_j(X^{(j)}) =$

$(g_j(x_{1j}), ..., g_j(x_{nj}))'$, and $e = (e_1, ..., e_n)'$ is the vector of disturbance terms, with e_i's being independent with mean zero and variance σ^2. For proving the asymptotic optimality of the model average estimators, we need no distributional assumption on e, but for the purpose of developing the weight choice criterion we assume that e follows a multivariate normal distribution. The same approach was taken by Zhang et al. (2015), who examined model averaging in another context. Now, assume that there are M candidate models each corresponding to a different covariate set and degree of smoothing. We let $\hat{\mu}_m = P_m Y$ be the estimator of μ under the m^{th} candidate model, where P_m is a hat matrix. The form of P_m depends on the estimation method. As discussed in Section 1, backfitting, smoothing splines and marginal integration are some of the common estimation methods that have been extensively studied in the literature.

The FMA estimator of μ can be expressed as $\hat{\mu}(w) = \sum_{m=1}^{M} w_m \widehat{\mu}_m = P(w)Y$, where $w \in \mathcal{W} = \{w \in [0, 1]^M : \sum_{m=1}^{M} w_m = 1\}$ is the weight vector, and $P(w) = \sum_{m=1}^{M} w_m P_m$. We assume that the M^{th} model has the largest dimension among all candidate models.

Our weight choice criterion is based on a minimisation of a plug-in estimator of the risk of $\hat{\mu}(w)$ under a squared error loss function. Define the squared error loss function of $\hat{\mu}(w)$ as $L(w) = \|\mu - \hat{\mu}(w)\|^2$. Note that

$$\begin{aligned} L(w) &= \|Y - \hat{\mu}(w)\|^2 - 2e'(Y - \hat{\mu}(w)) + e'e \\ &= \|Y - \hat{\mu}(w)\|^2 - 2e'(I - P(w))(\mu + e) + e'e \\ &= \|Y - \hat{\mu}(w)\|^2 + 2e'P(w)e - 2e'(I - P(w))\mu - e'e. \qquad (2.2) \end{aligned}$$

From (2.2), we obtain the following scaled risk function:

$$\begin{aligned} R^0(w) &= E(L(w)/\sigma^2) \\ &= E\left[\|Y - \hat{\mu}(w)\|^2/\sigma^2 + 2\sum_{m=1}^{M} w_m e'P_m e/\sigma^2\right] - n. \qquad (2.3) \end{aligned}$$

Although an ideal approach would be to choose w by minimising $R^0(w)$ directly, this would not result in a solution as $R^0(w)$ involves unknown expectations. Hence we consider minimising a plug-in estimator of $R^0(w)$. Ignoring the constant n in (2.3), our plug-in estimator of $R^0(w)$ takes the form

$$\frac{\|Y - \hat{\mu}(w)\|^2}{\widehat{\sigma}_M^2} + 2\sum_{m=1}^{M} w_m E\left(\frac{e'P_m e}{\widetilde{\sigma}_m^2}\right), \qquad (2.4)$$

where $\widehat{\sigma}_M^2 = \widehat{e}_M'\widehat{e}_M/(n - tr(P_M))$ is the least squares estimator of σ^2 based on the largest model, $\widehat{e}_M = Y - \hat{\mu}_M$, and $\widetilde{\sigma}_m^2 = \|Y - \hat{\mu}_m\|^2/n = Y'(I_n - P_m)'(I_n - P_m)Y/n$ is the maximum likelihood estimator of σ^2 based on the m^{th} candidate model.

Our substitution of σ^2 by $\widehat{\sigma}_M^2$ in the first term on the right-hand-side of (2.3) follows Mallows (1973), who used the same approach to derive the

Mallows' Cp criterion. On the other hand, $e'P_m e/\sigma^2$ can be thought of as the "penalty" for the m^{th} model. Hence it makes sense to estimate σ^2 in each of $e'P_m e/\sigma^2$ in (2.3) by $\widetilde{\sigma}_m^2$ so that there are different penalties for different models. This is similar to the idea of using different tuning parameters for different coefficients in LASSO (Wang et al., 2007).

To utilise (2.4) as a weight choice criterion, an evaluation of $E(e'P_m e/\widetilde{\sigma}_m^2)$ is required. By the Stein's lemma (Stein, 1981), we have

$$\begin{aligned}
& E(e'P_m e/\widetilde{\sigma}_m^2) \\
= \;& \sigma^2 Etr\left(\frac{\partial P_m e\widetilde{\sigma}_m^{-2}}{\partial Y'}\right) = \sigma^2 Etr\left(P_m \frac{\partial e\widetilde{\sigma}_m^{-2}}{\partial Y'}\right) \\
= \;& \sigma^2 tr\left\{P_m E\left(\widetilde{\sigma}_m^{-2}\right) + P_m E\left(e\frac{\partial \widetilde{\sigma}_m^{-2}}{\partial Y'}\right)\right\} \\
= \;& tr\left\{P_m E\left(\sigma^2\widetilde{\sigma}_m^{-2}\right) + \sigma^4 P_m E\left(\frac{\partial^2 \widetilde{\sigma}_m^{-2}}{\partial YY'}\right)\right\} \\
= \;& tr\left\{P_m E\left(\sigma^2\widetilde{\sigma}_m^{-2}\right) + \sigma^4 P_m E\left(2\widetilde{\sigma}_m^{-6}\frac{\partial \widetilde{\sigma}_m^2}{\partial Y}\frac{\partial \widetilde{\sigma}_m^2}{\partial Y'} - \widetilde{\sigma}_m^{-4}\frac{\partial^2 \widetilde{\sigma}_m^2}{\partial YY'}\right)\right\}.
\end{aligned}$$

Noting that $\partial\widetilde{\sigma}_m^2/\partial Y = 2n^{-1}(I-P_m)'(I-P_m)Y$ and $\partial^2\widetilde{\sigma}_m^2/\partial YY' = 2n^{-1}(I-P_m)'(I-P_m)$, we can write

$$\begin{aligned}
E(e'P_m e/\widetilde{\sigma}_m^2) \;=\;& E\{\sigma^2\widetilde{\sigma}_m^{-2}tr(P_m) + 8n^{-2}\sigma^4\widetilde{\sigma}_m^{-6}Y'(I-P_m)'(I-P_m)P_m(I-P_m)' \\
& (I-P_m)Y - 2n^{-1}\sigma^4\widetilde{\sigma}_m^{-4}tr\left((I-P_m)P_m(I-P_m)'\right)\}. \qquad (2.5)
\end{aligned}$$

Substituting (2.5) in (2.4), we obtain the following weight choice criterion:

$$\left\|Y - \hat{\mu}(w)\right\|^2/\hat{\sigma}_M^2 + 2\sum_{m=1}^{M} w_m \mathrm{df}_m \qquad (2.6)$$

where

$$\begin{aligned} \mathrm{df}_m &= \sigma^2\widetilde{\sigma}_m^{-2}tr(P_m) \\ &\quad +8n^{-2}\sigma^4\widetilde{\sigma}_m^{-6}Y'(I-P_m)'(I-P_m)P_m(I-P_m)'(I-P_m)Y \\ &\quad -2n^{-1}\sigma^4\widetilde{\sigma}_m^{-4}tr\left((I-P_m)P_m(I-P_m)'\right). \end{aligned}$$

For the special case of a symmetric and idempotent P_m, the weight choice criterion simplifies to

$$\|Y-\hat{\mu}(w)\|^2/\hat{\sigma}_M^2+2\sum_{m=1}^{M}w_m\frac{\sigma^2tr(P_m)}{\widetilde{\sigma}_m^2}. \tag{2.7}$$

The unknown σ^2 in (2.7) may be replaced by the least squares estimator $\widehat{\sigma}_m^2=\|Y-\hat{\mu}_m\|^2/(n-tr(P_m))$ based on the m^{th} candidate model, or its maximum likelihood counterpart $\widetilde{\sigma}_m^2$, both are commonly used for estimating σ^2. We first consider replacing σ^2 in (2.7) by $\widehat{\sigma}_m^2$. This yields

$$\|Y-\hat{\mu}(w)\|^2/\hat{\sigma}_M^2+2\sum_{m=1}^{M}w_m\frac{ntr(P_m)}{n-tr(P_m)}. \tag{2.8}$$

