OPTIMAL MODEL AVERAGING ESTIMATION FOR PARTIALLY LINEAR MODELS

Xinyu Zhang[1,2] and Wendun Wang[3]

[1]*School of Mathematics and Statistics, Qingdao University*

[2]*Academy of Mathematics and Systems Science, Chinese Academy of Sciences*

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

Supplementary Material

This online supplement contains two parts: the first part is Appendices which provide the proofs of Theorems 1 and 2; and the second part consists of four sections with additional theoretical and simulation results. Section S3 provides further explanations on the conditions for the theorems. Section S4 provides the detailed proof of Theorem 1. Section S5 presents additional simulation studies, including the comparison with linear model averaging, data generation process with a diverging number of candidate models, and the case with autoregressive errors. Section S6 presents additional figures and tables for the empirical application.

Part 1: Appendices

S1 Proof of Theorem 1

For convenience purposes, similar to the proof in Zhang et al. (2013), we treat $\mathbf{X}$ and $\mathbf{Z}$ as non-random throughout the Appendix. Allowing for randomness would not invalidate the proof, because all our technical conditions hold almost surely.

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

$$\lambda_{\max}(\boldsymbol{\Omega}) = O(1). \tag{S1.1}$$

The proof of (2.6) is similar to that of Theorem 1' in Wan et al. (2010). Following their steps and using (S1.1), transformation $\boldsymbol{\epsilon}^* = \boldsymbol{\Omega}^{-1/2}\boldsymbol{\epsilon}$, and Condition 2, the only argument that we need to verify to complete the proof is that

$$\max_s\{\lambda_{\max}(\mathbf{P}_s)\} = O(1) \quad \text{and} \quad \max_s\{\lambda_{\max}(\mathbf{P}_{(s)}\mathbf{P}_{(s)}^{\mathrm{T}})\} = O(1). \tag{S1.2}$$

Therefore, in this Appendix, we only focus on the proof of (S1.2), but we provide a detailed proof of Theorem 1 in Section S.2 of this supplement.

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

$$\lambda_{\max}^2(\mathbf{K}_{(s)}) \le \max_i \sum\nolimits_{j=1}^n |K_{(s),ij}| \max_j \sum\nolimits_{i=1}^n |K_{(s),ij}|. \tag{S1.3}$$

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))

$$\left\{\begin{array}{c} \lambda_{\max}(\mathbf{B}_1\mathbf{B}_2) \leq \lambda_{\max}(\mathbf{B}_1)\lambda_{\max}(\mathbf{B}_2) \\ \\ \lambda_{\max}(\mathbf{B}_1+\mathbf{B}_2) \leq \lambda_{\max}(\mathbf{B}_1)+\lambda_{\max}(\mathbf{B}_2) \end{array}\right. . \tag{S1.4}$$

From (S1.4) and $\lambda_{\max}(\widetilde{\mathbf{P}}_{(s)}) = 1$, we obtain that for $1 \leq s \leq S_n$

$$\begin{aligned} \lambda_{\max}(\mathbf{P}_{(s)}\mathbf{P}_{(s)}^{\mathrm{T}}) &\leq \lambda_{\max}^2(\mathbf{P}_{(s)}) \\ &= \lambda_{\max}^2\{\widetilde{\mathbf{P}}_{(s)}(\mathbf{I}_n - \mathbf{K}_{(s)}) + \mathbf{K}_{(s)}\} \\ &\leq [\lambda_{\max}(\widetilde{\mathbf{P}}_{(s)})\{1+\lambda_{\max}(\mathbf{K}_{(s)})\} + \lambda_{\max}(\mathbf{K}_{(s)})]^2 \\ &= [\{1+\lambda_{\max}(\mathbf{K}_{(s)})\} + \lambda_{\max}(\mathbf{K}_{(s)})]^2, \end{aligned} \tag{S1.5}$$

which, together with (S1.3) and Condition 1, implies (S1.2). This completes the proof.

S2 Proof of Theorem 2

Note that

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

Hence, from the proof of Theorem 1, in order to prove (2.6), we only need to verify that

$$\sup_{\mathbf{w}\in\mathcal{W}} [|\mathrm{trace}\{\mathbf{P}(\mathbf{w})\widehat{\mathbf{\Omega}}_{(s^*)}\} - \mathrm{trace}\{\mathbf{P}(\mathbf{w})\mathbf{\Omega}\}|/R_n(\mathbf{w})] = o_p(1). \tag{S2.1}$$

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 (2.7), we have

$$\begin{aligned}
&\sup_{\mathbf{w}\in\mathcal{W}}[|\mathrm{trace}\{\mathbf{P}(\mathbf{w})\widehat{\mathbf{\Omega}}_{(s^*)}\}-\mathrm{trace}\{\mathbf{P}(\mathbf{w})\mathbf{\Omega}\}|/R_n(\mathbf{w})]\\
=&\sup_{\mathbf{w}\in\mathcal{W}}[|(\mathbf{y}-\mathbf{P}_{(s^*)}\mathbf{y})^{\mathrm{T}}\mathbf{Q}(\mathbf{w})(\mathbf{y}-\mathbf{P}_{(s^*)}\mathbf{y})-\mathrm{trace}\{\mathbf{Q}(\mathbf{w})\mathbf{\Omega}\}|/R_n(\mathbf{w})]\\
=&\sup_{\mathbf{w}\in\mathcal{W}}[|(\boldsymbol{\epsilon}+\boldsymbol{\mu}-\mathbf{P}_{(s^*)}\boldsymbol{\mu}-\mathbf{P}_{(s^*)}\boldsymbol{\epsilon})^{\mathrm{T}}\mathbf{Q}(\mathbf{w})(\boldsymbol{\epsilon}+\boldsymbol{\mu}-\mathbf{P}_{(s^*)}\boldsymbol{\mu}-\mathbf{P}_{(s^*)}\boldsymbol{\epsilon})\\
&-\mathrm{trace}\{\mathbf{Q}(\mathbf{w})\mathbf{\Omega}\}|/R_n(\mathbf{w})]\\
\leq&\sup_{\mathbf{w}\in\mathcal{W}}[|\boldsymbol{\epsilon}^{\mathrm{T}}(\mathbf{I}_n-\mathbf{P}_{(s^*)})^{\mathrm{T}}\mathbf{Q}(\mathbf{w})(\mathbf{I}_n-\mathbf{P}_{(s^*)})\boldsymbol{\epsilon}\\
&-\mathrm{trace}\{(\mathbf{I}_n-\mathbf{P}_{(s^*)})^{\mathrm{T}}\mathbf{Q}(\mathbf{w})(\mathbf{I}_n-\mathbf{P}_{(s^*)})\mathbf{\Omega}\}|/R_n(\mathbf{w})]\\
&+2\sup_{\mathbf{w}\in\mathcal{W}}[|\boldsymbol{\epsilon}^{\mathrm{T}}(\mathbf{I}_n-\mathbf{P}_{(s^*)})^{\mathrm{T}}\mathbf{Q}(\mathbf{w})(\mathbf{I}_n-\mathbf{P}_{(s^*)})\boldsymbol{\mu}|/R_n(\mathbf{w})]\\
&+\sup_{\mathbf{w}\in\mathcal{W}}[|\boldsymbol{\mu}^{\mathrm{T}}(\mathbf{I}_n-\mathbf{P}_{(s^*)})^{\mathrm{T}}\mathbf{Q}(\mathbf{w})(\mathbf{I}_n-\mathbf{P}_{(s^*)})\boldsymbol{\mu}|/R_n(\mathbf{w})]\\
&+\sup_{\mathbf{w}\in\mathcal{W}}[|\mathrm{trace}(\mathbf{P}_{(s^*)}^{\mathrm{T}}\mathbf{Q}(\mathbf{w})\mathbf{P}_{(s^*)}\mathbf{\Omega})|/R_n(\mathbf{w})]\\
&+2\sup_{\mathbf{w}\in\mathcal{W}}[|\mathrm{trace}(\mathbf{P}_{(s^*)}^{\mathrm{T}}\mathbf{Q}(\mathbf{w})\mathbf{\Omega})|/R_n(\mathbf{w})]\\
\equiv&\ \Xi_1+\Xi_2+\Xi_3+\Xi_4+\Xi_5. \qquad\qquad (\mathrm{S}2.2)
\end{aligned}$$

