Optimal Model Averaging Estimation for Partially Linear Models

Xinyu Zhang*

Chinese Academy of Sciences, and Capital University of Economics and Business

Wendun Wang

Econometric Institute, Erasmus University Rotterdam, and Tinbergen Institute

SUMMARY: 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 commonly-used model selection and averaging methods. The proposed procedure is further applied to study Japan's sovereign credit default swap spreads.

Keywords: 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 a clear interpretation of the estimation results. However, an increasing number of studies have noted that the relationship between the response variable

*Corresponding author. E-mail addresses: xinyu@amss.ac.cn. Zhang's work is supported by National Natural Science Foundation of China (Grant no. 71522004).

Electronic copy available at: https://ssrn.com/abstract=2948380

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) showed that the HIV viral load depends nonlinearly on treatment time when studying the effectiveness of antiretroviral medicines. Ignoring nonlinearity can cause incorrect estimates and inference, which further result in misleading explanations and decisions. For example, ignoring the nonlinear effect of global stock markets on the local market may lead to unawareness 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 an extensive attention in theoretical and applied statistics due to their flexible specification. It allows for both linear and nonparametric relation 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 more preferred than the fully nonparametric models since it still preserves the advantages of linear models, e.g., an easy interpretation of the linear covariates, and suffers less from the dimensionality curse. There exists a wide range of applications using PLMs 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 are proposed to estimate PLMs, for example, smoothing splines (Engle et al., 1986; Heckman, 1986), kernel smoothing (Speckman, 1988; Robinson, 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 the correctly specified model is given. In practice, however, researchers are ignorant of the true model. One needs to decide which covariates

Electronic copy available at: https://ssrn.com/abstract=2948380

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 fundamentally important, as it greatly influence the estimation and prediction results. These two types of uncertainty is generally referred to as model uncertainty.

Typical methods to address model uncertainty is to test and/or select the best model using some data-driven approaches. The most popular might be to use the information criteria, such as Akaike information criterion (AIC) and 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, proposed several variable selection methods. To further determine the structure of the model (which covariates in the (non)linear function), a commonly used method is to test 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 handle the 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., forthcoming). 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 interest is to identify the right model structure. Nevertheless, if the research purpose is to estimate the parameters or to make prediction, it seems more plausible to take into account all (potentially) useful models, while the model selection approaches can be rather "risky" since they all force us to end up "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 estimates from different models. As an alternative to model selection, model averaging can substantially reduce risk (Hansen, 2014). It is an integrated process that takes both the model uncertainty and estimation uncertainty into account. 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 all 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 Mallow-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 is asymptotically optimal under some regularity conditions. Next, we replace the unknown covariance matrix by 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 model averaging for linear regression models. 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 our best 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.

We compare the proposed model averaging estimator with popular model selection and averaging estimators. Our simulation study considers two cases. In the first case, only the linear component is uncertain, and candidate models differ in the 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 be 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-criteria-based methods in the second case. We also apply our method to examine 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 ones, and we generally find a better prediction performance of 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 are given in the Appendix.

2 Model Averaging Estimation

2.1 Model and estimators

We consider the partially linear model (PLM)

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

where $(x_{i1}, x_{i2}, \ldots)$ is a countably infinite *non-random* vector, $\mathbf{Z}_i = (z_{i1}, \ldots, z_{iq})^{\mathrm{T}}$ is a *non-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 independent and (possibly) heteroscedastic random errors with $\mathrm{E}(\epsilon_i) = 0$ and $\mathrm{E}(\epsilon_i^2) = \sigma_i^2$. We denote the expectation of the response variable as $\mu_i = \mathrm{E}(y_i) = \sum_{j=1}^{\infty} x_{ij}\beta_j + g(\mathbf{Z}_i)$.

Our purpose 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, 2014).[1] For this purpose, we use S_n candidate PLMs to approximate (1), where S_n is allowed to diverge to infinity as $n \to \infty$. The s^{th} approximation (or candidate) PLM is

$$y_i = \mathbf{X}_{(s),i}^{\mathrm{T}} \boldsymbol{\beta}_{(s)} + g_{(s)}(\mathbf{Z}_{(s),i}) + b_{(s),i} + \epsilon_i, \quad i = 1 \ldots, n \tag{2}$$

where $\mathbf{X}_{(s),i}$ is a p_s-dimensional sub-vector of $(x_{i1}, x_{i2}, \ldots)^{\mathrm{T}}$ used in the linear component, $\mathbf{Z}_{(s),i}$ is a vector in the nonparametric component which can be different from $\mathbf{Z}_i$, $g_s(\cdot)$ 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 we consider two sources of uncertainty: the uncertainty of which variable to include in the model; and the uncertainy 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 $\mathbf{X}_{(s)} = (\mathbf{X}_{(s),1}, \ldots, \mathbf{X}_{(s),n})^{\mathrm{T}}$, $\mathbf{Z}_{(s)} = (\mathbf{Z}_{(s),1}, \ldots, \mathbf{Z}_{(s),n})^{\mathrm{T}}$, and $\mathbf{g}_{(s)} = \{g(\mathbf{Z}_{(s),1}), \ldots, g(\mathbf{Z}_{(s),n})\}^{\mathrm{T}}$.

To provide an optimal weighting scheme, we first need to estimate each candidate model. We follow Speckman (1988) to use kernel smoothing estimation. One of the advantages of this method is its light computation 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 $k_{h_s}(\cdot) = k(\cdot/h_s)/h_s$. Also, 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

$$\widehat{\boldsymbol{\beta}}_{(s)} = (\widetilde{\mathbf{X}}_{(s)}^{\mathrm{T}} \widetilde{\mathbf{X}}_{(s)})^{-1} \widetilde{\mathbf{X}}_{(s)}^{\mathrm{T}} (\mathbf{I}_n - \mathbf{K}_{(s)}) \mathbf{y}, \quad \widehat{\mathbf{g}}_{(s)} = \mathbf{K}_{(s)} (\mathbf{y} - \mathbf{X}_{(s)} \widehat{\boldsymbol{\beta}}_{(s)}),$$

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 $\boldsymbol{\mu}$ then follows as

$$\widehat{\boldsymbol{\mu}}_{(s)} = \mathbf{X}_{(s)} \widehat{\boldsymbol{\beta}}_{(s)} + \widehat{\mathbf{g}}_{(s)} = \widetilde{\mathbf{X}}_{(s)} (\widetilde{\mathbf{X}}_{(s)}^{\mathrm{T}} \widetilde{\mathbf{X}}_{(s)})^{-1} \widetilde{\mathbf{X}}_{(s)}^{\mathrm{T}} (\mathbf{I}_n - \mathbf{K}_{(s)}) \mathbf{y} + \mathbf{K}_{(s)} \mathbf{y}.$$

[1] Since the purpose of this paper is *not* to estimate the coefficients of linear component and unknown function of non-parametric component, 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).

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 curse of dimensionality, q_s (the dimension of $\mathbf{Z}_{(s)}$) cannot be large.