In order to utilise (2.8) one needs an explicit form of the hat matrix P_m. As discussed earlier, the form of P_m depends on the method of estimation. Here, we consider the spline smoothing estimation method which has the advantage of simplicity (Huang and Yang, 2004). Let a_m, b_m and c_m represent some sequences varying with m. Denote the knot sequence as $\{a=\zeta_{j,0}<\zeta_{j,1}<\cdots<\zeta_{j,N_j^{a_m}}<\zeta_{j,N_j^{a_m}+1}=b\}$ for g_j, where $a<b$ are

finite numbers and $N_j^{a_m}$ is the number of interior knots. Let φ be the polynomial spline space consisting of functions that are, first, polynomials of degree $l_j^{b_m}$ on intervals $[\zeta_{j,s},\zeta_{j,s+1})(s=0,...,N_j^{a_m}-1)$ and $[\zeta_{j,N_j^{a_m}},\zeta_{j,N_j^{a_m}+1}]$, and second, $l_j^{b_m}-1$ times continuously differentiable on $[a,b]$ (Stone, 1985; de Boor, 2001; Xue, 2009; Huang et al., 2010). We assume that there exists a basis $B_j^m(x)=(B_{j1}(x),...,B_{jq_j^m}(x))'$ ($q_j^m=N_j^{a_m}+l_j^{b_m}$ and $x\in[a,b]$) for the spline space φ. Some examples include the truncated power basis and B-spline basis (de Boor, 2001). Without loss of generality, assume that $x_{ij}\in[a,b]$. Denote $B_j^m=(B_j^m(x_{1j}),...,B_j^m(x_{nj}))'$. Let s_{c_m} be the index set of covariates under the m^{th} model, i.e., s_{c_m} is a subset of $\{1,...,d\}$. As the candidate models may be misspecified, d is not necessarily equal to d_0 given in (2.1). For the m^{th} candidate model, we can write the spline estimator of μ as $\hat{\mu}_m=\sum_{j\in s_{c_m}}B_j^m\hat{\theta}_j^m$, where $\hat{\theta}^m=\left[\hat{\theta}_j^{m\prime}\right]'_{j\in s_{c_m}}$ represents the $\sum_{j\in s_{c_m}}q_j^m$ dimensional vector satisfying

$$\hat{\theta}^m=\mathrm{argmin}_{\theta^m}\|Y-\sum_{j\in s_{c_m}}B_j^m\theta_j^m\|^2, \tag{2.9}$$

with $\theta^m=\left[\theta_j^{m\prime}\right]'_{j\in s_{c_m}}$. This yields $P_m=B^m(B^{m\prime}B^m)^{-1}B^{m\prime}$, which is symmetric and idempotent, where $B^m=[B_j^m]_{j\in s_{c_m}}$ is the $n\times\sum_{j\in s_{c_m}}q_j^m$ basis matrix. Substituting P_m in (2.8) results in the criterion

$$\phi(w)=\|Y-\hat{\mu}(w)\|^2+2\widehat{\sigma}_M^2\sum_{m=1}^M w_m\frac{nr_m}{n-r_m}, \tag{2.10}$$

where $r_m = \sum_{j \in s_{c_m}} q_j^m$ and $q_j^m = N_j^{a_m} + l_j^{b_m}$. Write $\hat{w} = \text{argmin}_{w \in \mathcal{W}} \phi(w)$, and label $\hat{\mu}(\hat{w})$, the FMA estimator of μ that uses $\hat{w}$, the additive model average (AMA) estimator.

Remark 1. The way in which a_m, b_m and c_m vary with m determines their explicit expressions. Let h_1 and h_2 be the numbers of knot sequences and degrees respectively. If, for example, a_m vary in $\{1, 2, ..., h_1\}$ and b_m vary in $\{1, 2, ..., h_2\}$, then we can write $a_m = m - h_1[(m-1)/h_1]$, $b_m = 1 + \left[\frac{m-1-h_1h_2\left[\frac{m-1}{h_1h_2}\right]}{h_1}\right]$, and $c_m = 1 + [(m-1)/(h_1h_2)]$, where $[A]$ denotes the integral part of A.

Remark 2. If the unknown σ^2 in the penalty term of (2.7) is replaced by the maximum likelihood estimator $\widetilde{\sigma}_m^2$ instead of $\widehat{\sigma}_m^2$, then (2.8) is changed to

$$\|Y - \hat{\mu}(w)\|^2 / \hat{\sigma}_M^2 + 2 \sum_{m=1}^{M} w_m tr(P_m). \tag{2.11}$$

Substituting P_m in (2.11) yields the alternative criterion

$$\phi_H(w) = \|Y - \hat{\mu}(w)\|^2 + 2\widehat{\sigma}_M^2 \sum_{m=1}^{M} w_m r_m, \tag{2.12}$$

which has the same form as Hansen's (2007) Mallows weight choice criterion for linear regression. Denote $\tilde{w} = \text{argmin}_{w \in \mathcal{W}} \phi_H(w)$. We label the model average estimator $\hat{\mu}(\tilde{w})$ arising from $\phi_H(w)$ the AMAH estimator.

Clearly, $\phi(w)$ and $\phi_H(w)$ are plug-in versions of the same criterion, and they differ only in the estimator of σ^2 being used in the penalty term of the criterion. Like Hansen (2007), $\phi_H(w)$ can actually be derived without the normality assumption on the error term. The estimators resulting from $\phi(w)$ and $\phi_H(w)$ have the same asymptotic properties but different finite sample properties. See Section 4 for details. In fact, since $\phi(w)$ imposes heavier penalties on the larger models, it may have some merits over $\phi_H(w)$ in small samples.

Remark 3. Our weight choice scheme is formulated as a quadratic programming problem. For small to moderate values of M (say $M \leq 400$), the computation is manageable and can be performed efficiently via, for example, the function 'solve.QP' in R. When M is large, one can first subset the models into a smaller candidate set via "model screening" before applying model averaging to the smaller set. Some well-known model screening methods include the "top M" method based on AIC and/or BIC scores (Yuan and Yang, 2005) and the forward and backward stepwise procedure developed specifically for additive models when only the covariates are subject to uncertainty (Huang and Yang, 2004).

Now, denote the risk function and the minimum possible risk as $R(w) = E\left[\|\mu - \hat{\mu}(w)\|^2\right]$ and $\xi_n = \inf_{w \in \mathcal{W}} R(w)$ respectively. Let C be a generic positive constant that can take on different values in different contexts. Consider the following regularity conditions:

Condition (C.1). $E(e_i^{4G}) \leq C < \infty$, and $M\xi_n^{-2G}\sum_{m=1}^{M}(R(w_{m0}))^G \to 0$ for some fixed integer $1 \leq G < \infty$, where w_{m0} is an $M \times 1$ vector with the m^{th} element taking on the value of unity and the others zeros.

Condition (C.2). $\mu'\mu/n = O(1)$.

Condition (C.3). $r_M^2/n \leq C < \infty$.

Conditions (C.1) - (C.3) are similar to those used by Wan et al. (2010). The first part of Condition (C.1) places a constraint on the moment of $\{e_i\}$, while the second part restricts the rate of increase of M. As an example, assume that ξ_n has the order of n^β with $\beta > 1/2$, which is a reasonable assumption for nonparametric regressions (Ando and Li, 2014), and $\max_{1\leq m\leq M} R(w_{m0}) = O(n)$. Then the second part of (C.1) holds if $M^2/n^{G(2\beta-1)} \to 0$. It has been shown in other contexts that if the second part of (C.1) is removed, then the asymptotic optimality of FMA estimators can only be established under a somewhat restricted weight set (Cheng et al., 2015). Condition (C.2) is mild and commonly used. Condition (C.3) restricts the rate at which r_M increases.

Theorem 1. *Under Conditions (C.1)-(C.3),*

$$L(\hat{w})/\inf_{w\in\mathcal{W}} L(w) \xrightarrow{p} 1, \tag{2.13}$$

and

$$L(\tilde{w})/\inf_{w\in\mathcal{W}} L(w) \xrightarrow{p} 1, \tag{2.14}$$

where $\hat{w} = argmin_{w\in\mathcal{W}}\phi(w)$ *and* $\tilde{w} = argmin_{w\in\mathcal{W}}\phi_H(w)$.

Proof. See the Appendix.

Theorem 1 shows that both the AMA and AMAH estimators are asymptotically optimal in the sense of achieving the smallest possible squared error.

Remark 4. For the FMA estimator to approach the true nonparametric component of the model, it is necessary for the number of knots to increase with the sample size. Stone (1985) and Huang and Yang (2004) showed that the estimator of the nonparametric function can attain the optimal rate of convergence if the number of knots is of the order of $n^{1/5}$. In the following, we demonstrate that if the candidate model set includes the models associated with the spline estimators with the optimal rate of convergence given in Stone (1985), the asymptotic optimality stated in Theorem 1 remains valid provided that some mild regularity conditions are fulfilled. Let us fix the polynomial degree and the number of covariates, allow the number of knots to vary, and assume that the candidate set comprises models of dimensions (r_m) in the order of $n^{1/5}$ (i.e., the number of knots has the order of $n^{1/5}$). Then the mean squared error of the spline estimator (divided by n) has the order $n^{-4/5}$, which is the optimal rate of convergence for spline estimators (Stone, 1985). For the above example, Condition (C.3) clearly holds, and Theorem 1 remains valid if Condition (C.2) and the first part of Condition (C.1) are true and $M^2/n^{G/5} = o(1)$. The latter holds for appropriate M and G and is a sufficient condition for the second part of Condition (C.1) to hold.

Remark 5. Like all model selection and averaging procedures, the performance of the proposed procedure depends heavily on the construction of the candidate set. We need to impose some conditions on the candidate model set in order for the proposed procedure to be asymptotically optimal. See Section 2 for a discussion of the conditions. In practice, one often looks to the context of the investigation for guidance in constructing candidate models, e.g., econometric models are usually based on economic theories. In the absence of any context-specific information, one can consider a large number of candidate models with varying degrees of smoothness and different covariates in the initial stage, then apply model screening to reduce the candidate set to a manageable scale. See Remark 3 above.