Define $\rho=\max_s\max_i|\rho_{ii}^{(s)}|$. From (S1.3), (S1.4), and Conditions 4-5, we have

$$\begin{aligned}
\rho\ \leq&\ cn^{-1}\max_s\{|\mathrm{trace}(\mathbf{P}_{(s)})|\}\\
\leq&\ cn^{-1}\max_s\{|\mathrm{trace}(\widetilde{\mathbf{P}}_{(s)})-\mathrm{trace}(\widetilde{\mathbf{P}}_{(s)}\mathbf{K}_{(s)})|\}+cn^{-1}\max_s|\mathrm{trace}(\mathbf{K}_{(s)})|\\
\leq&\ cn^{-1}\max_s|\mathrm{trace}(\widetilde{\mathbf{P}}_{(s)})|+cn^{-1}\max_s|\mathrm{trace}(\widetilde{\mathbf{P}}_{(s)}\mathbf{K}_{(s)})|+cn^{-1}\max_s|\mathrm{trace}(\mathbf{K}_{(s)})|
\end{aligned}$$

$$\begin{aligned}
&= cn^{-1}\widetilde{p} + cn^{-1}2^{-1}\max_s|\mathrm{trace}(\widetilde{\mathbf{P}}_{(s)}\mathbf{K}_{(s)} + \mathbf{K}_{(s)}^{\mathrm{T}}\widetilde{\mathbf{P}}_{(s)})| + cn^{-1}\max_s|\mathrm{trace}(\mathbf{K}_{(s)})| \\
&\leq cn^{-1}\widetilde{p} + cn^{-1}2^{-1}\max_s\{\lambda_{\max}(\widetilde{\mathbf{P}}_{(s)}\mathbf{K}_{(s)} + \mathbf{K}_{(s)}^{\mathrm{T}}\widetilde{\mathbf{P}}_{(s)})\mathrm{rank}(\widetilde{\mathbf{P}}_{(s)}\mathbf{K}_{(s)} + \mathbf{K}_{(s)}^{\mathrm{T}}\widetilde{\mathbf{P}}_{(s)})\} \\
&\qquad + cn^{-1}\max_s|\mathrm{trace}(\mathbf{K}_{(s)})| \\
&\leq cn^{-1}\widetilde{p} + cn^{-1}2\max_s\{p_s\lambda_{\max}(\widetilde{\mathbf{P}}_{(s)})\lambda_{\max}(\mathbf{K}_{(s)})\} + cn^{-1}\max_s|\mathrm{trace}(\mathbf{K}_{(s)})| \\
&= O(n^{-1}\widetilde{p} + n^{-1}h^{-1}). \qquad \text{(S2.3)}
\end{aligned}$$

It follows from (2.3) and Condition 2 that

$$\xi_n \to \infty,\ S_n\xi_n^{-2G} = o(1), \quad \text{and} \quad \xi_n^{-2}\|\mathbf{P}_{(s^*)}\boldsymbol{\mu} - \boldsymbol{\mu}\|^2 = o(1). \qquad \text{(S2.4)}$$

Using (S1.1), (S1.2), (S2.3), Chebyshev's inequality, and Theorem 2 of Whittle (1960), we can obtain that, for any $\delta > 0$,

$$\begin{aligned}
\Pr(\Xi_1 > \delta) &\leq \sum\nolimits_{s=1}^{S_n}\Pr[|\boldsymbol{\epsilon}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{P}_{(s^*)})^{\mathrm{T}}\mathbf{Q}_{(s)}(\mathbf{I}_n - \mathbf{P}_{(s^*)})\boldsymbol{\epsilon} \\
&\qquad - \mathrm{trace}\{(\mathbf{I}_n - \mathbf{P}_{(s^*)})^{\mathrm{T}}\mathbf{Q}_{(s)}(\mathbf{I}_n - \mathbf{P}_{(s^*)})\boldsymbol{\Omega}\}| > \delta\xi_n] \\
&\leq \delta^{-2G}\xi_n^{-2G}\sum\nolimits_{s=1}^{S_n}E[\boldsymbol{\epsilon}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{P}_{(s^*)})^{\mathrm{T}}\mathbf{Q}_{(s)}(\mathbf{I}_n - \mathbf{P}_{(s^*)})\boldsymbol{\epsilon} \\
&\qquad - \mathrm{trace}\{(\mathbf{I}_n - \mathbf{P}_{(s^*)})^{\mathrm{T}}\mathbf{Q}_{(s)}(\mathbf{I}_n - \mathbf{P}_{(s^*)})\boldsymbol{\Omega}\}]^{2G} \\
&\leq c_1\delta^{-2G}\xi_n^{-2G}\sum\nolimits_{s=1}^{S_n}\mathrm{trace}^G\{\boldsymbol{\Omega}^{1/2}(\mathbf{I}_n - \mathbf{P}_{(s^*)})^{\mathrm{T}}\mathbf{Q}_{(s)}(\mathbf{I}_n - \mathbf{P}_{(s^*)})\boldsymbol{\Omega} \\
&\qquad \times(\mathbf{I}_n - \mathbf{P}_{(s^*)})^{\mathrm{T}}\mathbf{Q}_{(s)}(\mathbf{I}_n - \mathbf{P}_{(s^*)})\boldsymbol{\Omega}^{1/2}\} \\
&\leq c_1\delta^{-2G}\xi_n^{-2G}\lambda_{\max}^{4G}(\mathbf{I}_n - \mathbf{P}_{(s^*)})\lambda_{\max}^{2G}(\boldsymbol{\Omega})n^G\rho^{2G}S_n \\
&= \xi_n^{-2G}S_n\{O(n^{-1}\widetilde{p}^2 + n^{-1}h^{-2})\}^G, \qquad \text{(S2.5)}
\end{aligned}$$

where c_1 is a positive constant and G is the integer defined in Condition 2.

