OPTIMAL MODEL AVERAGING ESTIMATION FOR PARTIALLY LINEAR MODELS

Xinyu Zhang[1] and Wendun Wang[2]

[1]*Chinese Academy of Sciences, and Capital University of Economics and Business*

[2]*Econometric Institute, Erasmus University Rotterdam, and Tinbergen Institute*

Abstract: This article studies optimal model averaging for partially linear models with heteroscedasticity. A Mallows-type criterion is proposed to choose the weight. The resulting model averaging estimator is proved to be asymptotically optimal under some regularity conditions. Simulation experiments show that the proposed model averaging method is superior to other commonly used model selection and averaging methods. The proposed procedure is further applied to study Japan's sovereign credit default swap spreads.

Key words and phrases: Asymptotic optimality, Heteroscedasticity, Model averaging, Partially linear model

1. Introduction

Linear regression models have been predominantly popular in a variety of applications, including biology, economics, psychology, and machine learning. One important reason may be its simplicity and the clear interpretation of the estimation results. However, an increasing number of studies have noted that the relationship between the response variable and covariates is not always linear. To list a few examples, Barro (1996) found that democracy can influence economic development in a nonlinear pattern. Henderson et al. (2012) and Su & Lu (2013) found a nonlinear effect of initial state on the economic growth rate. Liang et al.

(2007), in a study on the effectiveness of antiretroviral medicines, showed that the HIV viral load depends nonlinearly on treatment time. Ignoring nonlinearity can result in incorrect estimates and inferences, further resulting in misleading explanations and bad decisions. For example, ignoring the nonlinear effect of global stock markets on the local market may lead to a lack of awareness of financial contagion; Simply estimating a linear relationship between inflation and economic growth may lead to inappropriate inflation-targeting policies.

To avoid potential ignorance of nonlinearity, partially linear models (PLMs) have received extensive attention in theoretical and applied statistics due to their flexible specification. It allows for both linear and nonparametric relations between covariates and the response variable. This type of specification is also frequently used when the primary interest is in the linear component, whereas the relation between the mean response and additional covariates is not easily parameterized. The superiority of the partially linear model over the standard linear models is that it does not require the parametric assumption for all covariates and allows us to capture potential nonlinear effects. This model is sometimes preferred over the fully nonparametric models since it preserves the advantages of linear models, e.g., an easy interpretation of the linear covariates, and suffering less from the dimensionality curse. PLMs are used in a wide range of applications in the literature; see, for example, Engle et al. (1986) for an economic application and Liang et al. (2007) for a medical application.

Various methods have been proposed to estimate PLMs, for example, smoothing splines (Engle et al., 1986; Heckman, 1986), kernel smoothing (Robinson, 1988; Speckman, 1988), local polynomial estimation (Hamilton & Truong, 1997), and penalized splines (Ruppert et al., 2003). See Härdle et al. (2000) for a comprehensive survey. These estimation methods are all based on the assumption that a correctly specified model is given. In practice, however,

researchers are ignorant of the true model. One needs to decide which covariates are in the model (covariate uncertainty), and further whether to assign a covariate in the linear or nonparametric component given that it is in the model (structure uncertainty). The specification of covariates and the model structure is of fundamental importance as it greatly influences the estimation and prediction results. These two types of uncertainty are generally referred to as model uncertainty.

Typical methods to address model uncertainty involve testing and/or selecting the best model using data-driven approaches. The most popular method may be to use an information criterion (IC), such as the Akaike information criterion (AIC) or Bayesian information criterion (BIC). To decide which variables to include in the PLMs, Ni et al. (2009), Bunea (2004), and Xie & Huang (2009), among others, have proposed several variable selection methods. To further determine the structure of the model (which covariates to include in the (non)linear function), a commonly used method is to test the linear null hypotheses against nonlinear alternatives for each covariate. Such tests, however, often have low power when the number of covariates is large (Zhang et al., 2011). In addition, these testing and selection methods perform model selection and estimation in two separate steps. Thus the uncertainty in the model selection procedure is ignored in the estimation step, making it difficult to study the properties of the final estimator (Danilov & Magnus, 2004; Magnus et al., 2016). Zhang et al. (2011) provided a model selection approach based on smoothing spline ANOVA to automatically and consistently distinguish linear and nonlinear component. This method is useful if the goal is to identify the correct model structure, but if the research purpose is to estimate the parameters or to make predictions, it seems more plausible to take into account all (potentially) useful models. However, the model selection approaches can be rather "risky"

since they require "putting all our inferential eggs in one unevenly woven basket" (Longford, 2005).

In this paper we follow a different approach. Instead of selecting one model, we address model uncertainty by appropriately averaging the estimates from different models. As an alternative to model selection, model averaging can substantially reduce risk (Hansen, 2014). It is an integrated process that accounts for both the model uncertainty and estimation uncertainty. Model averaging has long been a popular approach within the Bayesian paradigm; see, for example, Hoeting et al. (1999) for a comprehensive review. In recent years, optimal model averaging methods have been actively developed, for instance, Mallows model averaging (Hansen, 2007), OPT method (Liang et al., 2011), jackknife model averaging (JMA) (Hansen & Racine, 2012), heteroskedasticity-robust model averaging (Liu & Okui, 2013), optimal averaging method for linear mixed-effects models (Zhang et al., 2014), and optimal averaging quantile estimators Lu & Su (2015). These methods are asymptotically optimal in the sense that they minimize the predictive squared error in the large sample case, but they mainly focus on the linear models. To the best of our knowledge, there are no optimal model averaging estimators for PLMs. The main purpose of this paper is to fill this gap.

Our model averaging approach can simultaneously incorporate the covariate and structure uncertainty in PLMs, which is not much studied in the PLM literature. Heteroscedastic random errors are also allowed. To show the optimality of our method, we first assume that the covariance matrix of errors is known, and propose a Mallows-type weight choice criterion, which is an unbiased estimator of the expected predictive squared error up to a constant. We prove that the weights obtained by minimizing this criterion are asymptotically optimal under some regularity conditions. Next, we replace the unknown covariance matrix with its

estimated counterpart, and show that the plugged-in criterion still leads to asymptotically optimal weights.

One may naturally formulate this study as an extension of linear regression model averaging. However, we emphasize that such an extension is by no means straightforward and routine because the existing methods, such as Mallows model averaging, typically do not involve kernel smoothing. To the best of our knowledge, our work is the first to study the optimal averaging that involves kernels. One of our main technical contributions is to provide an optimal weight choice in a kernel smoothing framework.

Our work is also related to Xu et al. (2014), which considered frequentist model averaging and post-model-selection inference in an additive partially linear model. Under the local misspecification setup, their averaging estimator is consistent but *may not be* optimal. We differ from this study by relaxing the local misspecification assumption, thereby allowing all candidate models to be possibly misspecified, and we study the optimal averaging estimator. Moreover, they focus on parameter estimation, while we are interested in prediction. Another related work is Zhao et al. (2016), which modeled massive heterogeneous data in a partially linear framework. To estimate the commonality parameter, they proposed to average commonality estimators obtained from heterogeneous sub-populations. While the averaging idea is similar, our candidate estimators are obtained from the same sample but different models, whereas theirs are from the same model but different sub-populations.

We compare the proposed model averaging estimator with popular model selection and averaging estimators for PLMs. Our simulation study considers two cases. In the first case, only the linear component is uncertain, and the candidate models differ in their inclusion of linear variables. In addition to linear component uncertainty, the second case considers

the situation where there is also uncertainty in choosing which covariates to include in the (non)linear function. In both cases, the proposed estimator performs best in most of the cases, especially when R^2 is moderate and low. Only when R^2 is particularly high, our model averaging estimator is not as good as information-criterion-based methods in the second case. We also apply our method to study Japan's sovereign credit default swap spreads. We find that allowing for nonlinearity indeed provides several new insights. For example, the effect of the global stock market performance on the local market is strengthened in the volatile period, suggesting the existence of financial contagion. The out-of-sample prediction exercise further illustrates the advantage of partially linear models over the linear models, and we generally find better prediction performance for our estimator compared to other partially linear model estimators.

The remainder of this paper is organized as follows. Section 2 introduces our model averaging estimator and presents its asymptotic optimality. Section 3 investigates the finite sample performance of the proposed estimator. A real data example is studied in Section 4 and Section 5 provides some concluding remarks. Technical proofs, additional simulation results and additional tables and figure for the real data example can be found in our online supplement.