Now, denote the (infeasible) optimal weight obtained from a direct minimisation of $R(w)$ as $w^0 = \text{argmin}_{w \in \mathcal{W}} R(w)$. It is assumed that w^0 is an interior point of $\mathcal{W}$. Let $\lambda_{\min}(A)$ and $\lambda_{\max}(A) = \|A\|$ be the minimum and maximum singular values of a general real matrix A respectively. Denote $\Lambda_1 = (\hat{\mu}_1, ..., \hat{\mu}_M)$ and $\Lambda = \Lambda_1' \Lambda_1$. Consider the following regularity conditions:

<u>Condition</u> (C.4). There are two positive constants κ_1 and κ_2, such that $Pr(0 < \kappa_1 < \lambda_{\min}(\Lambda/n) \leq \lambda_{\max}(\Lambda/n) < \kappa_2 < \infty)$ tends to 1 as $n \to \infty$.

<u>Condition</u> (C.5). $\lambda_{\max}\{(B^{m\prime} B^m/n)^{-1}\} = O(1)$ uniformly in m.

<u>Condition</u> (C.6). $r_M/n = o(1)$ and $M r_M/(n^{2\delta} \xi_n) = o(1)$, where δ is a positive constant.

Condition (C.4) requires both the minimum and maximum singular values of Λ/n to be bounded away from zero and infinity. Other studies that have used similar conditions include Fan and Peng (2004) and Bickel and Levina (2008). Condition (C.5) implies that the maximum singular value of $(B^{m\prime} B^m/n)^{-1}$ is bounded. Ravikumar et al. (2009) used a similar condition. Condition (C.6) allows M and r_M to increase with n, but also places a restriction on their diverging rates.

Theorem 2. *If Conditions (C.2), and (C.4) - (C.6) are satisfied, then there exist local minimisers* $\hat{w}$ *and* $\tilde{w}$ *of* $\phi(w)$ *and* $\phi_H(w)$*, respectively, such that*

$$\|\hat{w} - w^0\| = O_p\left(\xi_n^{1/2} n^{-1/2+\delta}\right), \tag{2.15}$$

and

$$\|\tilde{w} - w^0\| = O_p\left(\xi_n^{1/2} n^{-1/2+\delta}\right), \tag{2.16}$$

where δ *is a positive constant defined under Condition (C.6).*

Proof. See Appendix.

Theorem 2 shows that the weights obtained by minimising $\phi(w)$ and $\phi_H(w)$ approach the optimal weights at the rate of $\xi_n^{1/2} n^{-1/2+\delta}$.

Remark 6. Condition (C.4) is mild when the dimension M of Λ is fixed; it is a strong condition when M diverges, although similar conditions are frequently used in literature, as explained above. In fact, in the latter event, one can consider substituting $\lambda_{\max}(\Lambda/n) = O_p(1)$ by the weaker condition $\lambda_{\max}(\Lambda/n) = O_p(M)$, which is a reasonable alternative because $\lambda_{\max}(\Lambda/n) \leq trace(\Lambda/n) = O_p(M)$ when the diagonal elements of Λ/n are uniformly $O_p(1)$. Now, assuming that $\lambda_{\max}(\Lambda/n) = O_p(M)$, all other things being equal, by following the steps of deriving Theorem 2, we can show that $\|\hat{w} - w^0\| = O_p\left(M^{1/2}\xi_n^{1/2} n^{-1/2+\delta}\right)$, replacing the original conclusion of $\|\hat{w} - w^0\| = O_p\left(\xi_n^{1/2} n^{-1/2+\delta}\right)$ obtained under Condition (C.4). In other words, the use of $\lambda_{\max}(\Lambda/n) = O_p(M)$ in lieu of $\lambda_{\max}(\Lambda/n) = O_p(1)$ implied by the stronger Condition (C.4) when M diverges will result in a slower convergence rate of $\hat{w}$.

3. Weight choice for additive autoregressive models

The purpose of this section is to extend the preceding analysis to an additive autoregressive model, which has the same structure as (2.1), except that $X^{(j)} = (y_{1-j}, ..., y_{n-j})'$ so that $g_j(X^{(j)}) = (g_j(y_{1-j}), ..., g_j(y_{n-j}))'$ and $B_j^m = (B_j^m(y_{1-j}), ..., B_j^m(y_{n-j}))'$, and the subscript i which indexes the observation number is replaced by the time index t ($1 \le t \le n$), i.e., x_{ij} is replaced by y_{t-j} everywhere. We assume that $\{y_t\}$ is a stationary time series process. Like the additive models, the additive autoregressive models are attractive alternatives to traditional nonparametric time series models due to their ability to alleviate the curse of dimensionality. The additive autoregressive models have been extensively studied in the literature. See Chen and Tsay (1993), Huang and Yang (2004), Li and Yang (2007), among others.

Denote $A(w) = I - P(w)$, and let $\tilde{R}(w) = \|A(w)\mu\|^2 + \sigma^2 tr\left\{P^2(w)\right\}$ and $\tilde{\xi}_n = \inf_{w\in\mathcal{W}} \tilde{R}(w)$, where $\tilde{R}(w)$ is an analogue of $R(w)$ (they are the same under the case of independent data in Section 2). Now, consider the following regularity conditions:

Condition (C.7). $\max_{m\in\{1,\ldots,M\},j\in s_{cm},i\in\{1,\ldots,q_j^m\}} E\left|B_{ji}^2(y_{t-j})\right| < \infty$.

Condition (C.8). $\lambda_{\max}\left\{(B^{m\prime}B^m/n)^{-1}\right\} = O_p(1)$ uniformly for $1 \le m \le M$.

Condition (C.9). $\mu'\mu/n = O_p(1)$.

Condition (C.10). $r_M/n = o(1)$ and $r_M^{1/2}n^{1/2}\tilde{\xi}_n^{-1} = o_p(1)$.

Condition (C.7) is a standard moment condition for establishing asymptotic results. Conditions (C.8) and (C.9) are the counterparts of Conditions (C.5) and (C.2) respectively in the context of a time series model. Condition (C.10) assumes that $\tilde{\xi}_n$ increases at a rate faster than $n^{1/2}$ for fixed r_M. The same assumption was used by Ando and Li (2014).

Theorem 3. *Provided that Conditions (C.7) - (C.10) hold and $\{e_t\}$ is mutually independent, Theorem 1 continues to hold.*

Proof. See the Appendix.

Theorem 4. *Provided that Conditions (C.4) and (C.6) - (C.9) hold, and $\{e_t\}$ is mutually independent, Theorem 2 continues to hold.*

Proof. See the Appendix.

Theorems 3 and 4 extend the results on the asymptotic optimality and consistency for the AMA and AMAH estimators from independent data to stationary time series data. One important assumption for Theorems 3 and 4 is that the disturbances are independent. Our results can be further extended to the case where $\{e_t\}$ is weakly correlated. Consider the following regularity conditions:

Condition (C.11). $\{y_t, e_t\}$ is α-mixing with size $-\gamma/(\gamma-2)$ with $\gamma > 2$.

Condition (C.12). $E\,|B_{ji}(y_{t-j})e_t|^{\gamma} < \infty$ uniformly for i and j, where γ is defined in (C.11).

Condition (C.13). $E\,\{B_{ji}(y_{t-j})e_t\} = O(1/\sqrt{n})$ uniformly for i and j.

Conditions (C.11) and (C.12) are frequently used to establish the Central Limit Theorem of estimators under dependent data (e.g., White, 1984). Condition (C.13) implies that $\{e_t\}$ is weakly correlated.

Theorem 5. *Provided that Conditions (C.8) - (C.13) are satisfied, Theorem 1 continues to hold.*

Proof. See the Appendix.

Theorem 6. *Provided that Conditions (C.4), (C.6), (C.8), (C.9), and (C.11) - (C.13) are satisfied, Theorem 2 continues to hold.*

Proof. See the Appendix.

4. Finite sample analysis

In this section, we focus on the comparison of the finite sample properties of the proposed AMA and AMAH estimators with a number of other estimators, including the AIC and BIC model selection estimators and the smoothed-AIC (SAIC) and smoothed-BIC (SBIC) model averaging estimators. The AIC and BIC scores for the m^{th} model are defined as $\mathrm{AIC}^{(m)} = n\mathrm{log}\widetilde{\sigma}_m^2 + 2r_m$ and $\mathrm{BIC}^{(m)} = n\mathrm{log}\widetilde{\sigma}_m^2 + (\mathrm{log}n)r_m$ respectively. The AIC (BIC) estimator selects the model with the smallest AIC (BIC) score. The SAIC estimator is an FMA estimator that assigns the weight

$$w_{\mathrm{AIC},m} = exp(-\mathrm{AIC}^{(m)}/2)/\sum_{m=1}^{M} exp(-\mathrm{AIC}^{(m)}/2)$$

to the m^{th} model, $1 \leq m \leq M$. The SBIC estimator is defined analogously.