It follows from (S2.4)–(S2.5) and Condition 6 that $\Xi_1 = o_p(1)$.

Using (S1.1), (S1.2), (S1.4), (S2.3) and (S2.4), we have

$$\begin{aligned}
\Xi_2 &\leq 2\xi_n^{-1}\|(\mathbf{I}_n - \mathbf{P}_{(s^*)})\boldsymbol{\mu}\| \sup_{\mathbf{w}\in\mathcal{W}} \|\mathbf{Q}(\mathbf{w})(\mathbf{I}_n - \mathbf{P}_{(s^*)})\boldsymbol{\epsilon}\| \\
&\leq 2\xi_n^{-1}\|(\mathbf{I}_n - \mathbf{P}_{(s^*)})\boldsymbol{\mu}\| \sup_{\mathbf{w}\in\mathcal{W}} \{\rho\|(\mathbf{I}_n - \mathbf{P}_{(s^*)})\boldsymbol{\epsilon}\| \\
&\leq 2\xi_n^{-1}\|(\mathbf{I}_n - \mathbf{P}_{(s^*)})\boldsymbol{\mu}\|\rho\{1 + \lambda_{\max}(\mathbf{P}_{(s^*)})\}\|\boldsymbol{\epsilon}\| \\
&= o(1)O_p(n^{-1/2}\widetilde{p} + n^{-1/2}h^{-1}), \qquad \text{(S2.6)}
\end{aligned}$$

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

Using (S1.2), (S1.4), (S2.3), (S2.4) and Condition 3, we have

$$\begin{aligned}
\Xi_3 &\leq \xi_n^{-1}\rho\|(\mathbf{I}_n - \mathbf{P}_{(s^*)})\boldsymbol{\mu}\|^2 \\
&\leq \xi_n^{-1}\|(\mathbf{I}_n - \mathbf{P}_{(s^*)})\boldsymbol{\mu}\|\rho\|\boldsymbol{\mu}\|\{1 + \lambda_{\max}(\mathbf{P}_{(s^*)})\} \\
&= o(1)O(n^{-1/2}\widetilde{p} + n^{-1/2}h^{-1}), \qquad \text{(S2.7)}
\end{aligned}$$

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

Using (S1.1), (S1.2) and (S1.4), we have

$$\begin{aligned}
\Xi_4 + \Xi_5 &\leq 2\xi_n^{-1}\mathrm{rank}(\mathbf{P}_{(s^*)}) \sup_{\mathbf{w}\in\mathcal{W}} [\lambda_{\max}\{\mathbf{P}_{(s^*)}^{\mathrm{T}}\mathbf{Q}(\mathbf{w})\mathbf{P}_{(s^*)}\boldsymbol{\Omega}\}] \\
&\quad + 4\xi_n^{-1}\mathrm{rank}(\mathbf{P}_{(s^*)}) \sup_{\mathbf{w}\in\mathcal{W}} [\lambda_{\max}\{\mathbf{P}_{(s^*)}^{\mathrm{T}}\mathbf{Q}(\mathbf{w})\boldsymbol{\Omega}\}] \\
&\leq 2\xi_n^{-1}\widetilde{p}\rho\lambda_{\max}^2(\mathbf{P}_{(s^*)})\lambda_{\max}(\boldsymbol{\Omega}) + 4\xi_n^{-1}\widetilde{p}\rho\lambda_{\max}(\mathbf{P}_{(s^*)})\lambda_{\max}(\boldsymbol{\Omega}) \\
&= \xi_n^{-1}O(n^{-1}\widetilde{p}^2 + n^{-1}h^{-1}\widetilde{p}), \qquad \text{(S2.8)}
\end{aligned}$$

which, along with (S2.4) and Condition 6, implies that $\Xi_4 + \Xi_5 = o(1)$.

Therefore, we can get (S2.1). This completes the proof.

Part 2: Additional theoretical and simulation results

S3 Further discussions on conditions

Following Appendix A of Speckman (1988), we first provide justifications of Conditions 4 and 1. We consider a simple case with $\mathbf{Z}_{(s),i}$ being a scalar $Z_{(s),i}$. Then, $K_{(s),ii} = k_{h_s}(0)/\sum_{j^*=1}^{n} k_{h_s}(Z_{(s),i} - Z_{(s),j^*})$, by which, we obtain

$$K_{(s),ii} = O(n^{-1}h_s^{-1}), \tag{S3.1}$$

following the assumption (g) of Speckman (1988), which we quote below

"There is a probability density $p(z)$ *of* $Z_{(s),i}$ *on [0,1] such that* $n^{-1}\sum_{i=1}^{n} c(Z_{(s),i}) \to \int_0^1 c(z)p(z)\,dz$ *as* $n \to \infty$ *for any continuous function* $c(z)$*."*

Condition 4 directly follows from (S3.1). If $K_{(s),ij}$ is non-negative, then the row sums of $\mathbf{K}_{(s)}$ are identically unity. Hence, the first bound of Condition 1 is trivial. Assumption (g) of Speckman (1988) implies that the column sums can be approximated by integrals and the dependence on h_s can be regarded to vanish as $n \to \infty$. Therefore, the second bound of Condition 1 is also reasonable.

Now we examine Condition 2. The first part of Condition 2 is a moment condition. To explain the second part of Condition 2, we define $\varrho_n = \max_{1\leq s\leq S} R_n(\mathbf{w}_s^o)$. The second part of Condition 2 is implied by $S_n^2\xi_n^{-2G}\varrho_n^G \to 0$, and it depends on the infimum risk of model averaging estimators (i.e., ξ_n) and the maximum risk of model selection estimators (i.e., ϱ_n). Both ξ_n and ϱ_n depend on the magnitude of model misspecification, which is difficult to quantify in practice.