With estimators of each model readily there, we can obtain the model averaging estimator of $\boldsymbol{\mu}$ by

$$\widehat{\boldsymbol{\mu}}(\mathbf{w}) = \sum\nolimits_{s=1}^{S_n} w_s \widehat{\boldsymbol{\mu}}_{(s)} = \mathbf{P}(\mathbf{w})\mathbf{y},$$

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 $\mathbf{P}(\mathbf{w}) = \sum_{s=1}^{S_n} \mathbf{P}_{(s)}$.

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

$$R_n(\mathbf{w}) = \mathrm{E}\{L_n(\mathbf{w})\} = \|\mathbf{P}(\mathbf{w})\boldsymbol{\mu} - \boldsymbol{\mu}\|^2 + \mathrm{trace}\{\mathbf{P}(\mathbf{w})\boldsymbol{\Omega}\mathbf{P}^{\mathrm{T}}(\mathbf{w})\}, \tag{3}$$

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

$$C_n(\mathbf{w}) = \|\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \mathbf{y}\|^2 + 2\mathrm{trace}\{\mathbf{P}(\mathbf{w})\boldsymbol{\Omega}\}, \tag{4}$$

as we can show

$$R_n(\mathbf{w}) = \mathrm{E}\{C_n(\mathbf{w})\} - \mathrm{trace}(\boldsymbol{\Omega}),$$

where $\mathrm{trace}(\boldsymbol{\Omega})$ is unrelated to $\mathbf{w}$. Therefore, if we know $\boldsymbol{\Omega}$, the weights can be obtained by

$$\widehat{\mathbf{w}} = \mathrm{argmin}_{\mathbf{w}\in\mathcal{W}} C_n(\mathbf{w}). \tag{5}$$

Averaging using this weight choice is named Mallows averaging of partially linear models (MAPLM). The optimality of such a weight choice holds under some regularity conditions. Before we provide these conditions, some notations are required. 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 $i \in \{1, \ldots, n\}$, and all limiting properties here and throughout the text are under $n \to \infty$.

Condition (C.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* $s \in \{1, \ldots, S_n\}$.

Condition (C.2) *For some integer* $G \geq 1$, $\max_i E(\epsilon_i^{4G}) < \infty$ *and*

$$S_n \xi_n^{-2G} \sum\nolimits_{s=1}^{S_n} \{R_n(\mathbf{w}_s^o)\}^G \to 0.$$

Condition (C.1) is the same as assumption (i) of Speckman (1988) that bounds the kernel. Condition (C.2) requires $\xi_n \to \infty$, meaning that 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 (C.1)-(C.2),*

$$\frac{L_n(\widehat{\mathbf{w}})}{\inf_{\mathbf{w} \in \mathcal{W}} L_n(\mathbf{w})} \to 1 \tag{6}$$

in probability as $n \to \infty$.

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 Appendix A.1) takes advantage of several inequalities involving kernels, and it provides a technical innovation on how to study the optimal model averaging in a kernel smoothing framework.

So far we have assumed that the covariance matrix $\mathbf{\Omega}$ is known. This is, of course, not the case in practice, and the criterion (4) is therefore computationally infeasible. To have 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

$$\widehat{\mathbf{\Omega}}_{(s^*)} = \mathrm{diag}(\widehat{\epsilon}_{s^*,1}^2, \ldots, \widehat{\epsilon}_{s^*,n}^2), \tag{7}$$

where $(\widehat{\epsilon}_{s^*,1}, \ldots, \widehat{\epsilon}_{s^*,n})^{\mathrm{T}} = \mathbf{y} - \widehat{\boldsymbol{\mu}}_{(s^*)} = \mathbf{y} - \mathbf{P}_{(s^*)}\mathbf{y}$. We shall 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. 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).

Replacing $\mathbf{\Omega}$ with its estimator $\widehat{\mathbf{\Omega}}$, the feasible criterion is thus

$$\widehat{C}_n(\mathbf{w}) = \|\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \mathbf{y}\|^2 + 2\text{trace}\{\mathbf{P}(\mathbf{w})\widehat{\mathbf{\Omega}}_{(s^*)}\}, \tag{8}$$

and the weights can be obtained by

$$\widetilde{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathcal{W}} \widehat{C}_n(\mathbf{w}). \tag{9}$$

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 of Matlab that generally work effectively and efficiently even when S_n is large.

We now show that the weights obtained by minimizing the feasible criterion (8) are still asymptotic 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 $h = \min_s h_s$. Following conditions are prerequisites.

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

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

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

Condition (C.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 (C.4) is a natural extension of the condition (h) of Speckman (1988). Condition (C.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 (C.6) regards the bandwidth and is less restrictive than $n^{-1}h^{-2} = o(1)$ required in Theorem 2 of Speckman (1988). The second part of (C.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.

Theorem 2 *Under Conditions (C.1)-(C.6),*

$$\frac{L_n(\widetilde{\mathbf{w}})}{\inf_{\mathbf{w}\in\mathcal{W}} L_n(\mathbf{w})} \to 1 \tag{10}$$

in probability as $n \to \infty$.

Remark 1. It is a question how to choose an optimal bandwidth h_s in each candidate model. While this question is of interest, it is especially difficult in our case, because each candidate model is just an approximation to the true one with the approximation error. In our numerical examples, the bandwidth h_s is chosen by minimizing generalized cross-validation criterion. We also try different choices of h_s, and the results are qualitatively similar.

Remark 2. Theorem 2 holds no matter $\boldsymbol{\Omega}$ 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 $\boldsymbol{\Omega}$ 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 putting too much confidence in a single model. Using 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 on 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. In particular, 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}. Corresponding coefficients are set as $\beta_j = 1/j$. For simplicity, we consider the 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 $g_2(\mathbf{Z}_i) = 2(z_{i1} - 0.5)^3 + \sin(z_{i2})$. 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 $R^2 = \mathrm{var}\left(\mu_1, \ldots, \mu_n\right) / \mathrm{var}\left(y_1, \ldots, y_n\right)$ 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 at $n = 100$, 200, and 400.

In 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. First, it is a priori which variable is in 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 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 which covariates should be chosen as parametric regressors, and which should enter the nonparametric component. Therefore, in addition to the uncertainty of which variable to include, we are also uncertain whether a covariate should be 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 computation, we assume that only four covariates $(x_{i1}, x_{i2}, x_{i3}, z_{i1})$ are observed, while others are omitted. Different from the first case, candidate models here allow a subset of $(x_{i1}, x_{i2}, x_{i3}, z_{i1})$ in the nonparametric function, and the remaining can be in the linear component or not in the model at all. Again, we require each candidate model containing 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.

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).