4.1 Simulations for the independent data case

We consider the model given in (2.1) with $\mu = (\mu_1, ..., \mu_n)'$, $e \sim N(0, \sigma^2 I_n)$, and $\sigma = 0.4, 1.0, 1.5$. The following process of μ is considered:

$$\begin{aligned} \mu_i &= x_{i1} + (2x_{i2} - 1)^2 + \sin(2\pi x_{i3})/(2 - \sin(2\pi x_{i3})) \\ &\quad + \big\{0.1\sin(2\pi x_{i4}) + 0.2\cos(2\pi x_{i4}) + 0.3(\sin(2\pi x_{i4}))^2 \\ &\quad +0.4(\cos(2\pi x_{i4}))^3 + 0.5(\sin(2\pi x_{i4}))^3\big\} + 0.5\alpha(\sin(2\pi U_i))^2. \quad (4.1) \end{aligned}$$

We let $x_{ij} = (V_{ij} + kU_i)/(1 + k)$ for $j = 1, ..., 7$, where V_{ij} and U_i are i.i.d. U[0,1] observations and $k = 0, 0.5, 1, 1.5, 2$. When $k \neq 0$, x_{ij}'s have common U_i for different j, i.e., $E(x_{i1}x_{i2}) \neq 0$, and the data are correlated. We use the parameter α in (4.1) to control the degree of model misspecification.

We first consider the uncertainty of choice of covariates. With 7 covariates, there are 2^7 candidate models. Note that (4.1) is similar to the simulation setup of Xue (2009), who considered the special case of $\alpha = 0$. In our simulations, we set $\alpha = 1$ so that all candidate models are misspecified. In addition, we set the polynomial degree to 3, and let the knots be equidistant. Following Huang and Yang (2004), we set the number of knots to be the smallest integer greater than or equal to $(2n)^{1/5} - 1$. We consider sample sizes of $n = 50, 70, 100, 150, 200$. Let $\tilde{\mu}$ be an estimator of μ. Our comparison of the performance of estimators is based on the squared error $||\mu - \tilde{\mu}||^2$, averaged over 1000 replications. Table 1 below and Tables S.1 and S.2 in the Supplementary Material present the results. While the following commentary applies to all values of k considered, to conserve space, we only report results corresponding to $k = 0, 1, 1.5$. Results for other values of k are available upon request.

Our results show that the AMA and the AMAH estimators are almost always the two best estimators with respect to averaged squared errors. In

Table 1: Averaged squared errors ($\times 10^{-1}$) under independent data when covariate selection is subject to uncertainty and $\sigma = 0.4$

	n	AIC	BIC	SAIC	SBIC	AMAH	AMA
$k = 0$	50	1.279	1.589	1.217	1.366	1.066	1.272
	70	1.031	1.178	0.976	1.087	0.921	1.024
	100	0.878	0.912	0.840	0.884	0.822	0.860
	150	0.634	0.602	0.615	0.601	0.597	0.605
	200	0.576	0.547	0.560	0.547	0.550	0.552
$k = 1$	50	1.206	1.259	1.128	1.088	0.875	0.970
	70	0.947	1.078	0.869	0.977	0.772	0.830
	100	0.771	0.941	0.719	0.872	0.678	0.710
	150	0.604	0.816	0.575	0.760	0.547	0.564
	200	0.534	0.691	0.514	0.647	0.497	0.504
$k = 1.5$	50	1.190	1.244	1.109	1.066	0.847	0.919
	70	0.921	1.061	0.838	0.951	0.732	0.782
	100	0.758	0.916	0.700	0.843	0.654	0.684
	150	0.605	0.777	0.572	0.727	0.534	0.549
	200	0.528	0.682	0.506	0.644	0.482	0.489

the rare cases where neither AMA nor AMAH produces the best estimates (e.g., when $\sigma = 0.4$, $k = 0$ and $n = 200$), they typically yield larger average squared errors than the best estimator only by a small margin. On the other hand, when they dominate other estimators, they usually do so by a large margin. When $\sigma = 1.5$ (high noise levels), AMA performs better than AMAH, while the converse is observed when $\sigma = 0.4$ (low noise level), and the two estimators exhibit comparable performance when $\sigma = 1$ (moderate noise level). Without exception, the SAIC and SBIC estimators outperform their model selection counterparts, although both the SAIC and SBIC frequently exhibit inferior performance to the proposed AMA and AMAH estimators. In general, the ordinal rankings of estimators are unaffected by the values of k.

We now consider the case where the covariates are certain but the s-moothing degree and number of knots are uncertain. Specifically, we let $\alpha = 0$, and select the degree of smoothing and the number of knots from $\{1, 2, 3\}$ and $\{2, 3, ..., ceiling\left((2n)^{1/5}\right) + 2\}$ respectively, where $ceiling\,(A)$ denotes the smallest integer greater than or equal to A. Thus, there are $3\left\{ceiling\left((2n)^{1/5}\right) + 1\right\}$ candidate models. Tables S.3 - S.5 in the Supplementary Material, where the simulation results are reported, show that in the overwhelming majority of cases, the AMA estimator is the best estimator; either the AMAH or the SBIC estimator produces the second best estimates, and all of the AMA, AMAH and SBIC estimators uniformly dominate the two model selection estimators. Although the SAIC estimator invariably yields more accurate estimates than the AIC estimator, it can have inferior performance to the BIC estimator.

In addition, we have also considered the cases where the error term follows non-normal distributions, such as t distributions, and the results (not reported here) show that the AMA often performs the best and AMAH the second best. As well, as suggested by a referee, we have computed the average weights and selection frequencies of models by each method. The results are reported in Section 2 of the Supplementary Material.

4.2 Simulation for the dependent data case

Our simulation study is based on the following autoregressive process $\{y_t, t = 0, \pm 1, ...\}$:

$$\begin{aligned} y_t &= -0.4(3 - y_{t-1}^2)/(1 + y_{t-1}^2) \\ &\quad +0.6(3 - (y_{t-2} - 0.5)^3)/(1 + (y_{t-2} - 0.5)^4) \\ &\quad +\alpha y_{t-10} + 0.1 e_t, \end{aligned} \tag{4.2}$$

where $e_t = \rho e_{t-1} + \varepsilon_t$, and $\varepsilon_t \sim N(0, 1)$. As in Subsection 4.1, we first consider the uncertainty of choice of covariates. We set y_{t-1}, y_{t-2}, y_{t-3} and

y_{t-4} as potential covariates, and assume that all candidate models are in the form of the additive autoregressive model given in Section 3. This yields 2^4 candidate additive autoregressive models, with $y_t = b + e_t$ being the null model, where b is an intercept. When $\alpha = \rho = 0$, the process (4.2) reduces to the process considered by Huang and Yang (2004). In our simulations, we set $n = 80, 100, 200$ and $\alpha = 0.3, 0.4$ so that all candidate models are misspecified. As in Subsection 4.1, we set the polynomial degree to 3 and let the knots be equidistant. To assess the predictive performance of the methods, we calculate the squared prediction errors $(y_{n+1} - \tilde{y}_{n+1})^2$ averaged over 50000 simulation trials, where $\tilde{y}_{n+1}$ is an estimator of y_{n+1}. Unlike the independent data case in Subsection 4.1 for which the results are based on 1000 replications, a substantially larger number of replication trials are required here in order to obtain stable results.

The results, reported in Table 2 and Table S.6 in the Supplementary Material, show that in the overwhelming majority of cases, the AMA estimator results in the best performance with the AMAH estimator coming in a close second. Generally speaking, the values of α, ρ and n have little effect on the ordinal ranking of estimators but they do have some bearings on actual magnitudes of the squared prediction errors. As expected, as α increases, the prediction squared errors of all estimators increase, *ceteris* *paribus*. For most cases, the AIC (BIC) estimator is dominated by its model averaging counterpart. The BIC (SBIC) estimator is often preferred to the AIC (SAIC) estimator; exceptions occur, for example, when $n = 200$ and ρ is small, the AIC-based estimators can have an edge over their BIC-based counterparts.

We next consider uncertainty in the degree of smoothing within (4.2). We set $\alpha = 0$ and $\rho = 0, 0.2, 0.4$. The simulation results, reported in Table S.7 of the Supplementary Material that are again based on 50000 simulation trials, show that AMA is uniformly the best estimator, followed by either the AMAH or SBIC estimator. The worst estimates are always produced by the AIC estimator.

5. Empirical data applications

5.1 Example 1

In this subsection, we apply the proposed method to the theophylline concentration data as part of the 'Theoph' dataset from the R 'datasets' package. Oral doses of theophylline are given to twelve individuals for eleven times each, resulting in 132 observations. The response variable of interest is the theophylline concentration in the individual, labelled as (CON). The following are the covariates expected to influence CON: the amount of o-

Table 2: Averaged squared prediction errors ($\times 10^{-1}$) under dependent data when covariate selection is subject to uncertainty and $\alpha = 0.3$

	n	AIC	BIC	SAIC	SBIC	AMAH	AMA
$\rho = 0$	80	1.210	1.038	1.170	1.022	1.009	0.978
	100	0.921	0.873	0.903	0.858	0.825	0.819
	200	0.736	0.782	0.735	0.769	0.708	0.707
$\rho = 0.2$	80	1.136	0.997	1.100	0.978	0.953	0.928
	100	0.945	0.893	0.929	0.876	0.842	0.833
	200	0.793	0.840	0.793	0.827	0.758	0.756
$\rho = 0.4$	80	1.417	1.100	1.367	1.076	1.137	1.074
	100	1.034	0.947	1.019	0.932	0.921	0.908
	200	0.901	0.870	0.900	0.862	0.837	0.829

ral doses of theophylline administered to the individual, the time interval from drug administration to the time of sampling, and the weight of the individual. There are 2^3 combinations of these covariates resulting in 2^3 candidate models. The basis of our analysis is the additive model (2.1). We set the polynomial degree to be 3, and let the knots be equidistant with the number of knots equal to $ceiling\left((2n)^{1/5}\right) - 1$.

