Model Averaging in Predictive Regressions*

Chu-An Liu† and Biing-Shen Kuo‡

November 2015

Abstract

This paper considers forecast combination in a predictive regression. We construct the point forecast by combining predictions from all possible linear regression models given a set of potentially relevant predictors. We derive the asymptotic risk of least squares averaging estimators in a local asymptotic framework. We then develop a frequentist model averaging criterion, an asymptotically unbiased estimator of the asymptotic risk, to select forecast weights. Monte Carlo simulations show that our averaging estimator compares favorably with alternative methods such as weighted AIC, weighted BIC, Mallows model averaging, and jackknife model averaging. The proposed method is applied to stock return predictions.

JEL Classification: C52, C53

Keywords: Forecast combination, Local asymptotic theory, Plug-in estimators.

*We thank the co-editor, two referees, Bruce Hansen, and Jack Porter for many constructive comments and suggestions. We are also grateful to Sheng-Kai Chang, Jau-er Chen, Serena Ng, Tatsushi Oka, Liangjun Su, Denis Tkachenko, Aman Ullah, and conference participants of CFE-ERCIM 2013, EEA-ESEM 2013, AMES 2013, SETA 2013, the Tsinghua International Conference in Econometrics, and CMES 2013 for their discussions and suggestions. All errors remain the authors'.

†Institute of Economics, Academia Sinica. Email: caliu@econ.sinica.edu.tw.

‡Department of International Business, National Chengchi University. Email: bsku@nccu.edu.tw.

1 Introduction

The challenge of empirical studies on forecasting practice is that one does not know exactly what predictors should be included in the true model. In order to address the model uncertainty, forecast combination has been widely used in economics and statistics; see Granger (1989), Clemen (1989), Timmermann (2006), and Stock and Watson (2006) for literature reviews. Although there is plenty of empirical evidence to support the success of forecast combination, there is no unified view on selecting the forecast weights in a general framework.

The main goal of this paper is to provide a data-driven approach to weight selection for forecast combination. Building on the idea of the focused information criterion (FIC) proposed by Claeskens and Hjort (2003), we introduce a frequentist model averaging criterion to select the weights for candidate models and study its properties. More recently, FIC has been extended to several models, including the general semiparametric model (Claeskens and Carroll, 2007), the generalized additive partial linear model (Zhang and Liang, 2011), the Tobin model with a nonzero threshold (Zhang, Wan, and Zhou, 2012), the generalized empirical likelihood estimation (Sueishi, 2013), the generalized method of moments estimation (DiTraglia, 2014), and the propensity score weighted estimation of the treatment effects (Kitagawa and Muris, 2013; Lu, 2015). Despite the growing literature on FIC, little work has been done on forecast combination.

Following Hjort and Claeskens (2003), Hansen (2014), and Liu (2015), we examine the asymptotic risk of least squares averaging estimators in a local asymptotic framework where the regression coefficients of potentially relevant predictors are in a local $T^{-1/2}$ neighborhood of zero. This local-to-zero framework ensures the consistency of the averaging estimator while in general presents an asymptotic bias. The local asymptotic framework has an advantage of yielding the same stochastic order of squared biases and variances. Thus, the optimal forecast combination is the one that achieves the best trade-off between bias and variance in this context.

For a given set of potentially relevant predictors, we construct the point forecast by combining predictions from all possible linear regression models. Under the local-to-zero assumption, we derive the asymptotic distribution of the averaging estimator for a predictive regression model. We show that the averaging estimator with fixed weights is asymptotically normal and then derive a representation for the asymptotic risk of least squares averaging estimators without the i.i.d. normal assumption. This result allows us to decompose the asymptotic risk into the bias and variance components.

Hence, the proposed model averaging criterion can be used to address the trade-off between bias and variance of forecast combination. The proposed model averaging criterion is an estimate of the asymptotic risk. Therefore, the data-driven weights that minimize the model averaging criterion are expected to close to the optimal weights that minimize the asymptotic risk.

To illustrate the proposed forecast combination approach, we study the predictability of U.S. stock returns. Following Welch and Goyal (2008) and Rapach, Strauss, and Zhou (2010), we use U.S. quarterly data to investigate the out-of-sample equity premium. We find strong evidence that the performance of the proposed approach is better than the historical average benchmark. In particular, our forecast combination approach achieves lower cumulative squared prediction error than those produced by other averaging methods such as weighted AIC, weighted BIC, and jackknife model averaging. Our results support the findings of Rapach, Strauss, and Zhou (2010) and Elliott, Gargano, and Timmermann (2013) that forecast combinations consistently achieve significant gains on out-of-sample predictions.

We now discuss the related literature. There is a large body of literature on forecast combination, including both Bayesian and frequentist model averaging. Since the seminal work of Bates and Granger (1969), many forecast combination methods are proposed, including Granger and Ramanathan (1984), Min and Zellner (1993), Raftery, Madigan, and Hoeting (1997), Buckland, Burnham, and Augustin (1997), Yang (2004), Zou and Yang (2004), Hansen (2008), Hansen (2010), Elliott, Gargano, and Timmermann (2013), and Cheng and Hansen (2015). There are also many alternative approaches to combine or shrink forecasts, for example, bagging (Breiman, 1996; Inoue and Kilian, 2008), the LASSO (Tibshirani, 1996), the adaptive LASSO (Zou, 2006), and the model confidence set (Hansen, Lunde, and Nason, 2011), among others.

Our paper is closely related to Hansen (2008), who proposes to select the forecast weights by minimizing the Mallows model averaging (MMA) criterion. The MMA criterion approximates the mean squared forecast error (MSFE) by the sum of squared errors and a penalty term. Hence, the MMA criterion addresses the trade-off between the model fit and model complexity. Hansen (2008) shows that the MMA criterion is an asymptotically unbiased estimator of the MSFE in a homoskedastic linear regression model. Like the MMA criterion, our model averaging criterion is also asymptotically unbiased for the MSFE. We, however, employ a drifting asymptotic framework to approximate the MSFE, and do not restrict model errors to be homoskedastic. In this paper, we show that the proposed plug-in averaging estimator is a generalized Mallows'

C_p-type averaging estimator for predictive regression models with heteroskedastic errors. The plug-in averaging estimator is equivalent to the MMA estimator in the homoskedastic framework. Numerical comparisons show that our estimator achieves lower relative risk than the MMA estimator in most simulations.

One popular model averaging approach is the simple equal-weighted average. The simple equal-weighted average is appropriate to use if all the candidate models have similar prediction powers. Recently, Elliott, Gargano, and Timmermann (2013) extend the idea of the simple equal-weighted average to complete subset regressions. They construct the forecast combination by using equal-weighted combination based on all possible models that keep the number of predictors fixed. Instead of choosing the weights, the subset regression combinations have to choose the number of predictors κ, and the data-driven method for κ still needs further investigation. Monte Carlo shows that the performance of complete subset regressions is sensitive to the choice of κ, while the performance of our model averaging criterion is relatively robust in most simulations.

There is a large literature on the asymptotic optimality of model selection. Shibata (1980) and Ing and Wei (2005) demonstrate that model selection estimators based on the Akaike information criterion or the final prediction criterion asymptotically achieve the lowest possible MSFE in homoskedastic autoregressive models. Li (1987) shows the asymptotic optimality of the Mallows criterion in homoskedastic linear regression models. Andrews (1991a) extends the asymptotic optimality to the heteroskedastic linear regression models. Shao (1997) provides a general framework to discuss the asymptotic optimality of various model selection procedures.

The existing literature on the asymptotic optimality of model averaging is comparatively small. Hansen (2007) introduces the MMA estimator and demonstrates the asymptotic optimality of the MMA estimator for nested and homoskedastic linear regression models. Wan, Zhang, and Zou (2010) extend the asymptotic optimality of the MMA estimator for continuous weights and a non-nested setup. Hansen and Racine (2012) propose the jackknife model averaging estimator and demonstrate the asymptotic optimality in heteroskedastic linear regression models. Liu and Okui (2013) propose the heteroskedasticity-robust C_p estimator and demonstrate its optimality in the linear regression models with heteroskedastic errors. Zhang, Zou, and Liang (2014) propose a Mallows-type model averaging estimator for the linear mixed-effects models and establish the asymptotic optimality. These asymptotic theories, however, are limited to the random sample and hence are not directly applicable to forecast combination for dependent data. In a recent paper, Zhang, Wan, and Zou (2013) show

the asymptotic optimality of the jackknife model averaging estimator in the presence of lagged dependent variables. They assume that the dependent variable follows the stationary AR(∞) process. A more general theory needs to be developed in a future study.

