Econometric Theory, **37**, 2021, 388–407.
doi:10.1017/S0266466620000055

A NEW STUDY ON ASYMPTOTIC OPTIMALITY OF LEAST SQUARES MODEL AVERAGING*

XINYU ZHANG
Academy of Mathematics and Systems Science
Chinese Academy of Sciences

In this article, we present a comprehensive study of asymptotic optimality of least squares model averaging methods. The concept of asymptotic optimality is that in a large-sample sense, the method results in the model averaging estimator with the smallest possible prediction loss among all such estimators. In the literature, asymptotic optimality is usually proved under specific weights restriction or using hardly interpretable assumptions. This article provides a new approach to proving asymptotic optimality, in which a general weight set is adopted, and some easily interpretable assumptions are imposed. In particular, we do not impose any assumptions on the maximum selection risk and allow a larger number of regressors than that of existing studies.

1. INTRODUCTION

Least squares model averaging has been explored in depth from the theoretical perspective and has been used widely in empirical applications since the seminal work of Hansen (2007). One key property possessed by many least squares model averaging estimators is asymptotic optimality, whereby the model averaging estimator results in the smallest possible prediction loss among all such estimators as the sample size increases to infinity.

In the least squares model averaging literature, there are two major approaches to proving asymptotic optimality based on the restrictions on the weight set. The first uses a general weight set

$$\mathcal{W} \equiv \left\{ w \in [0,1]^{M_n} : \sum_{m=1}^{M_n} w_m = 1 \right\},$$

where $w = (w_1, \ldots, w_{M_n})^{\mathrm{T}}$ is the weight vector and M_n is the number of candidate models. In this approach, however, certain hardly interpretable or strong

*We thank two anonymous referees, the Co-Editor, the Editor Peter C.B. Phillips, Tie Xie, and Jiahui Zou for many constructive comments and suggestions. Zhang gratefully acknowledges the research support from National Natural Science Foundation of China (grant numbers 71925007, 71631008, and 11688101). All errors remain the author. Address correspondence to Xinyu Zhang, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China; e-mail: xinyu@amss.ac.cn.

© Cambridge University Press 2020

https://doi.org/10.1017/S0266466620000055 Published online by Cambridge University Press

assumptions on the number of candidate models, the minimum averaging risk, the maximum selection risk, and the number of regressors are required. See, for example, Wan, Zhang, and Zou (2010), Liu and Okui (2013), Zhang, Wan, and Zou (2013b), Ando and Li (2014), among others.

The other approach establishes asymptotic optimality under a subset of $\mathcal{W}$. A widely used subset of $\mathcal{W}$ is

$$\mathcal{W}(N) = \{w \in \mathcal{W} : w_m \in \{0, 1/N, 2/N, \ldots, 1\}\},$$

where N is a positive integer. See Hansen (2007), Kuersteiner and Okui (2010), Hansen and Racine (2012), Cheng, Ing, and Yu (2015), Liu, Okui, and Yoshimura (2016), and Liu and Tao (2016) for example. In certain cases, the candidate models must be sequentially nested. When establishing asymptotic optimality under subset $\mathcal{W}(N)$ and nested candidate models, Hansen (2007) did not impose any assumption on the number of candidate models and merely used a mild assumption on the minimum averaging risk.

However, regardless of the theoretical approach being used, weight set $\mathcal{W}$ is adopted for an application of a least squares model averaging method in practice. Furthermore, the nested candidate models depend on the ordering of regressors; however, there is often no clear justification for the ordering. Therefore, in this article, we consider a general weight set in the first approach and prove asymptotic optimality of least squares model averaging estimators according to an alternative but clearer method. The assumptions we make are more easily interpretable than those of many existing studies. In particular, our theorems do not impose any assumptions on the maximum selection risk and allow a larger number of regressors. Our article provides new support for establishing asymptotic optimality and sheds light on the theoretical foundation of least squares model averaging.

The remainder of this article begins with a description of models in Section 2. Section 3 presents asymptotic optimality of the Mallows model averaging (MMA) under homoscedasticity and compares the new assumptions with the existing ones. Section 4 analyzes the asymptotic optimality of jackknife model averaging (JMA) (Hansen and Racine, 2012) under heteroscedasticity and compares the new assumptions with the existing ones. Section 5 concludes. The proofs of theorems are detailed in the Appendix.

2. MODELS AND MODEL AVERAGING

Following Hansen (2007), Hansen and Racine (2012), and Zhang et al. (2013b), we consider a sample of independent observations (y_i, x_i) for $i = 1, \ldots, n$. Let the conditional mean be $\mu_i = \mathbb{E}(y_i|x_i)$, so

$$y_i = \mu_i + e_i, \tag{1}$$

where $\mathbb{E}(e_i|x_i) = 0$ and $\mathbb{E}(e_i^2|x_i) = \sigma_i^2$. Let $y = (y_1, \ldots, y_n)^{\mathrm{T}}$, $\mu = (\mu_1, \ldots, \mu_n)^{\mathrm{T}}$, and $e = (e_1, \ldots, e_n)^{\mathrm{T}}$.

Suppose that there are M_n candidate models and that the regression matrix of the mth model is denoted by $X_{(m)}$. We allow candidate models to be non-nested;

https://doi.org/10.1017/S0266466620000055 Published online by Cambridge University Press

that is, the variables in $X_{(m)}$ may not be included in $X_{(s)}$, and vice versa. We assume that for any $m \in \{1, \ldots, M_n\}$, $X_{(m)}$ has a full column rank k_m; thus under the mth model, the estimator of μ is

$$\widehat{\mu}_{(m)} = X_{(m)} \left(X_{(m)}^{\mathrm{T}} X_{(m)}\right)^{-1} X_{(m)}^{\mathrm{T}} y \equiv P_{(m)} y, \tag{2}$$

where $P_{(m)} = X_{(m)}(X_{(m)}^{\mathrm{T}} X_{(m)})^{-1} X_{(m)}^{\mathrm{T}}$. Let $w = (w_1, \ldots, w_{M_n})^{\mathrm{T}}$ be a weight vector in set

$$\mathcal{W} \equiv \left\{ w \in [0,1]^{M_n} : \sum_{m=1}^{M_n} w_m = 1 \right\},$$

then the model averaging estimator of μ is defined as

$$\widehat{\mu}(w) = \sum_{m=1}^{M_n} w_m \widehat{\mu}_{(m)} = \sum_{m=1}^{M_n} w_m P_{(m)} y \equiv P(w) y, \tag{3}$$

where $P(w) = \sum_{m=1}^{M_n} w_m P_{(m)}$. Define the squared prediction loss as $L_n(w) = \|\widehat{\mu}(w) - \mu\|^2$ and the corresponding prediction risk as $R_n(w) = \mathbb{E}\{L_n(w)|x_1, \ldots, x_n\}$, where $\|\cdot\|^2$ stands for the squared Euclidean norm.

3. HOMOSCEDASTIC ERROR AND MMA

In this section, we consider homoscedastic error setting, that is, $\sigma_1^2 = \cdots = \sigma_n^2 = \sigma^2$. Let $k = (k_1, \ldots, k_{M_n})^{\mathrm{T}}$, $A_{(m)} = I_n - P_{(m)}$, and $A(w) = I_n - P(w)$. Hansen (2007) proposed choosing weights by minimizing the Mallows' criterion

$$\mathcal{C}_n(w) = \|A(w)y\|^2 + 2\sigma^2 w^{\mathrm{T}} k. \tag{4}$$

Criterion $\mathcal{C}_n(w)$ is an unbiased estimator of the risk $R_n(w)$ up to a constant term. Denote $\widehat{w}_{\mathrm{MMA}} = \arg\min_{w \in \mathcal{W}} \mathcal{C}_n(w)$, so that the MMA estimator of μ is $\widehat{\mu}(\widehat{w}_{\mathrm{MMA}})$. We will establish optimality under clearer, more easily interpretable conditions than the existing papers. Let $\xi_n = \inf_{w \in \mathcal{W}} R_n(w)$. All limiting processes discussed in this article assume the limit of $n \to \infty$.

Assumption 1. $\mathbb{E}(e_i^4 | x_i) \le \upsilon < \infty$ almost surely for $i = 1, \ldots, n$, where υ is a positive constant.

Assumption 2. $M_n^2 \xi_n^{-1} \to 0$ almost surely.

THEOREM 1. *If Assumptions 1 and 2 hold, then*

$$\frac{L_n(\widehat{w}_{\mathrm{MMA}})}{\inf_{w \in \mathcal{W}} L_n(w)} \to 1 \tag{5}$$

in probability.

The proof of Theorem 1 is provided in Appendix A.2. Based on Theorem 1, the MMA estimator $\widehat{\mu}(\widehat{w}_{\mathrm{MMA}})$ is called asymptotically optimal in the sense that its

https://doi.org/10.1017/S0266466620000055 Published online by Cambridge University Press

squared prediction loss is asymptotically identical to that of the infeasible best model averaging estimator.

3.1. Discussions of Assumptions 1 and 2

Assumption 1 places a bound on the conditional moments. Assumption 2 restricts the rates of increase in the number of candidate models M_n and the minimum averaging risk ξ_n.

We compare our assumptions with the assumptions (15) and (16) of Hansen (2007), which are

$$\xi_n \to \infty \tag{6}$$

and

$$\mathbb{E}\left(e_i^{4(N+1)}|x_i\right) \le \upsilon < \infty \tag{7}$$

almost surely, respectively, where υ is a positive constant and N is a positive integer determined by a discrete weight set

$$\mathcal{W}(N) = \{w \in \mathcal{W} : w_m \in \{0, 1/N, 2/N, \ldots, 1\}\}.$$

Obviously, our assumptions are different from these existing ones and (6) is milder than Assumption 2, but (7) is stronger than Assumption 1. It is worth noting that the asymptotic optimality in Hansen (2007) and other related works such as Liu et al. (2016) was built under the discrete weight set $\mathcal{W}(N)$, that is,

$$\frac{L_n(\widehat{w}_{\text{MMA}})}{\inf_{w\in\mathcal{W}(N)} L_n(w)} \to 1,$$

and $\inf_{w\in\mathcal{W}(N)} L_n(w) \ge \inf_{w\in\mathcal{W}} L_n(w)$. Additionally, in these works, the candidate models are nested, that is, the mth candidate model uses the first k_m regressors. Furthermore, Cheng et al. (2015) developed the asymptotic optimality under a continuous weight set, which improves the result of Hansen (2007), but the weight set of Cheng et al. (2015) is still a subset of $\mathcal{W}$ and the number of nonzero weights in the set equals or is smaller than N and thus is finite. The nested candidate models are also needed in Cheng et al. (2015).