Finally, we discuss Condition 5. As explained in Hansen & Racine (2012), this assumption excludes only extremely unbalanced designs, for example, where a single observation remains relevant asymptotically. Therefore, Condition 5 is also a reasonable assumption.

S4 Detailed proof of Theorem 1

Let $\mathbf{A}(\mathbf{w}) = \mathbf{I}_n - \mathbf{P}(\mathbf{w})$. Note that

$$C_n(\mathbf{w}) = L_n(\mathbf{w}) + \|\boldsymbol{\epsilon}\|^2 + 2\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{A}(\mathbf{w})\boldsymbol{\mu} + 2\left[\mathrm{trace}\{\mathbf{P}(\mathbf{w})\boldsymbol{\Omega}\} - \boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{P}(\mathbf{w})\boldsymbol{\epsilon}\right].$$

Similar to the proof of Theorem 2.1 of Li (1987), Theorem 1 is valid if the following equations hold:

$$\sup_{\mathbf{w}\in\mathcal{W}}[|\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{A}(\mathbf{w})\boldsymbol{\mu}|/R_n(\mathbf{w})] = o_p(1), \tag{S4.1}$$

$$\sup_{\mathbf{w}\in\mathcal{W}}[|\mathrm{trace}\{\mathbf{P}(\mathbf{w})\boldsymbol{\Omega}\} - \boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{P}(\mathbf{w})\boldsymbol{\epsilon}|/R_n(\mathbf{w})] = o_p(1), \tag{S4.2}$$

and

$$\sup_{\mathbf{w}\in\mathcal{W}}[|L_n(\mathbf{w})|/R_n(\mathbf{w})-1|]=o_p(1). \tag{S4.3}$$

Therefore, the main task of the proof is to verify (S4.1)–(S4.3).

First, from (2.3) and Condition 2, we can obtain (S4.1) using exactly the same proving steps as those in (A.1) of Wan et al. (2010). Second, to verify (S4.2), we let $\boldsymbol{\epsilon}^*=\boldsymbol{\Omega}^{-1/2}\boldsymbol{\epsilon}$. From (2.3), (A.1) and Theorem 2 of Whittle (1960), we have that for any $\delta>0$,

$$\begin{aligned}
& \Pr\left\{\sup_{\mathbf{w}\in\mathcal{W}}[|\mathrm{trace}\{\mathbf{P}(\mathbf{w})\boldsymbol{\Omega}\}-\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{P}(\mathbf{w})\boldsymbol{\epsilon}|/R_n(\mathbf{w})]>\delta\right\}\\
= & \Pr\left\{\sup_{\mathbf{w}\in\mathcal{W}}[|\mathrm{trace}\{\mathbf{P}(\mathbf{w})\boldsymbol{\Omega}\}-\boldsymbol{\epsilon}^{*\mathrm{T}}\boldsymbol{\Omega}^{1/2}\mathbf{P}(\mathbf{w})\boldsymbol{\Omega}^{1/2}\boldsymbol{\epsilon}^*|/R_n(\mathbf{w})]>\delta\right\}\\
\le & \sum\nolimits_{s=1}^{S_n}\Pr\left\{|\mathrm{trace}\{\mathbf{P}(\mathbf{w}_s^o)\boldsymbol{\Omega}\}-\boldsymbol{\epsilon}^{*\mathrm{T}}\boldsymbol{\Omega}^{1/2}\mathbf{P}(\mathbf{w}_s^o)\boldsymbol{\Omega}^{1/2}\boldsymbol{\epsilon}^*|>\delta\xi_n\right\}\\
\le & \sum\nolimits_{s=1}^{S_n}E\left\{[\mathrm{trace}\{\mathbf{P}(\mathbf{w}_s^o)\boldsymbol{\Omega}\}-\boldsymbol{\epsilon}^{*\mathrm{T}}\boldsymbol{\Omega}^{1/2}\mathbf{P}(\mathbf{w}_s^o)\boldsymbol{\Omega}^{1/2}\boldsymbol{\epsilon}^*]^{2G}\delta^{-2G}\xi_n^{-2G}\right\}\\
\le & C\delta^{-2G}\xi_n^{-2G}\sum\nolimits_{s=1}^{S_n}\mathrm{trace}^G\{\boldsymbol{\Omega}^{1/2}\mathbf{P}(\mathbf{w}_s^o)\boldsymbol{\Omega}\mathbf{P}(\mathbf{w}_s^o)^{\mathrm{T}}\boldsymbol{\Omega}^{1/2}\}\\
\le & C\delta^{-2G}\xi_n^{-2G}\lambda_{\max}^G(\boldsymbol{\Omega})\sum\nolimits_{s=1}^{S_n}\mathrm{trace}^G\{\boldsymbol{\Omega}^{1/2}\mathbf{P}(\mathbf{w}_s^o)\mathbf{P}(\mathbf{w}_s^o)^{\mathrm{T}}\boldsymbol{\Omega}^{1/2}\}\\
\le & C\delta^{-2G}\xi_n^{-2G}\lambda_{\max}^G(\boldsymbol{\Omega})\sum\nolimits_{s=1}^{S_n}R_n^G(\mathbf{w}_s^o), \qquad \text{(S4.4)}
\end{aligned}$$

where C is a positive constant. Combining (S4.4) and Condition 2, we can obtain (S4.2).

Finally, we prove (S4.3). For this purpose, we first note that Equation (2.3) implies that

$$\begin{aligned}&L_n(\mathbf{w})-R_n(\mathbf{w})\\=\;&\|\mathbf{P}(\mathbf{w})\mathbf{y}-\boldsymbol{\mu}\|^2-\|\mathbf{P}(\mathbf{w})\boldsymbol{\mu}-\boldsymbol{\mu}\|^2-\mathrm{trace}\{\mathbf{P}(\mathbf{w})\boldsymbol{\Omega}\mathbf{P}^{\mathrm{T}}(\mathbf{w})\}\\=\;&\|\mathbf{P}(\mathbf{w})\boldsymbol{\epsilon}\|^2-\mathrm{trace}\{\mathbf{P}(\mathbf{w})\boldsymbol{\Omega}\mathbf{P}^{\mathrm{T}}(\mathbf{w})\}-2\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{P}(\mathbf{w})^{\mathrm{T}}\mathbf{A}(\mathbf{w})\boldsymbol{\mu}.\qquad\text{(S4.5)}\end{aligned}$$