The outline of the paper is as follows. Section 2 presents the forecasting model and describes the averaging estimator. Section 3 presents the asymptotic framework and the plug-in averaging estimator for forecast combination and discusses the relationship between the plug-in averaging estimator and the Mallows' C_p-type averaging estimator. Section 4 evaluates the finite sample performance of the plug-in averaging estimator and other averaging estimators in two simulation experiments. Section 5 applies the plug-in forecast combination to the predictability of U.S. stock returns. Section 6 concludes the paper. Proofs and figures are included in the Appendix.

2 Model and Estimation

Suppose we have observations $(y_t, \mathbf{x}_t, \mathbf{z}_t)$ for $t = 1, ..., T$. The goal is to construct a point forecast of y_{T+1} given $(\mathbf{x}_T, \mathbf{z}_T)$ using the one-step-ahead forecasting model

$$y_{t+1} = \mathbf{x}_t'\boldsymbol{\beta} + \mathbf{z}_t'\boldsymbol{\gamma} + e_{t+1}, \tag{2.1}$$

$$\mathrm{E}(\mathbf{h}_t e_{t+1}) = 0, \tag{2.2}$$

where y_{t+1} is a scalar dependent variable, $\mathbf{h}_t = (\mathbf{x}_t', \mathbf{z}_t')'$, $\mathbf{x}_t$ $(p \times 1)$ and $\mathbf{z}_t$ $(q \times 1)$ are vectors of predictors, and e_t is an unobservable error term. Here, $\mathbf{x}_t$ is a set of "must-have" predictors, which must be included in the model based on theoretical grounds, while $\mathbf{z}_t$ is a set of "potentially relevant" predictors, which may or may not be included in the model. Note that $\mathbf{x}_t$ is allowed to be an empty matrix or include only a constant term. The potentially relevant predictors could be lags of y_t, deterministic terms, any nonlinear transformations of the original predictors, or the interaction terms between the predictors. The error term is allowed to be heteroskedastic, and there is no further assumption on the distribution of the error term. We assume throughout that $1 \leq p + q \leq T - 1$, and we do not let the number of predictors p and q increase with the sample size T.

We now consider a set of M approximating models indexed by $m = 1, ..., M$, where the mth model includes all must-have predictors $\mathbf{x}_t$ and a subset of potentially relevant predictors $\mathbf{z}_t$. The mth model has $p + q_m$ predictors. We do not place any restrictions on the model space. The set of models could be nested or non-nested. If we consider

a sequence of nested models, then $M = q + 1$. If we consider all possible subsets of potentially relevant predictors, then $M = 2^q$.

Let $\mathbf{y} = (y_1, y_2, ..., y_T)'$, $\mathbf{X} = (\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_{T-1})'$, $\mathbf{Z} = (\mathbf{z}_0, \mathbf{z}_1, ..., \mathbf{z}_{T-1})'$, and $\mathbf{e} = (e_1, e_2, ..., e_T)'$. In matrix notation, the model (2.1) can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{e} = \mathbf{H}\boldsymbol{\theta} + \mathbf{e}, \tag{2.3}$$

where $\mathbf{H} = (\mathbf{X}, \mathbf{Z})$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\gamma}')'$.

Let $\boldsymbol{\Pi}_m$ be a $q_m \times q$ selection matrix that selects the included potentially relevant predictors in the mth model. Let $\mathbf{I}$ denote an identity matrix and $\mathbf{0}$ a zero matrix. Also,

$$\mathbf{S}_m = \begin{pmatrix} \mathbf{I}_p & \mathbf{0}_{p\times q_m} \\ \mathbf{0}_{q\times p} & \boldsymbol{\Pi}_m' \end{pmatrix}$$

is a selection matrix of dimension $(p + q) \times (p + q_m)$.

The unconstrained least squares estimator of $\boldsymbol{\theta}$ in the full model is $\widehat{\boldsymbol{\theta}} = (\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\mathbf{y}$, and the least squares estimator in the mth submodel is $\widehat{\boldsymbol{\theta}}_m = (\mathbf{H}_m'\mathbf{H}_m)^{-1}\mathbf{H}_m'\mathbf{y}$, where $\mathbf{H}_m = (\mathbf{X}, \mathbf{Z}_m) = (\mathbf{X}, \mathbf{Z}\boldsymbol{\Pi}_m') = \mathbf{H}\mathbf{S}_m$. The predicted value is $\widehat{\mathbf{y}}(m) = \mathbf{H}_m\widehat{\boldsymbol{\theta}}_m = \mathbf{H}\mathbf{S}_m\widehat{\boldsymbol{\theta}}_m$. Thus, the one-step-ahead forecast given information up to period T from this mth model is

$$\widehat{y}_{T+1|T}(m) = \mathbf{h}_T'\mathbf{S}_m\widehat{\boldsymbol{\theta}}_m. \tag{2.4}$$

Let $\mathbf{w} = (w_1, ..., w_M)'$ be a weight vector with $w_m \geq 0$ and $\sum_{m=1}^M w_m = 1$. That is, $\mathbf{w} \in \mathcal{H}^M$ where $\mathcal{H}^M = \left\{\mathbf{w} \in [0, 1]^M : \sum_{m=1}^M w_m = 1\right\}$. The one-step-ahead combination forecast is

$$\widehat{y}_{T+1|T}(\mathbf{w}) = \sum_{m=1}^M w_m\widehat{y}_{T+1|T}(m) = \sum_{m=1}^M w_m\mathbf{h}_T'\mathbf{S}_m\widehat{\boldsymbol{\theta}}_m = \mathbf{h}_T'\widehat{\boldsymbol{\theta}}(\mathbf{w}), \tag{2.5}$$

where $\widehat{\boldsymbol{\theta}}(\mathbf{w}) = \sum_{m=1}^M w_m\mathbf{S}_m\widehat{\boldsymbol{\theta}}_m$ is an averaging estimator of $\boldsymbol{\theta}$.

3 Forecast Combinations

In the previous section we defined the one-step-ahead combination forecast with fixed weights. Our goal is to select the forecast weights to minimize the asymptotic risk

over the set of all possible forecast combinations. In this section, we first describe the connection between the asymptotic risk, in-sample mean squared error (MSE), and one-step-ahead mean squared forecast error (MSFE). We then characterize the optimal weights of forecast combinations in a local asymptotic framework and present a plug-in method to estimate the infeasible optimal weights. In the last subsection, we show the equivalence between the plug-in averaging estimator and the Mallows' C_p-type averaging estimator.

3.1 MSE and MSFE

We first show that the one-step-ahead MSFE approximately equals the in-sample MSE when the observations are strictly stationary. Thus, the weight vector that minimizes the in-sample MSE is expected to minimize the one-step-ahead MSFE.

Let $\sigma^2 = \mathrm{E}(e_t^2)$ and $\mu_t = \mathbf{x}_t'\boldsymbol{\beta} + \mathbf{z}_t'\boldsymbol{\gamma}$ be the conditional mean. Then, we rewrite the model (2.1) as $y_{t+1} = \mu_t + e_{t+1}$. Similarly, for any fixed-weight vector, we write $\widehat{\mu}_t(\mathbf{w}) = \sum_{m=1}^{M} w_m \mathbf{h}_t'\mathbf{S}_m\widehat{\boldsymbol{\theta}}_m = \mathbf{h}_t'\widehat{\boldsymbol{\theta}}(\mathbf{w})$.

Following common practice, we consider the quadratic loss function and define the in-sample MSE as

$$MSE(\mathbf{w}) = \mathrm{E}\left(\frac{1}{T}\sum_{t=1}^{T}\left(\mu_t - \widehat{\mu}_t(\mathbf{w})\right)^2\right). \tag{3.1}$$

The in-sample MSE measures the global fit of the averaging estimator since it is constructed using the entire sample. Following a similar argument in Cheng and Hansen (2015), we have

$$\begin{aligned} MSFE(\mathbf{w}) &= \mathrm{E}\left(y_{T+1} - \widehat{y}_{T+1|T}(\mathbf{w})\right)^2 \\ &= \mathrm{E}\left(e_{T+1}^2 + (\mu_T - \widehat{\mu}_T(\mathbf{w}))^2\right) \\ &\simeq \mathrm{E}\left(e_{T+1}^2 + (\mu_t - \widehat{\mu}_t(\mathbf{w}))^2\right) \\ &= \sigma^2 + MSE(\mathbf{w}), \end{aligned} \tag{3.2}$$

where the second equality holds by the fact that e_{T+1} is uncorrelated with $\widehat{\mu}_T(\mathbf{w})$ and the approximation in the third line is valid for stationary $(y_t, \mathbf{h}_t)$.[1]

Let the optimal weight vector be the value that minimizes $MSE(\mathbf{w})$ over $\mathbf{w} \in \mathcal{H}^M$.