To see how much harm it can cause by ignoring the nonlinearity, we compare our methods with linear model averaging and four alternatives for partially linear models. The linear model averaging considers all candidate models to be fully linear and with different observed covariates, and we average them by minimizing a standard heteroscedastic-robust Mallows criterion (HRCp, Liu & Okui, 2013). Four alternative estimation methods for PLMs including two selection methods and two averaging methods. Two model selection methods are based on AIC and BIC. They select the model with the smallest information criterion, defined respectively as

$$\text{AIC}_s = \log(\widehat{\sigma}_s^2) + 2n^{-1}\text{trace}(\mathbf{P}_{(s)}) \quad \text{and} \quad \text{BIC}_s = \log(\widehat{\sigma}_s^2) + n^{-1}\text{trace}(\mathbf{P}_{(s)})\log(n),$$

where $\widehat{\sigma}_s^2 = n^{-1}\|\mathbf{y} - \widehat{\boldsymbol{\mu}}_{(s)}\|^2$. Two model averaging methods are smoothed AIC (SAIC) and smoothed BIC (SBIC) (Buckland et al., 1997). The weight of model s is constructed by $\exp(-\text{AIC}_s/2)/\sum_{s=1}^{S}\exp(-\text{AIC}_s/2)$ for SAIC and $\exp(-\text{BIC}_s/2)/\sum_{s=1}^{S}\exp(-\text{BIC}_s/2)$ for SBIC.

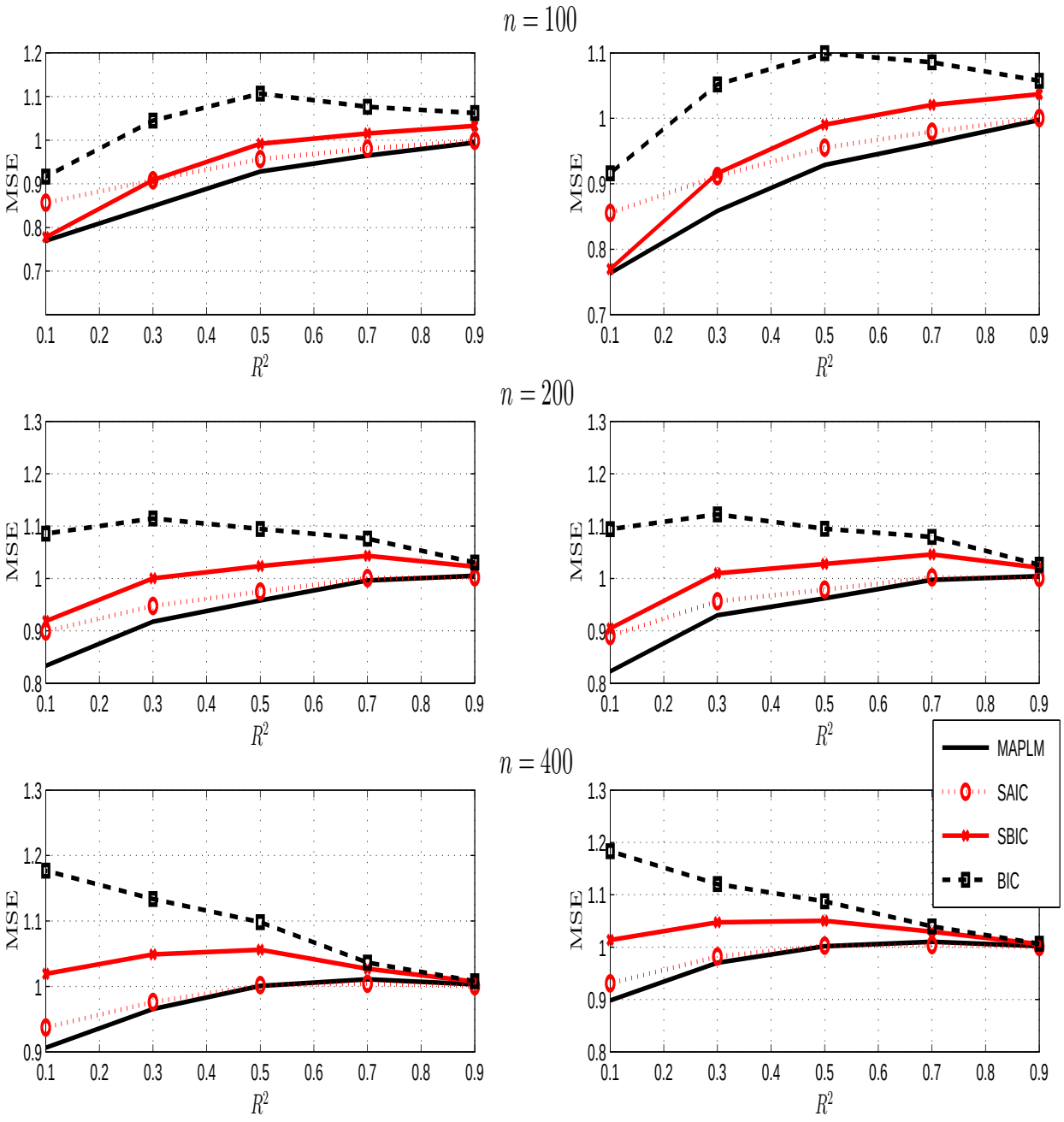
To evaluate these methods, we compute the mean squared error (MSE) of the predictive variable as $500^{-1}\sum_{r=1}^{500}\|\widehat{\boldsymbol{\mu}}^{(r)} - \boldsymbol{\mu}\|^2$, where 500 is the number of replications and $\widehat{\boldsymbol{\mu}}^{(r)}$ denotes the estimator of $\boldsymbol{\mu}$ in the r^{th} replication. For comparison convenience, all MSEs are normalized by dividing 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 we see that model averaging methods outperform selection ones. The superiority of averaging methods is particularly obvious when R^2 is small. As R^2 increases, the difference between model selection and averaging becomes smaller. The especially good performance of the averaging methods when R^2 is low and moderate is because identifying the best model is difficult in the presence of large noise. In that case the model chosen by a selection procedure can be far away from the best, which unsurprisingly leads to inaccurate estimates. On the contrary, model averaging does not rely on a single model, and thus shields 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 could be sometimes more preferred because little noise in the data allows the selection criterion to correctly pick up the right model.

Figure 1 presents the results when there is only uncertainty in the linear component specification. Our method yields the smallest MSE in almost all cases, except that information-criterion model averaging sometimes have a marginal advantage over ours when R^2 is very large. Most figures show that the advantage of our method becomes more prominent as R^2 decreases. The good performance of MAPLM is partly because