Hence, (S4.3) holds if we can verify that

$$\sup_{\mathbf{w}\in\mathcal{W}}[|\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{P}(\mathbf{w})^{\mathrm{T}}\mathbf{A}(\mathbf{w})\boldsymbol{\mu}|/R_n(\mathbf{w})]=o_p(1),\qquad\text{(S4.6)}$$

and

$$\sup_{\mathbf{w}\in\mathcal{W}}[|\|\mathbf{P}(\mathbf{w})\boldsymbol{\epsilon}\|^2-\mathrm{trace}\{\mathbf{P}(\mathbf{w})\boldsymbol{\Omega}\mathbf{P}^{\mathrm{T}}(\mathbf{w})\}|/R_n(\mathbf{w})]=o_p(1).\qquad\text{(S4.7)}$$

From (A.2), we know that for $\mathbf{w}^*\in\mathcal{W}$ and $\mathbf{w}\in\mathcal{W}$, we have that

$$\begin{aligned}&\|\mathbf{P}(\mathbf{w}^*)^{\mathrm{T}}\mathbf{A}(\mathbf{w})\boldsymbol{\mu}\|^2\\\leq\;&\lambda_{\max}\left\{\mathbf{P}(\mathbf{w}^*)\mathbf{P}(\mathbf{w}^*)^{\mathrm{T}}\right\}\|\mathbf{A}(\mathbf{w})\boldsymbol{\mu}\|^2\\\leq\;&\lambda_{\max}^2\{\mathbf{P}(\mathbf{w}^*)\}\|\mathbf{A}(\mathbf{w})\boldsymbol{\mu}\|^2\\=\;&O(1)\|\mathbf{A}(\mathbf{w})\boldsymbol{\mu}\|^2,\qquad\text{(S4.8)}\end{aligned}$$

and

$$\begin{aligned}\mathrm{trace}\left\{\mathbf{P}(\mathbf{w})\mathbf{P}(\mathbf{w}^*)^{\mathrm{T}}\mathbf{P}(\mathbf{w}^*)\mathbf{P}^{\mathrm{T}}(\mathbf{w})\right\}&\leq\lambda_{\max}\left\{\mathbf{P}(\mathbf{w}^*)^{\mathrm{T}}\mathbf{P}(\mathbf{w}^*)\right\}\mathrm{trace}\left\{\mathbf{P}(\mathbf{w})\mathbf{P}^{\mathrm{T}}(\mathbf{w})\right\}\\&\leq\lambda_{\max}^2\{\mathbf{P}(\mathbf{w}^*)\}\,\mathrm{trace}\left\{\mathbf{P}(\mathbf{w})\mathbf{P}^{\mathrm{T}}(\mathbf{w})\right\}\\&=O(1)\mathrm{trace}\left\{\mathbf{P}(\mathbf{w})\mathbf{P}^{\mathrm{T}}(\mathbf{w})\right\}.\qquad\text{(S4.9)}\end{aligned}$$

Now with Condition 2, (2.3), (S4.8), and (S4.9) readily there, we can follow exactly the same steps as those in (A.4) and (A.5) of Wan et al. (2010) to obtain (S4.6) and (S4.7). This completes the proof.

S5 Additional simulation studies

S5.1 Comparison with linear model averaging

To see how much harm it can cause by ignoring the nonlinearity, we first compare our method with linear model averaging (LMA) that considers all candidate models to be fully linear. Theoretically, LMA should work better if 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. As the degree of nonlinearity increases, better fit of nonparametric estimation dominates its efficiency loss and slow convergence, and thus MAPLM should outperform LMA. To confirm this theoretical argument, we make the comparison between MAPLM and LMA in three data generation processes with different nonlinear functions, i.e.

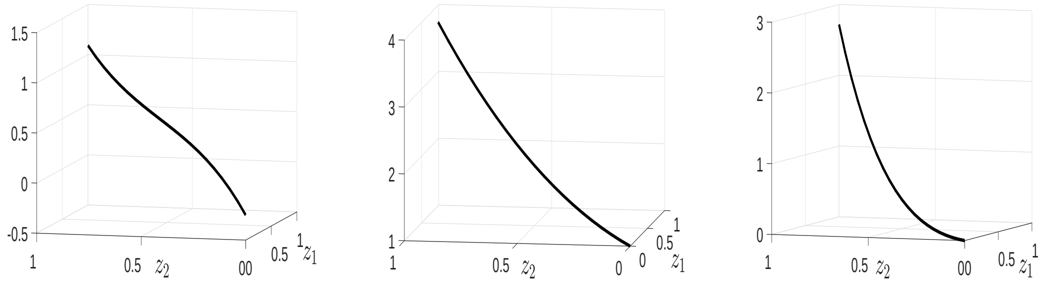
DGP1: $g_1(z_{i1}, z_{i2}) = 2(z_{i1} - 0.5)^3 + \sin(z_{i2})$,

DGP2: $g_2(z_{i1}, z_{i2}) = \exp(z_{i1}) + z_{i2}^2$,

DGP3: $g_3(z_{i1}, z_{i2}) = \exp(z_{i1}) * z_{i2}^2$.

The degree of nonlinearity of the three functions is shown in Figure S.1. Clearly, g_3 generates the most nonlinear relationship. g_2 is slightly more nonlinear than g_1, both of which are closer to linearity than g_3. Hence, we expect that the performance of MAPLM and LMA is comparable under g_1 and g_2, but MAPLM should demonstrate more superiority under g_3. Also, because of the above-mentioned tradeoff between better fit and efficiency, the relative performance of MAPLM and LMA depends on the signal-to-noise ratio, the sample size, and the degree of uncertainty in model specification. We shall examine the effect of these factors in turn.

Figure S.1: Nonlinear functions in three data generation processes



g_1 g_2 g_3

We first consider the case where there is uncertainty only in the linear component, and the results are given in Figures S.2. First, we see that MAPLM generally demonstrates its superiority over LMA when R^2 and sample size are moderate or large, except in DGP1 with small degree of nonlinearity. As the R^2 and sample size decrease, LMA performs better than MAPLM. This is not surprising because nonparametric estimation tends to fit the noise more than least squares when R^2 is small, and it suffers from the curse of dimensionality when the sample size is small. Next, we examine how the degree of nonlinearity influence the performance. As expected, the advantage of MAPLM becomes more prominent as the degree of nonlinearity increases. In particular, LMA slightly outperforms MAPLM in most of cases of DGP1 where nonlinearity is weak, but their discrepancy is small. As the degree of nonlinearity increases in DGP2, MAPLM outperforms LMA when $R^2 \geq 0.5$. With a large sample size $n = 400$, MAPLM beats LMA even when $R^2 = 0.3$. In DGP3 with the largest degree of nonlinearity, MAPLM outperforms LMA for a even wider range of R^2. It starts to dominate LMA under $R^2 = 0.3$ even when $n = 200$, and the disadvantage of LMA under large R^2 is magnified.