2. Model averaging estimation

2.1. Model and estimators

We consider the partially linear model (PLM)

$$y_i = \sum_{j=1}^{\infty} x_{ij}\beta_j + g(\mathbf{Z}_i) + \epsilon_i, \quad i = 1, \ldots, n, \tag{2.1}$$

where $\mathbf{X}_i = (x_{i1}, x_{i2}, \ldots)$ is a countably infinite random vector, $\mathbf{Z}_i = (z_{i1}, \ldots, z_{iq})^{\mathrm{T}}$ is a

random vector in some bounded domain $\mathcal{D} \subset \mathbb{R}^q$, $g(\cdot)$ is an unknown function from $\mathbb{R}^p$ to $\mathbb{R}^1$, and $\epsilon_1, \ldots, \epsilon_n$ are heteroscedastic random errors with $\mathrm{E}(\epsilon_i|\mathbf{X}_i, \mathbf{Z}_i) = 0$ and $\mathrm{E}(\epsilon_i^2|\mathbf{X}_i, \mathbf{Z}_i) = \sigma_i^2$. We denote the expectation of the response variable as $\mu_i = \mathrm{E}(y_i|\mathbf{X}_i, \mathbf{Z}_i) = \sum_{j=1}^{\infty} x_{ij}\beta_j + g(\mathbf{Z}_i)$.

Our goal is to estimate μ_i, which is of particular use for prediction, and this is also the typical goal in the optimal model averaging literature (e.g., Hansen, 2007; Lu and Su, 2015). For this purpose, we use S_n candidate PLMs to approximate (2.1), where S_n is allowed to diverge to infinity as $n \to \infty$. The s^{th} approximation (or candidate) PLM is

$$y_i = \mathbf{X}_{(s),i}^{\mathrm{T}}\boldsymbol{\beta}_{(s)} + g_{(s)}(\mathbf{Z}_{(s),i}) + b_{(s),i} + \epsilon_i, \quad i = 1 \ldots, n \tag{2.2}$$

where $\mathbf{X}_{(s),i}$ is a vector in the linear component, $\mathbf{Z}_{(s),i}$ is a vector in the nonparametric component, $g_s(\cdot)$ is an unknown function from $\mathbb{R}^{q_s}$ to $\mathbb{R}^1$, and $b_{(s),i} = \mu_i - \mathbf{X}_{(s),i}^{\mathrm{T}}\boldsymbol{\beta}_{(s)} - g_{(s)}(\mathbf{Z}_{(s),i})$ represents the approximation error in the s^{th} model. Here, the linear component $\mathbf{X}_{(s),i}$ is allowed to contain variables in $\mathbf{Z}_i$, and reversely the nonparametric covariate $\mathbf{Z}_{(s),i}$ could contain variables in $\mathbf{X}_i$. Hence, (2.2) permits two sources of uncertainty: the uncertainty of which variables to include in the model and the uncertainty of whether a covariate should be in the linear or nonparametric component given that it is in the model, i.e., the variables in the two components may mutually exchange. See, for example, the second case in Section 3. Let $\mathbf{X}_{(s)} = (\mathbf{X}_{(s),1}, \ldots, \mathbf{X}_{(s),n})^{\mathrm{T}}$, $\mathbf{Z}_{(s)} = (\mathbf{Z}_{(s),1}, \ldots, \mathbf{Z}_{(s),n})^{\mathrm{T}}$, $\mathbf{g}_{(s)} = \{g(\mathbf{Z}_{(s),1}), \ldots, g(\mathbf{Z}_{(s),n})\}^{\mathrm{T}}$, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^{\mathrm{T}}$, $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$, and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^{\mathrm{T}}$.

Remark 1. Since estimating the coefficients of the linear component and the non-parametric component is *not* the purpose of this paper, we do not need the conditions for consistency or asymptotic normality of the coefficient estimates, for example, the conditions in Section 1.3

of Härdle et al. (2000).

To provide an optimal weighting scheme, we first need to estimate each candidate model. We follow Speckman (1988) and use kernel smoothing estimation. One of the advantages of this method is its light computational burden, which is crucial in our case since the number of candidate models is typically substantial. To define Speckman's (1988) estimator, let $k(\cdot)$ be a kernel function, h_s be a bandwidth, and $k_{h_s}(\cdot) = k(\cdot/h_s)/h_s$. Furthermore, denote $\mathbf{K}_{(s)} = \{K_{(s),ij}\}$ as an $n \times n$ smoother matrix with $K_{(s),ij} = k_{h_s}(\mathbf{Z}_{(s),i} - \mathbf{Z}_{(s),j})/\sum_{j^*=1}^{n} k_{h_s}(\mathbf{Z}_{(s),i} - \mathbf{Z}_{(s),j^*})$. The kernel smoothing estimator of $\boldsymbol{\beta}_{(s)}$ and $\mathbf{g}_{(s)}$ can then be obtained by $\widehat{\boldsymbol{\beta}}_{(s)} = (\widetilde{\mathbf{X}}_{(s)}^{\mathrm{T}}\widetilde{\mathbf{X}}_{(s)})^{-1}\widetilde{\mathbf{X}}_{(s)}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{K}_{(s)})\mathbf{y}$ and $\widehat{\mathbf{g}}_{(s)} = \mathbf{K}_{(s)}(\mathbf{y} - \mathbf{X}_{(s)}\widehat{\boldsymbol{\beta}}_{(s)})$, where $\widetilde{\mathbf{X}}_{(s)} = (\mathbf{I}_n - \mathbf{K}_{(s)})\mathbf{X}_{(s)}$ and $\mathbf{I}_n$ is an $n \times n$ identity matrix. The estimator of $\boldsymbol{\mu}$ is then $\widehat{\boldsymbol{\mu}}_{(s)} = \mathbf{X}_{(s)}\widehat{\boldsymbol{\beta}}_{(s)} + \widehat{\mathbf{g}}_{(s)} = \widetilde{\mathbf{X}}_{(s)}(\widetilde{\mathbf{X}}_{(s)}^{\mathrm{T}}\widetilde{\mathbf{X}}_{(s)})^{-1}\widetilde{\mathbf{X}}_{(s)}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{K}_{(s)})\mathbf{y} + \mathbf{K}_{(s)}\mathbf{y}$. Letting $\widetilde{\mathbf{P}}_{(s)} = \widetilde{\mathbf{X}}_{(s)}(\widetilde{\mathbf{X}}_{(s)}^{\mathrm{T}}\widetilde{\mathbf{X}}_{(s)})^{-1}\widetilde{\mathbf{X}}_{(s)}^{\mathrm{T}}$ and $\mathbf{P}_{(s)} = \widetilde{\mathbf{P}}_{(s)}(\mathbf{I}_n - \mathbf{K}_{(s)}) + \mathbf{K}_{(s)}$, we can write $\widehat{\boldsymbol{\mu}}_{(s)} = \mathbf{P}_{(s)}\mathbf{y}$. Note that because of the curse of dimensionality, q_s (the dimension of $\mathbf{Z}_{(s)}$) cannot be large.

With the estimators of each model readily available, we can obtain the model averaging estimator of $\boldsymbol{\mu}$ by $\widehat{\boldsymbol{\mu}}(\mathbf{w}) = \sum_{s=1}^{S_n} w_s\widehat{\boldsymbol{\mu}}_{(s)} = \mathbf{P}(\mathbf{w})\mathbf{y}$, where $\mathbf{w} = (w_1, \ldots, w_{S_n})^{\mathrm{T}}$ is the weight vector belonging to the set $\mathcal{W} = \{\mathbf{w} \in [0,1]^{S_n} : \sum_{s=1}^{S_n} w_s = 1\}$ and $\mathbf{P}(\mathbf{w}) = \sum_{s=1}^{S_n} w_s\mathbf{P}_{(s)}$.

Remark 2. We point out that although heteroscedasticity is allowed in the data generating process (2.1), we do not immediately take it into account when estimating each candidate model (using kernel smoothing). Instead, we incorporate heteroscedasticity when estimating the unknown variance-covariance matrix (for the weight estimation). This is a typical treatment in the literature on model averaging under heteroscedasticity, such as Hansen & Racine

(2012), Liu & Okui (2013), and Zhang et al. (2015). The main reason is that an estimator that incorporates heteroscedasticity for each candidate model is not necessarily more efficient than an estimator that fails to do so, and the latter is computationally much simpler and faster.

2.2. Weight choice criterion and asymptotic optimality

Define the predictive squared loss $L_n(\mathbf{w}) = \|\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}\|^2$ and expected loss

$$R_n(\mathbf{w}) = \mathrm{E}\{L_n(\mathbf{w})\} = \|\mathbf{P}(\mathbf{w})\boldsymbol{\mu} - \boldsymbol{\mu}\|^2 + \mathrm{trace}\{\mathbf{P}(\mathbf{w})\boldsymbol{\Omega}\mathbf{P}^{\mathrm{T}}(\mathbf{w})\}, \tag{2.3}$$

where $\boldsymbol{\Omega} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$. To select the optimal weights in the sense of minimizing L_n, we propose to minimize the following Mallows-type criterion

$$C_n(\mathbf{w}) = \|\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \mathbf{y}\|^2 + 2\mathrm{trace}\{\mathbf{P}(\mathbf{w})\boldsymbol{\Omega}\}, \tag{2.4}$$

as we can show that $R_n(\mathbf{w}) = \mathrm{E}\{C_n(\mathbf{w})\} - \mathrm{trace}(\boldsymbol{\Omega})$, where $\mathrm{trace}(\boldsymbol{\Omega})$ is unrelated to $\mathbf{w}$. Therefore, if we know $\boldsymbol{\Omega}$, the weights can be obtained as

$$\widehat{\mathbf{w}} = \mathrm{argmin}_{\mathbf{w}\in\mathcal{W}} C_n(\mathbf{w}). \tag{2.5}$$

Averaging using this weight choice is called Mallows averaging of partially linear models (MAPLM). The optimality of such a weight choice holds under some regularity conditions. Define $\xi_n = \inf_{\mathbf{w}\in\mathcal{W}} R_n(\mathbf{w})$ and $\mathbf{w}_s^o$ as a weight vector with the s^{th} element taking on the value of unity and other elements zeros (model selection weight). Let $\max_i$ indicate maximization over $i \in \{1, \ldots, n\}$, where all limiting properties here and throughout the text are under $n \to \infty$.