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

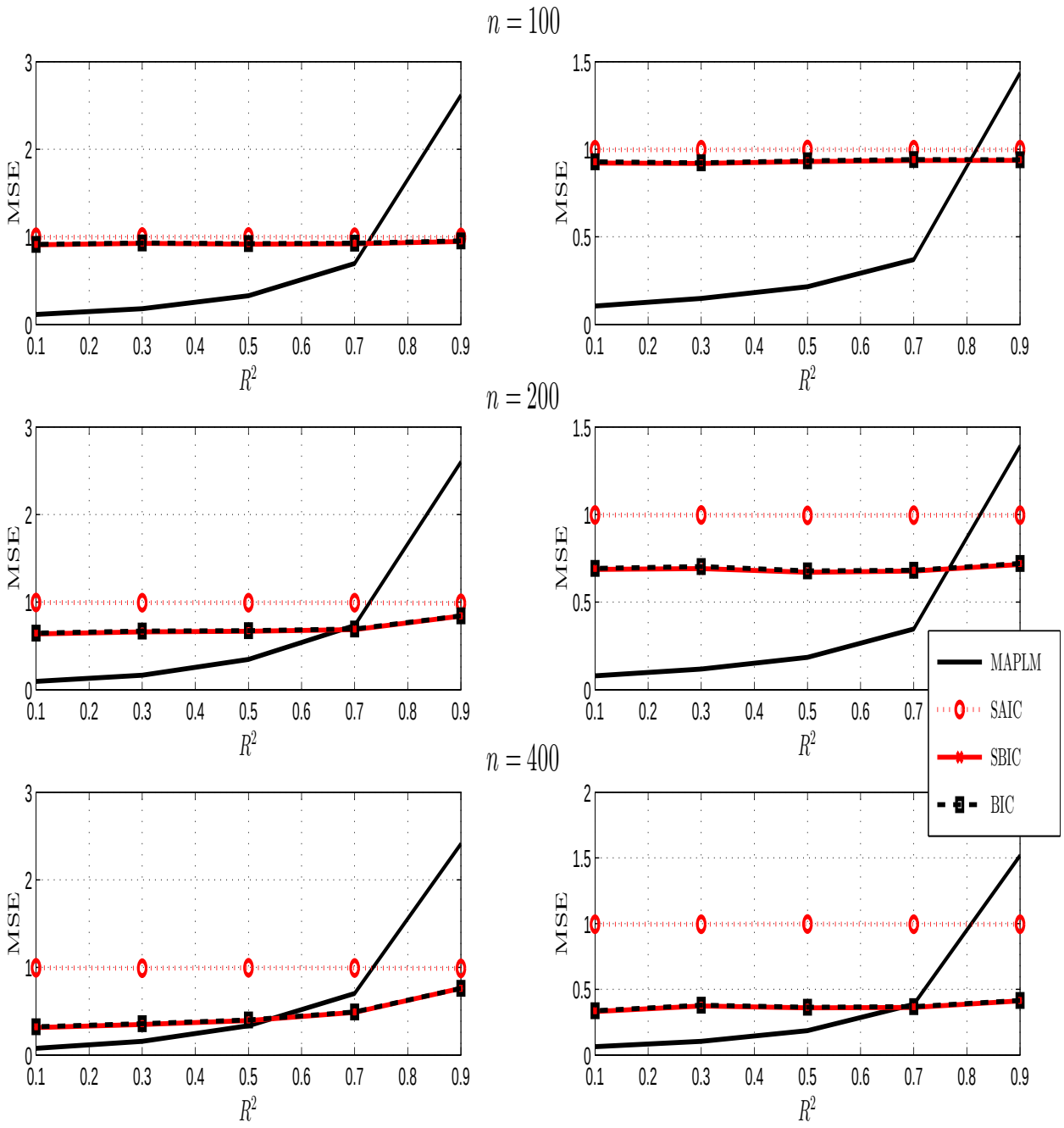


Notes: Figures in the left column are under $g_1(z_1, z_2) = \exp(z_1) + z_2^2$ and figures in the right column are under $g_2(z_1, z_2) = 2(z_1 - 0.5)^3 + \sin(z_2)$.

the optimality of MAPLM does not rely on the correct specification of candidate models. Comparing methods in different sample sizes, we find that when we have a relatively small and moderate sample ($n = 100$ and 200), only SAIC marginally outperforms MAPLM when R^2 is particularly large (over 0.9). When the sample size is large ($n = 400$), MAPLM still dominates other methods for a wide range of R^2, but the difference between MAPLM and SAIC becomes smaller. The latter even produces the least MSE when R^2 is larger than 0.7. We also note that all methods almost perform equally well when the sample size is large and R^2 is 0.9. Further examination suggests that all 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 can be partly explained by the fact that the bias-variance tradeoff is not so significant in this situation such that model selection is able to pick up the right model.

Next, Figure 2 compares estimation results when there is structure uncertainty, in addition to uncertainty in covariate inclusion. In this case both linear and nonparametric components vary over candidate models. We see that MAPLM produces lower MSE than its rivals in all cases when R^2 is less than 0.7. This again demonstrates that our model averaging approach is preferred when the model is characterized by much noise and identifying the best model is difficult, as in most practical applications. Model selection and averaging using AIC and SAIC lead to largely similar results, and so do BIC and SBIC. Further examination shows that there is always a dominant model (usually the model with only one nonparametric component) receiving much lower AIC than other candidate models, and thus selection and averaging are almost equivalent. This is also true for BIC. The nearly constant relationship between four information-criteria based methods is due to the fact that the variation in difference is relatively small compared to the size of MSE.

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



Notes: Figures in the left column are under $g_1(z_1, z_2) = \exp(z_1) + z_2^2$ and figures in the right column are under $g_2(z_1, z_2) = 2(z_1 - 0.5)^3 + \sin(z_2)$.

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 on 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 over the world. Also, Japan is the world's third largest economy with its financial market playing an important role in international finance, and a crisis in Japan can damage investors' confidence on 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 we study the predictability of these indicators. We focus on the CDS contract written on the credit event "complete restructuring", as this 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 the domestic economic performance: the domestic stock market return (measured by Dow Jones Japan Total Stock Market Total Return Index) and its volatility, and the nominal Yen-US Dollar exchange rate. We also follow Longstaff et al. (2011) to consider three global-market determinants: the global stock market return (measured by 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. (2014) for details of variable construction. We focus on the post-earthquake sample from March 12, 2011 (one day after 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.

4.1 Linear model specification

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

Table 1: Estimation results of linear models

	OLS	HRCp
Domestic stock returns	-1.5182^{***}	-1.2790
	(0.1752)	
Domestic stock volatility	0.6165^{***}	0.0576
	(0.1758)	
Foreign exchange rate	-0.3250^{*}	-0.3777
	(0.1727)	
Global returns	1.0107^{***}	0.9842
	(0.1733)	
US treasury yield	-0.3672^{**}	-0.3649
	(0.1750)	
global default risk premium	-0.0774	-0.0230
	(0.1689)	

Notes: Standard errors in parentheses. ***, **, * denote significance at 1%, 5%, and 10%, respectively.

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

the least square estimation results. The least square estimates show that the domestic stock return, its volatility, and the global stock return are the three most important determinants with a significant effect on Japan's CDS spread. More particularly, the domestic stock return, as a measure of local economic performance, has a strongly negative effect. It can affect the CDS spread by influencing government's willingness to take fiscal reforms, and an effective fiscal reform is typically regarded as an important tool of reducing default risk. Therefore, when the domestic economy is weak, policy makers are less willing to implement the reforms, because reforms can impose extra pressure on the distressed economy. This 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. This is in line with the economic theory that higher volatility indicates a less stable economic status and thus the probability of default is larger. 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 the sovereign CDS spreads. The negative effect is due to the fact that good global economic performance can positively influence Japanese government's willingness to repay, and thus lower the sovereign CDS spreads. On the other hand, a good global economy would also encourage investment in general, and hence increase the CDS spreads. The overall impact of the global stock return depends on which effect is dominant. It is very likely that one effect is more prominent in some situation but dominated by the other effect in a different situation. Such potential heterogeneity cannot be captured by the 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 the weakness of Japan's current economic situation and less external demand, which further leads to higher sovereign CDS spreads. 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 US, which can positively influence Japan's economy and further encourage Japanese government to repay.