Hansen and Racine (2012) adopted the following assumption

$$\bar{q}_n = o\left(\xi_n^{1/N}\right), \tag{8}$$

where $\bar{q}_n = \max_{1\le m\le M_n} q_{mn}$, q_{mn} is the number of models that have exactly m parameters, and N is the positive integer in the discrete weight set $\mathcal{W}(N)$. In the special case of nested models, $\bar{q}_n = 1$; then assumption (8) holds automatically due to (6). For a general non-nested framework with $M_n = 2^{p_n} - 1$ (i.e., all possible combinations of regressors are used), where p_n is the number of observed regressors, assumption (8) guarantees that $M_n p_n^{-1} = o(\xi_n^{1/N})$, which further implies Assumption 2 when $N \ge 3$ since $M_n p_n^{-1} = (2^{p_n} - 1)p_n^{-1} >> (2^{p_n} - 1)^{0.9} = M_n^{0.9}$. Hence, both (8) and Assumption 2 constrain the number of non-nested models,

https://doi.org/10.1017/S0266466620000055 Published online by Cambridge University Press

but unlike the case of Hansen and Racine (2012), the asymptotic optimality in this article is established under a general continuous weight set $\mathcal{W}$.

Wan et al. (2010) and related papers such as Liu and Okui (2013), Xie (2015), and Zhang, Zou, and Carroll (2015) established asymptotic optimality under the assumptions that for some fixed positive integer G and positive constant υ,

$$\mathbb{E}\left(e_i^{4G}|x_i\right) \leq \upsilon < \infty \tag{9}$$

and

$$M_n \xi_n^{-2G} \sum_{m=1}^{M_n} R_n^G\left(w_m^0\right) \to 0 \tag{10}$$

almost surely, where w_m^0 is an M_n-dimensional vector, in which the mth element is one and the others are zeros. Similar to (7), when $G > 1$, assumption (9) is stronger than Assumption 1 and excludes random errors from heavy-tailed distributions, such as a t-distribution with the degrees of freedom smaller than or equal to 5 and a Pareto distribution with the shape parameter smaller than or equal to 5. When $G = 1$, $M_n \xi_n^{-2G} \sum_{m=1}^{M_n} R_n^G(w_m^0) = M_n \xi_n^{-2} \sum_{m=1}^{M_n} R_n(w_m^0) \geq M_n \xi_n^{-2} \sum_{m=1}^{M_n} \xi_n = M_n^2 \xi_n^{-1}$, which means that assumption (10) is stronger than Assumption 2.

Let $\eta_n = \max_{1 \leq m \leq M_n} R_n(w_m^0)$, which is the maximum selection risk. Assumption (10) implies that

$$\xi_n^{-2} \eta_n \to 0, \tag{11}$$

which imposes a restriction on the maximum selection risk η_n. In contrast, our Assumption 2 does not impose any restriction on η_n. Assume that $\eta_n \sim \xi_n^{15/8}$, then assumption (10) implies $M_n = o(\xi_n^{G/8})$; thus if $G < 4$, assumption (10) is stronger than Assumption 2. Assume that $\eta_n \sim \xi_n^{9/8}$, then assumption (10) is implied by $M_n = o(\xi_n^{7G/16})$; thus if $G \geq 2$, then assumption (10) is milder than Assumption 2. Therefore, neither assumption (10) nor Assumption 2 is stronger than the other. However, Assumption 2 is clearer, simpler, and more easily interpretable than assumption (10).

In eqn. (11) of Ando and Li (2014), the following condition is imposed:

$$M_n^{4G+2} n^G \xi_n^{-2G} \to 0, \tag{12}$$

which is stronger than Assumption 2 for any positive integer G.

Now, we consider an example with $R_n(w_m^0) = n k_m^{-a} + k_m$ for some $a \geq 1$. This example was also considered in Cheng et al. (2015). We observe that $R_n(w_m^0)$ is minimized if $k_m \sim (an)^{1/(1+a)}$ and that the minimized order of $R_n(w_m^0)$ is $n^{1/(1+a)}$; then, $\xi_n = O(n^{1/(1+a)})$. If we use a nested candidate model set with the mth model using the first m regressors, then Assumption 2 needs $M_n = o(\xi_n^{1/2}) = o(n^{1/(2+2a)}) << (an)^{1/(1+a)}$; that is, the model achieving the optimal rate of $R_n(w_m^0)$

https://doi.org/10.1017/S0266466620000055 Published online by Cambridge University Press

TABLE 1. The comparison of assumptions in three papers

Assumptions on	Hansen (2007)	Wan et al. (2010)	This article
Models	Nested	**Nested or non-nested**	**Nested or non-nested**
Weights	$w \in \mathcal{W}(N)$	$\boldsymbol{w \in \mathcal{W}}$	$\boldsymbol{w \in \mathcal{W}}$
Moments	$\mathbb{E}(e_i^{4(N+1)}\|x_i) \le \upsilon$	$\mathbb{E}(e_i^{4G}\|x_i) \le \upsilon$	$\mathbb{E}(\boldsymbol{e_i^4}\|\boldsymbol{x_i}) \le \boldsymbol{\upsilon}$
η_n	**None**	$\xi_n^{-2}\eta_n \to 0$	**None**
ξ_n and M_n	$\boldsymbol{\xi_n^{-1} \to 0}$	$M_n\xi_n^{-2G}\sum_{m=1}^{M_n} R_n^G(w_m^0) \to 0$	$\xi_n^{-1}M_n^2 \to 0$

Note: In each row, the mildest assumption is in bold.

cannot be included in the candidate model set.[1] This is a problem with the theory in this article and warrants future research.

Last, we use Table 1 to summarize and compare the assumptions used in Hansen (2007), Wan et al. (2010), and this article. In each row of the table, we show in bold the mildest assumption in that row.

4. HETEROSKEDASTIC ERROR AND JMA

In the heteroskedastic error setting, Hansen and Racine (2012) proposed choosing weights by minimizing the jackknife criterion

$$\mathcal{J}_n(w) = \|\widetilde{\mu}(w) - y\|^2, \tag{13}$$

where $\widetilde{\mu}(w) = \sum_{m=1}^{M_n} w_m \widetilde{\mu}_{(m)}$ and $\widetilde{\mu}_{(m)}$ is the jackknife estimator of μ under model m. The ith component of $\widetilde{\mu}_{(m)}$ is

$$\widetilde{\mu}_{(m),i} = X_{(m),i}^{\mathrm{T}}\left(X_{(m),[-i]}^{\mathrm{T}}X_{(m),[-i]}\right)^{-1}X_{(m),[-i]}^{\mathrm{T}}y_{[-i]},$$

where $X_{(m),i}$ is the ith column of $X_{(m)}^{\mathrm{T}}$, and $X_{(m),[-i]}$ and $y_{[-i]}$ are $X_{(m)}$ and y with the ith rows deleted. Let $\widehat{w}_{\mathrm{JMA}} = \arg\min_{w\in\mathcal{W}} \mathcal{J}_n(w)$; then, the JMA estimator of μ is $\widehat{\mu}(\widehat{w}_{\mathrm{JMA}})$.

Assumption 3. $\inf_{1\le i\le n}\sigma_i^2 \ge \underline{\sigma}^2 > 0$ almost surely.

Let $h_{m,ii} = X_{(m),i}^{\mathrm{T}}(X_{(m)}^{\mathrm{T}}X_{(m)})^{-1}X_{(m),i}$ which is the ith diagonal element of $P_{(m)}$, $\bar{h} = \max_{1\le m\le M_n}\max_{1\le i\le n} h_{m,ii}$, and $\bar{k} = \max_{1\le m\le M_n} k_m$.

Assumption 4. $\bar{h} \to 0$ and $\bar{h}\bar{k}\xi_n^{-1} \to 0$ almost surely.

THEOREM 2. *If Assumptions 1 and 4 hold, then*

$$\frac{L_n(\widehat{w}_{\mathrm{JMA}})}{\inf_{w\in\mathcal{W}} L_n(w)} \to 1 \tag{14}$$

in probability.

[1] In fact, if one knew the risk is minimized by $k_m \sim (an)^{1/(1+a)}$, then the sizes of candidate models can be set around $(an)^{1/(1+a)}$, and using all $(an)^{1/(1+a)}$ nested candidate models may not be necessary.

https://doi.org/10.1017/S0266466620000055 Published online by Cambridge University Press

The proof of Theorem 2 is provided in Appendix A.3. According to Theorem 2, the JMA estimator $\widehat{\mu}(\widehat{w}_{\text{JMA}})$ is asymptotically optimal under the heteroskedastic error setting.

4.1. Discussions of Assumptions 3 and 4

Assumption 3 excludes the case of degenerate heteroskedasticity and is exactly the same as Cond. (A.1) of Hansen and Racine (2012).

The first part of Assumption 4 is exactly the same as Cond. (A.9) of Hansen and Racine (2012) and is commonly used in the literature on cross-validation.

To discuss the second part of Assumption 4, we introduce the following assumption:

$$h_{m,ii} \leq \Lambda k_m n^{-1} \tag{15}$$

almost surely for some positive constant Λ. This assumption is widely used; see, for example, Li (1987) and Andrews (1991). Under assumption (15), the second part of Assumption 4 is equivalent to

$$\bar{k} = o\left(n^{1/2}\xi_n^{1/2}\right) \tag{16}$$

almost surely. Hence, the largest number of regressors $\bar{k}$ is constrained.