[1]Hansen (2008) shows that the MSFE approximately equals MSE in a homoskedastic linear regression model with stationary time series data. Elliott, Gargano, and Timmermann (2013) also have a similar argument for complete subset regressions.

Since σ^2 is a constant and not related to the weight vector $\mathbf{w}$, Equation (3.2) implies that the optimal weight vector that minimizes the $MSE(\mathbf{w})$ is expected to minimize the $MSFE(\mathbf{w})$.

Let $\mathbf{Q} = \mathrm{E}\left(\mathbf{h}_t\mathbf{h}_t'\right) > 0$. We follow Hansen (2014) to define the asymptotic trimmed risk or weighted MSE of an estimator $\tilde{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ as

$$R(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \lim_{\zeta\to\infty} \liminf_{T\to\infty} \mathrm{E} \min\{T(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})'\mathbf{Q}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}), \zeta\}. \tag{3.3}$$

Note that $\mathrm{E}((\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})'\mathbf{Q}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}))$ is the risk of an estimator $\tilde{\boldsymbol{\theta}}$ under the weighted squared error loss function, which may not be finite unless $\tilde{\boldsymbol{\theta}}$ has sufficient finite moments. Thus, we use ζ to bound the expectation when the risk does not exist for finite T. We choose the covariance matrix $\mathbf{Q}$ as a weight matrix, so that the weighted MSE function (3.3) plus σ^2 corresponds to one-step-ahead MSFE.[2] Thus, it is natural to use the asymptotic risk to approximate the MSE. The asymptotic risk is well-defined and straightforward to calculate when the estimator $\tilde{\boldsymbol{\theta}}$ has an asymptotic distribution. In order to obtain a good approximation to the finite sample behavior, we study the asymptotic trimmed risk in a local asymptotic framework, which we will describe in the following section.

3.2 Local Asymptotic Framework

In a constant parameter model, i.e., nonzero and fixed values of $\boldsymbol{\gamma}$, the least squares estimators for all possible models except the full model have omitted variable bias. The risk of these models tends to infinity with the sample size, and hence the asymptotic approximations break down. To obtain a useful approximation, we follow Hjort and Claeskens (2003), Hansen (2014), and Liu (2015), and use a local-to-zero asymptotic framework to approximate the in-sample MSE. More precisely, the parameters $\boldsymbol{\gamma}$ are modeled as being in a local $T^{-1/2}$ neighborhood of zero. This local-to-zero framework is similar to that used in weak instrument theory (Staiger and Stock, 1997).

Assumption 1. $\boldsymbol{\gamma} = \boldsymbol{\gamma}_T = \boldsymbol{\delta}/\sqrt{T}$, where $\boldsymbol{\delta}$ is an unknown vector.

Assumption 2. $\{y_{t+1}, \mathbf{h}_t\}$ is a strictly stationary and ergodic time series with finite $r > 4$ moments and $\mathrm{E}(e_{t+1}|\mathcal{F}_t) = 0$, where $\mathcal{F}_t = \sigma(\mathbf{h}_t, \mathbf{h}_{t-1}, ...; e_t, e_{t-1}, ...)$.

[2] As mentioned by Hansen (2014), the weight matrix $\mathbf{Q}$ induces invariance to parameter scaling and rotation, and the trimming ζ is introduced to avoid the requirement of the uniform integrability condition.

Assumption 1 assumes that $\boldsymbol{\gamma}$ is local to zero, and it ensures that the asymptotic mean squared error of the averaging estimator remains finite. The local asymptotic framework is a technical device commonly used to analyze the asymptotic and finite sample properties of the model selection and averaging estimator, for example, Claeskens and Hjort (2003), Leeb and Pötscher (2005), Pötscher (2006), and Elliott, Gargano, and Timmermann (2013). Note that the $O(T^{-1/2})$ framework gives squared model biases of the same order $O(T^{-1})$ as estimator variances. Hence, in this context the optimal forecast combination is the one that achieves the best trade-off between bias and variance. Alternatively, Assumption 1 could be replaced by imposing the i.i.d. normal assumption on the error term; see Hansen (2014) for a discussion.

Assumption 2 states that data is strictly stationary, and it implies that e_{t+1} is conditionally unpredictable at time t. Assumption 2 is similar to Assumption 1 of Hansen (2014) and Assumption R(i)-(ii) of Cheng and Hansen (2015). Assumption 2 is sufficient to imply that $T^{-1}\mathbf{H}'\mathbf{H} \xrightarrow{p} \mathbf{Q}$ and $T^{-1/2}\mathbf{H}'\mathbf{e} \xrightarrow{d} \mathbf{R} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\Omega})$ where $\boldsymbol{\Omega} = \mathrm{E}\left(\mathbf{h}_t\mathbf{h}_t'e_{t+1}^2\right)$. Note that if the error term is i.i.d. and homoskedastic, then $\boldsymbol{\Omega}$ can be simplified as $\boldsymbol{\Omega} = \sigma^2\mathbf{Q}$. Since the selection matrix $\mathbf{S}_m$ is nonrandom with elements either 0 or 1, for the mth model we have $T^{-1}\mathbf{H}_m'\mathbf{H}_m \xrightarrow{p} \mathbf{Q}_m$ where $\mathbf{Q}_m = \mathbf{S}_m'\mathbf{Q}\mathbf{S}_m$ is nonsingular. The following theorem establishes the asymptotic distribution of the averaging estimator with fixed weights.

Theorem 1. *Suppose that Assumptions 1–2 hold. As $T \to \infty$, we have*

$$\sqrt{T}\left(\widehat{\boldsymbol{\theta}}(\mathbf{w}) - \boldsymbol{\theta}\right) \xrightarrow{d} \mathbf{N}\big(\mathbf{A}(\mathbf{w})\boldsymbol{\delta}, \mathbf{V}(\mathbf{w})\big)$$

$$\mathbf{A}(\mathbf{w}) = \sum_{m=1}^{M} w_m \left(\mathbf{P}_m\mathbf{Q} - \mathbf{I}_{p+q}\right)\mathbf{S}_0$$

$$\mathbf{V}(\mathbf{w}) = \sum_{m=1}^{M} w_m^2\mathbf{P}_m\boldsymbol{\Omega}\mathbf{P}_m + 2\sum\sum_{m\neq\ell} w_m w_\ell \mathbf{P}_m\boldsymbol{\Omega}\mathbf{P}_\ell$$

where $\mathbf{P}_m = \mathbf{S}_m\mathbf{Q}_m^{-1}\mathbf{S}_m'$ and $\mathbf{S}_0 = (\mathbf{0}_{q\times p}, \mathbf{I}_q)'$.

Theorem 1 shows the asymptotic normality of the averaging estimator with nonrandom weights. We use this result to compute the asymptotic trimmed risk of the averaging estimator. If we assign the whole weight to the full model, i.e., all predictors are included in the model, it is easy to see that we have a conventional asymptotic distribution with mean zero (zero bias) and sandwich-form variance $\mathbf{Q}^{-1}\boldsymbol{\Omega}\mathbf{Q}^{-1}$. Note that $\mathbf{A}(\mathbf{w})\boldsymbol{\delta}$ represents the asymptotic bias term of the averaging estimator $\widehat{\boldsymbol{\theta}}(\mathbf{w})$. The

magnitude of the asymptotic bias is determined by the covariance matrix $\mathbf{Q}$ and the local parameter $\boldsymbol{\delta}$. The asymptotic variance of the averaging estimator $\mathbf{V}(\mathbf{w})$ has two components. The first component is the weighted average of the variance of each model, and the second component is the weighted average of the covariance between any two models.

3.3 Weighted Focused Information Criterion

The model selection estimator is a special case of the model averaging estimator. If we consider the unit weight vector $\mathbf{w}_{1,m}$, where the mth element is one and the others are zeros, then the averaging estimator simplifies to a selection estimator. Let $\widehat{\boldsymbol{\theta}}(m) = \mathbf{S}_m\widehat{\boldsymbol{\theta}}_m = \widehat{\boldsymbol{\theta}}(\mathbf{w}_{1,m})$ be the least squares estimator of θ in the mth submodel.