Condition 1 $\max_i \sum_{j=1}^n |K_{(s),ij}| = O(1)$ *and* $\max_j \sum_{i=1}^n |K_{(s),ij}| = O(1)$ *uniformly for* $s \in \{1, \ldots, S_n\}$, *almost surely.*

Condition 2 *For some integer* $G \geq 1$, $\max_i E(\epsilon_i^{4G}|\mathbf{X}_i, \mathbf{Z}_i) < \infty$ *and* $S_n\xi_n^{-2G}\sum_{s=1}^{S_n}\{R_n(\mathbf{w}_s^o)\}^G \to 0$ *almost surely.*

Condition 1 is the same as assumption (i) of Speckman (1988), which bounds the kernel. Condition 2 requires $\xi_n \to \infty$, i.e., there is no finite approximating model whose bias is zero (Hansen & Racine, 2012 and Liu & Okui, 2013). This condition also constrains the rates of S_n and $R_n(\mathbf{w}_s^o)$ going to the infinity, and is widely used in other model averaging studies; see, for example, Wan et al. (2010), Liu & Okui (2013), and Ando & Li (2014).

Theorem 1 *Under Conditions 1-2, we have that as* $n \to \infty$,

$$L_n(\widehat{\mathbf{w}})/\inf_{\mathbf{w}\in\mathcal{W}} L_n(\mathbf{w}) \to 1 \; in \; probability. \tag{2.6}$$

Theorem 1 shows that the model averaging procedure using $\widehat{\mathbf{w}}$ is asymptotically optimal in the sense that the resulting squared loss is asymptotically identical to that of the infeasible best possible model averaging estimator. The proof of Theorem 1 (see the online supplement) takes advantage of several inequalities involving kernels, and it provides a technical innovation for studying the optimal model averaging in a kernel smoothing framework.

So far we have assumed that the covariance matrix $\mathbf{\Omega}$ is known. This is not the case in practice, and the criterion (2.4) is therefore computationally infeasible. To obtain a feasible criterion, we estimate $\mathbf{\Omega}$ based on the residues from the largest model indexed by $s^* = \arg\max_{s\in\{1,\ldots,S_n\}}(p_s + q_s)$, that is

$$\widehat{\mathbf{\Omega}}_{(s^*)} = \text{diag}(\widehat{\epsilon}_{s^*,1}^2, \ldots, \widehat{\epsilon}_{s^*,n}^2), \tag{2.7}$$

where $(\widehat{\epsilon}_{s^*,1}, \ldots, \widehat{\epsilon}_{s^*,n})^{\mathrm{T}} = \mathbf{y} - \widehat{\boldsymbol{\mu}}_{(s^*)} = \mathbf{y} - \mathbf{P}_{(s^*)}\mathbf{y}$. The idea of using the largest model to estimate the variance parameter or covariance matrix is also advocated by Hansen (2007) and

Liu & Okui (2013). We distinguish between two cases here. First, if the candidate models have the same nonparametric component but only differ in the inclusion of linear covariates, the largest model is unambiguously the one with all linear covariates included. In the more general case with uncertainty in both linear and nonparametric components, the model with the largest dimension is not uniquely defined since the models with the same dimension can differ in the structure of linear and nonparametric components. Therefore, we propose to use the the largest *linear* model to estimate $\mathbf{\Omega}$ in this case. Although the largest linear model is nested in the largest nonlinear model, including a large number of covariates in the nonlinear component leads to a highly inaccurate estimate of this component due to the curse of dimensionality. The inaccurate estimate further results in a poor estimator of error variance. Moreover, in most applications, the dimension of the nonlinear component is typically low; see, for example, Yatchew and No (2001) and Liang (2006). Hence, the estimated error variance obtained from the largest linear model is a good approximate in practice. Nevertheless, we point out that when the total number of covariates is particularly small, such that the largest nonlinear model is of low dimension, it might make more sense to use the largest nonlinear model to estimate the error variance.

By replacing $\mathbf{\Omega}$ with its estimator $\widehat{\mathbf{\Omega}}$, the feasible criterion becomes

$$\widehat{C}_n(\mathbf{w}) = \|\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \mathbf{y}\|^2 + 2\text{trace}\{\mathbf{P}(\mathbf{w})\widehat{\mathbf{\Omega}}_{(s^*)}\}, \tag{2.8}$$

and the weights can be obtained by

$$\widetilde{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathcal{W}} \widehat{C}_n(\mathbf{w}). \tag{2.9}$$

Let $\mathbf{H} = (\widehat{\boldsymbol{\mu}}_{(1)} - \mathbf{y}, \ldots, \widehat{\boldsymbol{\mu}}_{(S_n)} - \mathbf{y})$ and $\mathbf{b} = \{\text{trace}(\mathbf{P}_{(1)}\widehat{\mathbf{\Omega}}_{(s^*)}), \ldots, \text{trace}(\mathbf{P}_{(S_n)}\widehat{\mathbf{\Omega}}_{(s^*)})\}^{\mathrm{T}}$. We can rewrite $\widehat{C}_n(\mathbf{w})$ as $\widehat{C}_n(\mathbf{w}) = \mathbf{w}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{H}\mathbf{w} + 2\mathbf{w}^{\mathrm{T}}\mathbf{b}$, which is a quadratic function of

$\mathbf{w}$, and the optimization can be done by standard software packages, such as quadprog in Matlab, which are generally effective and efficient even when S_n is large.

We now show that the weights obtained by minimizing the feasible criterion (2.8) are still asymptotically optimal. Denote $\rho_{ii}^{(s)}$ as the i^{th} diagonal element of $\mathbf{P}_{(s)}$. Let $\max_s(\min_s)$ represent maximization(minimization) over $s \in \{1, \ldots, S_n\}$, $\widetilde{p} = \max_s p_s$, and $h = \min_s h_s$. Assume the following conditions hold almost surely.

Condition 3 $\|\boldsymbol{\mu}\|^2 = O(n)$.

Condition 4 *trace*$(\mathbf{K}_{(s)}) = O(h^{-1})$ *uniformly for* $s \in \{1, \ldots, S_n\}$.

Condition 5 *There exists a constant* c *such that* $|\rho_{ii}^{(s)}| \leq cn^{-1}|$*trace*$(\mathbf{P}_{(s)})|$ *for all* $s \in \{1, \ldots, S_n\}$.

Condition 6 $n^{-1}h^{-2} = O(1)$ *and* $n^{-1}\widetilde{p}^2 = O(1)$.

Condition 3 concerns the sum of n elements of $\boldsymbol{\mu}$ and is commonly used in linear regression models; see, for example, Wan et al. (2010) and Liang et al. (2011). Condition 4 is a natural extension of Condition (h) of Speckman (1988). Condition 5 is commonly used to ensure the asymptotic optimality of cross-validation; see, for example, Andrews (1991) and Hansen & Racine (2012). The first part of Condition 6 regards the bandwidth and is less restrictive than the $n^{-1}h^{-2} = o(1)$ required in Theorem 2 of Speckman (1988). The second part of Condition 6, which is the same as Condition (12) of Wan et al. (2010), allows p_s's to increase as $n \to \infty$, but restricts their increasing rates. Further explanations of these conditions are provided in the online supplement.

Theorem 2 *Under Conditions 1-6, we have that as* $n \to \infty$,

$$L_n(\widetilde{\mathbf{w}}) / \inf_{\mathbf{w} \in \mathcal{W}} L_n(\mathbf{w}) \to 1 \; in \; probability. \tag{2.10}$$

The proof of Theorem 2 is provided in the online supplement.

Remark 3. The question of how to choose the optimal bandwidth h_s in each candidate model remains. While this question is of interest, it is especially difficult in our case because each candidate model is just an approximation of the true model and therefore includes approximation error. In our numerical examples, the bandwidth h_s is chosen by minimizing the generalized cross-validation criterion. As an alternative, we also consider bandwidth selection using cross-validation, a popular criterion in the presence of heteroscedasticity. The simulation results show that the two criteria lead to almost identical relative performance of their competing methods, but cross-validation is computationally much more expensive than generalized cross-validation.