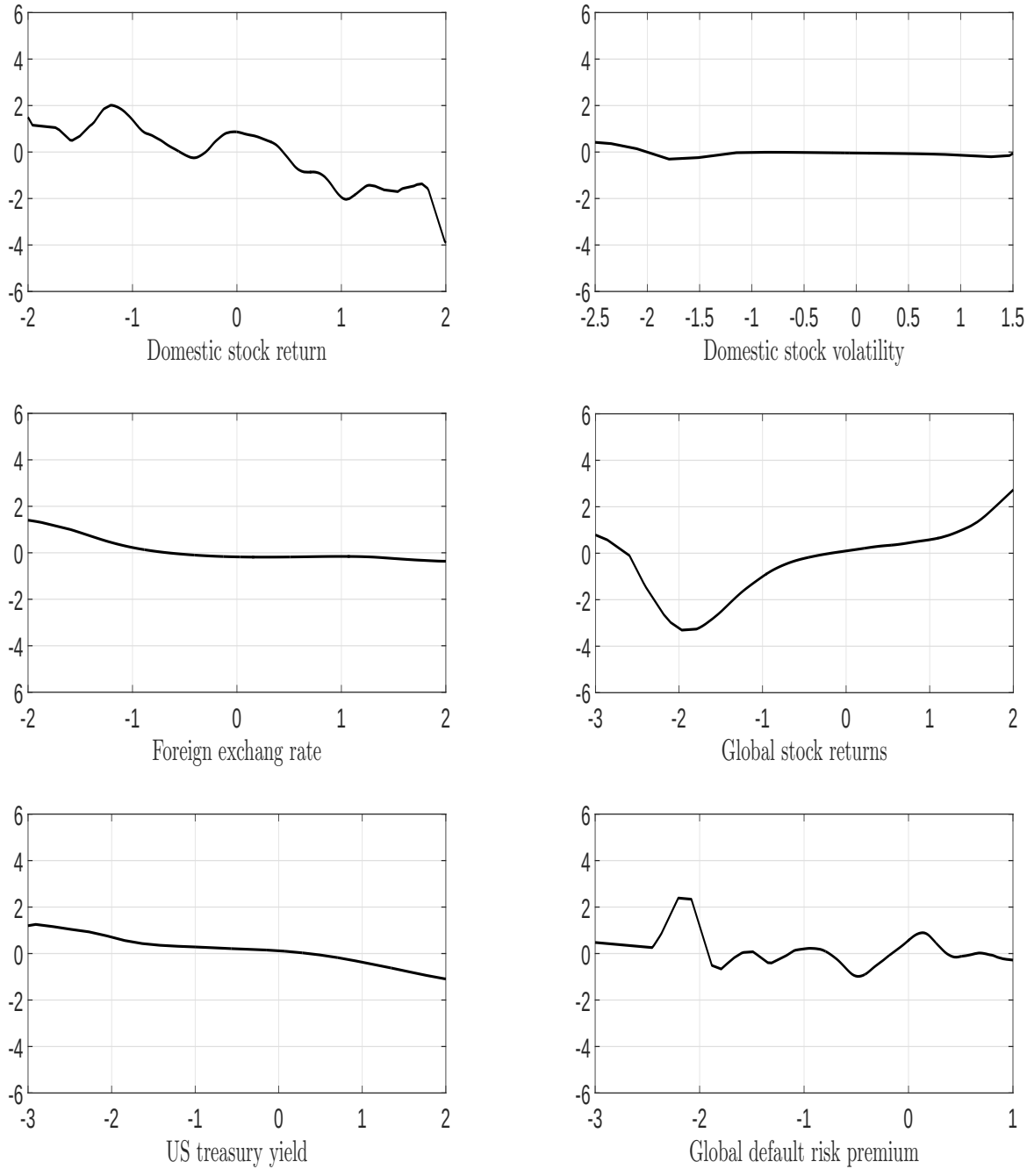
We then compare the estimates obtained from the least square and model averaging. We find that the signs of all estimated coefficients are the same for both methods. Nevertheless, model averaging produces quite different estimates for some determinants, such as the domestic stock return, its volatility, 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. We verify the linearity of each determinant by assigning it in 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 facilitates us to clearly verify whether each determinant has a nonlinear effect on Japan's CDS spreads, and also avoids the dimensionality and computational issue of having many nonparametric covariates at the same time.

Figure 3 presents the nonparametric estimates of each determinant using the proposed MAPLM. We see that the effects of the domestic stock market return, its 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, the foreign exchange rate, global stock returns, and US treasury yield show different degrees of nonlinearity. We also 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 linearity of the nonparametric component by the fiducial method. 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 are in the linear component. Therefore, no averaging takes place in this testing procedure. The p-values of the tests are reported in Table 2. We see that the test accepts the null hypothesis of linearity for domestic stock return, its volatility, and global default risk premium. The reported p-values also confirm that the effects of the foreign

Figure 3: Nonparametric estimation for each macroeconomic determinant



exchange rate and global stock returns cannot be well approximated by linear functions. The test statistic for US treasury yield is not available because this variable only takes a few discrete values. Thus it is less clear whether one can assume a linear effect of US treasury yield.

Table 2: Linearity test for each determinant

	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 results of nonparametric estimation and regression diagnostics, we carefully 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 of exchange rates is below average, but the curve is relatively flat and close to zero as the change increases. The negative relationship between the exchange rate and Japan's CDS spread is in line with the findings of 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 the typical "U-shape". We see that the change of Japan's CDS spreads is particularly high when global returns are in the extreme, either a large positive change or a large negative change. This suggests 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 boom. 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 the "crisis" period. This result provides evidence of financial contagion from the global stock market to Japan's sovereign credit

market. 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 made under different situations. Such financial contagion cannot be captured by the linear models. Last but not the least, the curve of US treasury yield is similar but less nonlinear compared to that of the foreign exchange rate. We generally observe a negative effect of 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 of treasury yield is extreme.

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 (MAPLM, SAIC, and SBIC) and two model selection (AIC and BIC) methods for partially linear models, and one linear model averaging method.

The linear model averaging is based on the HRCp as above. It considers candidate models with at least one determined included, so that it averages over $2^6 - 1$ candidate models. For PLM averaging, the most general specification is of course to consider all possibilities that a determinant can be in the linear component, in the nonlinear component, or not in the model. However, this may cause a dimensionality problem by including too many determinants in the nonlinear component. Thus we assign determinants in 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, domestic stock market return and its volatility. It is also clear that the foreign exchange rate and global stock return have a nonlinear impact on Japan's CDS spread, and thus it seems necessary to include these two determinants in the nonlinear component if they are in the model. As for US treasury yield, since its effect only exhibits a moderate degree of nonlinearity and the formal linearity test is not informative for it, we are less certain whether to assign this variable in the linear or nonlinear component.