We consider all of the six estimators in the last section, and an alternative AMA estimator obtained with B^m in $P_m = B^m(B^{m\prime}B^m)^{-1}B^{m\prime}$ replaced by the regressor matrix of the m^{th} model. This essentially reduces the additive model to a linear regression model, and as such, the corresponding model average estimator combines least squares estimators from linear regressions using (2.8) as the weight choice criterion. We label this estimator as AMAli, to distinguish it from the AMA and AMAH estimators that combine additive models. We randomly select $n_1 = 80, 100, 120$ observations from the sample as training data and use the remaining $132 - n_1$ observations as test data. The following mean squared prediction error (MSPE) is used to gauge the performance of the estimators:

$$\frac{1}{(132 - n_1)} \sum_{i=1}^{132-n_1} \left(\text{CON}_i - \widehat{\text{CON}}_i\right)^2,$$

where CON_i and $\widehat{\text{CON}}_i$ are the i^{th} actual and predicted values of CON in the test sample. We repeat the data splitting and estimation process

Table 3: Results for Data Example 1

n_1	AIC	BIC	SAIC	SBIC	AMAH	AMA	AMAli
80	0.458	0.430	0.442	0.430	0.417	0.421	0.953
100	0.413	0.409	0.405	0.410	0.385	0.390	0.942
120	0.386	0.402	0.391	0.403	0.375	0.377	0.941

1000 times and compute the average of MSPEs across the replications. The results are reported in Table 3.

The results show that regardless of the values of n_1, the AMAH and AMA estimators invariably deliver the best and second best estimates. In all cases, the AMAli estimator yields prediction outcomes that are inferior to those of other estimators by a large margin, which suggests that the nonlinear additive models is a more appropriate analytical framework than the linear model for the data in hand.

5.2 Example 2

Our second data example is based on the 'LakeHuron' dataset from the R package 'datasets'. The objective is to forecast the level of Lake Huron, one of the five largest lakes in North America, by an additive autoregressive model. Data are available on the level (in feet) of the lake between 1875

Table 4: Results for Data Example 2

n_1	AIC	BIC	SAIC	SBIC	AMAH	AMA	AMAli
60	1.148	0.699	0.969	0.699	0.691	0.619	0.697
70	1.077	0.779	1.137	0.777	0.741	0.664	0.778
80	0.799	0.575	0.720	0.572	0.551	0.529	0.762

and 1972, totalling 98 annual observations. We label the first difference of this time series as $\{y_t\}_{t=1}^{97}$. We let the maximum lag order be 4. Thus there are 2^4 candidate models, with the largest model being

$$y_t = g_1(y_{t-1}) + g_2(y_{t-2}) + g_3(y_{t-3}) + g_4(y_{t-4}) + e_t.$$

$\{y_t\}_{t=g}^{n_1+k+g}(n_1 = 60, 70, 80, g = 1, ..., D)$ is used for model training and $y_{n_1+k+g+1}$ for prediction evaluation. We choose the same degree of smoothing and number of knots as in Subsection 5.1. We conduct $D = 97 - (n_1 + k + 1)$ one-step ahead predictions with the forecast window being moved ahead by one observation each time. Table 4 presents the mean squared prediction error (MSPE), defined as $\sum_{t=n_1+k+2}^{97} (y_t - \hat{y}_t)^2 / D$, with $\hat{y}_t$ being the one-step ahead model average predictor of y_t.

We consider the seven estimators as in the previous data example. Table 4 shows that in all cases the AMA estimator yields the best predictions, followed by the AMAH estimator, whereas the AIC or SAIC estimators often deliver the worst forecasts. The AMAli estimator invariably performs worse than the AMA estimator; however, occasionally it outperforms the other selection and averaging estimators. This indicates that while the nonlinear additive autogressive model is a more appropriate analytical framework than the linear autoregressive model for this data set, functional form misspecification may be compensated by a superior estimation technique.

6. Concluding remarks

In this paper, we have proposed a plug-in model averaging approach for the nonparametric additive model and the additive autoregressive model, and developed two estimators, the AMA and AMAH estimators. The numerical results from the paper support the use of model averaging in these models, and show that AMA is often a superior alternative to AMAH.

One aspect of model specification that is not questioned throughout our analysis is the assumption of a constant error variance. Extending the weight choice procedure and associated theories to the context of heteroscedastic disturbances is an area for future research. As well, this paper emphasises the development of a weight choice method oriented towards an improvement in efficiency with respect to point estimation. There is clearly a need to develop inference procedures based on the model averaging estimator. In this regard, the asymptotic distributions of some FMA estimators have been derived; see, for example, Hjort and Claeskens (2003), Liu (2015), and Zhang and Liu (2019). The asymptotic distribution theory for our proposed model averaging estimator deserves to be further studied. In addition, it is also interesting to study the adaptive estimation for unknown smoothness by model averaging (see Yang (2001) and Zhang et al. (2013) for related results), which is left for our future research.

Supplementary Material

The Supplementary Material contains additional simulation results (Tables S.1-S.15).

Acknowledgements

We thank the Editor, Associate Editor and two referees for helpful comments and suggestions. This work was partially supported by the National Natural Science Foundation of China (Grant Nos. 12001534, 71973116, 11971323 and 12031016). Wan's work was also supported by a General Research Fund from the Hong Kong Research Grants Council (Grant No. CityU 11500419).

References

Ando, T. and Li, K. C. (2014). A model-averaging approach for high-dimensional regression. *Journal of the American Statistical Association*, 109:254–265.

Belitz, C. and Lang, S. (2008). Simultaneous selection of variables and smoothing parameters in structured additive regression models. *Computational Statistics and Data Analysis*, 53:61–81.

Bickel, P. and Levina, E. (2008). Regularized estimation of large covariance matrices. *Annals of Statistics*, 36:199–227.

Bontemps, C., Simioni, M., and Surry, Y. (2008). Semiparametric hedonic price models: assessing the effects of agricultural nonpoint source pollution. *Journal of Applied Econometrics*, 23:825–842.

Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997). Model selection: an integral part of inference. *Biometrics*, 53:603–618.

Buja, A., Hastie, T., and Tibshirani, R. (1989). Linear smoothers and additive models (with discussion). *Annals of Statistics*, 17:453–555.

Cantoni, E., Mills Flemming, J., and Ronchetti, E. (2011). Variable selection in additive models by nonnegative garrote. *Statistical Modelling*, 11:237–252.

Chen, J., Li, D., Linton, O., and Lu, Z. (2018a). Semiparametric ultra-high dimensional model averaging of nonlinear dynamic time series. *Journal of the American Statistical Association*, 113:919–932.

Chen, R., Liang, H., and Wang, J. (2011). Determination of linear components in additive models. *Journal of Nonparametric Statistics*, 23:367–383.

Chen, R. and Tsay, R. (1993). Nonlinear additive ARX models. *Journal of the American Statistical Association*, 88:955–967.

Chen, Z., Fan, J., and Li, R. (2018b). Error variance estimation in ultrahigh dimensional additive models. *Journal of the American Statistical Association*, 113:315–327.

Cheng, T., Ing, C., and Yu, S. (2015). Toward optimal model averaging in regression models with time series errors. *Journal of Econometrics*, 189:321–334.

Claeskens, G. (2016). Statistical model choice. *Annual Review of Statistics and Its Application*, 3:233–256.

Davydov, Y. A. (1968). Convergence of distributions generated by stationary stochastic processes. *Theory of Probability and Its Applications*, 13:691–696.

de Boor, C. (2001). *A Practical Guide to Splines*. Springer, New York.

Doksum, K. and Koo, J. Y. (2000). On spline estimators and prediction intervals in nonparametric regression. *Computational Statistics and Data Analysis*, 35:76–82.

Eyto, E. and Irvine, K. (2007). Assessing the status of shallow lakes using an additive model of biomass size spectra. *Aquatic Conservation Marine and Freshwater Ecosystems*, 17:724–736.

Fan, J., Feng, Y., and Song, R. (2011). Nonparametric independence screening in sparse ultra-high dimensional additive models. *Journal of the American Statistical Association*, 106:544–557.

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

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics*, 32:928–961.

Gao, Y. (2015). *Model Averaging for Nonlinear Complex Data and Its Applications*. PhD thesis, Chinese Academy of Sciences.

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

Hansen, B. E. (2014). Nonparametric sieve regression: Least squares, averaging least squares, and cross-validation. In Racine, J. S., Su, L., and Ullah, A., editors, *Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics*. Oxford University Press.

Hansen, B. E. and Racine, J. (2012). Jackknife model averaging. *Journal of Econometrics*, 167:38–46.

Härdle, W. and Korostelev, A. (1996). Search for significant variables in nonparametric additive regression. *Biometrika*, 83:541–549.