Remark 4. Theorem 2 holds no matter $\mathbf{\Omega}$ is estimated by the largest partially linear model (in the case with only linear component uncertainty) or the largest linear model (in the case with structure uncertainty), as long as the number of covariates is fixed. An alternative strategy to estimate $\mathbf{\Omega}$ is based on the *averaged* residuals $\widehat{\boldsymbol{\epsilon}}(\mathbf{w}) = \{\widehat{\epsilon}_1(\mathbf{w}), \ldots, \widehat{\epsilon}_n(\mathbf{w})\}^{\mathrm{T}} = \mathbf{y} - \widehat{\boldsymbol{\mu}}(\mathbf{w})$. The motivation of this strategy is to avoid placing too much confidence in a single model. The use of the averaged residuals does not affect the validity of Theorem 2 and produces similar numerical results. Detailed results of this alternative estimation strategy and proofs of this remark are available upon request.

3. Simulation study

3.1. Data generation process

Our setting is similar to the infinite-order regression by Hansen (2007) except that we have a nonlinear function in addition to the linear component. Specifically, we generate the data by $y_i = \mu_i + \epsilon_i = \sum_{j=1}^{500} \beta_j x_{ij} + g(\mathbf{Z}_i) + \epsilon_i$, where $\mathbf{X}_i = \{x_{i1}, \ldots, x_{i500}\}^{\mathrm{T}}$ is drawn from a multivariate normal distribution with mean 0 and covariance $0.5^{|j_1 - j_2|}$ between x_{ij_1} and x_{ij_2}. The corresponding coefficients are set as $\beta_j = 1/j$. For simplicity, we consider a nonlinear function of two *correlated* variables, i.e., $g(\mathbf{Z}_i) = g(z_{i1}, z_{i2})$, and we generate $z_{i1} = 0.3u_1 + 0.7u_2$ and $z_{i2} = 0.7u_1 + 0.3u_2$ where u_1 and u_2 are independent and uniformly distributed. Two variants of nonlinear functions are studied: $g_1(\mathbf{Z}_i) = \exp(z_{i1}) + z_{i2}^2$ and $g_2(\mathbf{Z}_i) = 2(z_{i1} - 0.5)^3 + \sin(z_{i2})$. The errors are normally distributed and heteroscedastic as $\epsilon_i \sim \mathrm{N}(0, \eta^2 x_{i2}^2)$. We change the value of η, so that $R^2 = \mathrm{var}\,(\mu_1, \ldots, \mu_n)\,/\mathrm{var}\,(y_1, \ldots, y_n)$ varies from 0.1 to 0.9, where $\mathrm{var}(\cdot)$ denotes the sample variance. Since all covariates are correlated with each other, R^2 cannot be easily written as a function of η. We therefore *numerically* compute R^2 based on each chosen η. The sample size is set to $n = 50$, 100, and 200, and the results of $n = 400$ are given in Section S.3 of the online supplement.

In real applications, the model is typically a simplified version of the data generating process with a number of variables omitted, either because of ignorance or because of data limitations. To mimic this situation, we omit z_{i2} and some components of $\mathbf{X}_i$ for every candidate model. We consider two cases with different types of model uncertainty. In the first case, it is a priori which variables belong to the nonparametric component (based on existing theory or the research question of interest), but the specification of the linear component is uncertain. In this case, all candidate models share a common nonparametric function of z_{i1} (with z_{i2} being omitted), and their linear components are a subset of $\{x_{i1}, \ldots, x_{i5}\}^{\mathrm{T}}$ (with the remaining x_{ij}'s being omitted). We require each candidate model to include at least one

linear covariate, leading to $2^5 - 1 = 31$ candidate models.

In the second case, there is no a priori knowledge about which covariates should be chosen as parametric regressors and which belong to the nonparametric component. Therefore, in addition to the uncertainty of which variables to include, we are also uncertain about whether a covariate should be included in the linear or nonparametric component. As the number of covariates increases, the number of candidate models now increases even more dramatically than in the first case. To facilitate the computation, we assume that only four covariates $(x_{i1}, x_{i2}, x_{i3}, z_{i1})$ are observed, whereas the others are omitted. In contrast to the first case, the candidate models here allow a subset of $(x_{i1}, x_{i2}, x_{i3}, z_{i1})$ in the nonparametric function, and the remaining can be included in the linear component or not in the model at all. Again, we require each candidate model to contain at least one linear and one nonparametric covariate. This leads to $\binom{4}{3}(2^3 - 1) + \binom{4}{2}(2^2 - 1) + \binom{4}{1} = 50$ candidate models. More simulation designs, such as a diverging number of candidate models, data with a larger degree of nonlinearity and autoregressive errors, are presented in the supplement. The results are essentially the same.

3.2. Estimation and comparison

We estimate each candidate model using the quadric kernel $k(v) = 15/16(1-v^2)^2 I(|v| \leq 1)$, where $I(\cdot)$ is an indicator function. In the first case with only linear component uncertainty, the covariance matrix $\mathbf{\Omega}$ is estimated using the largest candidate model, i.e., the partially linear model containing all observable linear covariates, and in the second case it is estimated from the largest *linear* model (with all observable variables included linearly and no nonparametric component). We mainly compare MAPLM with four alternative estimation methods for PLMs including two selection methods and two averaging methods. The two

model selection methods are based on AIC and BIC, and they select the model with the smallest information criterion, defined respectively, as $\text{AIC}_s = \log(\widehat{\sigma}_s^2) + 2n^{-1}\text{trace}(\mathbf{P}_{(s)})$ and $\text{BIC}_s = \log(\widehat{\sigma}_s^2) + n^{-1}\text{trace}(\mathbf{P}_{(s)})\log(n)$, where $\widehat{\sigma}_s^2 = n^{-1}\|\mathbf{y} - \widehat{\boldsymbol{\mu}}_{(s)}\|^2$. The two model averaging methods are smoothed AIC (SAIC) and smoothed BIC (SBIC) (Buckland et al., 1997). The weight of model s is constructed by $w_s^{\text{AIC}} = \exp(-\text{AIC}_s/2)/\sum_{s=1}^{S}\exp(-\text{AIC}_s/2)$ and $w_s^{\text{BIC}} = \exp(-\text{BIC}_s/2)/\sum_{s=1}^{S}\exp(-\text{BIC}_s/2)$.

To evaluate these methods, we compute the mean squared error (MSE) of the predictive variable as $500^{-1}\sum_{r=1}^{500}\|\widehat{\boldsymbol{\mu}}^{(r)} - \boldsymbol{\mu}\|^2$, where 500 is the number of replications and $\widehat{\boldsymbol{\mu}}^{(r)}$ denotes the estimator of $\boldsymbol{\mu}$ in the r^{th} replication. For convenient comparison, all MSEs are normalized by dividing by the MSE produced by AIC model selection.

3.3. Results

We first describe some general observations from the results, and then discuss each case in detail. In general, the model averaging methods outperform the selection methods. The superiority of the averaging methods is particularly obvious when R^2 is small. As R^2 increases, the difference between model selection and averaging decreases. The performance of the averaging methods when R^2 is small and moderate is especially good because identifying the best model is difficult in the presence of substantial noise. In that case, the model chosen by a selection procedure can be far from ideal, which unsurprisingly leads to inaccurate estimates. By contrast, model averaging does not rely on a single model and thus provides protection against choosing a poor model. This observation is also in line with Yuan & Yang (2005) and Zhang et al. (2012). When R^2 is large, model selection is sometimes preferred because the minimal noise in the data allows the selection criterion to choose the correct model.