Wan et al. (2010) used the conditions $\bar{k}^2 n^{-1} = O(1)$ and (10) that, together, are stronger than (16). Liu et al. (2016) used the condition

$$\bar{k} = o\left(n^{1/2}\right) \tag{17}$$

almost surely, which is stronger than (16). Ando and Li (2014) used the conditions

$$M_n^{4G+2} n^G \xi_n^{-2G} \to 0 \text{ and } \bar{k} n^{-3/4} = O(1) \tag{18}$$

almost surely, that imply $M_n^4 n \xi_n^{-2} \bar{k}^4 n^{-3} = o(1)$, and thus

$$\bar{k} = o\left(n^{1/2}\xi_n^{1/2} M_n^{-1}\right)$$

almost surely, which is stronger than (16).

As discussed in Section 3.1, for a general non-nested framework with $M_n = 2^{p_n} - 1$ candidate models, where p_n is the number of observed regressors, assumption (8) used in Hansen and Racine (2012) needs $2^{p_n} p_n^{-1} = o(\xi_n^{1/N})$, which is stronger than $p_n = o(\xi_n^{1/2})$ if $N \geq 3$. Thus, based on $\bar{k} \leq p_n$, we obtain that when $N \geq 3$, assumption (8) is also stronger than (16) under the general non-nested framework.

In summary, under the widely used assumption (15), the second part of Assumption 4 is milder than the existing conditions. In particular, our new assumption places a milder constraint on the largest number of regressors $\bar{k}$ than do the existing conditions.

https://doi.org/10.1017/S0266466620000055 Published online by Cambridge University Press

5. CONCLUSIONS

In this article, we study asymptotic optimality of least squares model averaging under non-nested candidate models with a general weight set $\mathcal{W}$. We establish asymptotic optimality with new theorems under easily interpretable assumptions on the number of candidate models (M_n) and the minimum averaging risk (ξ_n). The restrictions on the number of regressors (k_m) and the maximum selection risk (η_n) we impose are substantially milder than those in the existing literature. The proof techniques demonstrated in the Appendix illustrate a new path for establishing asymptotic optimality of least squares model averaging.

It is worth noting that there are model averaging studies such as Juditsky and Nemirovski (2000), Yang (2004), and Zhang, Lu, and Zou (2013a) that study the risk bound under a finite sample size. The framework and the assumptions used to derive the risk bound in the above studies are very different from those in our article. The risk bound property does not imply asymptotic optimality in this article unless several restrictive conditions are imposed.

As a final note, our Assumption 2, that is, $\xi_n^{-1}M_n^2 \to 0$, still constrains the rate of increase in M_n as $n \to \infty$. In practice, model screening is usually performed before the model averaging estimation to control the magnitude of M_n. Popular model screening methods include but are not limited to those of Yuan and Yang (2005), Claeskens, Croux, and van Kerckhoven (2006), Zhang et al. (2013a), Hansen (2014), and Xie (2017). The topic of relaxing this constraint is beyond the scope of this article and is left to future research.

REFERENCES

Ando, T. & K.C. Li (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.

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

Claeskens, G., C. Croux, & J. van Kerckhoven (2006) Variable selection for logistic regression using a prediction-focused information criterion. *Biometrics* 62, 972–979.

Gao, Y., X. Zhang, S. Wang, T.T.-L. Chong, & G. Zou (2019) Frequentist model averaging for threshold models. *Annals of the Institute of Statistical Mathematics* 71, 275–306.

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

Juditsky, A. & A. Nemirovski (2000) Functional aggregation for nonparametric regression. *Annals of Statistics* 28, 681–712.

Kuersteiner, G. & R. Okui (2010) Constructing optimal instruments by first-stage prediction averaging. *Econometrica* 78, 697–718.

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

Liu, C.-A. & J. Tao (2016) Model selection and model averaging in nonparametric instrumental variables models. Working paper.

https://doi.org/10.1017/S0266466620000055 Published online by Cambridge University Press

Liu, Q. & R. Okui (2013) Heteroskedasticity-robust C_p model averaging. *The Econometrics Journal* 16, 463–472.
Liu, Q., R. Okui, & A. Yoshimura (2016) Generalized least squares model averaging. *Econometric Reviews* 35, 1692–1752.
Racine, J. (1997) Feasible cross-validatory model selection for general stationary processes. *Journal of Applied Econometrics*, 12 169–179.
Rao, C.R. (1973) *Linear Statistical Inference and its Applications*, vol. 2. Wiley, New York, NY.
Wan, A.T.K., X. Zhang, & G. Zou (2010) Least squares model averaging by Mallows criterion. *Journal of Econometrics* 156, 277–283.
Wiens, D.P. (1992) On moments of quadratic forms in non-spherically distributed variables. *Statistics* 23, 265–270.
Xie, T. (2015) Prediction model averaging estimator. *Economics Letters* 131, 5–8.
Xie, T. (2017) Heteroscedasticity-robust model screening: A useful toolkit for model averaging in big data analytics. *Economics Letters* 151, 119–122.
Yang, Y. (2004) Combining forecasting procedures: Some theoretical results. *Econometric Theory* 20, 176–222.
Yuan, Z. & Y. Yang (2005) Combining linear regression models: When and how? *Journal of the American Statistical Association* 100, 1202–1214.
Zhang, X. (2010) Model averaging and its applications. Ph.D. Thesis, Academy of Mathematics and Systems Science, Chinese Academy of Sciences.
Zhang, X., Z. Lu, & G. Zou (2013a) Adaptively combined forecasting for discrete response time series. *Journal of Econometrics* 176, 80–91.
Zhang, X., A.T. Wan, & G. Zou (2013b) Model averaging by jackknife criterion in models with dependent data. *Journal of Econometrics* 174, 82–94.
Zhang, X., G. Zou, and R.J. Carroll (2015) Model averaging based on Kullback-Leibler distance. *Statistica Sinica* 25, 1583.

APPENDIXES

A.1. NOTATIONS AND PRELIMINARY RESULTS

LEMMA A.1. *(Rao, 1973). For a $p \times p$ positive-definite and symmetric matrix U and a p-dimensional vector u,*

$$\sup_z \frac{\left(z^{\mathrm{T}}u\right)^2}{z^{\mathrm{T}}Uz} = u^{\mathrm{T}}U^{-1}u. \tag{A.1}$$

LEMMA A.2. *Assume that $\mathbb{E}(e_i) = 0$, $\mathbb{E}(e_i^2) = \sigma^2$ and $\mathbb{E}(e_i^4)$ exists. Let $\kappa = E(e_i^4) - 3\sigma^4$. For two $n \times n$ matrices O_1 and O_2,*

$$\begin{aligned}\mathbb{E}(e^{\mathrm{T}}O_1ee^{\mathrm{T}}O_2e) = \sigma^4 &\left\{tr(O_1)tr(O_2) + tr\left(O_1O_2 + O_1^{\mathrm{T}}O_2\right)\right\} \\ &+ \kappa\, tr(O_1 * O_2),\end{aligned} \tag{A.2}$$

*where * denotes the Hadamard product, that is, the ijth component of O_1 * O_2 equals the product of the ijth components of O_1 and O_2.*

https://doi.org/10.1017/S0266466620000055 Published online by Cambridge University Press

Proof of Lemma A.2. From Wiens (1992), we know that if O_1 and O_2 are symmetric, then

$$\mathbb{E}\left(e^{\mathrm{T}}O_1ee^{\mathrm{T}}O_2e\right)=\sigma^4\{\mathrm{tr}(O_1)\mathrm{tr}(O_2)+2\mathrm{tr}(O_1O_2)\}+\kappa\mathrm{tr}(O_1*O_2). \tag{A.3}$$

When O_1 and O_2 are asymmetric, from (A.3), we know that

$$\begin{aligned}\mathbb{E}(e^{\mathrm{T}}O_1ee^{\mathrm{T}}O_2e)&=4^{-1}\mathbb{E}\left\{e^{\mathrm{T}}\left(O_1+O_1^{\mathrm{T}}\right)ee^{\mathrm{T}}\left(O_2+O_2^{\mathrm{T}}\right)e)\right\}\\&=4^{-1}\sigma^4\left[\mathrm{tr}\left(O_1+O_1^{\mathrm{T}}\right)\mathrm{tr}\left(O_2+O_2^{\mathrm{T}}\right)+2\mathrm{tr}\left\{\left(O_1+O_1^{\mathrm{T}}\right)\left(O_2+O_2^{\mathrm{T}}\right)\right\}\right]\\&\quad+4^{-1}\kappa\mathrm{tr}\left\{\left(O_1+O_1^{\mathrm{T}}\right)*\left(O_2+O_2^{\mathrm{T}}\right)\right\}\\&=\sigma^4\left\{\mathrm{tr}(O_1)\mathrm{tr}(O_2)+\mathrm{tr}\left(O_1O_2+O_1^{\mathrm{T}}O_2\right)\right\}+\kappa\mathrm{tr}(O_1*O_2).\end{aligned}$$

This completes the proof. □

LEMMA A.3. *(Zhang, 2010; Gao et al., 2019). Let*

$$\widetilde{w}=argmin_{w\in\mathcal{W}}\{L_n(w)+a_n(w)+b_n\},$$

where $a_n(w)$ is a term related to w and b_n is a term unrelated to w. If

$$\sup_{w\in\mathcal{W}}|a_n(w)|/R_n(w)=o_p(1),\quad\sup_{w\in\mathcal{W}}|R_n(w)-L_n(w)|/R_n(w)=o_p(1),$$

and there exists a constant c and a positive integer N^* so that when $n\geq N^*$, $\inf_{w\in\mathcal{W}}R_n(w)\geq c>0$ almost surely, then $L_n(\widetilde{w})/\inf_{w\in\mathcal{W}}L_n(w)\to 1$ in probability.

Let $\lambda_{\max}(B_1)$ denote the largest eigenvalue of a symmetric matrix B_1 and $\zeta_{\max}(B_2)$ denote the largest singular value of a matrix B_2.