Theorem 2. *Suppose that Assumptions 1–2 hold. We have*

$$R(\widehat{\boldsymbol{\theta}}(m), \boldsymbol{\theta}) = \operatorname{tr}\left(\mathbf{Q}\mathbf{C}_m\boldsymbol{\delta}\boldsymbol{\delta}'\mathbf{C}_m'\right) + \operatorname{tr}\left(\mathbf{Q}\mathbf{P}_m\boldsymbol{\Omega}\mathbf{P}_m\right) \tag{3.4}$$

where $\mathbf{C}_m = \left(\mathbf{P}_m\mathbf{Q} - \mathbf{I}_{p+q}\right)\mathbf{S}_0$.

Theorem 2 presents the asymptotic trimmed risk of the least squares estimators in the mth model under the local asymptotic framework. We can use (3.4) to select a best approximating model, and this is the idea of the weighted focused information criterion (wFIC) proposed by Claeskens and Hjort (2008). Let $\widehat{m}$ be the model that minimizes (3.4). Combining Theorem 2 with (3.2), we deduce that $\widehat{m}$ is expected to be the model that minimizes the MSFE.

To use (3.4) for model selection, we need to estimate the unknown parameters $\mathbf{Q}$, $\boldsymbol{\Omega}$, $\mathbf{C}_m$, $\mathbf{P}_m$, and $\boldsymbol{\delta}$. Let $\widehat{\mathbf{Q}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{h}_t\mathbf{h}_t'$. Then we have $\widehat{\mathbf{Q}} \xrightarrow{p} \mathbf{Q}$ under Assumption 2. The covariance matrix $\boldsymbol{\Omega}$ can also be consistently estimated by the method of moments estimator.[3] Note that both $\widehat{\mathbf{C}}_m$ and $\widehat{\mathbf{P}}_m$ are functions of $\mathbf{Q}$ and selection matrices, which can also be consistently estimated by the sample analogue under Assumption 2.

[3]Let $\widehat{e}_{t+1} = y_{t+1} - \mathbf{h}_t'\widehat{\boldsymbol{\theta}}$ be the least squares residual for the full model. The heteroskedasticity and autocorrelation consistent covariance matrix estimator is $\widehat{\boldsymbol{\Omega}} = \sum_{j=-T}^{T} K(j/S_T)\widehat{\boldsymbol{\Gamma}}(j)$, where $K(\cdot)$ is a kernel function, S_T is the bandwidth, $\widehat{\boldsymbol{\Gamma}}(j) = \frac{1}{T}\sum_{t=1}^{T-j}\mathbf{h}_t\mathbf{h}_{t+j}'\widehat{e}_{t+1}\widehat{e}_{t+1+j}$ for $j \geq 0$, and $\widehat{\boldsymbol{\Gamma}}(j) = \widehat{\boldsymbol{\Gamma}}(-j)'$ for $j < 0$. Under some regularity conditions, it follows that $\widehat{\boldsymbol{\Omega}} \xrightarrow{p} \boldsymbol{\Omega}$; see Newey and West (1987) and Andrews (1991b). If the error term is serially uncorrelated and identically distributed, then $\boldsymbol{\Omega}$ can be consistently estimated by $\widehat{\boldsymbol{\Omega}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{h}_t\mathbf{h}_t'\widehat{e}_{t+1}^2$, the heteroskedasticity-consistent covariance matrix estimator proposed by White (1980).

Unlike other unknown parameters, the consistent estimator for the local parameter $\boldsymbol{\delta}$ is not available due to the local asymptotic framework. We can, however, construct an asymptotically unbiased estimator of $\boldsymbol{\delta}$ by using the estimator from the full model. That is, $\widehat{\boldsymbol{\delta}} = \sqrt{T}\widehat{\boldsymbol{\gamma}}$. Theorem 1 and the delta method show that

$$\widehat{\boldsymbol{\delta}} = \sqrt{T}\widehat{\boldsymbol{\gamma}} \xrightarrow{d} \mathbf{R}_{\boldsymbol{\delta}} = \boldsymbol{\delta} + \mathbf{S}_0'\mathbf{Q}^{-1}\mathbf{R} \sim N(\boldsymbol{\delta}, \mathbf{S}_0'\mathbf{Q}^{-1}\boldsymbol{\Omega}\mathbf{Q}^{-1}\mathbf{S}_0). \tag{3.5}$$

As shown above, $\widehat{\boldsymbol{\delta}}$ is an asymptotically unbiased estimator for $\boldsymbol{\delta}$. Since the mean of $\mathbf{R}_{\boldsymbol{\delta}}\mathbf{R}_{\boldsymbol{\delta}}'$ is $\boldsymbol{\delta}\boldsymbol{\delta}' + \mathbf{S}_0'\mathbf{Q}^{-1}\boldsymbol{\Omega}\mathbf{Q}^{-1}\mathbf{S}_0$, we construct the asymptotically unbiased estimator of $\boldsymbol{\delta}\boldsymbol{\delta}'$ as

$$\widehat{\boldsymbol{\delta}\boldsymbol{\delta}'} = \widehat{\boldsymbol{\delta}}\widehat{\boldsymbol{\delta}}' - \mathbf{S}_0'\widehat{\mathbf{Q}}^{-1}\widehat{\boldsymbol{\Omega}}\widehat{\mathbf{Q}}^{-1}\mathbf{S}_0. \tag{3.6}$$

Following Claeskens and Hjort (2008), we define the wFIC of the mth submodel as

$$\text{wFIC}(m) = \text{tr}\left(\widehat{\mathbf{Q}}\widehat{\mathbf{C}}_m\left(\widehat{\boldsymbol{\delta}}\widehat{\boldsymbol{\delta}}' - \mathbf{S}_0'\widehat{\mathbf{Q}}^{-1}\widehat{\boldsymbol{\Omega}}\widehat{\mathbf{Q}}^{-1}\mathbf{S}_0\right)\widehat{\mathbf{C}}_m'\right) + \text{tr}\left(\widehat{\mathbf{Q}}\widehat{\mathbf{P}}_m\widehat{\boldsymbol{\Omega}}\widehat{\mathbf{P}}_m\right), \tag{3.7}$$

which is an asymptotically unbiased estimator of $R(\widehat{\boldsymbol{\theta}}(m), \boldsymbol{\theta})$. We then select the model with the lowest wFIC.

3.4 Plug-In Averaging Estimator

We now extend the idea of the weighted focused information criterion to model averaging.[4] The following theorem presents the asymptotic trimmed risk of the averaging estimator in the local asymptotic framework.

Theorem 3. *Suppose that Assumptions 1–2 hold. We have*

$$R(\widehat{\boldsymbol{\theta}}(\mathbf{w}), \boldsymbol{\theta}) = \mathbf{w}'\boldsymbol{\psi}\mathbf{w} \tag{3.8}$$

where $\boldsymbol{\psi}$ is an $M \times M$ matrix with the (m, ℓ)th element

$$\psi_{m,\ell} = \text{tr}\left(\mathbf{Q}\mathbf{C}_m\boldsymbol{\delta}\boldsymbol{\delta}'\mathbf{C}_\ell'\right) + \text{tr}\left(\mathbf{Q}\mathbf{P}_m\boldsymbol{\Omega}\mathbf{P}_\ell\right). \tag{3.9}$$

[4]Claeskens and Hjort (2008) propose a smoothed wFIC averaging estimator, which assigns the weights of each candidate model by using the exponential wFIC. The simulations show that the performance of the smoothed wFIC averaging estimator is sensitive to the choice of the nuisance parameter. Furthermore, there is no data-driven method available for the nuisance parameter.

Note that the mth diagonal element of $\boldsymbol{\psi}$ characterizes the bias and variance of the mth model while the off-diagonal elements measure the product of biases and covariance between different models. Theorem 3 is a more general statement than Theorem 2 of Elliott, Gargano, and Timmermann (2013). First, we do not restrict the setup to i.i.d. data. Second, we allow any arbitrary combination between models. Third, we do not restrict the weights to be equal.

From Theorem 3, we define the optimal weight vector as the value that minimizes the asymptotic risk over $\mathbf{w} \in \mathcal{H}^M$:

$$\mathbf{w}^o = \underset{\mathbf{w}\in\mathcal{H}^M}{\operatorname{argmin}}\, \mathbf{w}'\boldsymbol{\psi}\mathbf{w}. \tag{3.10}$$

Combining Theorem 3 with (3.2), we deduce that $\mathbf{w}^o$ is nearly equivalent to the optimal weight vector that minimizes the MSFE. Note that the objection function is linear-quadratic in $\mathbf{w}$, which means the optimal weight vector can be computed numerically via quadratic programming.