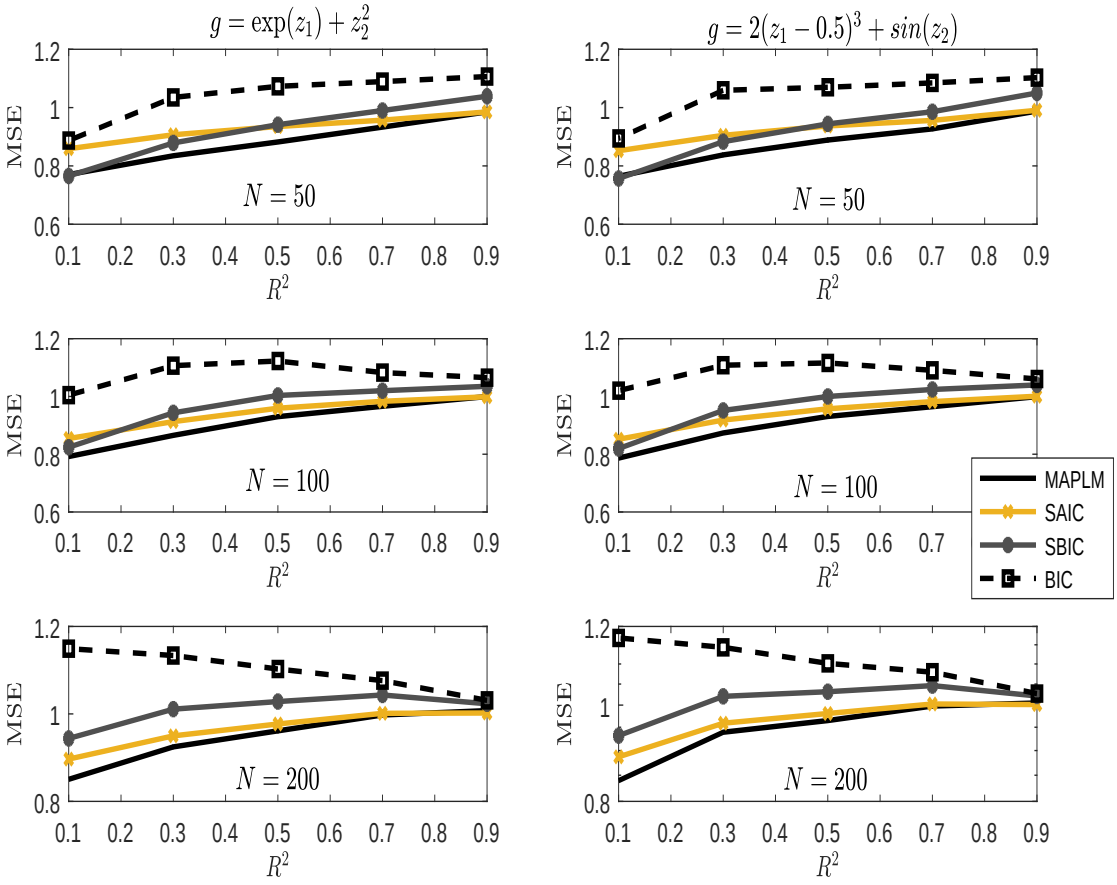
Figure 1 presents the results when there is uncertainty in only the linear component

specification. Our method yields the smallest MSE in almost all cases, but the information-criterion model averaging sometimes has a marginal advantage when R^2 is very large. Most of the figures show that the advantage of our method increases as R^2 decreases. The good performance of MAPLM is partly because the optimality of MAPLM does not rely on the correct specification of candidate models. The comparison of the methods for different sample sizes shows that when we have a relatively small or moderate sample size ($n = 50$ and 100), MAPLM outperforms all the competing methods over the whole range of R^2. When the sample size is relatively large ($n = 200$), MAPLM still dominates the other methods for a wide range of R^2, but the difference between MAPLM and SAIC decreases. We also note that all the methods perform almost equally well when the sample size is large and R^2 is 0.9. Further examination suggests that the methods tend to select or impose a large weight on the same model when there is little noise in the model and the sample size is large. This similarity can be partly explained by the fact that the bias-variance tradeoff is not so significant in this situation, so model selection is able to choose the correct model.

Figure 2 compares the estimation results when there is structure uncertainty in addition to uncertainty in covariate inclusion. In this case, both linear and nonparametric components vary over the candidate models. MAPLM produces much lower MSE than its rivals in all cases when R^2 is equal to or less than 0.7, which again demonstrates that our model averaging approach is preferred when the model is characterized by substantial noise and identifying the best model is difficult, as in most practical applications. The poorer performance of MAPLM under particularly large R^2 is mainly a result of allowing for far more uncertainty than necessary in this case, which prevents MAPLM from assigning a very large weight to the best model. Specifically, on one hand, allowing for more uncertainty (in both the linear

and nonlinear components) than in the first case causes MAPLM to average over a larger model space, which generates a larger number of weight parameters to estimate. On the other hand, when the data are highly informative (with large R^2), there often exists a best model, and IC are capable of selecting this model. By contrast, simultaneously estimating a large number of weights clearly prevents MAPLM from assigning a very large weight to the best model, resulting in the poorer performance of MAPLM in this case.

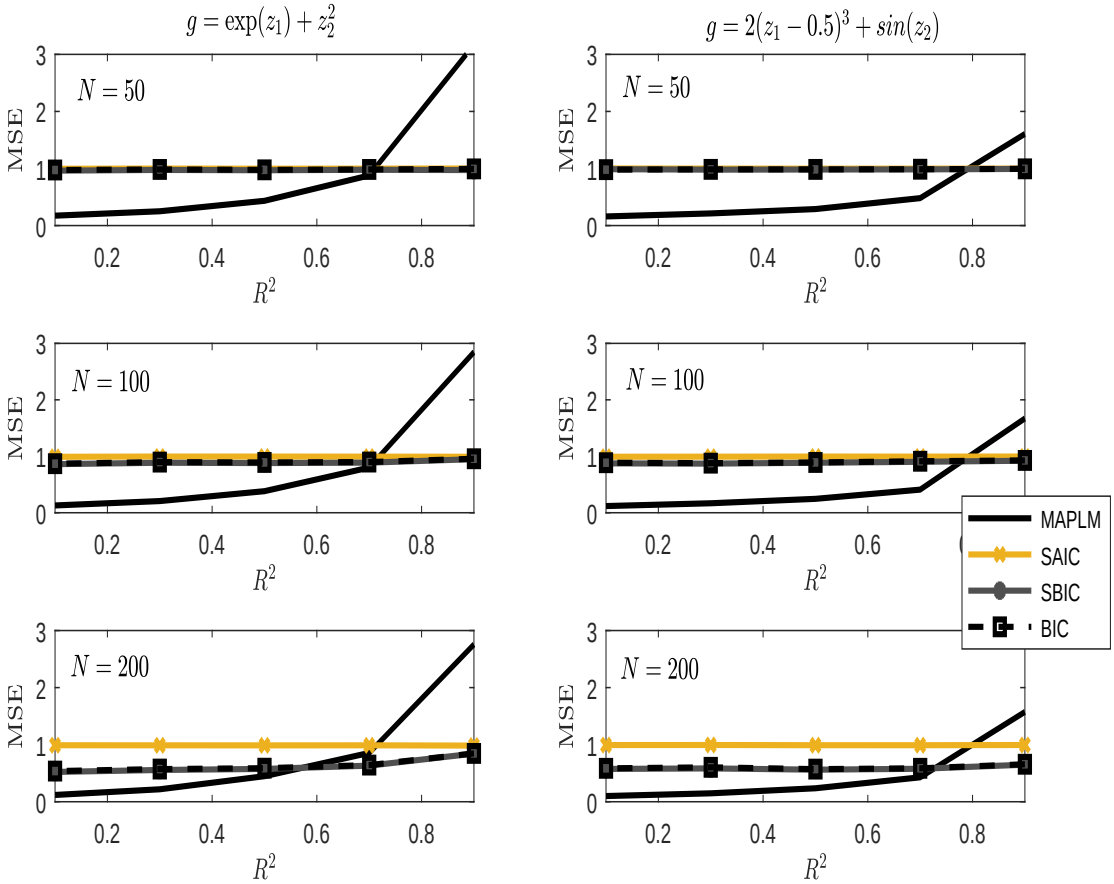
Figure 1: Mean square error comparison: Uncertainty only in the linear component



Moreover, model selection and averaging using AIC and SAIC lead to largely similar

results, as do BIC versus SBIC. These results indicate that there is a dominant model that significantly outperforms the others, and this dominant model is often the one with the most covariates in the nonparametric component. This further suggests that IC tend to select the most general model whenever possible, because nonparametric estimation typically fits better than least squares estimation. However, the dominant model is not necessarily the best in all cases. When the data are characterized by substantial noise, a large nonparametric model mainly fits the noise; thus, the IC-based methods perform much worse than MAPLM. When the data are highly informative, i.e., $R^2 = 0.9$, the dominant model coincides with the best model, leading to the better performance of the IC-based methods than MAPLM.

Figure 2: Mean square error comparison: Uncertainty in both components



To see how much harm can be caused by ignoring nonlinearity, we also compare our method with linear model averaging (LMA) that considers all candidate models to be fully linear. Theoretically, LMA should work better when the model is linear or the degree of nonlinearity is small since nonparametric estimation converges much slower and is generally less efficient than least squares estimation. As the degree of nonlinearity increases, the better fit achieved by nonparametric estimation dominates its efficiency loss and slow convergence; thus MAPLM should outperform LMA under these conditions. Our simulation results (presented in the supplement) under different degrees of nonlinearity precisely confirm this theoretical argument. Moreover, we find that LMA slightly outperforms MAPLM when R^2 and the sample size are small. As R^2 and the sample size increase, MAPLM quickly demonstrates its significant superiority over LMA. Detailed simulation designs, results, and explanations are provided in the online supplement.

4. Empirical application

We apply our method to study Japan's sovereign credit default swap (CDS) spreads. A CDS contract is an insurance contract against the credit event specified in the contract. Its spread is the insurance premium that the buyer under protection has to pay, and it reflects investors' expectations about a country's sovereign credit risk. The likelihood of default typically depends on the country's willingness (rather than ability) to repay, and the government often makes the repayment decision based on a cost-benefit analysis using the information of the country's macroeconomic fundamentals. Japan's sovereign CDS spreads are of worldwide interest since Japan has long been characterized by its high government debt. The ratio

of gross government debt to GDP even reached 237.9% in 2012, the highest in the world. Furthermore, Japan is the world's third largest economy, and its financial market plays an important role in international finance. A crisis in Japan could damage investors' confidence in the government debt of many other heavily indebted industrial countries.