Table 3: 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

Allowing this ambiguous determinant to possibly enter the nonlinear component leads to a more complete model space, but may also suffer from the dimensionality curse. There is not a *priori* how one makes an appropriate tradeoff between a more complete model space or the dimensionality curse. Therefore, we compare the prediction performance of six methods in two scenarios. In Scenario I, we only allow the foreign exchange rate and global stock return to possibly be in the nonlinear component. In other words, the foreign exchange rate and global stock return can either be not included by the model or 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 only in that we also allow US treasury yield to possibly enter the nonlinear component, in addition to the foreign exchange rate and global stock return. Hence there are three possibilities for this uncertain determinant: not included by the model, in the linear component, or the nonlinear component of the model. We consider prediction samples ranging from 20% to 5% of the entire sample.

Table 3 presents the mean square prediction error (MSPE) of five PLM methods. All values are normalized by dividing 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 of the cases except when the prediction sample is 15%. In all cases, MAPLM outperforms the linear Mallows averaging. This demonstrates that incorporating the necessary nonlinearity does improve the prediction performance. Since the performance of linear model averaging is invariant to scenarios, we can also compare the predictability of MAPLM in two scenarios. Interesting, we see that allowing US treasury yield to possibly enter the nonlinear component improves the prediction performance for all methods if the prediction sample is larger than 5%. However, when we have a small prediction sample, considering a smaller model space is indeed better. One possible explanation is that averaging over a larger model space may offset more noise by better diversifying. When the prediction sample is large, the diversification gain from averaging over a larger model space is sizable, which 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 become more accurate and closer with each, and thus the diversification gain is less.

5 Concluding remarks

Partially linear models have become widely popular in applied econometrics since they allow a more flexible specification compared to the linear models and provide more interpretable estimates compared to the fully nonparametric models. Estimation of partially linear models is subjected to at least two types of uncertainty, the uncertainty on which variables to include in the model and the uncertainty on whether a covariate should be in the linear or nonlinear component given that it is in the model. Typical model testing or selection methods cannot properly address these two types of uncertainty simultaneously, especially if the research interest is to estimate the parameters or to make prediction. In this paper, we propose an optimal model averaging procedure for PLMs that can jointly incorporate the two types of model uncertainty. 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 does exist a large degree of nonlinearity in the effect of macroeconomic determinants, such as the global stock return and exchange rate. Conventional linear models cannot capture such nonlinearity, and ignoring the nonlinearity can cause unawareness of financial contagion, which may further lead to inappropriate policies and investment decisions.

At least two issues deserve future research. First, the computational burden of our model averaging method would be substantial when the number of candidate models is large. In this regard, 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 has to be smaller than n and its increasing rate is also restricted by the second part of Condition (C.6). How to develop optimal model averaging method for high or ultrahigh dimensional PLMs is an interesting open question for future studies.

Appendices

A.1 Proof of Theorem 1

Denote the largest singular value of a matrix $\mathbf{A}$ by $\lambda_{\max}(\mathbf{A})$. From the first part of Condition (C.2), we have

$$\lambda_{\max}(\mathbf{\Omega}) = O(1). \tag{A.1}$$

Using (A.1), transformation $\boldsymbol{\epsilon}^* = \mathbf{\Omega}^{-1/2}\boldsymbol{\epsilon}$,Condition (C.2) and the proof Theorem 1' of Wan et al. (2010), in order to prove (6), we need only to further verify that

$$\max_s\{\lambda_{\max}(\mathbf{P}_s)\} = O(1) \quad \text{and} \quad \max_s\{\lambda_{\max}(\mathbf{P}_{(s)}\mathbf{P}_{(s)}^{\mathrm{T}})\} = O(1). \tag{A.2}$$

By an inequality of Reisz (see Hardy et al. (1952) or Speckman (1988)), we know that

$$\lambda_{\max}^2(\mathbf{K}_{(s)}) \leq \max_i \sum\nolimits_{j=1}^{n} |K_{(s),ij}| \max_j \sum\nolimits_{i=1}^{n} |K_{(s),ij}|. \tag{A.3}$$

In addition, it is well known that for any two $n \times n$ matrices $\mathbf{B}_1$ and $\mathbf{B}_2$ (see, for example, Li (1987))

$$\lambda_{\max}(\mathbf{B}_1\mathbf{B}_2) \leq \lambda_{\max}(\mathbf{B}_1)\lambda_{\max}(\mathbf{B}_2) \quad \text{and} \quad \lambda_{\max}(\mathbf{B}_1+\mathbf{B}_2) \leq \lambda_{\max}(\mathbf{B}_1)+\lambda_{\max}(\mathbf{B}_2). \tag{A.4}$$

From (A.4) and $\lambda_{\max}(\widetilde{\mathbf{P}}_{(s)}) = 1$, we obtain that for $1 \leq s \leq S_n$

$$\begin{aligned}
\lambda_{\max}(\mathbf{P}_{(s)}\mathbf{P}_{(s)}^{\mathrm{T}}) &\leq \lambda_{\max}^2(\mathbf{P}_{(s)}) \\
&= \lambda_{\max}^2\{\widetilde{\mathbf{P}}_{(s)}(\mathbf{I}_n - \mathbf{K}_{(s)}) + \mathbf{K}_{(s)}\} \\
&\leq [\lambda_{\max}(\widetilde{\mathbf{P}}_{(s)})\{1+\lambda_{\max}(\mathbf{K}_{(s)})\} + \lambda_{\max}(\mathbf{K}_{(s)})]^2 \\
&= [\{1+\lambda_{\max}(\mathbf{K}_{(s)})\} + \lambda_{\max}(\mathbf{K}_{(s)})]^2, \qquad \text{(A.5)}
\end{aligned}$$

which, together with (A.3) and Condition (C.1), implies (A.2). This completes the proof.

A.2 Proof of Theorem 2

Note that

$$\widehat{C}_n(\mathbf{w}) = C_n(\mathbf{w}) + \mathrm{trace}\{\mathbf{P}(\mathbf{w})\widehat{\boldsymbol{\Omega}}_{(s^*)}\} - \mathrm{trace}\{\mathbf{P}(\mathbf{w})\boldsymbol{\Omega}\}.$$

Hence, from the proof of Theorem 1, in order to prove (6), we only need to verify that

$$\sup_{\mathbf{w}\in\mathcal{W}}[|\mathrm{trace}\{\mathbf{P}(\mathbf{w})\widehat{\boldsymbol{\Omega}}_{(s^*)}\} - \mathrm{trace}\{\mathbf{P}(\mathbf{w})\boldsymbol{\Omega}\}|/R_n(\mathbf{w})] = o_p(1). \tag{A.6}$$