The optimal weights, however, are infeasible, since they depend on the unknown parameter $\boldsymbol{\psi}$. Similar to Liu (2015), we propose a plug-in estimator to estimate the optimal weights for the forecasting model. We first estimate the asymptotic risk by plugging in an asymptotically unbiased estimator. We then choose the data-driven weights by minimizing the sample analog of the asymptotic risk and use these estimated weights to construct the one-step-ahead forecast combination.

Let $\widehat{\boldsymbol{\psi}}$ be a sample analog of $\boldsymbol{\psi}$ with the (m,ℓ)th element

$$\widehat{\psi}_{m,\ell} = \operatorname{tr}(\widehat{\mathbf{Q}}\widehat{\mathbf{C}}_m\widehat{\boldsymbol{\delta}\boldsymbol{\delta}'}\widehat{\mathbf{C}}_\ell') + \operatorname{tr}(\widehat{\mathbf{Q}}\widehat{\mathbf{P}}_m\widehat{\boldsymbol{\Omega}}\widehat{\mathbf{P}}_\ell). \tag{3.11}$$

The data-driven weights based on the plug-in estimator are defined as

$$\widehat{\mathbf{w}} = \underset{\mathbf{w}\in\mathcal{H}^M}{\operatorname{argmin}}\, \mathbf{w}'\widehat{\boldsymbol{\psi}}\mathbf{w}, \tag{3.12}$$

where $\mathbf{w}'\widehat{\boldsymbol{\psi}}\mathbf{w}$ is an asymptotically unbiased estimator of $\mathbf{w}'\boldsymbol{\psi}\mathbf{w}$.[5] Similar to the optimal weight vector, the data-driven weights can also be found numerically via quadratic

[5]Claeskens and Hjort (2008) suggest estimating the first term of $\psi_{m,\ell}$ by $\max\{0, \operatorname{tr}(\widehat{\mathbf{Q}}\widehat{\mathbf{C}}_m\widehat{\boldsymbol{\delta}\boldsymbol{\delta}'}\widehat{\mathbf{C}}_\ell')\}$ to avoid the negative estimate for the squared bias term. However, our simulations show that this modified estimator is not a stable estimator for $\psi_{m,\ell}$. Therefore, we focus on the estimator (3.11) in this paper.

programming.[6] The plug-in one-step-ahead combination forecast is

$$\widehat{y}_{T+1|T}(\widehat{\mathbf{w}}) = \mathbf{h}_T'\widehat{\boldsymbol{\theta}}(\widehat{\mathbf{w}}). \tag{3.13}$$

As mentioned by Hjort and Claeskens (2003), we can also estimate $\boldsymbol{\psi}$ by inserting $\widehat{\boldsymbol{\delta}}$ for $\boldsymbol{\delta}$. The alternative estimator of $\psi_{m,\ell}$ is

$$\widetilde{\psi}_{m,\ell} = \operatorname{tr}\left(\widehat{\mathbf{Q}}\widehat{\mathbf{C}}_m\widehat{\boldsymbol{\delta}}\widehat{\boldsymbol{\delta}}'\widehat{\mathbf{C}}_\ell'\right) + \operatorname{tr}(\widehat{\mathbf{Q}}\widehat{\mathbf{P}}_m\widehat{\boldsymbol{\Omega}}\widehat{\mathbf{P}}_\ell). \tag{3.14}$$

Although $\widetilde{\psi}_{m,\ell}$ is not an asymptotically unbiased estimator, the simulation shows that the estimator (3.14) has better finite sample performance than the estimator (3.11) in most ranges of the parameter space.[7]

It is quite easy to model the heteroskedasticity by the plug-in method since the estimated weights depend on the covariance matrix estimator $\widehat{\boldsymbol{\Omega}}$. Another advantage of the plug-in method is that the correlations between different models are taken into account in the data-driven weights.

The proposed forecast combination method is the prediction counterpart to the plug-in averaging estimator proposed in Liu (2015). Similar to Liu (2015), we employ a drifting asymptotic framework and use the asymptotic risk to approximate the finite sample MSE. We, however, focus attention on the global fit of the model instead of a scalar function of parameters. Furthermore, we characterize the optimal weights under the weighted quadratic loss function instead of a pointwise loss function as in Liu (2015).[8]

[6]Note that when $M > 2$, there is no closed-form solution to (3.12). When $M = 2$, the closed-form solution to (3.12) is $\widehat{w}_1 = \widetilde{w}$ and $\widehat{w}_2 = 1 - \widetilde{w}$ where $\widetilde{w} = (\widehat{\psi}_{2,2} - \widehat{\psi}_{1,2})/(\widehat{\psi}_{1,1} + \widehat{\psi}_{2,2} - 2\widehat{\psi}_{1,2})$ if $\widehat{\psi}_{1,2} < \min\{\widehat{\psi}_{1,1}, \widehat{\psi}_{2,2}\}$, $\widetilde{w} = 1$ if $\widehat{\psi}_{1,1} \leq \widehat{\psi}_{1,2} < \widehat{\psi}_{2,2}$, or $\widetilde{w} = 0$ if $\widehat{\psi}_{2,2} \leq \widehat{\psi}_{1,2} < \widehat{\psi}_{1,1}$.

[7]As pointed out by Hansen (2014), the averaging estimator is the classic James-Stein estimator, which is a biased estimator. Hansen (2014) shows that the nested least squares averaging estimator has lower asymptotic risk than the unrestricted estimator. We might follow Hansen (2014) and apply Stein's Lemma to investigate the asymptotic risk of the estimators (3.11) and (3.14). A rigorous demonstration is beyond the scope of this paper and is left for future research.

[8]Liu (2015) considers a smooth real-valued function $\mu(\boldsymbol{\beta}, \boldsymbol{\gamma})$ as the parameter of interest. Suppose that we set $\mu(\boldsymbol{\beta}, \boldsymbol{\gamma}) = y_{T+1|T} = \mathbf{x}_T'\boldsymbol{\beta} + \mathbf{z}_T'\boldsymbol{\gamma}$. Let $\mathbf{h}_T = (\mathbf{x}_T', \mathbf{z}_T')'$. Then the plug-in averaging estimator proposed by Liu (2015) is $\widehat{\mu}(\widehat{\mathbf{w}}) = \sum_{m=1}^{M} \widehat{w}_m\widehat{\mu}_m = \widehat{y}_{T+1|T}(\widehat{\mathbf{w}})$, where $\widehat{\mathbf{w}} = \operatorname{argmin}_{\mathbf{w}\in\mathcal{H}^M} \mathbf{w}'\widehat{\boldsymbol{\Psi}}\mathbf{w}$ and the (m,ℓ)th element of $\widehat{\boldsymbol{\Psi}}$ is $\widehat{\Psi}_{m,\ell} = \mathbf{h}_T'(\widehat{\mathbf{C}}_m\widehat{\boldsymbol{\delta}}\widehat{\boldsymbol{\delta}}'\widehat{\mathbf{C}}_\ell' + \widehat{\mathbf{P}}_m\widehat{\boldsymbol{\Omega}}\widehat{\mathbf{P}}_\ell)\mathbf{h}_T$, which is different from the proposed estimator defined in Equation (3.11). Note that the above estimator depends heavily on the covariate values $\mathbf{h}_T$, and simulations show that the above estimator is not a stable estimate.