In this section, we first examine how macroeconomic indicators affect Japan's CDS spreads and then study the predictability of these indicators. We focus on the CDS contract written on the credit event "complete restructuring", which is the most popular credit event insured by a sovereign CDS contract, and we consider the contract maturity of five years, following Longstaff et al. (2011). Our potential macroeconomic determinants include three domestic variables that reflect domestic economic performance: the domestic stock market return (measured by the Dow Jones Japan Total Stock Market Total Return Index), its volatility, and the nominal Yen-US Dollar exchange rate. We also follow Longstaff et al. (2011) and consider three global-market determinants: the global stock market return (measured by the Morgan Stanley Capital International US Total Return Index), US treasury yield (with the constant maturity of five years), and the global default risk premium (approximated by US investment-grade corporate bond spreads). See Longstaff et al. (2011) and Qian et al. (2017) for details of the variable construction. We focus on the post-earthquake sample from March 12, 2011 (one day after the Tohuku earthquake) to October 10, 2012 to avoid significant structural breaks, and the number of observations is 388. All data are first-differenced based on a preliminary unit root analysis and then normalized. The change in Japan's sovereign CDS spreads (before normalization) is plotted in the left panel of Figure S.6 of the online supplement, and its sample autocorrelation function is plotted in the right panel. These two plots show that the differenced series does not exhibit strong serial correlation. Table S.1

of the online supplement provides the descriptive statistics of the first-differenced sovereign CDS spreads and potential determinants.

4.1. Linear model specification

The existing literature on sovereign CDS spreads mostly considers linear models in which all the determinants are assumed to have a linear effect on the spreads; see, for example, Longstaff et al. (2011) and Dieckmann & Plank (2011). We initially follow this convention to estimate the effect of our six potential determinants using linear models. We consider ordinary least squares (OLS) estimation and linear model averaging using the heteroscedastic-robust Mallows criterion (HRC_p). Linear model averaging treats all determinants linearly, but it takes into account the uncertainty of whether a determinant is included in the model.

Table 1: Estimation results of linear models

	OLS	LMA		OLS	LMA
Domestic stock return	-1.5182***	-1.2790***	*Domestic stock volatility*	0.6165***	0.0576
	(0.1752)	(0.3189)		(0.1758)	(0.7208)
Foreign exchange rate	-0.3250*	-0.3777***	*Global stock return*	1.0107***	0.9842***
	(0.1727)	(0.1839)		(0.1733)	(0.2188)
US treasury yield	-0.3672**	-0.3649**	*Global default risk premium*	-0.0774	-0.0230
	(0.1750)	(0.2349)		(0.1689)	(0.0948)

Notes: Standard errors are in parentheses. ***, **, and * denote significance at 1%, 5%, and 10%, respectively. The significance of LMA is based on bootstrap confidence intervals with 200 random samplings.

Table 1 presents the estimation results of the linear models. Since all determinants are normalized, the size of the coefficients reflects the relative importance. We first focus on the

least squares estimation results. The least squares estimates show that the domestic stock return, its volatility, and the global stock return are the three most important determinants and have a significant effect on Japan's CDS spread. More specifically, the domestic stock return, as a measure of local economic performance, has a strongly negative effect. The domestic stock return can affect the CDS spread by influencing the government's willingness to implement fiscal reforms, and effective fiscal reform is typically regarded as an important tool to reduce default risk. Therefore, when the domestic economy is weak, policy makers are less willing to implement reforms because the reforms can impose extra pressure on the distressed economy. This failure to enact reforms thus increases the sovereign CDS spreads. The strong and negative effect of domestic stock returns is in line with the literature (see, e.g., Longstaff et al., 2011 and Dieckmann & Plank, 2011). The domestic stock market volatility is positively associated with sovereign CDS spreads, which is in line with the economic theory that higher volatility indicates a less stable economic status and thus a higher probability of default. The other important determinant is the global stock return, which has a positive effect on Japan's sovereign CDS spread. Theoretically, the global stock market return may impose two opposite impacts on sovereign CDS spreads. The negative effect is due to the fact that good global economic performance can positively influence the Japanese government's willingness to repay, thus lowering the sovereign CDS spread. On the other hand, a good global economy would also encourage investment in general, thereby increasing the CDS spread. The overall impact of the global stock return depends on which effect is dominant. It is likely that one effect is more prominent in some situations but dominated by the other effect in other situations. This potential heterogeneity cannot be captured by linear models.

Less significant but still important determinants include the foreign exchange rate and

US treasury yield. The negative effect of the foreign exchange rate is expected because a low Yen-US Dollar exchange rate reflects weakness in Japan's current economic situation and less external demand, which leads to higher sovereign CDS spread. The negative relationship between US treasury yield and Japan's CDS spread is also intuitive because a high treasury yield signals good economic performance in the US, which can positively influence Japan's economy and encourage repayment by the Japanese government.

We also compare the estimates obtained from least squares and model averaging and find that the signs of all the estimated coefficients are the same for both methods. Nevertheless, model averaging produces quite different estimates for some determinants, such as the domestic stock return, its volatility, and the global default risk premium, which suggests that there is a large degree of model uncertainty.

4.2. Partially linear specification

Next, we examine whether the widely used linearity assumption is appropriate here. The verification is based on both economic theory and statistical tools. First, from the theoretical perspective, the literature of sovereign CDS spreads generally does not provide firm theory about nonlinear effects for most covariates, except the domestic and global stock returns. Qian et al. (2017) found that these two covariates play different roles in tranquil and turbulent periods. Specifically, global stock returns are more prominent during turbulent periods, and domestic stock returns are more prominent during tranquil periods. The nonlinear effect of global stock returns is also supported by extensive literature on financial contagion, which suggests that the link between domestic markets and the global market is often strengthened during periods of crisis; see, e.g. Eichengreen et al. (1996) and Bae et al. (2003). Therefore, it is reasonable to consider the potential nonlinear effect of global stock returns.

Next, we verify the linearity of each determinant by assigning it to the nonparametric component of partially linear models. We include *one* determinant in the nonparametric component each time while keeping others in the linear component. This process enables us to verify whether each determinant has a nonlinear effect on Japan's CDS spreads and also avoids the dimensionality and computational issue of simultaneously considering many nonparametric covariates. Figure S.7 of the online supplement presents the nonparametric estimates of each determinant using the proposed MAPLM, i.e., $\widehat{\mathbf{g}}(\widehat{\mathbf{w}}) = \sum_{s=1}^{S_n} \widehat{w}_s \widehat{\mathbf{g}}_{(s)}$, where $\widehat{\mathbf{g}}_{(s)}$ is the nonparametric estimate obtained from each candidate model and $\widehat{\mathbf{w}} = (\widehat{w}_1, \ldots, \widehat{w}_{S_n})^{\mathrm{T}}$ is weights estimated by MAPLM. We see that the effects of domestic stock market volatility and global default risk premium do not exhibit a clear nonlinear pattern. They either have a relatively flat curve or fluctuate around zero, suggesting that these effects are almost linear or highly insignificant. In contrast, domestic stock returns, the foreign exchange rate, global stock returns, and US treasury yield show different degrees of nonlinearity.

Finally, we formally test the linearity for each determinant using the test statistic suggested by Li et al. (2010). This test statistic verifies the null hypothesis of the linearity of the nonparametric component by the fiducial method. Specifically, Li et al. (2010) proposed to first approximate the nonparametric component by a piecewise linear function. Thus, testing for linearity is transformed into testing for a linear restriction on the coefficient. The p-value of the test is then derived by the classic fiducial method (c.f., Xu & Li (2006)). To validate this test in our case, we implement the test in the *fixed* full model, where only one determinant is included in the nonparametric component each time and the remaining determinants are included in the linear component. Therefore, no averaging is performed in this testing

procedure. Although it appears to be more general to test the linearity of a covariate while assigning others to the nonparametric component, it is difficult in practice because a large number of covariates in the nonparametric leads to the the curse of dimensionality. The curse results in poor estimates and unreliable test statistics in finite samples. The p-values of the tests are reported in Table 2. We see that the test fails to reject the null hypothesis of linearity for the domestic stock return, its volatility, and global default risk premium. The reported p-values also confirm that the effects of the foreign exchange rate and global stock returns cannot be accurately approximated by linear functions. The test statistic for the US treasury yield is not available because this variable takes only a few discrete values. Thus, it is less clear whether one can assume a linear effect of the US treasury yield.