We then consider the second case where there is uncertainty in both linear and nonlinear components (structure uncertainty), and the results are presented in Figures S.3. In this case, MAPLM shows a dominant superiority over LMA over the whole range of R^2 in all DGPs and all sample sizes.

In general, we observe a tradeoff between better fit and efficiency in choosing between MAPLM and LMA. MAPLM containing a nonparametric component can better capture the nonlinear pattern, but converges slower and is less efficient than least squares. Therefore, when the degree of nonlinearity is large, better fit of nonparametric estimation dominates its efficiency loss and slow convergence, and this results in better performance of MAPLM than LMA. A large signal-to-noise ratio and a large sample size also favour MAPLM as both of these two situations help improve efficiency. Moreover, we find MAPLM more robust than LMA. Under weak nonlinearity when LMA works especially well, MAPLM is only slightly inferior to LMA. However, LMA can perform much worse than MAPLM and other methods in some situations.

Figure S.2: Mean square error comparison: Uncertainty only in the linear component

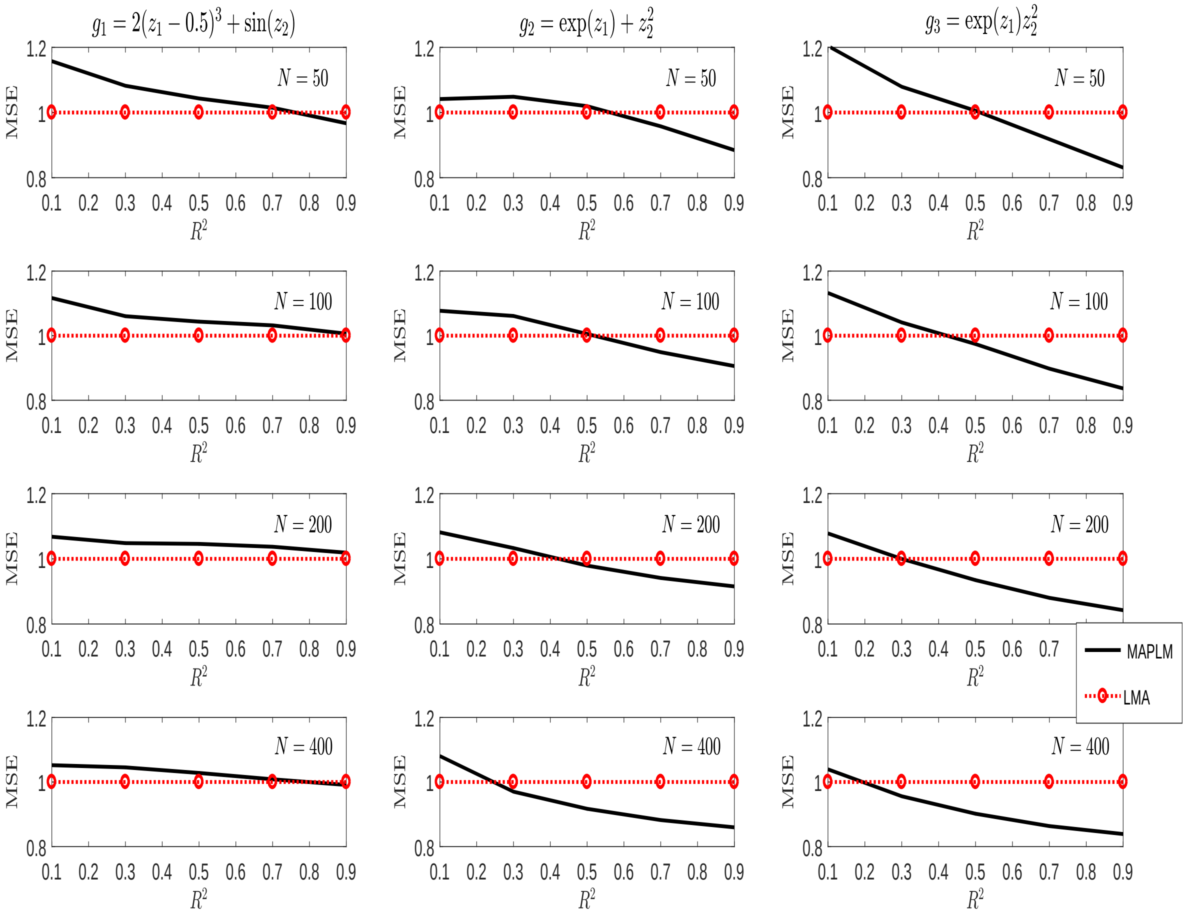
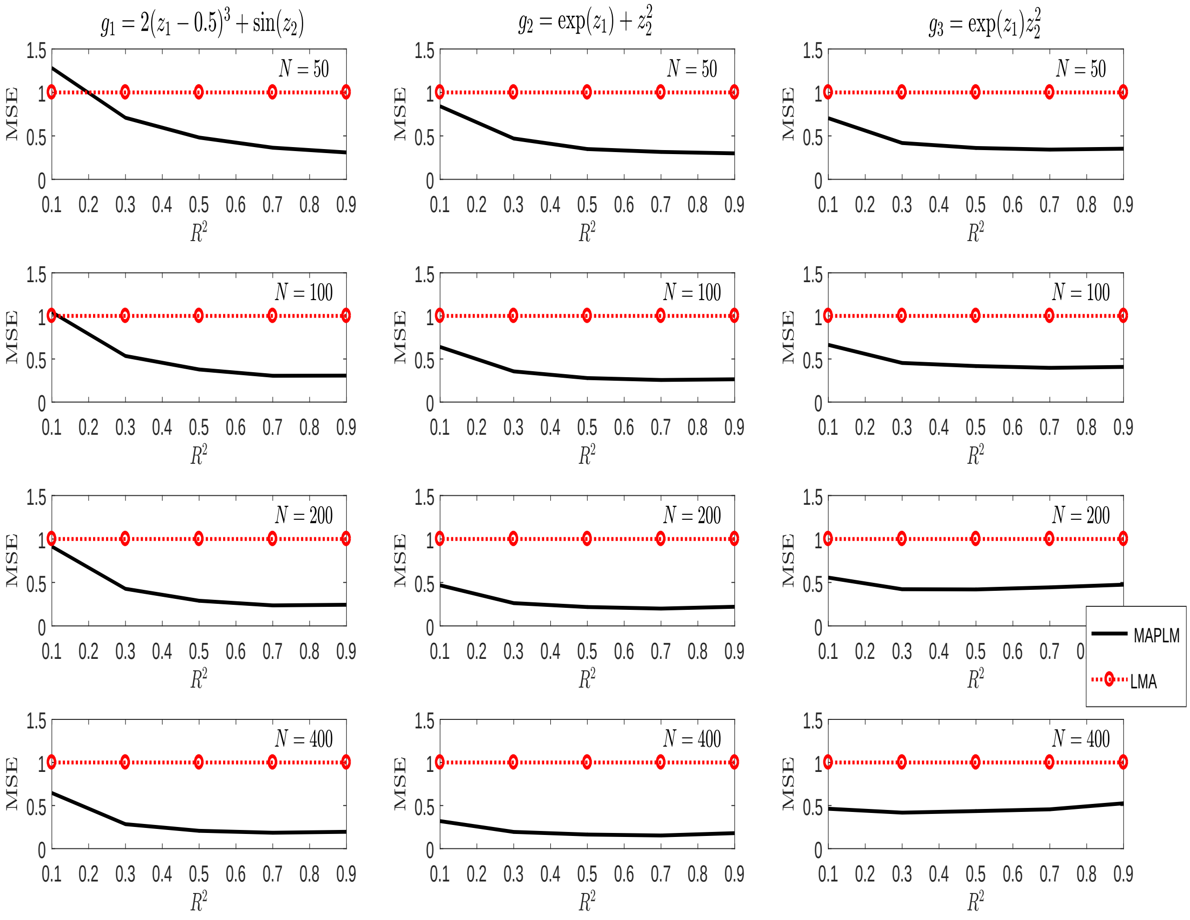