Let $\mathbf{Q}_{(s)} = \mathrm{diag}(\rho_{11}^{(s)}, \ldots, \rho_{nn}^{(s)})$ and $\mathbf{Q}(\mathbf{w}) = \sum_{s=1}^{S_n} w_s \mathbf{Q}_{(s)}$. Then, from (7), we have

$$\begin{aligned}
& \sup_{\mathbf{w}\in\mathcal{W}}[|\mathrm{trace}\{\mathbf{P}(\mathbf{w})\widehat{\boldsymbol{\Omega}}_{(s^*)}\} - \mathrm{trace}\{\mathbf{P}(\mathbf{w})\boldsymbol{\Omega}\}|/R_n(\mathbf{w})] \\
= & \sup_{\mathbf{w}\in\mathcal{W}}[|(\mathbf{y}-\mathbf{P}_{(s^*)}\mathbf{y})^{\mathrm{T}}\mathbf{Q}(\mathbf{w})(\mathbf{y}-\mathbf{P}_{(s^*)}\mathbf{y}) - \mathrm{trace}\{\mathbf{Q}(\mathbf{w})\boldsymbol{\Omega}\}|/R_n(\mathbf{w})] \\
= & \sup_{\mathbf{w}\in\mathcal{W}}[|(\boldsymbol{\epsilon}+\boldsymbol{\mu}-\mathbf{P}_{(s^*)}\boldsymbol{\mu}-\mathbf{P}_{(s^*)}\boldsymbol{\epsilon})^{\mathrm{T}}\mathbf{Q}(\mathbf{w})(\boldsymbol{\epsilon}+\boldsymbol{\mu}-\mathbf{P}_{(s^*)}\boldsymbol{\mu}-\mathbf{P}_{(s^*)}\boldsymbol{\epsilon}) \\
& -\mathrm{trace}\{\mathbf{Q}(\mathbf{w})\boldsymbol{\Omega}\}|/R_n(\mathbf{w})] \\
\leq & \sup_{\mathbf{w}\in\mathcal{W}}[|\boldsymbol{\epsilon}^{\mathrm{T}}(\mathbf{I}_n-\mathbf{P}_{(s^*)})^{\mathrm{T}}\mathbf{Q}(\mathbf{w})(\mathbf{I}_n-\mathbf{P}_{(s^*)})\boldsymbol{\epsilon} \\
& -\mathrm{trace}\{(\mathbf{I}_n-\mathbf{P}_{(s^*)})^{\mathrm{T}}\mathbf{Q}(\mathbf{w})(\mathbf{I}_n-\mathbf{P}_{(s^*)})\boldsymbol{\Omega}\}|/R_n(\mathbf{w})] \\
& +2\sup_{\mathbf{w}\in\mathcal{W}}[|\boldsymbol{\epsilon}^{\mathrm{T}}(\mathbf{I}_n-\mathbf{P}_{(s^*)})^{\mathrm{T}}\mathbf{Q}(\mathbf{w})(\mathbf{I}_n-\mathbf{P}_{(s^*)})\boldsymbol{\mu}|/R_n(\mathbf{w})] \\
& +\sup_{\mathbf{w}\in\mathcal{W}}[|\boldsymbol{\mu}^{\mathrm{T}}(\mathbf{I}_n-\mathbf{P}_{(s^*)})^{\mathrm{T}}\mathbf{Q}(\mathbf{w})(\mathbf{I}_n-\mathbf{P}_{(s^*)})\boldsymbol{\mu}|/R_n(\mathbf{w})] \\
& +\sup_{\mathbf{w}\in\mathcal{W}}[|\mathrm{trace}(\mathbf{P}_{(s^*)}^{\mathrm{T}}\mathbf{Q}(\mathbf{w})\mathbf{P}_{(s^*)}\boldsymbol{\Omega})|/R_n(\mathbf{w})]
\end{aligned}$$

$$
\begin{aligned}
&+2\sup_{\mathbf{w}\in\mathcal{W}}[|\mathrm{trace}(\mathbf{P}_{(s^*)}^{\mathrm{T}}\mathbf{Q}(\mathbf{w})\boldsymbol{\Omega})|/R_n(\mathbf{w})] \\
\equiv\ & \Xi_1+\Xi_2+\Xi_3+\Xi_4+\Xi_5. \qquad (A.7)
\end{aligned}
$$

Define $\rho=\max_s\max_i|\rho_{ii}^{(s)}|$. From (A.3), (A.4), and Conditions (C.4)-(C.5), we have

$$
\begin{aligned}
\rho &\le cn^{-1}\max_s\{|\mathrm{trace}(\mathbf{P}_{(s)})|\} \\
&\le cn^{-1}\max_s\{|\mathrm{trace}(\widetilde{\mathbf{P}}_{(s)})-\mathrm{trace}(\widetilde{\mathbf{P}}_{(s)}\mathbf{K}_{(s)})|\}+cn^{-1}\max_s|\mathrm{trace}(\mathbf{K}_{(s)})| \\
&\le cn^{-1}\max_s|\mathrm{trace}(\widetilde{\mathbf{P}}_{(s)})|+cn^{-1}\max_s|\mathrm{trace}(\widetilde{\mathbf{P}}_{(s)}\mathbf{K}_{(s)})|+cn^{-1}\max_s|\mathrm{trace}(\mathbf{K}_{(s)})| \\
&= cn^{-1}\widetilde{p}+cn^{-1}2^{-1}\max_s|\mathrm{trace}(\widetilde{\mathbf{P}}_{(s)}\mathbf{K}_{(s)}+\mathbf{K}_{(s)}^{\mathrm{T}}\widetilde{\mathbf{P}}_{(s)})|+cn^{-1}\max_s|\mathrm{trace}(\mathbf{K}_{(s)})| \\
&\le cn^{-1}\widetilde{p}+cn^{-1}2^{-1}\max_s\{\lambda_{\max}(\widetilde{\mathbf{P}}_{(s)}\mathbf{K}_{(s)}+\mathbf{K}_{(s)}^{\mathrm{T}}\widetilde{\mathbf{P}}_{(s)})\mathrm{rank}(\widetilde{\mathbf{P}}_{(s)}\mathbf{K}_{(s)}+\mathbf{K}_{(s)}^{\mathrm{T}}\widetilde{\mathbf{P}}_{(s)})\} \\
&\quad +cn^{-1}\max_s|\mathrm{trace}(\mathbf{K}_{(s)})| \\
&\le cn^{-1}\widetilde{p}+cn^{-1}2\max_s\{p_s\lambda_{\max}(\widetilde{\mathbf{P}}_{(s)})\lambda_{\max}(\mathbf{K}_{(s)})\}+cn^{-1}\max_s|\mathrm{trace}(\mathbf{K}_{(s)})| \\
&= O(n^{-1}\widetilde{p}+n^{-1}h^{-1}). \qquad (A.8)
\end{aligned}
$$

It follows from (3) and Condition (C.2) that

$$
\xi_n\to\infty,\ S_n\xi_n^{-2G}=o(1),\quad \text{and}\quad \xi_n^{-2}\|\mathbf{P}_{(s^*)}\boldsymbol{\mu}-\boldsymbol{\mu}\|^2=o(1). \qquad (A.9)
$$