LEMMA A.4. *For any $n_1\times n_2$ matrices B_1 and B_2,*

$$\zeta_{\max}(B_1B_2)\leq\zeta_{\max}(B_1)\zeta_{\max}(B_2) \tag{A.4}$$

and

$$\zeta_{\max}(B_1+B_2)\leq\zeta_{\max}(B_1)+\zeta_{\max}(B_2). \tag{A.5}$$

A.2. PROOF OF THEOREM 1

It is seen that

$$\begin{aligned}C_n(w)-L_n(w)&=\|\widehat{\mu}(w)-y\|^2+2\sigma^2\mathrm{tr}\{P(w)\}-\|\widehat{\mu}(w)-\mu\|^2\\&=2e^{\mathrm{T}}A(w)\mu+2\sigma^2\mathrm{tr}\{P(w)\}-2e^{\mathrm{T}}P(w)e+\|e\|^2,\end{aligned}$$

where the last term is unrelated to w, and

$$\begin{aligned}R_n(w)-L_n(w)&=\|A(w)\mu\|^2+\sigma^2\mathrm{tr}\left\{P^2(w)\right\}-\|\widehat{\mu}(w)-\mu\|^2\\&=2e^{\mathrm{T}}P(w)A(w)\mu+\sigma^2\mathrm{tr}\{P^2(w)\}-e^{\mathrm{T}}P^2(w)e.\end{aligned}$$

https://doi.org/10.1017/S0266466620000055 Published online by Cambridge University Press

In addition, Assumption 2 implies that there exists a constant c and a positive integer N^* so that when $n \geq N^*$, $\xi_n \geq c > 0$ almost surely. Hence, from Lemma A.3, to prove Theorem 1, it is sufficient to verify that

$$\sup_{w \in \mathcal{W}} |e^{\mathrm{T}} A(w)\mu| / R_n(w) = o_p(1), \tag{A.6}$$

$$\sup_{w \in \mathcal{W}} |e^{\mathrm{T}} P(w) A(w)\mu| / R_n(w) = o_p(1), \tag{A.7}$$

$$\sup_{w \in \mathcal{W}} |\sigma^2 \mathrm{tr}\{P(w)\} - e^{\mathrm{T}} P(w) e| / R_n(w) = o_p(1), \tag{A.8}$$

$$\sup_{w \in \mathcal{W}} |\sigma^2 \mathrm{tr}\{P^2(w)\} - e^{\mathrm{T}} P^2(w) e| / R_n(w) = o_p(1). \tag{A.9}$$

For convenience, in all proofs we assume x_i be nonstochastic instead of stochastic. This alternative assumption will not invalidate our proof, because all of our technical assumptions related to x_i hold almost surely.

Proof of (A.6). Let $\Phi = (\mu^{\mathrm{T}} A_{(m)} A_{(s)} \mu)_{M_n \times M_n}$, that is, the msth component of Φ is $\mu^{\mathrm{T}} A_{(m)} A_{(s)} \mu$, $G_{n \times M_n} = (A_{(1)}\mu, \ldots, A_{(M_n)}\mu)$, $\Psi = \left\{\sigma^2 \mathrm{tr}(P_{(m)} P_{(s)})\right\}_{M_n \times M_n}$, and $\Psi_0 = \sigma^2 \mathrm{diag}(k_1, \ldots, k_{M_n})$. So, $\Phi = G^{\mathrm{T}} G$. For any $w \in \mathcal{W}$,

$$w^{\mathrm{T}} \Psi_0 w \leq w^{\mathrm{T}} \Psi w, \tag{A.10}$$

because for any $m, s \in \{1, \ldots, M_n\}$, $w_m \geq 0$, $w_s \geq 0$ and

$$\mathrm{tr}(P_{(m)} P_{(s)}) = \mathrm{tr}\left(P_{(s)}^{1/2} P_{(m)} P_{(s)}^{1/2}\right) \geq 0.$$

In addition,

$$\begin{aligned} R_n(w) &= \mathbb{E}\{L_n(w)\} = \mathbb{E}\|\widehat{\mu}(w) - \mu\|^2 = \mathbb{E}\|P(w)y - \mu\|^2 \\ &= \mathbb{E}\|P(w)\mu - \mu + P(w)e\|^2 \\ &= \|A(w)\mu\|^2 + \sigma^2 \mathrm{tr}\left\{P^2(w)\right\} \\ &= w^{\mathrm{T}}(\Phi + \Psi) w \\ &\geq w^{\mathrm{T}}(\Phi + \Psi_0) w, \end{aligned} \tag{A.11}$$

where the last step is from (A.10). We also have

$$\Phi + \Psi_0 > 0, \tag{A.12}$$

because $\Phi = G^{\mathrm{T}} G$ and $\Psi_0 > 0$. Let $\rho = (e^{\mathrm{T}} A_{(1)}\mu \ldots, e^{\mathrm{T}} A_{(M_n)}\mu)^{\mathrm{T}}$. It is straightforward to show that

$$\mathbb{E}(\rho) = 0 \tag{A.13}$$

and

$$\mathrm{var}(\rho) = \mathbb{E}(\rho\rho^{\mathrm{T}}) = \mathbb{E}\left\{(e^{\mathrm{T}} A_{(m)} \mu \mu^{\mathrm{T}} A_{(s)} e)_{M_n \times M_n}\right\} = \sigma^2 \Phi. \tag{A.14}$$

https://doi.org/10.1017/S0266466620000055 Published online by Cambridge University Press

It is seen that

$$\begin{aligned}\sup_{w\in\mathcal{W}}\frac{\left(e^{\mathrm{T}}A(w)\mu\right)^2}{R_n^2(w)} &= \sup_{w\in\mathcal{W}}\frac{\left(\sum_{m=1}^{M_n}w_m e^{\mathrm{T}}A_{(m)}\mu\right)^2}{R_n^2(w)}\\ &= \sup_{w\in\mathcal{W}}\frac{\left(w^{\mathrm{T}}\rho\right)^2}{R_n^2(w)}\\ &\le \sup_{w\in\mathcal{W}}\frac{\left(w^{\mathrm{T}}\rho\right)^2}{w^{\mathrm{T}}(\Phi+\Psi_0)w}\sup_{w\in\mathcal{W}}\frac{1}{R_n(w)}\\ &\le \xi_n^{-1}\rho^{\mathrm{T}}(\Phi+\Psi_0)^{-1}\rho, \end{aligned} \tag{A.15}$$

where the third step is from (A.11) and the last step is from (A.12) and Lemma A.1. By Markov inequality, we have that for any $\delta > 0$,

$$\begin{aligned}\Pr\left\{\xi_n^{-1}\rho^{\mathrm{T}}(\Phi+\Psi_0)^{-1}\rho > \delta\right\} &\le \delta^{-1}\xi_n^{-1}\mathbb{E}\left\{\rho^{\mathrm{T}}(\Phi+\Psi_0)^{-1}\rho\right\}\\ &= \delta^{-1}\xi_n^{-1}\sigma^2\mathrm{tr}\left\{(\Phi+\Psi_0)^{-1}\Phi\right\}\\ &\le \delta^{-1}\xi_n^{-1}\sigma^2\mathrm{tr}\left\{(\Phi+\Psi_0)^{-1}\Phi+\Psi_0^{1/2}(\Phi+\Psi_0)^{-1}\Psi_0^{1/2}\right\}\\ &= \delta^{-1}\xi_n^{-1}\sigma^2 M_n, \end{aligned} \tag{A.16}$$

where the second step is from (A.13) and (A.14). Combining (A.15), (A.16), and Assumption 2, we obtain (A.6). □

Proof of (A.7). First, we have

$$\mathbb{E}\left(G^{\mathrm{T}}P_{(m)}^{\mathrm{T}}e\right) = 0 \tag{A.17}$$

and

$$\mathrm{var}\left(G^{\mathrm{T}}P_{(m)}^{\mathrm{T}}e\right) = \mathbb{E}\left(G^{\mathrm{T}}P_{(m)}^{\mathrm{T}}ee^{\mathrm{T}}P_{(m)}G\right) = \sigma^2\mathrm{tr}\left(P_{(m)}GG^{\mathrm{T}}P_{(m)}^{\mathrm{T}}\right). \tag{A.18}$$

It is seen that

$$\begin{aligned}\left\{\sup_{w\in\mathcal{W}}\frac{|e^{\mathrm{T}}P(w)A(w)\mu|}{R_n(w)}\right\}^2 &= \left\{\sup_{w\in\mathcal{W}}\frac{|\sum_{m=1}^{M_n}w_m e^{\mathrm{T}}P_{(m)}A(w)\mu|}{R_n(w)}\right\}^2\\ &\le \left\{\sup_{w\in\mathcal{W}}\frac{\sum_{m=1}^{M_n}w_m|e^{\mathrm{T}}P_{(m)}A(w)\mu|}{R_n(w)}\right\}^2\\ &\le \left\{\sup_m\sup_{w\in\mathcal{W}}\frac{|e^{\mathrm{T}}P_{(m)}A(w)\mu|}{R_n(w)}\right\}^2\\ &= \sup_m\sup_{w\in\mathcal{W}}\frac{(e^{\mathrm{T}}P_{(m)}Gw)^2}{R_n^2(w)}\\ &\le \sup_m\sup_{w\in\mathcal{W}}\frac{(e^{\mathrm{T}}P_{(m)}Gw)^2}{w^{\mathrm{T}}(\Phi+\Psi_0)w}\sup_{w\in\mathcal{W}}\frac{1}{R_n(w)}\\ &\le \xi_n^{-1}\sup_m e^{\mathrm{T}}P_{(m)}G(\Phi+\Psi_0)^{-1}G^{\mathrm{T}}P_{(m)}^{\mathrm{T}}e, \end{aligned} \tag{A.19}$$

https://doi.org/10.1017/S0266466620000055 Published online by Cambridge University Press

where the third step is from $\sum_{m=1}^{M_n} w_m = 1$ and $w_m \geq 0$, the fifth step is from (A.11), and the last step is from (A.12) and Lemma A.1. By Markov inequality, we have that for any $\delta > 0$,