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

Hjort, N. L. and Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association*, 98:879–899.

Huang, J., Horowitz, J. L., and Wei, F. (2010). Variable selection in nonparametric additive models. *Annals of Statistics*, 38:2282–2313.

Huang, J. Z. and Yang, L. (2004). Identification of non-linear additive autoregressive models. *Journal of the Royal Statistical Society, Series B*, 66:463–477.

Li, C., Li, Q., Racine, J., and Zhang, D. (2018). Optimal model averaging of varying coefficient models. *Statistica Sinica*, 28:2795–2809.

Li, W. and Yang, L. (2007). Spline-backfitted kernel smoothing of nonlinear additive autoregression model. *Annals of Statistics*, 35:2474–2503.

Liang, H., Zou, G., Wan, A. T. K., and Zhang, X. (2011). Optimal weight choice for frequentist model average estimators. *Journal of the American Statistical Association*, 106:1053–1066.

Linton, O. B. and Nielsen, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, 82:93–101.

Liu, C.-A. (2015). Distribution theory of the least squares averaging estimator. *Journal of Econometrics*, 186:142–159.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, 15:661–675.

Mammen, E., Linton, O., and Nielsen, J. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Annals of Statistics*, 27:1443–1490.

Nielsen, J. P. and Sperlich, S. (2005). Smooth backfitting in practice. *Journal of the Royal Statistical Society, Series B*, 67:43–61.

Opsomer, J.-D. (2000). Asymptotic properties of backfitting estimators. *Journal of Multivariate Analysis*, 73:166–179.

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009). Sparse additive models. *Journal of the Royal Statistical Society, Series B*, 71:1009–1030.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Annals of Statistics*, 9:1135–1151.

Stone, C. J. (1985). Additive regression and other nonparametric models. *Annals of Statistics*, 13:689–705.

Tjostheim, D. and Auestad, B. (1994). Nonparametric identification of nonlinear time series: projections. *Journal of the American Statistical Association*, 89:1398–1409.

Wan, A. T. K., Zhang, X., and Zou, G. (2010). Least squares model averaging by Mallows criterion. *Journal of Econometrics*, 156:277–283.

Wang, H., Li, G., and Tsai, C. (2007). Regression coefficient and autoregressive order shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 69:63–78.

Wei, C. Z. (1987). Adaptive prediction by least squares predictors in stochastic regression models with applications to time series. *Annals of Statistics*, 15:1667–1682.

White, H. (1984). *Asymptotic Theory for Econometricians.* Academic Press, Orlando.

Xue, L. (2009). Consistent variable selection in additive models. *Statistica Sinica*, 19:1281–1296.

Yang, Y. (2001). Adaptive regression by mixing. *Journal of the American Statistical Association*, 96:574–588.

Yuan, Z. and Yang, Y. (2005). Combining linear regression models: When and how? *Journal of the American Statistical Association*, 100:1202–1214.

Zhang, X. and Liu, C. (2019). Inference after model averaging in linear regression models. *Econometric Theory*, 35:816–841.

Zhang, X., Lu, Z., and Zou, G. (2013). Adaptively combined forecasting for discrete response time series. *Journal of Econometrics*, 176:80–91.

Zhang, X. and Wang, W. (2019). Optimal model averaging estimation for partially linear models. *Statistica Sinica*, 29:693–718.

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*, 111:1775–1790.

Zhang, X., Zou, G., and Carroll, R. (2015). Model averaging based on Kullback-Leibler distance. *Statistica Sinica*, 25:1583–1598.

Zhu, R., Wan, A. T. K., Zhang, X., and Zou, G. (2019). A Mallows-type model averaging estimator for the varying-coefficient partially linear model. *Journal of the American Statistical Association*, 114:882–892.

Appendices

A.1 Lemma and its proof

Lemma 1. *If* $\{e_t\}$ *is mutually independent,*

$$\sup_{-\infty \le t \le \infty} E\{|e_t|^q\} < \infty \tag{A.1}$$

for some $q \ge 2$, *and*

$$\max_{m\in\{1,\ldots,M\}, j\in s_{c_m}, i\in\{1,\ldots,q_j^m\}} E\left|B_{ji}^q(y_{t-j})\right| < \infty, \tag{A.2}$$

then

$$E\left[\max_{m\in\{1,\ldots,M\}}\left\|\frac{1}{\sqrt{n}}B^{m\prime}e\right\|^q\right] = O(r_M^{q/2}). \tag{A.3}$$

Proof of Lemma 1

Denote the t^{th} column of $B^{m\prime}$ as $B^m_{j\in s_{c_m}}(y_{t-j})$. Note that under Condition (A.2), there exists an $m^* \in \{1,\ldots,M\}$ such that

$$\begin{aligned}
& E\left[\max_{m\in\{1,\ldots,M\}}\left\|\frac{1}{\sqrt{n}}B^{m\prime}e\right\|^q\right] \\
= \; & E\left[\max_{m\in\{1,\ldots,M\}}\left\|\frac{1}{\sqrt{n}}\sum_{t=1}^n B^m_{j\in s_{c_m}}(y_{t-j})e_t\right\|^q\right] \\
= \; & E\left\{\sum_{j\in s_{c_{m^*}}}\sum_{i=1}^{q_j^{m^*}}\left(\frac{1}{\sqrt{n}}\sum_{t=1}^n B_{ji}(y_{t-j})e_t\right)^2\right\}^{q/2} \\
\le \; & r_{m^*}^{q/2-1}\sum_{j\in s_{c_{m^*}}}\sum_{i=1}^{q_j^{m^*}} E\left|\frac{1}{\sqrt{n}}\sum_{t=1}^n B_{ji}(y_{t-j})e_t\right|^q \\
\le \; & r_{m^*}^{q/2-1}\sum_{j\in s_{c_{m^*}}}\sum_{i=1}^{q_j^{m^*}} E\left|\frac{1}{n}\sum_{t=1}^n B^2_{ji}(y_{t-j})\right|^{q/2} \\
\le \; & r_{m^*}^{q/2-1}\sum_{j\in s_{c_{m^*}}}\sum_{i=1}^{q_j^{m^*}} \frac{1}{n}\sum_{t=1}^n E\left|B^q_{ji}(y_{t-j})\right| \\
\le \; & r_{m^*}^{q/2}\max_{j\in s_{c_{m^*}}, i\in\{1,\ldots,q_j^{m^*}\}} E\left|B^q_{ji}(y_{t-j})\right|
\end{aligned}$$

$$= \quad O(r_M^{q/2}), \tag{A.4}$$

where the second inequality (on the fifth line of (A.4)) follows Condition (A.1) and Lemma 2 of Wei (1987) since $\{e_t, \varpi_t\}$ is a sequence of martingale differences with ϖ_t being the σ-algebra generated by $\{y_t, y_{t-1}, ...\}$. This completes the proof of Lemma 1.

A.2 Proof of Theorem 1

Note that

$$\phi_H(w) = \|Y - \hat{\mu}(w)\|^2 + 2\widehat{\sigma}_M^2 \sum_{m=1}^M w_m r_m. \tag{A.5}$$

Hence we have

$$\phi(w) \quad = \quad \phi_H(w) + 2\widehat{\sigma}_M^2 \sum_{m=1}^M w_m \left(\frac{nr_m}{n - r_m} - r_m \right). \tag{A.6}$$

From results of Wan et al. (2010), to prove (2.13) of Theorem 1, it suffices to show that

$$\sup_{w \in \mathcal{W}} \left[R^{-1}(w) \left| \phi_H(w) - R(w) \right| \right] = o_p(1) \tag{A.7}$$

and

$$\sup_{w \in \mathcal{W}} \left[R^{-1}(w) \left| \widehat{\sigma}_M^2 \sum_{m=1}^M w_m \left(\frac{nr_m}{n - r_m} - r_m \right) \right| \right] = o_p(1). \tag{A.8}$$

Note that (A.7) can be verified from the proof of Theorem 2 in Wan et al. (2010). Now, let us consider (A.8). First, from Condition (C.2) and $E\|e\|^2 = n\sigma^2 = O(n)$, we have

$$\|Y\| \leq \|\mu\| + \|e\| = O_p(n^{1/2}). \tag{A.9}$$

Hence, by Condition (C.3),

$$\widehat{\sigma}_M^2 = Y'(I_n - P_M)Y/(n - r_M) \leq \lambda_{\max}(I_n - P_M)\|Y\|^2/(n - r_M) = O_p(1). \tag{A.10}$$

Furthermore, by Condition (C.1),

$$\sup_{w \in \mathcal{W}} \left[R^{-1}(w) \left| \widehat{\sigma}_M^2 \sum_{m=1}^M w_m \left\{ \frac{nr_m}{n - r_m} - r_m \right\} \right| \right]$$

$$\leq \quad \xi_n^{-1}\widehat{\sigma}_M^2 \frac{r_M^2}{n-r_M} \to 0. \tag{A.11}$$

Therefore, (2.13) of Theorem 1 is true. In addition, by Wan et al. (2010) and (A.7), it is straightforward to show that (2.14) of Theorem 1 holds under Conditions (C.1) - (C.3). This completes the proof of Theorem 1.