Theorem 4. *Let $\widehat{\mathbf{w}}$ be the plug-in weights defined in (3.11) and (3.12). Assume $\hat{\mathbf{\Omega}} \xrightarrow{p} \mathbf{\Omega}$. Suppose that Assumptions 1–2 hold. We have*

$$R(\widehat{\boldsymbol{\theta}}(\widehat{\mathbf{w}}), \boldsymbol{\theta}) = \mathrm{E}\left((\mathbf{A}(\mathbf{w}^*)\boldsymbol{\delta} + \mathbf{P}(\mathbf{w}^*)\mathbf{R})'\mathbf{Q}(\mathbf{A}(\mathbf{w}^*)\boldsymbol{\delta} + \mathbf{P}(\mathbf{w}^*)\mathbf{R})\right), \tag{3.15}$$

where

$$\mathbf{A}(\mathbf{w}^*) = \sum_{m=1}^{M} w_m^* \left(\mathbf{P}_m\mathbf{Q} - \mathbf{I}_{p+q}\right)\mathbf{S}_0, \tag{3.16}$$

$$\mathbf{P}(\mathbf{w}^*) = \sum_{m=1}^{M} w_m^* \mathbf{P}_m, \tag{3.17}$$

$$\mathbf{w}^* = \underset{\mathbf{w}\in\mathcal{H}^M}{\operatorname{argmin}}\, \mathbf{w}'\boldsymbol{\psi}^*\mathbf{w}, \tag{3.18}$$

and $\boldsymbol{\psi}^*$ is an $M \times M$ matrix with the (m,ℓ)th element

$$\psi_{m,\ell}^* = \operatorname{tr}\left(\mathbf{Q}\mathbf{C}_m(\mathbf{R}_{\boldsymbol{\delta}}\mathbf{R}_{\boldsymbol{\delta}}' - \mathbf{S}_0'\mathbf{Q}^{-1}\mathbf{\Omega}\mathbf{Q}^{-1}\mathbf{S}_0)\mathbf{C}_\ell'\right) + \operatorname{tr}\left(\mathbf{Q}\mathbf{P}_m\mathbf{\Omega}\mathbf{P}_\ell\right) \tag{3.19}$$

with $\mathbf{R}_{\boldsymbol{\delta}} = \boldsymbol{\delta} + \mathbf{S}_0'\mathbf{Q}^{-1}\mathbf{R}$ and $\mathbf{R} \sim \mathbf{N}(\mathbf{0}, \mathbf{\Omega})$.

Theorem 4 presents the asymptotic trimmed risk of the plug-in averaging estimator. Unlike the averaging estimator with fixed weights, the asymptotic trimmed risk of the plug-in averaging estimator depends on the normal random vector $\mathbf{R}$. Note that the limiting distribution of the plug-in averaging estimator is a nonlinear function of $\mathbf{R}$ instead of a normal distribution. For the alternative estimator of $\psi_{m,\ell}$ defined in (3.14), the asymptotic trimmed risk of the plug-in averaging estimator is the same except (3.19) is replaced by $\psi_{m,\ell}^* = \operatorname{tr}\left(\mathbf{Q}\mathbf{C}_m\mathbf{R}_{\boldsymbol{\delta}}\mathbf{R}_{\boldsymbol{\delta}}'\mathbf{C}_\ell'\right) + \operatorname{tr}\left(\mathbf{Q}\mathbf{P}_m\mathbf{\Omega}\mathbf{P}_\ell\right)$.

3.5 Relationship between the Plug-In Averaging Estimator and the Mallows' C_p-type Averaging Estimator

In this section we discuss the relationship between the plug-in averaging estimator and the Mallows' C_p-type averaging estimator. Suppose that there is no must-have predictor, i.e., $\mathbf{x}_t$ is an empty matrix. Then we have $\mathbf{S}_m = \mathbf{\Pi}_m'$, $\mathbf{S}_0 = \mathbf{I}_q$, and $\widehat{\mathbf{C}}_m =$

$\widehat{\mathbf{P}}_m\widehat{\mathbf{Q}} - \mathbf{I}_q$. Thus, the equation (3.11) can be rewritten as

$$\begin{aligned}\widehat{\psi}_{m,\ell} &= \text{tr}(\widehat{\mathbf{Q}}\widehat{\mathbf{C}}_m(\widehat{\boldsymbol{\delta}}\widehat{\boldsymbol{\delta}}' - \widehat{\mathbf{Q}}^{-1}\widehat{\boldsymbol{\Omega}}\widehat{\mathbf{Q}}^{-1})\widehat{\mathbf{C}}_\ell') + \text{tr}(\widehat{\mathbf{Q}}\widehat{\mathbf{P}}_m\widehat{\boldsymbol{\Omega}}\widehat{\mathbf{P}}_\ell) \\ &= \text{tr}\Big(\widehat{\mathbf{Q}}(\widehat{\mathbf{P}}_m\widehat{\mathbf{Q}} - \mathbf{I}_q)\widehat{\boldsymbol{\delta}}\widehat{\boldsymbol{\delta}}'(\widehat{\mathbf{Q}}\widehat{\mathbf{P}}_\ell - \mathbf{I}_q)\Big) \\ &\quad - \text{tr}\Big(\widehat{\mathbf{Q}}(\widehat{\mathbf{P}}_m\widehat{\mathbf{Q}} - \mathbf{I}_q)\widehat{\mathbf{Q}}^{-1}\widehat{\boldsymbol{\Omega}}\widehat{\mathbf{Q}}^{-1}(\widehat{\mathbf{Q}}\widehat{\mathbf{P}}_\ell - \mathbf{I}_q) - \widehat{\mathbf{Q}}\widehat{\mathbf{P}}_m\widehat{\boldsymbol{\Omega}}\widehat{\mathbf{P}}_\ell\Big) \\ &= (\widehat{\mathbf{e}}_m'\widehat{\mathbf{e}}_\ell - \widehat{\mathbf{e}}'\widehat{\mathbf{e}}) + \text{tr}(\widehat{\mathbf{Q}}_m^{-1}\widehat{\boldsymbol{\Omega}}_m) + \text{tr}(\widehat{\mathbf{Q}}_\ell^{-1}\widehat{\boldsymbol{\Omega}}_\ell) - \text{tr}(\widehat{\mathbf{Q}}^{-1}\widehat{\boldsymbol{\Omega}}), \end{aligned} \tag{3.20}$$

where $\widehat{\mathbf{e}} = \mathbf{y} - \mathbf{H}\widehat{\boldsymbol{\theta}}$, $\widehat{\mathbf{e}}_m = \mathbf{y} - \mathbf{H}_m\widehat{\boldsymbol{\theta}}_m$, $\widehat{\mathbf{Q}}_m = \mathbf{S}_m'\widehat{\mathbf{Q}}\mathbf{S}_m$, and $\widehat{\boldsymbol{\Omega}}_m = \mathbf{S}_m'\widehat{\boldsymbol{\Omega}}\mathbf{S}_m$; see the appendix for the derivation of (3.20). Therefore, the criterion function for the plug-in averaging estimator is

$$\mathbf{w}'\widehat{\boldsymbol{\psi}}\mathbf{w} = \mathbf{w}'\widetilde{\boldsymbol{\psi}}\mathbf{w} - \widehat{\mathbf{e}}'\widehat{\mathbf{e}} - \text{tr}(\widehat{\mathbf{Q}}^{-1}\widehat{\boldsymbol{\Omega}}) \tag{3.21}$$

where the (m,ℓ)th element of $\widetilde{\boldsymbol{\psi}}$ is

$$\widetilde{\psi}_{m,\ell} = \widehat{\mathbf{e}}_m'\widehat{\mathbf{e}}_\ell + \text{tr}(\widehat{\mathbf{Q}}_m^{-1}\widehat{\boldsymbol{\Omega}}_m) + \text{tr}(\widehat{\mathbf{Q}}_\ell^{-1}\widehat{\boldsymbol{\Omega}}_\ell). \tag{3.22}$$

Since $\widehat{\mathbf{e}}'\widehat{\mathbf{e}}$ and $\text{tr}(\widehat{\mathbf{Q}}^{-1}\widehat{\boldsymbol{\Omega}})$ are not related to the weight vector $\mathbf{w}$, minimizing $\mathbf{w}'\widehat{\boldsymbol{\psi}}\mathbf{w}$ over $\mathbf{w} = (w_1, ..., w_M)$ is equivalent to minimizing $\mathbf{w}'\widetilde{\boldsymbol{\psi}}\mathbf{w}$.

Let $\widehat{\mathbf{e}}(\mathbf{w}) = \mathbf{y} - \mathbf{H}\widehat{\boldsymbol{\theta}}(\mathbf{w})$ be the averaging residuals vector. Let $\mathbf{k} = (k_1, ..., k_M)'$ and $k_m = p + q_m$. If the error term is i.i.d. and homoskedastic, then the covariance matrix $\boldsymbol{\Omega}$ can be consistently estimated by $\widehat{\boldsymbol{\Omega}} = \widehat{\sigma}^2\widehat{\mathbf{Q}}$ where $\widehat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^T \widehat{e}_{t+1}^2$ and $\widehat{e}_{t+1} = y_{t+1} - \mathbf{h}_t'\widehat{\boldsymbol{\theta}}$. In this case, $\text{tr}(\widehat{\mathbf{Q}}_m^{-1}\widehat{\boldsymbol{\Omega}}_m) = \widehat{\sigma}^2 k_m$. Define $\boldsymbol{\Sigma}$ as an $M \times M$ matrix whose (m,ℓ)th element is $k_m + k_\ell$. Then, the criterion function for the plug-in averaging estimator is

$$\mathbf{w}'\widetilde{\boldsymbol{\psi}}\mathbf{w} = \widehat{\mathbf{e}}(\mathbf{w})'\widehat{\mathbf{e}}(\mathbf{w}) + \widehat{\sigma}^2\mathbf{w}'\boldsymbol{\Sigma}\mathbf{w} = \widehat{\mathbf{e}}(\mathbf{w})'\widehat{\mathbf{e}}(\mathbf{w}) + 2\widehat{\sigma}^2\mathbf{k}'\mathbf{w}, \tag{3.23}$$

which is the Mallows criterion proposed by Hansen (2007). Note that the last equality of (3.23) holds by the fact that $\mathbf{w}'\boldsymbol{\Sigma}\mathbf{w} = \mathbf{w}'(\mathbf{k}\mathbf{1}' + \mathbf{1}\mathbf{k}')\mathbf{w} = 2\mathbf{k}'\mathbf{w}$ where $\mathbf{1} = (1, ..., 1)'$ is an $M \times 1$ vector.