Table 2: Linearity test for each determinant

	p-value		p-value
Domestic stock returns	0.1651	*Domestic stock volatility*	0.4810
Foreign exchange rate	0.0265	*Global stock returns*	0.0042
US treasury yield	NA	*Global default risk premium*	0.9548

Based on the nonparametric estimation results and the regression diagnostics, we discuss the (potentially) nonlinear determinants and their economic implications. First, the estimated effect of the foreign exchange rate has a steeply downward trend when the change in exchange rates is below average, but the curve becomes relatively flat and close to zero as the change in exchange rates increases. The negative relationship between the exchange rate and Japan's CDS spread is in line with the findings of the linear models. Nevertheless, the nonparametric estimate shows that this relationship becomes much weaker when the exchange rate is high.

Second, the estimated effect of global returns is characterized by a typical "U-shape". We see that the change in Japan's CDS spreads is particularly high when global returns are at the extremes, either a large positive change or a large negative change, suggesting that the negative effect of global stock returns plays a more prominent role in a bear market while the positive effect is more important when the global financial market is in a bull market. We also observe that the curve is much steeper when the global stock market is in a slump, suggesting that the correlation between Japan's credit market and the global stock market is much stronger during periods of crisis. This result is in contrast to Longstaff et al. (2011), who reported a weak and insignificant effect of global stock returns on Japan. We argue that the insignificance is possibly a result of ignoring nonlinearity, such that the positive and negative effects offset each other, leading to an ambiguous overall effect. Such a nonlinear effect of global stock returns provides evidence of financial contagion from the global stock market to Japan's sovereign credit market, which cannot be captured by linear models. The finding of financial contagion is of particular importance for both policy makers and investors since it implies that adapted policies and investment strategies should be implemented under different situations. Finally, the curve of the US treasury yield is similar but less nonlinear than that of the foreign exchange rate. We generally observe a negative effect of the US treasury yield on Japan's sovereign CDS spreads, in line with the literature and our linear model estimates, but the effect is relatively stronger when the change in treasury yield is extreme.

Table 3 reports the estimates of the linear coefficients of the partially linear model with foreign exchange rate and global stock returns in the nonlinear component. To compute these estimates, let $\mathbf{X}$ be the matrix of the linear covariates of the full model (which contains the

domestic stock return and its volatility, US treasury yield, and global default risk premium) and $\Pi_{(s)}$ be a projection matrix such that $\mathbf{X}_{(s)} = \mathbf{X}\Pi_{(s)}$. Then, the model averaging estimates of the linear coefficients can be obtained by $\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}}) = \sum_{s=1}^{S_n} \widehat{w}_s \Pi_{(s)} \widehat{\boldsymbol{\beta}}_{(s)}$, similar to Hansen (2007)'s model averaging estimator. Since no standard inference theories are available for the optimal model averaging estimates, we provide 99% bootstrap confidence intervals for the model averaging estimates. The confidence intervals of AIC and BIC are computed with the selected model based on Theorem 4 of Speckman (1988), ignoring the uncertainty in the selection procedure. We see that the domestic stock return has the strongest negative association with the change in sovereign CDS spreads, as in the linear model; however, compared to the linear models, the estimated effect of the US treasury yield is weak and less significant in the PLMs.

To check whether our empirical results are sensitive to the predetermination of nonlinear covariates, we perform estimation and prediction (discussed in the next section) under different specifications of nonlinear covariates, and the results are generally quite robust.

4.3. Out-of-sample prediction

Finally, we examine the pseudo out-of-sample predicability of Japan's CDS spreads using six alternative methods: three model averaging methods (MAPLM, SAIC, and SBIC) and two model selection methods (AIC and BIC) for the partially linear models, and one linear model averaging method.

The linear model averaging is based on HRC_p, as above. It considers candidate models with at least one determinant included, so it averages over $2^6 - 1$ candidate models. For PLM averaging, the most general specification is to consider all possibilities, i.e., that a determinant can be in the linear component, in the nonlinear component, or not in the model. However,

Table 3: Estimates of the linear coefficients in the partially linear models

	MAPLM	SAIC	SBIC	AIC	BIC
Domestic stock returns	-1.2771	-1.5595	-1.5635	-15658	-1.5938
	$(-2.37, -0.62)$	$(-2.53, -0.65)$	$(-2.52, -0.00)$	$(-1.83, -1.30)$	$(-1.87, -1.32)$
Domestic stock vol.	0.0000	0.5330	0.5135	0.5325	0.5928
	$(-0.56, 2.23)$	$(-0.62, 2.28)$	$(-0.68, 2.26)$	$(0.23, 0.83)$	$(0.28, 0.91)$
US treasury yield	-0.3246	-0.2441	-0.0858	-0.3292	
	$(-0.78, 0.24)$	$(-0.86, 0.29)$	$(-0.88, 0.03)$	$(-0.65, -0.01)$	
Global risk premium	-0.1284	-0.0876	-0.0144		
	$(-0.47, 0.05)$	$(-0.52, 0.07)$	$(-0.65, 0.02)$		

this consideration may cause a dimensionality problem by including too many determinants in the nonlinear component. Thus, we assign determinants to the nonlinear component only when necessary. Based on the PLM analysis in the previous subsection, it seems reasonable to presume a linear relationship between Japan's CDS spreads and the global default risk premium and the domestic stock market return and its volatility. It is also clear that the foreign exchange rate and global stock returns have a nonlinear impact on Japan's CDS spread; thus it is necessary to include these two determinants in the nonlinear component when they are included in the model. As for the US treasury yield, since its effect only exhibits a moderate degree of nonlinearity and the formal linearity test is not informative, we are less certain whether to assign this variable to the linear or nonlinear component. Allowing this ambiguous determinant to enter the nonlinear component leads to a more complete model space but may also result in the dimensionality curse. There is no a *priori* knowledge of how to make

Table 4: Mean square prediction error of Japan's CDS spreads

	Prediction sample	MAPLM	SAIC	SBIC	AIC	BIC
Scenario I	5%	**0.8608**	0.9360	0.9278	0.9403	0.9253
	10%	**0.8490**	1.0162	1.0181	1.0256	1.0190
	15%	**0.9708**	1.0950	1.0830	1.1007	1.1106
	20%	**0.9927**	1.0933	1.1111	1.0751	1.1107
Scenario II	5%	**0.8865**	0.9723	0.9264	0.9673	0.9253
	10%	**0.7903**	0.9410	1.0175	0.9308	1.0190
	15%	0.8119	0.9814	0.8542	0.9652	**0.7770**
	20%	**0.8697**	0.9695	1.1073	0.9530	1.1107

an appropriate tradeoff between a more complete model space and the dimensionality curse. Therefore, we compare the prediction performance of six methods in two scenarios. In Scenario I, we allow only the foreign exchange rate and global stock return to be in the nonlinear component. In other words, the foreign exchange rate and global stock return can either not be included in the model or be in the nonlinear component of the model. The remaining determinants are either not in the model or in the linear component. Scenario II differs from Scenario I in that we also allow the US treasury yield to enter the nonlinear component. Hence, there are three possibilities for the uncertain determinant of the US treasury yield: not included in the model, included in the linear component, or included in the nonlinear component. We split the sample into two sub-samples, one for estimation and the other for prediction and evaluation. We consider the estimation sample varying from 80% to 95% of the whole period; thus the prediction sample ranging from 20% to 5% correspondingly.

Table 4 presents the mean square prediction error (MSPE) of five PLM methods. All

values are normalized by dividing by the MSPE of the linear model averaging method. We see that our MAPLM produces the lowest MSPE for all prediction samples in Scenario I. In Scenario II, MAPLM is the best in most cases, except when the prediction sample is 15%. In all cases, MAPLM outperforms the linear Mallows averaging, demonstrating that incorporating the necessary nonlinearity improves the prediction performance. Since the performance of linear model averaging is invariant to the scenario, we can also compare the predictability of MAPLM in the two scenarios. Interestingly, we observe that allowing the US treasury yield to enter the nonlinear component improves the prediction performance for all methods when the prediction sample is larger than 5%. However, when we have a small prediction sample, a smaller model space is better. One possible explanation is that averaging over a larger model space may offset the additional noise by better diversification. When the prediction sample is large, the diversification gain from averaging over a larger model space is substantial and dominates the estimation inaccuracy due to the dimensionality curse. This is, however, not the case when the prediction sample is small (or, equivalently, when the training sample is large) because the predicted values obtained from different candidate models are more accurate and more similar to each other; thus, the diversification gain is smaller.

5. Concluding remarks

Partially linear models have become popular in applied econometrics and statistics because they allow a more flexible specification compared to linear models and provide more interpretable estimates compared to fully nonparametric models. Estimation of partially linear models is subjected to at least two types of uncertainty: the uncertainty of which variables to include in the model and the uncertainty of whether a covariate should be in the linear or nonlinear component given that it is in the model. Typical model testing and selection meth-

ods do not appropriately address these two types of uncertainty simultaneously, especially when the research interest is to estimate the parameters or to make predictions. In this paper, we propose an optimal model averaging procedure for PLMs that jointly incorporates the two types of model uncertainty. The extension from linear model averaging to partially linear models is by no means straightforward and routine because it involves kernel smoothing, which complicates the proof of optimality. We demonstrate the advantages of our methods by examining the determinants of Japan's sovereign CDS spreads. Our empirical study suggests that there exists a large degree of nonlinearity in the effects of macroeconomic determinants, such as the global stock return and exchange rate. Conventional linear models do not capture such nonlinearity, and ignoring the nonlinearity can result in a lack of awareness of financial contagion, which may further lead to inappropriate policies and investment decisions.