A.3 Proof of Theorem 2

Denote $\epsilon_n = \xi_n^{1/2} n^{-1/2+\delta}$. By results of Fan and Peng (2004) and Chen et al. (2018a), to prove (2.15) of Theorem 2, it suffices to show that there exists a constant C_0 such that for the $M \times 1$ vector $u = (u_1, ..., u_M)'$,

$$\lim_{n\to\infty} P\left(\inf_{\|u\|=C_0, (w^0+\epsilon_n u)\in\mathcal{W}} \phi(w^0+\epsilon_n u) > \phi(w^0) \right) = 1. \tag{A.12}$$

This means that there exists a minimiser $\hat{w}$ in the set $\{w^0 + \epsilon_n u : \|u\| \leq C_0, (w^0 + \epsilon_n u) \in \mathcal{W}\}$ such that $\|\hat{w} - w^0\| = O_p(\epsilon_n)$.

Denote $\Omega_1 = (\mu - \hat{\mu}_1, ..., \mu - \hat{\mu}_M)$. Let $\bar{\pi} = (\pi_1, \cdots, \pi_M)'$ with $\pi_m = nr_m/(n - r_m)$ for $1 \leq m \leq M$. It is noted that

$$\begin{aligned} & \phi(w^0+\epsilon_n u) - \phi(w^0) \\ = \;& \epsilon_n^2 u'\Lambda u - 2\epsilon_n w^{0\prime}\Omega_1'\Lambda_1 u - 2e'P(\epsilon_n u)\mu - 2e'P(\epsilon_n u)e + 2\epsilon_n \widehat{\sigma}_M^2 u'\bar{\pi}. \end{aligned} \tag{A.13}$$

As $\lambda_{\min}(\Lambda/n) > \kappa_1$ under Condition (C.4), we have

$$\epsilon_n^2 u'\Lambda u > \kappa_1 n\epsilon_n^2 \|u\|^2 > 0, \tag{A.14}$$

with probability approaching 1.

Recognising that $\|\Omega_1 w^0\| = O_p(\xi_n^{1/2})$ since $E\|\Omega_1 w^0\|^2 = E\|\mu - \hat{\mu}(w^0)\|^2 = \xi_n$, and $\|\Lambda_1\| = \lambda_{\max}^{1/2}(\Lambda) = O_p(n^{1/2})$ by Condition (C.4), we have

$$\begin{aligned} |\epsilon_n w^{0\prime}\Omega_1'\Lambda_1 u| &\leq \epsilon_n \|\Lambda_1\| \|\Omega_1 w^0\| \|u\| \\ &= O_p(n^{1/2}\xi_n^{1/2}\epsilon_n)\|u\|. \end{aligned} \tag{A.15}$$

Hence, $\epsilon_n w^{0\prime}\Omega_1'\Lambda_1 u$ is dominated asymptotically by $\epsilon_n^2 u'\Lambda u$.

From Condition (C.5) and Lemma 1 for independent data cases,

$$\begin{aligned}
& \max_{m\in\{1,\ldots,M\}} e'P_m e \\
= & \max_{m\in\{1,\ldots,M\}} \left\{ e'B^m (B^{m\prime} B^m)^{-1} B^{m\prime} e \right\} \\
\leq & \max_{m\in\{1,\ldots,M\}} \lambda_{\max} \left\{ (B^{m\prime} B^m/n)^{-1} \right\} \max_{m\in\{1,\ldots,M\}} \left\| \frac{1}{\sqrt{n}} B^{m\prime} e \right\|^2 \\
= & O_p(r_M), \qquad (A.16)
\end{aligned}$$

which, together with (A.9), implies that

$$\max_{m\in\{1,\ldots,M\}} |e'P_m Y| \quad \leq \quad \|Y\| \max_{m\in\{1,\ldots,M\}} \left(e'P_m e\right)^{1/2} = O_p(n^{1/2} r_M^{1/2}).$$

Hence,

$$\begin{aligned}
|e'P(\epsilon_n u)\mu + e'P(\epsilon_n u)e| &= |e'P(\epsilon_n u)Y| \\
&\leq \epsilon_n \|u\| \left(M \max_{m\in\{1,\ldots,M\}} |e'P_m Y|^2 \right)^{1/2} \\
&\leq O_p(n^{1/2} r_M^{1/2} M^{1/2} \epsilon_n) \|u\| . \qquad (A.17)
\end{aligned}$$

Using Condition (C.6), we have

$$\frac{n^{1/2} r_M^{1/2} M^{1/2} \epsilon_n}{n\epsilon_n^2} = \frac{n^{1/2} r_M^{1/2} M^{1/2}}{n n^{-1/2+\delta} \xi_n^{1/2}} = \frac{r_M^{1/2} M^{1/2}}{n^{\delta} \xi_n^{1/2}} = o(1), \qquad (A.18)$$

and using (A.10),

$$\begin{aligned}
|\epsilon_n \widehat{\sigma}_M^2 u'\bar{\pi}| &\leq \epsilon_n \widehat{\sigma}_M^2 \|(\pi_1, \cdots, \pi_M)'\| \|u\| \\
&= \epsilon_n \widehat{\sigma}_M^2 \left(\sum_{m=1}^{M} (\pi_m)^2 \right)^{1/2} \|u\| \\
&= O_p\left(\frac{n r_M M^{1/2}}{n - r_M} \epsilon_n \right) \|u\| = O_p\left(r_M M^{1/2} \epsilon_n \right) \|u\| . \qquad (A.19)
\end{aligned}$$

From (A.17), (A.19) and the first part of Condition (C.6), it is readily seen that $|\epsilon_n \widehat{\sigma}_M^2 u'\bar{\pi}|$ is dominated by $|e'P(\epsilon_n u)Y|$, and from (A.18), both of these terms are dominated asymptotically by $\epsilon_n^2 u'\Lambda u$. Thus, (2.15) of Theorem 2 is proved. Also, we see that (2.16) of Theorem 2 is true using the same proving steps. This completes the proof of Theorem 2.

A.4 Proof of Theorem 3

First, we consider (2.13). Note that

$$\begin{aligned}
\phi(w) &= \|Y-\hat{\mu}(w)\|^2+2\widehat{\sigma}_M^2 w'\bar{\pi} \\
&= \|\mu+e-\hat{\mu}(w)\|^2+2\widehat{\sigma}_M^2 w'\bar{\pi} \\
&= \|\mu-\hat{\mu}(w)\|^2+2e'(\mu-P(w)\mu-P(w)e)+\|e\|^2+2\widehat{\sigma}_M^2 w'\bar{\pi} \\
&= L(w)-2e'P(w)\mu-2e'P(w)e+\|e\|^2+2\mu'e+2\widehat{\sigma}_M^2 w'\bar{\pi}. \qquad \text{(A.20)}
\end{aligned}$$

To prove (2.13), in light of the results of Wan et al. (2010), it suffices to show that

$$\sup_{w\in\mathcal{W}}\left[\tilde{R}^{-1}(w)\left|e'P(w)e\right|\right]=o_p(1), \qquad \text{(A.21)}$$

$$\sup_{w\in\mathcal{W}}\left[\tilde{R}^{-1}(w)\left|e'P(w)\mu\right|\right]=o_p(1), \qquad \text{(A.22)}$$

$$\sup_{w\in\mathcal{W}}\left[\tilde{R}^{-1}(w)\left|\widehat{\sigma}_M^2 w'\bar{\pi}\right|\right]=o_p(1), \qquad \text{(A.23)}$$

and

$$\sup_{w\in\mathcal{W}}\left[\tilde{R}^{-1}(w)\left|L(w)-\tilde{R}(w)\right|\right]=o_p(1). \qquad \text{(A.24)}$$

Let us consider (A.21). Note that

$$\begin{aligned}
&\sup_{w\in\mathcal{W}}\left[\tilde{R}^{-1}(w)\left|e'P(w)e\right|\right] \\
\le\ & \tilde{\xi}_n^{-1}\max_{m\in\{1,\ldots,M\}} e'P_m e \\
=\ & \tilde{\xi}_n^{-1}\max_{m\in\{1,\ldots,M\}}\left\{e'B^m(B^{m\prime}B^m)^{-1}B^{m\prime}e\right\} \\
\le\ & \tilde{\xi}_n^{-1}\max_{m\in\{1,\ldots,M\}}\lambda_{\max}\left\{(B^{m\prime}B^m/n)^{-1}\right\}\max_{m\in\{1,\ldots,M\}}\left\|\frac{1}{\sqrt{n}}B^{m\prime}e\right\|^2 \\
=\ & O_p(r_M\tilde{\xi}_n^{-1}). \qquad \text{(A.25)}
\end{aligned}$$

The last line of (A.25) is due to Conditions (C.7) and (C.8) and Lemma 1. Hence (A.21) is true under Condition (C.10). Similarly, we observe that

$$\sup_{w\in\mathcal{W}}\left[\tilde{R}^{-1}(w)\left|e'P(w)\mu\right|\right]$$

$$\begin{aligned}
&\leq \tilde{\xi}_n^{-1} \max_{m\in\{1,\ldots,M\}} \left(e' P_m \mu\mu' P_m e\right)^{1/2} \\
&\leq \|\mu\| \tilde{\xi}_n^{-1} \max_{m\in\{1,\ldots,M\}} \left(e' P_m e\right)^{1/2} \\
&= O_p(r_M^{1/2} n^{1/2} \tilde{\xi}_n^{-1}), \qquad \text{(A.26)}
\end{aligned}$$

where the equality on the last line of (A.26) is due to (A.25) and Condition (C.9). Hence (A.22) is also true by virtue of (A.26) and Condition (C.10).