$$\begin{aligned}
&\Pr\left\{\xi_n^{-1}\sup_m e^{\mathrm{T}}P_{(m)}G(\Phi+\Psi_0)^{-1}G^{\mathrm{T}}P_{(m)}^{\mathrm{T}}e > \delta\right\} \\
&\leq \sum_{m=1}^{M_n}\Pr\left\{\xi_n^{-1}e^{\mathrm{T}}P_{(m)}G(\Phi+\Psi_0)^{-1}G^{\mathrm{T}}P_{(m)}^{\mathrm{T}}e > \delta\right\} \\
&\leq \sum_{m=1}^{M_n}\delta^{-1}\xi_n^{-1}\mathbb{E}\left\{e^{\mathrm{T}}P_{(m)}G(\Phi+\Psi_0)^{-1}G^{\mathrm{T}}P_{(m)}^{\mathrm{T}}e\right\} \\
&= \sum_{m=1}^{M_n}\delta^{-1}\xi_n^{-1}\sigma^2\mathrm{tr}\left\{P_{(m)}G(\Phi+\Psi_0)^{-1}G^{\mathrm{T}}P_{(m)}^{\mathrm{T}}\right\} \\
&= \sum_{m=1}^{M_n}\delta^{-1}\xi_n^{-1}\sigma^2\mathrm{tr}\left\{(\Phi+\Psi_0)^{-1/2}G^{\mathrm{T}}P_{(m)}^{\mathrm{T}}P_{(m)}G(\Phi+\Psi_0)^{-1/2}\right\} \\
&\leq \sum_{m=1}^{M_n}\delta^{-1}\xi_n^{-1}\sigma^2\lambda_{\max}\left(P_{(m)}^{\mathrm{T}}P_{(m)}\right)\mathrm{tr}\left\{(\Phi+\Psi_0)^{-1/2}G^{\mathrm{T}}G(\Phi+\Psi_0)^{-1/2}\right\} \\
&\leq \sum_{m=1}^{M_n}\delta^{-1}\xi_n^{-1}\sigma^2\mathrm{tr}\left\{(\Phi+\Psi_0)^{-1/2}\Phi(\Phi+\Psi_0)^{-1/2}\right\} \\
&\leq \sum_{m=1}^{M_n}\delta^{-1}\xi_n^{-1}M_n\sigma^2 \\
&= \delta^{-1}\xi_n^{-1}M_n^2\sigma^2, \qquad \textbf{(A.20)}
\end{aligned}$$

where the third step is from (A.17) and (A.18), and the sixth step is from that $G^{\mathrm{T}}G = \Phi$ and that $P_{(m)}$ is symmetric and idempotent. Combining (A.19), (A.20), and Assumption 2, we obtain (A.7). □

Proof of (A.8). Let $\tau = \{\sigma^2\mathrm{tr}(P_{(1)}) - e^{\mathrm{T}}P_{(1)}e, \ldots, \sigma^2\mathrm{tr}(P_{(M_n)}) - e^{\mathrm{T}}P_{(M_n)}e\}^{\mathrm{T}}$. Then, we have

$$\mathbb{E}(\tau) = 0 \qquad \textbf{(A.21)}$$

and

$$\begin{aligned}
\mathrm{var}(\tau) &= \mathbb{E}\left(\tau\tau^{\mathrm{T}}\right) \\
&= \mathbb{E}\left[\left\{\left(\sigma^2\mathrm{tr}(P_{(m)}) - e^{\mathrm{T}}P_{(m)}e\right)\left(\sigma^2\mathrm{tr}(P_{(s)}) - e^{\mathrm{T}}P_{(s)}e\right)\right\}_{M_n\times M_n}\right] \\
&= \mathbb{E}\left[\left\{e^{\mathrm{T}}P_{(m)}ee^{\mathrm{T}}P_{(s)}e - \sigma^4\mathrm{tr}(P_{(m)})\mathrm{tr}(P_{(s)})\right\}_{M_n\times M_n}\right] \\
&= \left\{2\sigma^4\mathrm{tr}\left(P_{(m)}P_{(s)}\right) + \kappa\mathrm{tr}\left(P_{(m)} * P_{(s)}\right)\right\}_{M_n\times M_n}, \qquad \textbf{(A.22)}
\end{aligned}$$

https://doi.org/10.1017/S0266466620000055 Published online by Cambridge University Press

where the last step is from (A.3). In (A.22), Assumption 1 is actually used because we utilize κ here, but for simplicity, we do not list the assumption here and other places when it is used. It is seen that

$$\begin{aligned}\sup_{w\in\mathcal{W}}\frac{\left\{\sigma^2\mathrm{tr}\{P(w)\}-e^{\mathrm{T}}P(w)e\right\}^2}{R_n^2(w)} &= \sup_{w\in\mathcal{W}}\frac{\left(w^{\mathrm{T}}\tau\right)^2}{R_n^2(w)}\\ &\le \xi_n^{-1}\sup_{w\in\mathcal{W}}\frac{\left(w^{\mathrm{T}}\tau\right)^2}{w^{\mathrm{T}}\Psi_0 w}\\ &\le \xi_n^{-1}\tau^{\mathrm{T}}\Psi_0^{-1}\tau, \end{aligned} \tag{A.23}$$

where the second step is from (A.11) and the last step is from Lemma A.1. By Markov inequality, for any $\delta>0$,

$$\begin{aligned}&\Pr\left\{\xi_n^{-1}\tau^{\mathrm{T}}\Psi_0^{-1}\tau>\delta\right\}\\ &\quad\le \delta^{-1}\xi_n^{-1}\mathbb{E}\left(\tau^{\mathrm{T}}\Psi_0^{-1}\tau\right)\\ &\quad= \delta^{-1}\xi_n^{-1}\mathrm{tr}\left[\Psi_0^{-1}\left\{2\sigma^4\mathrm{tr}\left(P_{(m)}P_{(s)}\right)+\kappa\mathrm{tr}\left(P_{(m)}*P_{(s)}\right)\right\}_{M_n\times M_n}\right]\\ &\quad= 2\sigma^4\delta^{-1}\xi_n^{-1}\mathrm{tr}\left[\Psi_0^{-1}\left\{\mathrm{tr}\left(P_{(m)}P_{(s)}\right)\right\}_{M_n\times M_n}\right]\\ &\qquad+\kappa\delta^{-1}\xi_n^{-1}\mathrm{tr}\left[\Psi_0^{-1}\left\{\mathrm{tr}\left(P_{(m)}*P_{(s)}\right)\right\}_{M_n\times M_n}\right]\\ &\quad= 2\sigma^2\delta^{-1}\xi_n^{-1}\mathrm{tr}\left[\Psi_0^{-1}\Psi_0\right]\\ &\qquad+\kappa\delta^{-1}\xi_n^{-1}\mathrm{tr}\left[\Psi_0^{-1}\mathrm{diag}\left\{\mathrm{tr}\left(P_{(m)}*P_{(m)}\right);m=1,\ldots,M_n\right\}\right]\\ &\quad\le 2\sigma^2\delta^{-1}\xi_n^{-1}M_n+\kappa\delta^{-1}\xi_n^{-1}\mathrm{tr}\left[\Psi_0^{-1}\mathrm{diag}\left\{\mathrm{tr}\left(P_{(m)}^2\right);m=1,\ldots,M_n\right\}\right]\\ &\quad= 2\sigma^2\delta^{-1}\xi_n^{-1}M_n+\kappa\sigma^{-2}\delta^{-1}\xi_n^{-1}M_n, \end{aligned} \tag{A.24}$$

where $\mathrm{diag}\left\{\mathrm{tr}(P_{(m)}*P_{(m)});m=1,\ldots,M_n\right\}$ denotes a diagonal matrix with the mth component being $\mathrm{tr}(P_{(m)}*P_{(m)})$, the second step is from (A.21) and (A.22), the fourth step is because Ψ_0^{-1} is a diagonal matrix, and the fifth step is from the truth that for any symmetric matrix O,

$$\mathrm{tr}(O*O)\le\mathrm{tr}\left(O^2\right). \tag{A.25}$$

Combining (A.23), (A.24), and Assumption 2, we obtain (A.8). □

Proof of (A.9). For $j\in\{1,\ldots,M_n\}$, let

$$\nu_{(j)}=\left(\sigma^2\mathrm{tr}\left(P_{(j)}P_{(1)}\right)-e^{\mathrm{T}}P_{(j)}P_{(1)}e,\ldots,\sigma^2\mathrm{tr}\left(P_{(j)}P_{(M_n)}\right)-e^{\mathrm{T}}P_{(j)}P_{(M_n)}e\right)^{\mathrm{T}}.$$