The first term of the Mallows criterion measures the model fit, while the second term of the criterion measures the effective number of parameters and serves as a penalty term. Therefore, we can interpret the MMA criterion as a measure of model fit and model complexity.

If the error term is serially uncorrelated and identically distributed, then $\boldsymbol{\Omega}$ can be

consistently estimated by $\widehat{\mathbf{\Omega}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{h}_t\mathbf{h}_t'\widehat{e}_{t+1}^2$. In this case,

$$\mathrm{tr}(\widehat{\mathbf{Q}}_m^{-1}\widehat{\mathbf{\Omega}}_m) = \mathrm{tr}\left(\left(\sum_{t=1}^{T}\mathbf{h}_{m,t}\mathbf{h}_{m,t}'\right)^{-1}\left(\sum_{t=1}^{T}\mathbf{h}_{m,t}\mathbf{h}_{m,t}'\widehat{e}_{t+1}^2\right)\right) \equiv \widetilde{k}_m. \tag{3.24}$$

Define $\widetilde{\mathbf{\Sigma}}$ as an $M \times M$ matrix whose (m,ℓ)th element is $\widetilde{k}_m + \widetilde{k}_\ell$. Then, the criterion function for the plug-in averaging estimator is

$$\mathbf{w}'\widetilde{\boldsymbol{\psi}}\mathbf{w} = \widehat{\mathbf{e}}(\mathbf{w})'\widehat{\mathbf{e}}(\mathbf{w}) + \mathbf{w}'\widetilde{\mathbf{\Sigma}}\mathbf{w} = \widehat{\mathbf{e}}(\mathbf{w})'\widehat{\mathbf{e}}(\mathbf{w}) + 2\widetilde{\mathbf{k}}'\mathbf{w}, \tag{3.25}$$

where $\widetilde{\mathbf{k}} = (\widetilde{k}_1, ..., \widetilde{k}_M)'$. Thus, the criterion function (3.25) is equivalent to the heteroskedasticity-robust C_p criterion proposed by Liu and Okui (2013).

As shown in (3.23) and (3.25), the proposed plug-in averaging estimator is a generalized Mallows' C_p-type averaging estimator. Note that both Hansen (2007) and Liu and Okui (2013) make no distinction between must-have and potentially relevant predictors, which is different from our framework. When there is no must-have predictor, the plug-in averaging estimator is equivalent to the Mallows model averaging estimator if the covariance matrix estimator $\widehat{\mathbf{\Omega}} = \frac{1}{T}\widehat{\sigma}^2\sum_{t=1}^{T}\mathbf{h}_t\mathbf{h}_t'$ is used, and is equivalent to the heteroskedasticity-robust C_p averaging estimator if the covariance matrix estimator $\widehat{\mathbf{\Omega}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{h}_t\mathbf{h}_t'\widehat{e}_{t+1}^2$ is used. However, if $\mathbf{x}_t$ is not an empty matrix then the equivalence does not hold in general.

4 Finite Sample Investigation

We now evaluate the finite sample performance of the plug-in forecast combination method in comparison with other forecast combination approaches in two simulation setups. The first design is the linear regression model, and we consider all possible models. The second design is a moving average model with exogenous inputs, and we consider a sequence of nested candidate models.

4.1 Six Forecast Combination Methods

In the simulations, we consider the following forecast combination approaches: (1) smoothed Akaike information criterion model averaging estimator (labeled S-AIC), (2) smoothed Bayesian information criterion model averaging estimator (labeled S-BIC), (3) Mallows model averaging estimator (labeled MMA), (4) jackknife model averaging

estimator (labeled JMA), (5) the complete subset regressions approach, (6) the plug-in averaging estimator based on (3.11) (labeled PIA(1)), and the plug-in averaging estimator based on (3.14) (labeled PIA(2)). We briefly discuss each method below.

The S-AIC estimator is proposed by Buckland, Burnham, and Augustin (1997), and suggests assigning the weights of each candidate model by using the exponential Akaike information criterion. The weight is proportional to the log-likelihood of the model and is defined as $\widehat{w}_m = \exp(-\frac{1}{2}\text{AIC}_m)/\sum_{j=1}^{M}\exp(-\frac{1}{2}\text{AIC}_j)$, where $\text{AIC}_m = T\log(\widehat{\sigma}_m^2)+2(p+q_m)$, $\widehat{\sigma}_m^2 = \frac{1}{T}\sum_{t=1}^{T}\widehat{e}_{m,t+1}^2$, and $\widehat{e}_{m,t}$ are the least squares residuals from the model m. The S-BIC estimator is a simplified form of Bayesian model averaging (BMA). By assuming diffuse priors, the BMA weights approximately equal $\widehat{w}_m = \exp(-\frac{1}{2}\text{BIC}_m)/\sum_{j=1}^{M}\exp(-\frac{1}{2}\text{BIC}_j)$, where $\text{BIC}_m = T\log(\widehat{\sigma}_m^2) + \log(T)(p+q_m)$.

Hansen (2007) proposes the MMA estimator for homoskedastic linear regression models. The MMA estimator selects the weights by minimizing a Mallows criterion defined in (3.23). The idea behind the Mallows criterion is to approximate the mean squared error by the sum of squared errors and a penalty term. Hansen (2008) shows that the MMA criterion is an unbiased estimate of the in-sample mean squared error plus a constant term for stationary dependent observations.

Hansen and Racine (2012) propose the JMA estimator for non-nested and heteroskedastic linear regression models. The weights of the JMA estimator are chosen by minimizing a leave-one-out cross-validation criterion $\text{CV}(\mathbf{w}) = \mathbf{w}'\tilde{\mathbf{e}}'\tilde{\mathbf{e}}\mathbf{w}$, where $\tilde{\mathbf{e}} = (\tilde{\mathbf{e}}_1, ..., \tilde{\mathbf{e}}_M)$ is the $T \times M$ matrix of leave-one-out least squares residuals and $\tilde{\mathbf{e}}_m$ are the residuals of the model m obtained by least squares estimation without the tth observation. The MMA and JMA estimators are asymptotically optimal in the sense of achieving the lowest possible expected squared error in homoskedastic and heteroskedastic settings, respectively. The optimality, however, is limited to the random sample and hence is not directly applicable to forecast combination for time series data.

For the above four averaging estimators and the plug-in averaging estimator, the one-step-ahead combination forecast is computed as

$$\widehat{y}_{T+1|T}(\widehat{\mathbf{w}}) = \sum_{m=1}^{M}\widehat{w}_m\widehat{y}_{T+1|T}(m), \tag{4.1}$$

where $\widehat{w}_m$ is determined by S-AIC, S-BIC, MMA, JMA, PIA(1), or PIA(2).

Unlike previous methods, the complete subset regression method proposed by Elliott, Gargano, and Timmermann (2013) assigns equal weights to a set of models.