From (A.9) and (A.10), we can see that (A.23) is valid because

$$\begin{aligned}
&\sup_{w\in\mathcal{W}} \left[\tilde{R}^{-1}(w) \left|\widehat{\sigma}_M^2 w'\bar{\pi}\right|\right] \\
\leq\ & \tilde{\xi}_n^{-1} \widehat{\sigma}_M^2 \max_{m\in\{1,\ldots,M\}} n r_m/(n - r_m) \\
=\ & O_p(r_M \tilde{\xi}_n^{-1}) = o_p(1) \qquad \text{(A.27)}
\end{aligned}$$

by virtue of Condition (C.10).

Now, note that

$$\begin{aligned}
L(w) - \tilde{R}(w) &= \|\mu - P(w)\mu - P(w)e\|^2 - \tilde{R}(w) \\
&= \|A(w)\mu - P(w)e\|^2 - \tilde{R}(w) \\
&= e'P^2(w)e - 2\mu' A(w)P(w)e - \sigma^2 tr\left\{P^2(w)\right\}. \qquad \text{(A.28)}
\end{aligned}$$

Furthermore, we see from (A.25) that

$$\begin{aligned}
&\sup_{w\in\mathcal{W}} \left[\tilde{R}^{-1}(w) e' P^2(w) e\right] \\
\leq\ & \sup_{w\in\mathcal{W}} \lambda_{\max}\left\{P(w)\right\} \sup_{w\in\mathcal{W}} \left[\tilde{R}^{-1}(w) e' P(w) e\right] \\
=\ & \left\{\max_{m\in\{1,\ldots,M\}} \lambda_{\max}\left(P_m\right)\right\} \sup_{w\in\mathcal{W}} \left[\tilde{R}^{-1}(w) e' P(w) e\right] \\
\leq\ & \sup_{w\in\mathcal{W}} \left[\tilde{R}^{-1}(w) e' P(w) e\right] = O_p(r_M \tilde{\xi}_n^{-1}), \qquad \text{(A.29)}
\end{aligned}$$

$$\begin{aligned}
&\sup_{w\in\mathcal{W}} \left[\tilde{R}^{-1}(w) \left|\mu' A(w) P(w) e\right|\right] \\
\leq\ & \tilde{\xi}_n^{-1/2} \sup_{w\in\mathcal{W}} \left[\tilde{R}^{-1}(w) \left|e' P(w) A(w) \mu\mu' A(w) P(w) e\right|\right]^{1/2} \\
=\ & \|\mu\| \tilde{\xi}_n^{-1/2} \sup_{w\in\mathcal{W}} \left[\lambda_{\max}^{1/2}\left\{P(w)\right\} \lambda_{\max}\left\{A(w)\right\}\right] \sup_{w\in\mathcal{W}} \left[\tilde{R}^{-1}(w) e' P(w) e\right]^{1/2}
\end{aligned}$$

$$
\begin{aligned}
&= O(n^{1/2}\tilde{\xi}_n^{-1/2}) \sup_{w\in\mathcal{W}} \left[\tilde{R}^{-1}(w)e'P(w)e\right]^{1/2} \\
&= O_p(n^{1/2}\tilde{\xi}_n^{-1/2}r_M^{1/2}\tilde{\xi}_n^{-1/2}) = O_p(n^{1/2}r_M^{1/2}\tilde{\xi}_n^{-1}),
\end{aligned} \tag{A.30}
$$

and

$$
\begin{aligned}
&\sup_{w\in\mathcal{W}} \left[\tilde{R}^{-1}(w)tr\left\{P^2(w)\right\}\right] \\
\leq\ & \tilde{\xi}_n^{-1} \max_{m,l\in\{1,\ldots,M\}} tr\left(P_m P_l\right) \\
\leq\ & \tilde{\xi}_n^{-1} \max_{m,l\in\{1,\ldots,M\}} \left\{\lambda_{\max}\left(P_m P_l\right) \operatorname{rank}\left(P_m P_l\right)\right\} \\
\leq\ & \tilde{\xi}_n^{-1} \max_{m,l\in\{1,\ldots,M\}} \left\{\lambda_{\max}\left(P_m\right) \lambda_{\max}\left(P_l\right) \operatorname{rank}\left(P_m\right)\right\} \\
=\ & O_p(r_M\tilde{\xi}_n^{-1}).
\end{aligned} \tag{A.31}
$$

Together with Condition (C.10), these results imply that (A.24) and hence (2.13) are correct. Following the above proving steps, it is readily seen that (2.14) is also true. This completes the proof of Theorem 3.

A.5 Proof of Theorem 4

By Condition (C.7), (A.3) holds for $q = 2$. The remaining steps for proving Theorem 4 are nearly identical to those for proving Theorem 2, and thus are omitted for brevity.

A.6 Proof of Theorem 5

To prove Theorem 5, we first show that

$$
\frac{1}{\sqrt{n}}\sum_{t=1}^{n} B_{ji}(y_{t-j})e_t = O_p(1), \tag{A.32}
$$

uniformly for i and j. Note that

$$
\begin{aligned}
&\left|\frac{1}{\sqrt{n}}\sum_{t=1}^{n} B_{ji}(y_{t-j})e_t\right| \\
\leq\ & \left|\frac{1}{\sqrt{n}}\sum_{t=1}^{n} \left[B_{ji}(y_{t-j})e_t - E\left\{B_{ji}(y_{t-j})e_t\right\}\right]\right| + \left|\sqrt{n}E\left\{B_{ji}(y_{t-j})e_t\right\}\right|.
\end{aligned} \tag{A.33}
$$

By Condition (C.13), in order to prove (A.32), it suffices to verify that

$$Var\left(\frac{1}{\sqrt{n}}\sum_{t=1}^{n} B_{ji}(y_{t-j})e_t\right) \le C, \tag{A.34}$$

uniformly for i and j. The proof of (A.34) is similar to that of Gao (2015). We first write

$$\begin{aligned} & Var\left(\frac{1}{\sqrt{n}}\sum_{t=1}^{n} B_{ji}(y_{t-j})e_t\right) \\ = \; & \frac{1}{n}\sum_{t=1}^{n} Var\left(B_{ji}(y_{t-j})e_t\right) \\ & +\frac{2}{n}\sum_{t=1}^{n-1}\sum_{s=1}^{n-t} Cov\left(B_{ji}(y_{t-j})e_t, B_{ji}(y_{t-j+s})e_{t+s}\right). \end{aligned} \tag{A.35}$$

Since $\{y_t, e_t\}$ is α-mixing with size $-\gamma/(\gamma-2)$, $\{B_{ji}(y_{t-j})e_t\}$ is also α-mixing with the same size (White, 1984). From results by Davydov (1968) and Gao (2015), we have

$$\begin{aligned} & \left|Cov\left(B_{ji}(y_{t-j})e_t, B_{ji}(y_{t-j+s})e_{t+s}\right)\right| \\ \le \; & 12\left[E\left|B_{ji}(y_{t-j})e_t\right|^{\gamma}\right]^{1/\gamma}\left[E\left|B_{ji}(y_{t-j+s})e_{t+s}\right|^{\gamma}\right]^{1/\gamma}\alpha(s)^{1-2/\gamma} \\ \le \; & C\alpha(s)^{1-2/\gamma}, \end{aligned} \tag{A.36}$$

uniformly for i and j under Condition (C.12), where the mixing coefficient $\alpha(s) = O(s^{-\gamma/(\gamma-2)-\delta})$ with $\delta > 0$ by Condition (C.11). Therefore,

$$\begin{aligned} & \sum_{s=1}^{n-t}\left|Cov\left(B_{ji}(y_{t-j})e_t, B_{ji}(y_{t-j+s})e_{t+s}\right)\right| \\ \le \; & C\sum_{s=1}^{\infty} s^{-1-\delta(\gamma-2)/\gamma} \le C, \end{aligned} \tag{A.37}$$

implying that the second term on the right-hand side of (A.35) is bounded. In addition, the boundedness of the first term on the right-hand side of (A.35) is implied by Condition (C.12). Hence, (A.34) and therefore (A.32) are true.

In light of (A.4) and (A.32), we have

$$\begin{aligned} & \max_{m\in\{1,\ldots,M\}}\left\|\frac{1}{\sqrt{n}}B^{m\prime}e\right\|^2 \\ = \; & \max_{m\in\{1,\ldots,M\}}\sum_{j\in s_{c_m}}\sum_{i=1}^{q_j^m}\left|\frac{1}{\sqrt{n}}\sum_{t=1}^{n} B_{ji}(y_{t-j})e_t\right|^2 \\ = \; & O_p(r_M). \end{aligned} \tag{A.38}$$

Together with the steps for proving Theorem 3, this implies that Theorem 5 is true.

A.7 Proof of Theorem 6

Theorem 6 follows from (A.38) and the proof of Theorem 2. The details are omitted for brevity.