At least three issues deserve future research. First, the computational burden of our method would be substantial when the number of candidate models is large; therefore, a model screening step prior to model averaging is desirable. Second, although the dimension p_s is allowed to increase with the sample size n, it must be smaller than n and its increasing rate is restricted by the second part of Condition 6. How to develop an optimal model averaging method for high- or ultrahigh-dimensional PLMs is an interesting open question. Finally, if the research interest is to consistently estimate the linear and/or nonlinear component rather than to make predictions, a consistent model averaging estimator and post-model-averaging inference are desired. See, for example, Hjort & Claeskens (2003), Zhang & Liang (2011) and Xu et al. (2014). In these studies, a crucial assumption of local misspecification is required, and the weights also need to have an explicit form. By contrast, we do not utilize the local misspecification framework, and our weight estimates do not have an explicit form.

Therefore, the development of model averaging estimators for the linear and nonlinear components without local misspecification and analytical weights warrants further investigation.

Acknowledgments The authors are grateful to Co-Editor Zhiliang Ying, the Associate Editor and two referees for their constructive comments, and to Dr. Na Li for providing codes for nonlinearity test. Zhang's research was supported by National Natural Science Foundation of China (Grant nos. 71522004 and 11471324).

Online Supplement OnlineSupp.pdf describes the technical proofs and provide more explanations on the conditions as well as additional simulation studies.

References

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

ANDREWS, D. (1991). Asymptotic optimality of generalized C_L, cross-validation, and generalized cross-validation in regression with heteroskedastic errors. *Journal of Econometrics* **47**, 359–377.

BAE, K.-H., KAROLYI G. A. & STULZ, R. M. (1996). A new approach to measuring financial contagion. *The Review of Financial Studies* **16**, 717–763.

BARRO, R. J. (1996). Democracy and growth. *Journal of Economic Growth* **1**, 1–27.

BUCKLAND, S. T., BURNHAM, K. P. & AUGUSTIN, N. H. (1997). Model selection: An integral part of inference. *Biometrics* **53**, 603–618.

BUNEA, F. (2004). Consistent covariate selection and post model selection inference in semiparametric regression. *The Annals of Statistics* **32**, 898–927.

DANILOV, D. & MAGNUS, J. R. (2004). On the harm that ignoring pretesting can cause. *Journal of Econometrics* **122**, 27–46.

DETTE, H. & MUNK, A. (1998). Validation of linear regression models. *The Annals of Statistics* **26**, 778–800.

DIECKMANN, S. & PLANK, T. (2012). Default risk of advanced economies: An empirical analysis of credit default swaps during the financial crisis. *Review of Finance* **16**, 903–934.

EICHENGREEN, B., ROSE, A. & WYPLOSZ, C. (1996). Contagious currency crises. *Scandinavian Journal of Economics* **98**, 463–484.

ENGLE, R. F., GRANGER, C. W., RICE, J. & WEISS, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association* **81**, 310–320.

HAMILTON, S. A. & TRUONG, Y. K. (1997). Local linear estimation in partly linear models. *Journal of Multivariate Analysis* **60**, 1–19.

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

HANSEN, B. E. (2014). Model averaging, asymptotic risk, and regressor groups. *Quantitative Economics* **5**, 495–530.

HANSEN, B. E. & RACINE, J. (2012). Jackknife model averaging. *Journal of Econometrics* **167**, 38–46.

HÄRDLE, W., LIANG, H. & GAO, J. (2000). *Partially linear models*. Springer.

HARDY, G. H., LITTLEWOOD, J. E. & POLYA, G. (1952). *Inequalities*. Cambridge university press.

HECKMAN, N. E. (1986). Spline smoothing in a partly linear model. *Journal of the Royal Statistical Society. Series B (Methodological)* **48**, 244–248.

HENDERSON, D. J., PAPAGEORGIOU, C. & PARMETER, C. F. (2012). Growth empirics without parameters. *The Economic Journal* **122**, 125–154.

HJORT, N. L. & CLAESKENS, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association* **98**, 879–899.

HOETING, J. A., MADIGAN, D., RAFTERY, A. E. & VOLINSKY, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science* **14**, 382–417.

LI, N., XU, X. & JIN, P. (2010). Testing the linearity in partially linear models. *Journal of Nonparametric Statistics* **23**, 99–114.

LIANG, H., WANG, S. & CARROLL, R. J. (2007). Partially linear models with missing response variables and error-prone covariates. *Biometrika* **94**, 185–198.

LIANG, H., ZOU, G., WAN, A. T. K. & ZHANG, X. (2011). Optimal weight choice for frequentist model average estimators. *Journal of the American Statistical Association* **106**, 1053–1066.

LIU, Q. & OKUI, R. (2013). Heteroskedasticity-robust C_p model averaging. *The Econometrics Journal* **16**, 463–472.

LONGFORD, N. T. (2005). Editorial: Model selection and efficiency—is 'which model ...?' the right question? *Journal of the Royal Statistical Society. Series A (Statistics in Society)* **168**, 469–472.

LONGSTAFF, F. A., PAN, J., PEDERSEN, L. H. & SINGLETON, K. J. (2011). How sovereign is sovereign credit risk? *American Economic Journal: Macroeconomics* **3**, 75–103.

LU, X. & SU, L. (2015). Jackknife model averaging for quantile regressions. *Journal of Econometrics* **188**, 40–58.

MAGNUS, J. R., WANG, W. & ZHANG, X. (2016). Weighted average least square prediction. *Econometric Reviews* **35**, 1040–1074.

NI, X., ZHANG, H. H. & ZHANG, D. (2009). Automatic model selection for partially linear models. *Journal of the American Statistical Association* **100**, 2100–2111.

QIAN, Z., WANG, W. & JI, K. (2017). Sovereign credit risk, macroeconomic dynamics, and financial contagion: Evidence from Japan. *Macroeconomic Dynamics*, forthcoming.

ROBINSON, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica* **56**, 931–954.

RUPPERT, D., WAND, M. P. & CARROLL, R. J. (2003). *Semiparametric Regression*. Cambridge, New York: Cambridge University Press.

SPECKMAN, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society. Series B (Methodological)* **50**, 413–436.

SU, L. & LU, X. (2013). Nonparametric dynamic panel data models: Kernel estimation and specification testing. *Journal of Econometrics* **176**, 112–133.

WAN, A. T. K., ZHANG, X. & ZOU, G. (2010). Least squares model averaging by Mallows criterion. *Journal of Econometrics* **156**, 277–283.

WHITTLE, P. (1960). Bounds for the moments of linear and quadratic forms in independent variables. *Theory of Probability & Its Applications* **5**, 302–305.

XIE, H. & HUANG, J. (2009). SCAD-penalized regression in high-dimensional partially linear models. *The Annals of Statistics* **37**, 673–696.

XU, G., WANG, S. & HUANG, J. (2014). Focused information criterion and model averaging based on weighted composite quantile regression. *Scandinavian Journal of Statistics* **41**, 365–381.

XU, X. & LI, G. (2006). Fiducial inference in the pivotal family of distributions. *Science in China: Series A* **49**, 410–432.

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

ZHANG, H. H., CHENG, G. & LIU, Y. (2011). Linear or nonlinear? Automatic structure discovery for partially linear models. *Journal of the American Statistical Association* **106**, 1099–1112.

ZHANG, X. & LIANG, H. (2011). Focused information criterion and model averaging for generalized additive partial linear models. *The Annals of Statistics* **39**, 174–200.

ZHANG, X., WAN, A. T. K. & ZHOU, S. Z. (2012). Focused information criteria, model selection and model averaging in a Tobit model with a non-zero threshold. *Journal of Business & Economic Statistics* **30**, 132–142.

ZHANG, X., ZOU, G. & CARROLL, R. (2015). Model averaging based on Kullback-Leibler distance. *Statistica Sinica* **25**, 1583–1598.

ZHANG, X., ZOU, G. & LIANG, H. (2014). Model averaging and weight choice in linear mixed-effects models. *Biometrika* **101**, 205–218.

ZHAO, T., CHENG, G. & LIU, H. (2016). A partially linear framework for massive heterogeneous data. *The Annals of Statistics* **44**, 1400–1437.

Xinyu Zhang, Academy of Mathematics and Systems Science, Chinese Academy of Sciences
E-mail: xinyu@amss.ac.cn
Wendun Wang, Econometric Institute, Erasmus University Rotterdam, and Tinbergen Institute
E-mail: wang@ese.eur.nl