Figure S.3: Mean square error comparison: Uncertainty in both components



S5.2 Simulation results of additional data generation processes

First, to examine whether the simulation results are affected by expanding model space, we consider additional experiment designs. In particular, in the first case with uncertainty only in the linear component, we let the number of candidate models increase from 7 ($= 2^3 - 1$) to 31 ($= 2^5 - 1$), and further to 127 ($= 2^7 - 1$) as the sample size increases from 50, 100, to 200. In the second case with uncertainty both in the linear and nonlinear component, we let the number of candidate models vary from 12 to 50, and further to 133 as the sample size increases. In both cases, we generate the nonlinear component using the function $g(z) = \exp(z) + z^2$. The results are presented in Figure S.4. We see that the results are highly robust. With diverging number of candidate models, the superiority of MAPLM over linear model averging is even more obvious, which suggests that MAPLM is particularly useful when the model space is large.

Second, to mimic possible serial correlation in the empirical data, we consider data generation process with autoregressive errors. In particular, we generate the error as $\epsilon_i = \sigma(0.75\epsilon_{i-1} + u_i)$, where u_i follows a normal distribution with zero mean and variance 0.75, and σ controls the signal-to-noise ratio. The results are provided in Figure S.5, and the relative performance of competitive methods is hardly affected.

Figure S.4: Mean square error comparison: Diverging number of candidate models ($g(z_1, z_2) = \exp(z_1) + z_2^2$)

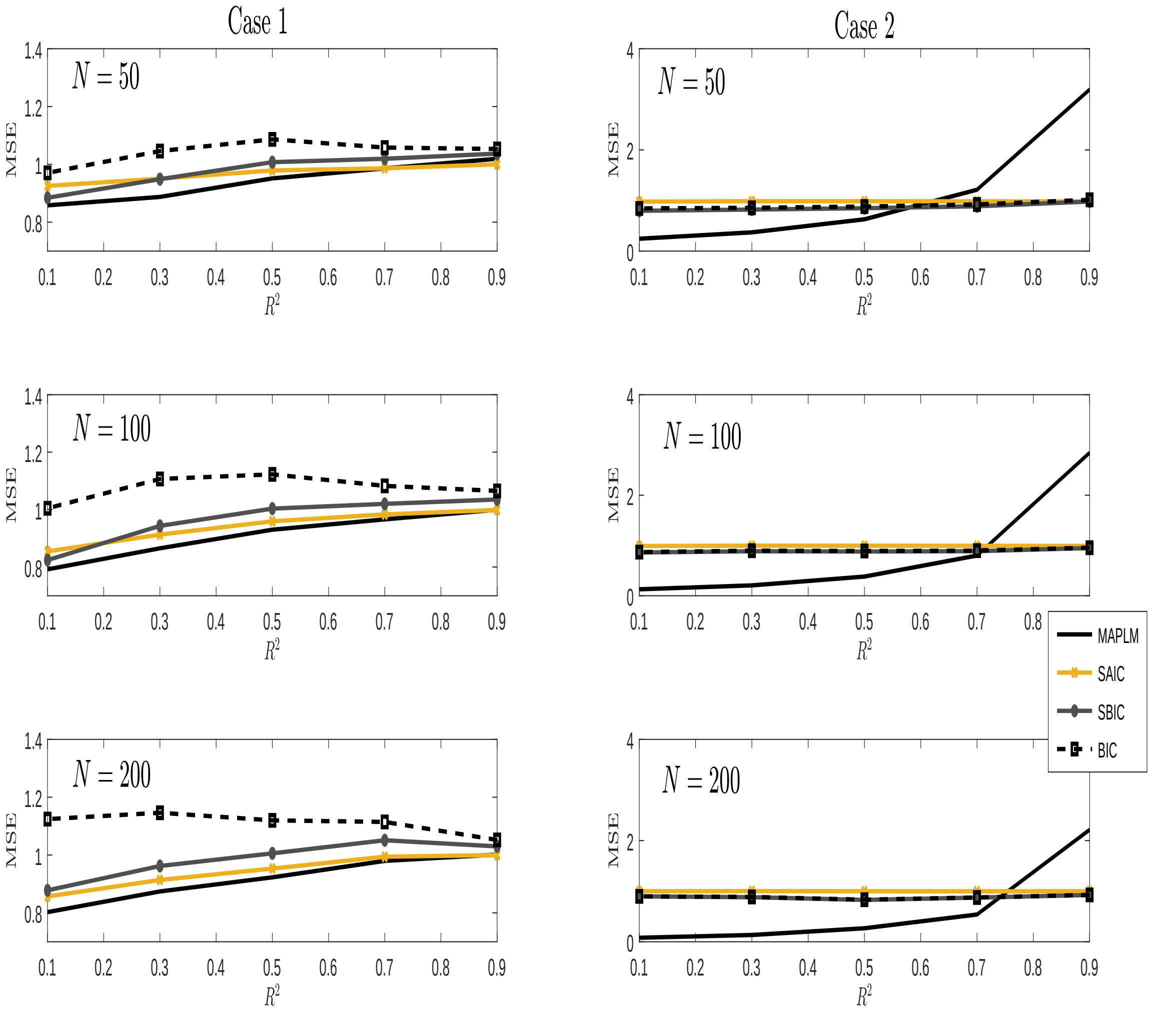
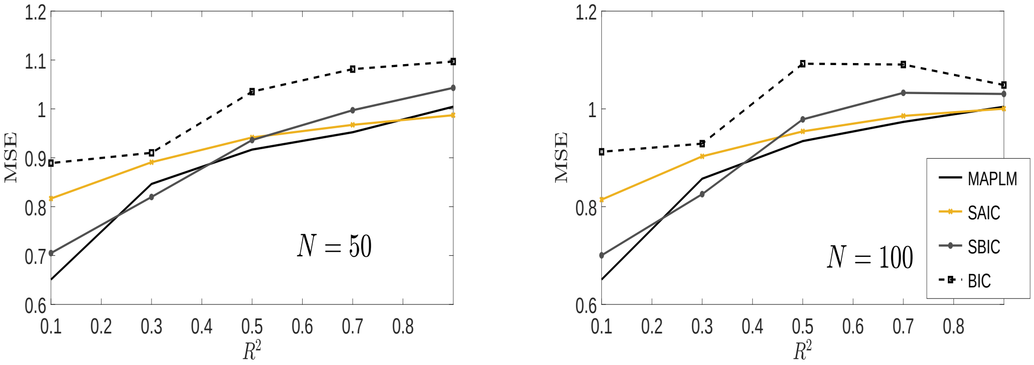