Then, we have

$$\mathbb{E}\left(\nu_{(j)}\right)=0 \tag{A.26}$$

https://doi.org/10.1017/S0266466620000055 Published online by Cambridge University Press

and

$$\begin{aligned}\operatorname{var}(\nu_{(j)}) &= \mathbb{E}\left(\nu_{(j)}\nu_{(j)}^{\mathrm{T}}\right)\\ &= \mathbb{E}\left[\left\{\left(\sigma^2\operatorname{tr}\left(P_{(j)}P_{(m)}\right)-e^{\mathrm{T}}P_{(j)}P_{(m)}e\right)\left(\sigma^2\operatorname{tr}(P_{(j)}P_{(s)})-e^{\mathrm{T}}P_{(j)}P_{(s)}e\right)\right\}_{M_n\times M_n}\right]\\ &= \mathbb{E}\left[\left\{e^{\mathrm{T}}P_{(j)}P_{(m)}ee^{\mathrm{T}}P_{(j)}P_{(s)}e-\sigma^4\operatorname{tr}\left(P_{(j)}P_{(m)}\right)\operatorname{tr}\left(P_{(j)}P_{(s)}\right)\right\}_{M_n\times M_n}\right]\\ &= \left[\sigma^4\operatorname{tr}\left(P_{(j)}P_{(m)}P_{(j)}P_{(s)}\right)+\sigma^4\operatorname{tr}\left(P_{(m)}P_{(j)}P_{(s)}\right)\right.\\ &\quad \left.+\kappa\operatorname{tr}\left\{\left(P_{(j)}P_{(m)}\right)*\left(P_{(j)}P_{(s)}\right)\right\}\right]_{M_n\times M_n}, \end{aligned} \tag{A.27}$$

where the last step is from Lemma A.2. For any $m, j \in \{1, \ldots, M_n\}$, we first show that

$$\begin{aligned}\operatorname{tr}\left(P_{(j)}P_{(m)}P_{(j)}P_{(m)}\right) &= \operatorname{tr}(P_{(m)}^{1/2}P_{(j)}P_{(m)}P_{(j)}P_{(m)}^{1/2})\\ &= \lambda_{\max}(P_{(m)})\operatorname{tr}\left(P_{(m)}^{1/2}P_{(j)}P_{(m)}^{1/2}\right)\\ &\le \operatorname{tr}\left(P_{(m)}^{1/2}P_{(j)}P_{(m)}^{1/2}\right)\\ &\le \operatorname{tr}(P_{(m)}), \end{aligned} \tag{A.28}$$

where the third step is from $\lambda_{\max}(P_{(m)}) \le 1$. Similarly, we have

$$\operatorname{tr}\left(P_{(m)}P_{(j)}P_{(m)}\right) \le \operatorname{tr}(P_{(m)}) \tag{A.29}$$

and

$$\operatorname{tr}\left\{\left(P_{(j)}P_{(m)}+P_{(m)}P_{(j)}\right)^2\right\} \le 4\operatorname{tr}\left(P_{(m)}\right). \tag{A.30}$$

It is seen that

$$\begin{aligned}&\left\{\sup_{w\in\mathcal{W}}\frac{\left|\sigma^2\operatorname{tr}\{P^2(w)\}-e^{\mathrm{T}}P(w)P(w)e\right|}{R_n(w)}\right\}^2\\ &= \left\{\sup_{w\in\mathcal{W}}\frac{\left|\sum_{j=1}^{M_n}w_j\sigma^2\operatorname{tr}\{P_{(j)}P(w)\}-\sum_{j=1}^{M_n}w_je^{\mathrm{T}}P_{(j)}P(w)e\right|}{R_n(w)}\right\}^2\\ &\le \left\{\sup_{w\in\mathcal{W}}\frac{\sum_{j=1}^{M_n}w_j|\sigma^2\operatorname{tr}\{P_{(j)}P(w)\}-e^{\mathrm{T}}P_{(j)}P(w)e|}{R_n(w)}\right\}^2\\ &\le \left\{\sup_j\sup_{w\in\mathcal{W}}\frac{|\sigma^2\operatorname{tr}\{P_{(j)}P(w)\}-e^{\mathrm{T}}P_{(j)}P(w)e|}{R_n(w)}\right\}^2\\ &= \sup_j\sup_{w\in\mathcal{W}}\frac{\left\{\sigma^2\operatorname{tr}\{P_{(j)}P(w)\}-e^{\mathrm{T}}P_{(j)}P(w)e\right\}^2}{R_n^2(w)}\\ &= \sup_j\sup_{w\in\mathcal{W}}\frac{\left(w^{\mathrm{T}}\nu_{(j)}\right)^2}{R_n^2(w)}\end{aligned}$$

https://doi.org/10.1017/S0266466620000055 Published online by Cambridge University Press

$$\begin{aligned}
&\le \xi_n^{-1} \sup_j \sup_{w\in\mathcal{W}} \frac{\left(w^{\mathrm{T}}\nu_{(j)}\right)^2}{w^{\mathrm{T}}\Psi_0 w} \\
&\le \xi_n^{-1} \sup_j \nu_{(j)}^{\mathrm{T}}\Psi_0^{-1}\nu_{(j)} \\
&\le \xi_n^{-1} \sum_{j=1}^{M_n} \nu_{(j)}^{\mathrm{T}}\Psi_0^{-1}\nu_{(j)}, \qquad \textbf{(A.31)}
\end{aligned}$$

where the sixth step is from (A.11) and the seventh step is from Lemma A.1. By Markov inequality, for any $\delta > 0$ and any $j \in \{1, \ldots, M_n\}$,

$$\begin{aligned}
&\Pr\left\{\xi_n^{-1}\nu_{(j)}^{\mathrm{T}}\Psi_0^{-1}\nu_{(j)} > \delta\right\} \\
&\le \delta^{-1}\xi_n^{-1}\mathbb{E}\left\{\nu_{(j)}^{\mathrm{T}}\Psi_0^{-1}\nu_{(j)}\right\} \\
&= \delta^{-1}\xi_n^{-1}\mathrm{tr}\Big[\Psi_0^{-1}\Big\{\sigma^4\mathrm{tr}\left(P_{(j)}P_{(m)}P_{(j)}P_{(s)}\right)+\sigma^4\mathrm{tr}\left(P_{(m)}P_{(j)}P_{(s)}\right) \\
&\qquad\qquad +\kappa\mathrm{tr}\left(\left(P_{(j)}P_{(m)}\right)*\left(P_{(j)}P_{(s)}\right)\right)\Big\}_{M_n\times M_n}\Big] \\
&= \delta^{-1}\xi_n^{-1}\sigma^4\mathrm{tr}\Big[\Psi_0^{-1}\left\{\mathrm{tr}\left(P_{(j)}P_{(m)}P_{(j)}P_{(s)}\right)+\mathrm{tr}\left(P_{(m)}P_{(j)}P_{(s)}\right)\right\}_{M_n\times M_n}\Big] \\
&\quad +\delta^{-1}\xi_n^{-1}\kappa\mathrm{tr}\Big[\Psi_0^{-1}\left\{\mathrm{tr}\left(\left(P_{(j)}P_{(m)}\right)*\left(P_{(j)}P_{(s)}\right)\right)\right\}_{M_n\times M_n}\Big] \\
&= \delta^{-1}\xi_n^{-1}\sigma^4\mathrm{tr}\Big[\Psi_0^{-1}\mathrm{diag}\left\{\mathrm{tr}\left(P_{(j)}P_{(m)}P_{(j)}P_{(m)}\right)\right. \\
&\qquad\qquad \left.+\mathrm{tr}\left(P_{(m)}P_{(j)}P_{(m)}\right); m=1,\ldots,M_n\right\}\Big] \\
&\quad +\delta^{-1}\xi_n^{-1}\kappa\mathrm{tr}\Big[\Psi_0^{-1}\mathrm{diag}\left\{\mathrm{tr}\left(\left(P_{(j)}P_{(m)}\right)*\left(P_{(j)}P_{(m)}\right)\right); m=1,\ldots,M_n\right\}\Big] \\
&= \delta^{-1}\xi_n^{-1}\sigma^4\mathrm{tr}\Big[\Psi_0^{-1}\mathrm{diag}\left\{\mathrm{tr}\left(P_{(j)}P_{(m)}P_{(j)}P_{(m)}\right)\right. \\
&\qquad\qquad \left.+\mathrm{tr}\left(P_{(m)}P_{(j)}P_{(m)}\right); m=1,\ldots,M_n\right\}\Big] \\
&\quad +\delta^{-1}\xi_n^{-1}\kappa 4^{-1}\mathrm{tr}\Big[\Psi_0^{-1}\mathrm{diag}\left\{\mathrm{tr}\left(\left(P_{(j)}P_{(m)}+P_{(m)}P_{(j)}\right)*\left(P_{(j)}P_{(m)}\right.\right.\right. \\
&\qquad\qquad \left.\left.\left.+P_{(m)}P_{(j)}\right)\right); m=1,\ldots,M_n\right\}\Big] \\
&\le \delta^{-1}\xi_n^{-1}2\sigma^4\mathrm{tr}\Big[\Psi_0^{-1}\mathrm{diag}\left\{\mathrm{tr}(P_{(m)}); m=1,\ldots,M_n\right\}\Big] \\
&\quad +\delta^{-1}\xi_n^{-1}\kappa 4^{-1}\mathrm{tr}\Big[\Psi_0^{-1}\mathrm{diag}\left\{\mathrm{tr}\left(\left(P_{(j)}P_{(m)}+P_{(m)}P_{(j)}\right)^2\right); m=1,\ldots,M_n\right\}\Big] \\
&\le \delta^{-1}\xi_n^{-1}2\sigma^4\mathrm{tr}\Big[\Psi_0^{-1}\mathrm{diag}\left\{\mathrm{tr}(P_{(m)}); m=1,\ldots,M_n\right\}\Big] \\
&\quad +\delta^{-1}\xi_n^{-1}\kappa\mathrm{tr}\Big[\Psi_0^{-1}\mathrm{diag}\left\{\mathrm{tr}(P_{(m)}); m=1,\ldots,M_n\right\}\Big] \\
&= \delta^{-1}\xi_n^{-1}M_n\left(2\sigma^4+\kappa\right)\sigma^{-2}, \qquad \textbf{(A.32)}
\end{aligned}$$

where the second step is from (A.26) and (A.27), the sixth step is from (A.25), (A.28), and (A.29), and the seventh step is from (A.30). Combining (A.31), (A.32), and Assumption 2, we obtain (A.9). We have established (A.6)–(A.9), which are sufficient for (5) as stated above (A.6). □

https://doi.org/10.1017/S0266466620000055 Published online by Cambridge University Press

A.3. PROOF OF THEOREM 2

Define the prediction loss of $\widetilde{\mu}(w)$ as $\widetilde{L}_n(w) = \|\widetilde{\mu}(w) - \mu\|^2$ and the corresponding risk as $\widetilde{R}_n(w) = \mathbb{E}\{\widetilde{L}_n(w)\}$. First, from the proof of Thm. 2 in Hansen and Racine (2012) and the first part of Assumption 4, we have

$$\sup_{w\in\mathcal{W}} |\widetilde{R}_n(w)/R_n(w) - 1| = o_p(1). \tag{A.33}$$

Hence, as in the proofs of Thm. 1 in Hansen and Racine (2012) and Thm. 2.1 of Zhang et al. (2013b), to prove Theorem 2, we only need to verify the following two results:

$$\sup_{w\in\mathcal{W}} |L_n(w)/R_n(w) - 1| = o_p(1) \tag{A.34}$$

and

$$\widetilde{L}_n(\widehat{w}_{\text{JMA}})/\inf_{w\in\mathcal{W}} \widetilde{L}_n(w) \to 1, \qquad \text{in probability.} \tag{A.35}$$