Using (A.1), (A.2), (A.8), Chebyshev's inequality, and Theorem 2 of Whittle (1960), we can obtain that, for any $\delta>0$,

$$
\begin{aligned}
\Pr(\Xi_1>\delta) &\le \sum\nolimits_{s=1}^{S_n}\Pr[|\boldsymbol{\epsilon}^{\mathrm{T}}(\mathbf{I}_n-\mathbf{P}_{(s^*)})^{\mathrm{T}}\mathbf{Q}_{(s)}(\mathbf{I}_n-\mathbf{P}_{(s^*)})\boldsymbol{\epsilon} \\
&\qquad -\mathrm{trace}\{(\mathbf{I}_n-\mathbf{P}_{(s^*)})^{\mathrm{T}}\mathbf{Q}_{(s)}(\mathbf{I}_n-\mathbf{P}_{(s^*)})\boldsymbol{\Omega}\}|>\delta\xi_n] \\
&\le \delta^{-2G}\xi_n^{-2G}\sum\nolimits_{s=1}^{S_n}E[\boldsymbol{\epsilon}^{\mathrm{T}}(\mathbf{I}_n-\mathbf{P}_{(s^*)})^{\mathrm{T}}\mathbf{Q}_{(s)}(\mathbf{I}_n-\mathbf{P}_{(s^*)})\boldsymbol{\epsilon} \\
&\qquad -\mathrm{trace}\{(\mathbf{I}_n-\mathbf{P}_{(s^*)})^{\mathrm{T}}\mathbf{Q}_{(s)}(\mathbf{I}_n-\mathbf{P}_{(s^*)})\boldsymbol{\Omega}\}]^{2G} \\
&\le c_1\delta^{-2G}\xi_n^{-2G}\sum\nolimits_{s=1}^{S_n}\mathrm{trace}^G\{\boldsymbol{\Omega}^{1/2}(\mathbf{I}_n-\mathbf{P}_{(s^*)})^{\mathrm{T}}\mathbf{Q}_{(s)}(\mathbf{I}_n-\mathbf{P}_{(s^*)})\boldsymbol{\Omega} \\
&\qquad \times(\mathbf{I}_n-\mathbf{P}_{(s^*)})^{\mathrm{T}}\mathbf{Q}_{(s)}(\mathbf{I}_n-\mathbf{P}_{(s^*)})\boldsymbol{\Omega}^{1/2}\} \\
&\le c_1\delta^{-2G}\xi_n^{-2G}\lambda_{\max}^{4G}(\mathbf{I}_n-\mathbf{P}_{(s^*)})\lambda_{\max}^{2G}(\boldsymbol{\Omega})n^G\rho^{2G}S_n \\
&= \xi_n^{-2G}S_n\{O(n^{-1}\widetilde{p}^2+n^{-1}h^{-2})\}^G, \qquad (A.10)
\end{aligned}
$$

where c_1 is a positive constant and G is the integer defined in Condition (C.2). It follows from (A.9)–(A.10) and Condition (C.6) that $\Xi_1=o_p(1)$.

Using (A.1), (A.2), (A.4), (A.8) and (A.9), we have

$$
\Xi_2 \le 2\xi_n^{-1}\|(\mathbf{I}_n-\mathbf{P}_{(s^*)})\boldsymbol{\mu}\|\sup_{\mathbf{w}\in\mathcal{W}}\|\mathbf{Q}(\mathbf{w})(\mathbf{I}_n-\mathbf{P}_{(s^*)})\boldsymbol{\epsilon}\|
$$

$$
\begin{aligned}
&\leq 2\xi_n^{-1}\|(\mathbf{I}_n - \mathbf{P}_{(s^*)})\boldsymbol{\mu}\| \sup_{\mathbf{w}\in\mathcal{W}}\{\rho\|(\mathbf{I}_n - \mathbf{P}_{(s^*)})\boldsymbol{\epsilon}\| \\
&\leq 2\xi_n^{-1}\|(\mathbf{I}_n - \mathbf{P}_{(s^*)})\boldsymbol{\mu}\|\rho\{1+\lambda_{\max}(\mathbf{P}_{(s^*)})\}\|\boldsymbol{\epsilon}\| \\
&= o(1)O(n^{-1/2}\widetilde{p} + n^{-1/2}h^{-1}), \qquad \text{(A.11)}
\end{aligned}
$$

which, along with Condition (C.6), implies that $\Xi_2 = o_p(1)$.

Using (A.2), (A.4), (A.8), (A.9) and Condition (C.3), we have

$$
\begin{aligned}
\Xi_3 &\leq \xi_n^{-1}\rho\|(\mathbf{I}_n - \mathbf{P}_{(s^*)})\boldsymbol{\mu}\|^2 \\
&\leq \xi_n^{-1}\|(\mathbf{I}_n - \mathbf{P}_{(s^*)})\boldsymbol{\mu}\|\rho\|\boldsymbol{\mu}\|\{1+\lambda_{\max}(\mathbf{P}_{(s^*)})\} \\
&= o(1)O(n^{-1/2}\widetilde{p} + n^{-1/2}h^{-1}), \qquad \text{(A.12)}
\end{aligned}
$$

which, along with Condition (C.6), implies that $\Xi_3 = o(1)$.

Using (A.1), (A.2) and (A.4), we have

$$
\begin{aligned}
\Xi_4 + \Xi_5 &\leq \xi_n^{-1}\mathrm{rank}(\mathbf{P}_{(s^*)}) \sup_{\mathbf{w}\in\mathcal{W}}[\lambda_{\max}\{\mathbf{P}_{(s^*)}^{\mathrm{T}}\mathbf{Q}(\mathbf{w})\mathbf{P}_{(s^*)}\boldsymbol{\Omega}\}] \\
&\quad + 2\xi_n^{-1}\mathrm{rank}(\mathbf{P}_{(s^*)}) \sup_{\mathbf{w}\in\mathcal{W}}[\lambda_{\max}\{\mathbf{P}_{(s^*)}^{\mathrm{T}}\mathbf{Q}(\mathbf{w})\boldsymbol{\Omega}\}] \\
&\leq \xi_n^{-1}\widetilde{p}\rho\lambda_{\max}^2(\mathbf{P}_{(s^*)})\lambda_{\max}(\boldsymbol{\Omega}) + 2\xi_n^{-1}\widetilde{p}\rho\lambda_{\max}(\mathbf{P}_{(s^*)})\lambda_{\max}(\boldsymbol{\Omega}) \\
&= \xi_n^{-1}O(n^{-1}\widetilde{p}^2 + n^{-1}h^{-1}\widetilde{p}), \qquad \text{(A.13)}
\end{aligned}
$$

which, along with (A.9) and Condition (C.6), implies that $\Xi_4 + \Xi_5 = o(1)$. Therefore, we can get (A.6). This completes the proof.

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.

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.

Dieckmann, S. & Plank, T. (2011). Default risk of advanced economies: An empirical analysis of credit default swaps during the financial crisis. *Review of Finance* **0**, 1–32.

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, D., Linton, O. & Lu, Z. (forthcoming). A flexible semiparametric forecasting model for time series. *Journal of Econometrics* .

Li, K.-C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set. *The Annals of Statistics* **15**, 958–975.

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

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. (forthcoming). Weighted average least square prediction. *Econometric Reviews* .

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. (2014). Sovereign credit risk, macroeconomic dynamics, and financial contagion: Evidence from Japan. *Working Paper* .

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.

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

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