Figure S.5: Mean square error comparison: AR errors in Case 1



S6 Additional figures and tables for the empirical application

Figure S.6: Features of the change in Japan's CDS spreads

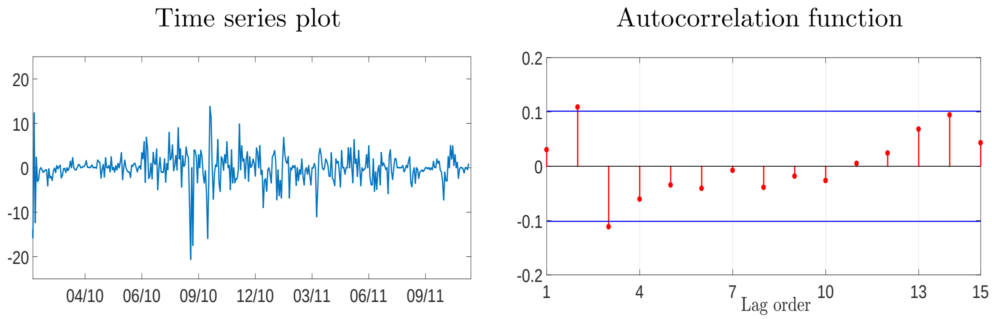
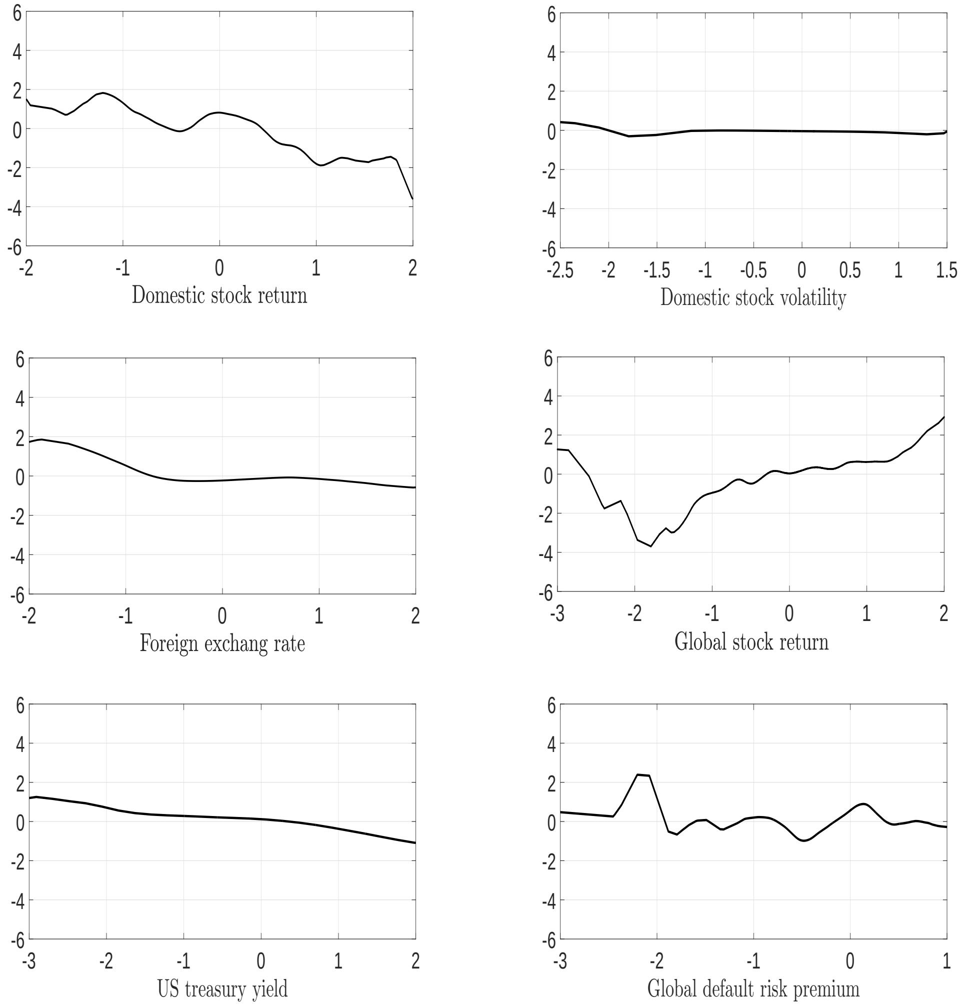


Table S.1: Descriptive statistics of the first-differenced data (before normalization)

	Mean	Variance	5% quantile	95% quantile
CDS spreads	−0.0001	14.3059	−4.9634	4.9031
Domestic stock return	0.0076	3.0897	−2.2639	2.3691
Domestic stock volatility	0.0002	0.4833	−0.2516	0.3563
Foreign exchange rate	−0.0128	0.1910	−0.6240	0.5940
Global stock return	−0.0028	3.5461	−2.9118	3.0539
US treasury yield	−0.0032	0.0022	−0.0900	0.0700
Global default risk premium	−0.0006	0.0031	−0.0890	0.0701

Figure S.7: Nonparametric estimation for each macroeconomic determinant



References

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

Hardy, G. H., Littlewood, J. E. & Polya, G. (1952). *Inequalities.* Cambridge university press.

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.

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

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.

Zhang, X., Wan, A. T. K. & Zou, G. (2013). Model averaging by jackknife criterion in models with dependent data. *Journal of Econometrics* **174**, 82–94.