Let $\Omega = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$. Similar to (A.6), under heteroscedasticity, we have

$$\begin{aligned} R_n(w) - L_n(w) &= \|A(w)\mu\|^2 + \text{tr}\left\{P^2(w)\Omega\right\} - \|\widehat{\mu}(w) - \mu\|^2 \\ &= 2e^{\text{T}}P(w)A(w)\mu + \sigma^2\text{tr}\{P^2(w)\Omega\} - e^{\text{T}}P^2(w)e. \end{aligned}$$

So, to prove (A.34), it is sufficient to verify (A.7) under heteroscedasticity and

$$\sup_{w\in\mathcal{W}} |\text{tr}\{P^2(w)\Omega\} - e^{\text{T}}P^2(w)e|/R_n(w) = o_p(1). \tag{A.36}$$

Proof of (A.7) under heteroscedasticity. Let $\widetilde{\Psi}_0 = \underline{\sigma}^2\text{diag}(k_1, \ldots, k_{M_n})$, where $\underline{\sigma}^2$ is defined in Assumption 3. From Assumption 3 and the derivation steps in (A.11), we know that under heteroscedasticity,

$$\begin{aligned} R_n(w) &= \|A(w)\mu\|^2 + \text{tr}\left\{P^2(w)\Omega\right\} \\ &\geq \|A(w)\mu\|^2 + \underline{\sigma}^2\text{tr}\left\{P^2(w)\right\} \\ &\geq w^{\text{T}}(\Phi + \widetilde{\Psi}_0)w. \end{aligned} \tag{A.37}$$

Using (A.37) and the steps in (A.19), we have

$$\left\{\sup_{w\in\mathcal{W}} \frac{|e^{\text{T}}P(w)A(w)\mu|}{R_n(w)}\right\}^2 \leq \xi_n^{-1}\sup_m e^{\text{T}}P_{(m)}G(\Phi + \widetilde{\Psi}_0)^{-1}G^{\text{T}}P_{(m)}^{\text{T}}e. \tag{A.38}$$

Using the steps in (A.20), we have

$$\begin{aligned} &\Pr\left\{\xi_n^{-1}\sup_m e^{\text{T}}P_{(m)}G\left(\Phi + \widetilde{\Psi}_0\right)^{-1}G^{\text{T}}P_{(m)}^{\text{T}}e > \delta\right\} \\ &= \sum_{m=1}^{M_n} \delta^{-1}\xi_n^{-1}\text{tr}\left\{P_{(m)}G\left(\Phi + \widetilde{\Psi}_0\right)^{-1}G^{\text{T}}P_{(m)}^{\text{T}}\Omega\right\} \end{aligned}$$

https://doi.org/10.1017/S0266466620000055 Published online by Cambridge University Press

$$\leq \max_{1\leq i\leq n}\sigma_i^2\sum_{m=1}^{M_n}\delta^{-1}\xi_n^{-1}\mathrm{tr}\left\{P_{(m)}G(\Phi+\widetilde{\Psi}_0)^{-1}G^{\mathrm{T}}P_{(m)}^{\mathrm{T}}\right\}$$
$$=\delta^{-1}\xi_n^{-1}M_n^2\max_{1\leq i\leq n}\sigma_i^2. \tag{A.39}$$

Combining (A.38) and(A.39) and Assumptions 1 and 2, we obtain (A.7) under heteroscedasticity. □

Proof of (A.36). For $j\in\{1,\ldots,M_n\}$, let

$$\widetilde{\nu}_{(j)}=\left(\mathrm{tr}\left(P_{(j)}P_{(1)}\Omega\right)-e^{\mathrm{T}}P_{(j)}P_{(1)}e,\ldots,\mathrm{tr}\left(P_{(j)}P_{(M_n)}\Omega\right)-e^{\mathrm{T}}P_{(j)}P_{(M_n)}e\right)^{\mathrm{T}}$$

and

$$\widetilde{\kappa}=E\left(e_i^4\right)/\sigma_i^4-3.$$

Then, similar to (A.26) and (A.27), we have

$$\mathbb{E}\left(\widetilde{\nu}_{(j)}\right)=0 \tag{A.40}$$

and

$$\begin{aligned}\mathrm{var}\left(\widetilde{\nu}_{(j)}\right)&=\mathbb{E}\left(\widetilde{\nu}_{(j)}\widetilde{\nu}_{(j)}^{\mathrm{T}}\right)\\&=\mathbb{E}\left[\left\{\left(\mathrm{tr}\left(P_{(j)}P_{(m)}\Omega\right)-e^{\mathrm{T}}P_{(j)}P_{(m)}e\right)\left(\mathrm{tr}\left(P_{(j)}P_{(s)}\Omega\right)-e^{\mathrm{T}}P_{(j)}P_{(s)}e\right)\right\}_{M_n\times M_n}\right]\\&=\mathbb{E}\Big[\Big\{e^{\mathrm{T}}\Omega^{-1/2}\Omega^{1/2}P_{(j)}P_{(m)}\Omega^{1/2}\Omega^{-1/2}ee^{\mathrm{T}}\Omega^{-1/2}\Omega^{1/2}P_{(j)}P_{(s)}\Omega^{1/2}\Omega^{-1/2}e\\&\quad-\mathrm{tr}\left(P_{(j)}P_{(m)}\Omega\right)\mathrm{tr}\left(P_{(j)}P_{(s)}\Omega\right)\Big\}_{M_n\times M_n}\Big]\\&=\Big\{\mathrm{tr}\left(P_{(j)}P_{(m)}\Omega P_{(j)}P_{(s)}\Omega\right)+\mathrm{tr}\left(P_{(m)}P_{(j)}\Omega P_{(j)}P_{(s)}\Omega\right)\\&\quad+\widetilde{\kappa}\mathrm{tr}\left(\left(\Omega^{1/2}P_{(j)}P_{(m)}\Omega^{1/2}\right)*\left(\Omega^{1/2}P_{(j)}P_{(s)}\Omega^{1/2}\right)\right)\Big\}_{M_n\times M_n},\end{aligned} \tag{A.41}$$

where the last step is from Lemma A.2. Using the steps in (A.31), we have

$$\left\{\sup_{w\in\mathcal{W}}\frac{|\mathrm{tr}\{P^2(w)\Omega\}-e^{\mathrm{T}}P^2(w)e|}{R_n(w)}\right\}^2\leq\xi_n^{-1}\sup_j\widetilde{\nu}_{(j)}^{\mathrm{T}}\widetilde{\Psi}_0^{-1}\widetilde{\nu}_{(j)}. \tag{A.42}$$

Using (A.40), (A.41), and the steps in (A.32), we have

$$\begin{aligned}\Pr\left\{\xi_n^{-1}\widetilde{\nu}_{(j)}^{\mathrm{T}}\widetilde{\Psi}_0^{-1}\widetilde{\nu}_{(j)}>\delta\right\}&=\delta^{-1}\xi_n^{-1}\mathrm{tr}\Big[\widetilde{\Psi}_0^{-1}\mathrm{diag}\left\{\mathrm{tr}\left(P_{(j)}P_{(m)}\Omega P_{(j)}P_{(m)}\Omega\right)\right.\\&\qquad\left.+\mathrm{tr}(P_{(m)}P_{(j)}\Omega P_{(j)}P_{(m)}\Omega);m=1,\ldots,M_n\right\}\Big]\\&\quad+\delta^{-1}\xi_n^{-1}\widetilde{\kappa}4^{-1}\mathrm{tr}\Big[\widetilde{\Psi}_0^{-1}\mathrm{diag}\left\{\mathrm{tr}\left((P_{(j)}P_{(m)}\Omega+P_{(m)}P_{(j)}\Omega)\right.\right.\\&\qquad\left.\left.*(P_{(j)}P_{(m)}\Omega+P_{(m)}P_{(j)}\Omega)\right);m=1,\ldots,M_n\right\}\Big]\\&\leq\left(\max_{1\leq i\leq n}\sigma_i^2\right)^2\delta^{-1}\xi_n^{-1}2\mathrm{tr}\Big[\widetilde{\Psi}_0^{-1}\mathrm{diag}\left\{\mathrm{tr}(P_{(m)});m=1,\ldots,M_n\right\}\Big]\end{aligned}$$

https://doi.org/10.1017/S0266466620000055 Published online by Cambridge University Press

$$+\Big(\max_{1\le i\le n}\sigma_i^2\Big)^2\delta^{-1}\xi_n^{-1}\widetilde{\kappa}\mathrm{tr}\Big[\widetilde{\Psi}_0^{-1}\mathrm{diag}\left\{\mathrm{tr}(P_{(m)});m=1,\ldots,M_n\right\}\Big]$$
$$\le\Big(\max_{1\le i\le n}\sigma_i^2\Big)^2\delta^{-1}\xi_n^{-1}M_n(2+\widetilde{\kappa})\underline{\sigma}^{-2}. \tag{A.43}$$

Obviously, when $\max_{1\le i\le n}\sigma_i^2=\underline{\sigma}^2=\sigma^2$, we have $\kappa=\sigma^4\widetilde{\kappa}$ (κ is defined in Lemma A.2) and thus the conclusions of (A.32) and (A.43) are the same. Combining (A.42), (A.43), and Assumptions 1 and 2, we obtain (A.36). Hence, (A.34) is obtained by (A.7) under heteroscedasticity and (A.36). □

Proof of (A.35). As in Hansen and Racine (2012), we define a matrix $\widetilde{P}_{(m)}$ so that $\widetilde{\mu}_{(m)}=\widetilde{P}_{(m)}y$. Let $D_{(m)}$ be an $n\times n$ diagonal matrix with the ith diagonal component being $(1-h_{m,ii})^{-1}$. From Racine (1997), $\widetilde{P}_{(m)}=D_{(m)}(P_{(m)}-I_n)+I_n$. Let $Q_{(m)}$ be an $n\times n$ diagonal matrix with the ith diagonal component being $h_{m,ii}/(1-h_{m,ii})$. So

$$\begin{aligned}\widetilde{P}_{(m)}&=D_{(m)}(P_{(m)}-I_n)+I_n\\&=(I_n+Q_{(m)})(P_{(m)}-I_n)+I_n\\&=P_{(m)}-Q_{(m)}A_{(m)}.\end{aligned} \tag{A.44}$$

Let $\widetilde{P}(w)=\sum_{m=1}^{M_n}w_m\widetilde{P}_m$ and $\widetilde{A}(w)=I_n-\widetilde{P}(w)$. So

$$\begin{aligned}\mathcal{J}_n(w)-\widetilde{L}_n(w)&=\|\widetilde{\mu}(w)-y\|^2-\|\widetilde{\mu}(w)-\mu\|^2\\&=2e^{\mathrm{T}}\widetilde{A}(w)\mu-2e^{\mathrm{T}}\widetilde{P}(w)e+\|e\|^2\end{aligned}$$

and

$$\widetilde{R}_n(w)-\widetilde{L}_n(w)=2e^{\mathrm{T}}\widetilde{P}(w)^{\mathrm{T}}\widetilde{A}(w)\mu+\mathrm{tr}\left\{\widetilde{P}(w)\Omega\widetilde{P}(w)^{\mathrm{T}}\right\}-e^{\mathrm{T}}\widetilde{P}(w)^{\mathrm{T}}\widetilde{P}(w)e.$$

Let $\widetilde{\xi}_n=\inf_{w\in\mathcal{W}}\widetilde{R}_n(w)$. From (A.33) and Assumption 2, we have

$$M_n^2\widetilde{\xi}_n^{-1}\to 0 \tag{A.45}$$

almost surely, which will play the same role in the proof of (A.35) as Assumption 2 in the proof of Theorem 1. Equation (A.45) implies that there exists a constant c and a positive integer N^* so that when $n\ge N^*$, $\widetilde{\xi}_n\ge c>0$ almost surely. Therefore, from Lemma A.3, to prove (A.35), it is sufficient to verify that

$$\sup_{w\in\mathcal{W}}|e^{\mathrm{T}}\widetilde{A}(w)\mu|/\widetilde{R}_n(w)=o_p(1), \tag{A.46}$$
$$\sup_{w\in\mathcal{W}}|e^{\mathrm{T}}\widetilde{P}(w)^{\mathrm{T}}\widetilde{A}(w)\mu|/\widetilde{R}_n(w)=o_p(1), \tag{A.47}$$
$$\sup_{w\in\mathcal{W}}|e^{\mathrm{T}}\widetilde{P}(w)e|/\widetilde{R}_n(w)=o_p(1), \tag{A.48}$$
$$\sup_{w\in\mathcal{W}}|\mathrm{tr}\{\widetilde{P}(w)\Omega\widetilde{P}(w)^{\mathrm{T}}\}-e^{\mathrm{T}}\widetilde{P}(w)^{\mathrm{T}}\widetilde{P}(w)e|/\widetilde{R}_n(w)=o_p(1). \tag{A.49}$$

Their proof steps are similar to the proof steps of (A.6)–(A.9), respectively. It is seen that

$$\begin{aligned}&\mathrm{tr}\{\widetilde{P}(w)\Omega\widetilde{P}(w)^{\mathrm{T}}\}\\&\quad\ge\underline{\sigma}^2\mathrm{tr}\left\{\widetilde{P}(w)\widetilde{P}(w)^{\mathrm{T}}\right\}\end{aligned}$$

https://doi.org/10.1017/S0266466620000055 Published online by Cambridge University Press

$$
\begin{aligned}
&= \underline{\sigma}^2 w^{\mathrm{T}} \left\{ \mathrm{tr}\left(\widetilde{P}_{(m)} \widetilde{P}_{(s)}^{\mathrm{T}} \right) \right\}_{M_n \times M_n} w \\
&= \underline{\sigma}^2 w^{\mathrm{T}} \left\{ \mathrm{tr}\left(\left(P_{(m)} - Q_{(m)} A_{(m)} \right) \left(P_{(s)} - Q_{(s)} A_{(s)} \right)^{\mathrm{T}} \right) \right\}_{M_n \times M_n} w \\
&= \underline{\sigma}^2 w^{\mathrm{T}} \left\{ \mathrm{tr}\left(P_{(m)} P_{(s)} + Q_{(m)} A_{(m)} A_{(s)} Q_{(s)} \right. \right. \\
&\qquad \left. \left. - Q_{(m)} A_{(m)} P_{(s)} - P_{(m)} A_{(s)} Q_{(s)} \right) \right\}_{M_n \times M_n} w \\
&= \underline{\sigma}^2 \mathrm{tr}\left\{ P^2(w) \right\} + \underline{\sigma}^2 \mathrm{tr} \left\{ \sum_{m=1}^{M_n} w_m Q_{(m)} A_{(m)} \left(\sum_{m=1}^{M_n} w_m Q_{(m)} A_{(m)} \right)^{\mathrm{T}} \right\} \\
&\quad - \underline{\sigma}^2 w^{\mathrm{T}} \left\{ \mathrm{tr}\left(Q_{(m)} A_{(m)} P_{(s)} + P_{(m)} A_{(s)} Q_{(s)} \right) \right\}_{M_n \times M_n} w \\
&\geq \underline{\sigma}^2 \mathrm{tr}\left\{ P^2(w) \right\} - \underline{\sigma}^2 w^{\mathrm{T}} \left\{ \mathrm{tr}\left(Q_{(m)} A_{(m)} P_{(s)} + P_{(m)} A_{(s)} Q_{(s)} \right) \right\}_{M_n \times M_n} w \\
&= \underline{\sigma}^2 \mathrm{tr}\left\{ P^2(w) \right\} - 2^{-1} \underline{\sigma}^2 w^{\mathrm{T}} \left\{ \mathrm{tr}\left(\left(Q_{(m)} A_{(m)} P_{(s)} + P_{(m)} A_{(s)} Q_{(s)} \right) \right. \right. \\
&\quad \left. \left. + \left(Q_{(m)} A_{(m)} P_{(s)} + P_{(m)} A_{(s)} Q_{(s)} \right)^{\mathrm{T}} \right) \right\}_{M_n \times M_n} w \\
&\geq \underline{\sigma}^2 \mathrm{tr}\left\{ P^2(w) \right\} - 2^{-1} \underline{\sigma}^2 \sup_m \sup_s \left[\lambda_{\max} \left\{ \left(Q_{(m)} A_{(m)} P_{(s)} + P_{(m)} A_{(s)} Q_{(s)} \right) \right. \right. \\
&\quad \left. \left. + \left(Q_{(m)} A_{(m)} P_{(s)} + P_{(m)} A_{(s)} Q_{(s)} \right)^{\mathrm{T}} \right\} \times \mathrm{rank} \left\{ \left(Q_{(m)} A_{(m)} P_{(s)} \right. \right. \right. \\
&\quad \left. \left. \left. + P_{(m)} A_{(s)} Q_{(s)} \right) + \left(Q_{(m)} A_{(m)} P_{(s)} + P_{(m)} A_{(s)} Q_{(s)} \right)^{\mathrm{T}} \right\} \right] \\
&\geq \underline{\sigma}^2 \mathrm{tr}\left\{ P^2(w) \right\} - 2^{-1} \underline{\sigma}^2 \sup_m \sup_s \left[\left\{ \lambda_{\max} \left(Q_{(m)} A_{(m)} P_{(s)} \right) + \lambda_{\max} \left(P_{(m)} A_{(s)} Q_{(s)} \right) \right. \right. \\
&\quad \left. \left. + \lambda_{\max} \left(Q_{(m)} A_{(m)} P_{(s)} \right) + \lambda_{\max} \left(P_{(m)} A_{(s)} Q_{(s)} \right) \right\} \right. \\
&\quad \times \left\{ \mathrm{rank} \left(Q_{(m)} A_{(m)} P_{(s)} \right) + \mathrm{rank} \left(P_{(m)} A_{(s)} Q_{(s)} \right) \right. \\
&\quad \left. \left. + \mathrm{rank} \left(Q_{(m)} A_{(m)} P_{(s)} \right) + \mathrm{rank} \left(P_{(m)} A_{(s)} Q_{(s)} \right) \right\} \right] \\
&\geq \underline{\sigma}^2 \mathrm{tr}\left\{ P^2(w) \right\} - 8 \underline{\sigma}^2 \bar{h} (1 - \bar{h})^{-1} \bar{k}, \qquad \textbf{(A.50)}
\end{aligned}
$$

where the last two steps are from Lemma A.4. Let $\widetilde{\Phi} = (\mu^{\mathrm{T}} \widetilde{A}_{(m)} \widetilde{A}_{(s)} \mu)_{M_n \times M_n}$. Then, similar to (A.11), we have

$$
\begin{aligned}
\widetilde{R}_n(w) &= \| \widetilde{A}(w) \mu \|^2 + \mathrm{tr}\left\{ \widetilde{P}(w) \Omega \widetilde{P}(w)^{\mathrm{T}} \right\} \\
&\geq w^{\mathrm{T}} \left(\widetilde{\Phi} + \sigma^{-2} \underline{\sigma}^2 \Psi_0 \right) w - 8 \underline{\sigma}^2 \bar{h} (1 - \bar{h})^{-1} \bar{k}, \qquad \textbf{(A.51)}
\end{aligned}
$$

which will play the same role as (A.11) in the proof of Theorem 1. Now, by the above results, Assumption 4 and the proof steps of (A.6)–(A.9), we can obtain (A.46)–(A.49). □

https://doi.org/10.1017/S0266466620000055 Published online by Cambridge University Press