Let $k = p + q$ be the number predictors used in the full model and κ the number of predictors used in all subset regressions. For a given set of potential predictors, the complete subset regression method constructs the forecast combination by using equal-weighted combination based on all possible models that include κ predictors. Let $n_{\kappa,k} = k!/((k-\kappa)!\kappa!)$ be the number of models considered based on κ subset regressions. The one-step-ahead combination forecast based on the complete subset regression method is

$$\widehat{y}_{T+1|T}(\kappa) = \frac{1}{n_{\kappa,k}} \sum_{m=1}^{n_{\kappa,k}} \mathbf{h}_T' \mathbf{S}_m \widehat{\boldsymbol{\theta}}_m \quad \text{s.t. } \operatorname{tr}(\mathbf{S}_m \mathbf{S}_m') = \kappa. \tag{4.2}$$

Instead of choosing the weights $\mathbf{w}$, the complete subset regression method has to choose the number of predictors κ for all models.[9]

We follow Ng (2013) and compare these estimators based on the relative risk. Let $\widehat{y}_{T+1|T}(m)$ be the prediction based on the model m, where $m = 1, ..., M$. Let $\widehat{y}_{T+1|T}(\widehat{\mathbf{w}})$ be the prediction based on the S-AIC, S-BIC, MMA, JMA, complete subset regressions, and plug-in averaging estimators. The relative risk is computed as the ratio of the risk based on the forecast combination method relative to the lowest risk among the candidate models:

$$\frac{\frac{1}{S}\sum_{s=1}^{S}\left(y_{s,T+1|T} - \widehat{y}_{s,T+1|T}(\widehat{\mathbf{w}})\right)^2}{\min\limits_{m\in\{1,...,M\}} \frac{1}{S}\sum_{s=1}^{S}\left(y_{s,T+1|T} - \widehat{y}_{s,T+1|T}(m)\right)^2},$$

where S is the number of simulations. We set $S = 5000$ for all experiments. The lower relative risk means better performance on predictions. When the relative risk exceeds one, it indicates that the forecast combination method does not outperform the best possible prediction among the candidate models.

[9]One limitation of subset regression combinations is that the approach is not suitable for the nested models. Suppose that we consider AR models up to order p. The goal is to average different AR models to minimize the risk function. In this case, we are not able to apply complete subset regressions.

4.2 Linear Regression Models

The data generation process for the first design is

$$y_{t+1} = \sum_{j=1}^{k} \beta_j x_{jt} + e_{t+1}, \tag{4.3}$$

$$x_{jt} = \rho_x x_{jt-1} + u_{jt}, \text{ for } j \geq 2. \tag{4.4}$$

We set $x_{1t} = 1$ to be the intercept and the remaining x_{jt} are AR(1) processes with $\rho_x = 0.5$ and 0.9. The predictors x_{jt} are correlated and all are potentially relevant. We generate $(u_{2t}, ..., u_{kt})'$ from a joint normal distribution $N(\mathbf{0}, \mathbf{Q}_u)$, where the diagonal elements of $\mathbf{Q}_u$ are 1, and off-diagonal elements are ρ_u. We set $\rho_u = 0.25, 0.5, 0.75,$ and 0.9. The error term e_t has mean zero and variance one. For the homoskedastic simulation, the error term is generated from a standard normal distribution. For the heteroskedastic simulation, we first generate an AR(1) process $\epsilon_t = 0.5\epsilon_{t-1} + \eta_t$, where $\eta_t \sim N(0, 0.75)$. Then, the error term is constructed by $e_t = 3^{-1/2}(1 - \rho_x^2)x_{kt}^2\epsilon_t$.

The regression coefficients are determined by the rule

$$\boldsymbol{\beta} = \frac{c}{\sqrt{T}} \left(1, \frac{k-1}{k}, ..., \frac{1}{k}\right)',$$

and the local parameters are determined by $\delta_j = \sqrt{T}\beta_j = c(k - j + 1)/k$ for $j \geq 2$. The parameter c is selected to vary the population $R^2 = \tilde{\boldsymbol{\beta}}'\mathbf{Q}_x\tilde{\boldsymbol{\beta}}/(1 + \tilde{\boldsymbol{\beta}}'\mathbf{Q}_x\tilde{\boldsymbol{\beta}})$, where $\tilde{\boldsymbol{\beta}} = (\beta_2, ..., \beta_k)'$ and $\mathbf{Q}_x = (1 - \rho_x^2)^{-1}\mathbf{Q}_u$ and R^2 varies on a grid between 0.1 and 0.9. We set the sample size to $T = 200$ and set $k = 5$. We consider all possible models, and hence the number of models is $M = 32$ for S-AIC, S-BIC, MMA, JMA, PIA(1), and PIA(2). For the complete subset regression method, the numbers of models are 5, 10, 10, 5, and 1 for $\kappa = 1$, 2, 3, 4, and 5, respectively.

Figures 1–4 show the relative risk for the first simulation setup. In each figure, the relative risk is displayed for $\rho_u = 0.25, 0.5, 0.75,$ and 0.9, respectively. We first compare the relative risk when the AR(1) coefficient of the predictor equals 0.5. Figures 1 and 2 show that both plug-in averaging estimators perform well and PIA(2) dominates other estimators in most ranges of the parameter space. The relative risk of MMA and JMA estimators is indistinguishable in the homoskedastic simulation, but JMA has lower relative risk than MMA for $\rho_u = 0.25$ and 0.5 in the heteroskedastic simulation. The S-AIC and MMA estimators have quite similar relative risk for the homoskedastic simulation, but S-AIC has much larger relative risk than MMA for the heteroskedastic

simulation. The S-BIC estimator has poor performance in both homoskedastic and heteroskedastic simulations.

Figure 3 compares the relative risk between the plug-in averaging estimator and the complete subset regressions in heteroskedastic simulations. The performance of the subset regression approach is sensitive to the choice of κ, the number of the predictors included in the model. As R^2 increases, the optimal value of κ tends to be greater. Unlike the complete subset regressions, the performance of the plug-in averaging estimator is quite robust to different values of R^2. In most cases, the plug-in averaging estimator has much lower relative risk than the complete subset regressions with different κ.

Figure 4 displays the relative risk for the large AR(1) coefficient. The relative performance of six estimators depends strongly on R^2 and ρ_u. Overall, the ranking of estimators is quite similar to that for $\rho_x = 0.5$. However, PIA(1) performs slightly better than PIA(2) for the heteroskedastic simulation when R^2 is small.

4.3 Moving Average Model with Exogenous Inputs

The second design is similar to that of Ng (2013). The data generation process is a moving average model with exogenous inputs

$$y_t = x_t + 0.5x_{t-1} + e_t + \beta e_{t-1}, \tag{4.5}$$

$$x_t = 0.5x_{t-1} + u_t. \tag{4.6}$$

The exogenous regressor x_t is an AR(1) process, and u_t is generated from a standard normal distribution. The error term e_t is generated from a normal distribution $N(0, \sigma_t^2)$, where $\sigma_t^2 = 1 + x_t^2$. The parameter β is varied on a grid from -0.5 to 0.5. The sample size is varied between $T = 100$, 200, 500, and 1000.

We consider a sequence of nested models based on regressors:

$$\{1, y_{t-1}, x_t, y_{t-2}, x_{t-1}, y_{t-3}, x_{t-2}\},$$

where the constant term is included in all models. The number of models is $M = 7$ for S-AIC, S-BIC, MMA, JMA, PIA(1), and PIA(2). For $\beta \neq 0$, the true model is infinite dimensional, and there is no true model among these seven candidate models. For $\beta = 0$, the true model size, or the number of regressors of the data generation process, is two. However, all seven models are wrong. In this setup, we do not compute the

complete subset regression because it cannot be applied when the candidate models are nested.

In Figure 5, the four panels display the relative risk for $T = 100$, 200, 500, and 1000, respectively. In each panel, the relative risk is displayed for β between -0.5 and 0.5. All forecast combination approaches, except S-BIC, have similar relative risk in most ranges of the parameter space, but PIA(2) has lower relative risk than other estimators when T and $|\beta|$ are large. S-BIC has much lower relative risk for large T and small $|\beta|$. In most cases, however, S-BIC has quite poor performance relative to other methods. Similar results for the AIC and BIC model selection estimators are also found in Yang (2007) and Ng (2013).

Figure 6 compares the average model size of six estimators.[10] As we expected, the average model size of S-BIC is smaller than those of other estimators. S-AIC and PIA(2) have similar average model sizes, and they tend to select the larger models compared to MMA, JMA, and PIA(1). An interesting observation is that the average model size is not symmetric around zero nor monotone in β.

5 Empirical Application

In this section, we apply the forecast combination method to stock return predictions. The challenge of empirical research on equity premium prediction is that one does not know exactly what variables are the good predictors of the stock return. Different studies suggest different economic variables and models for the equity premium prediction; see Rapach and Zhou (2012) for a literature review. Results from some studies contradict the findings of others. Welch and Goyal (2008) argue that numerous economic variables have poor out-of-sample predictions and these forecasting models are unable to provide forecasting gain relative to the historical average consistently. In order to take into account the model uncertainty, Rapach, Strauss, and Zhou (2010) and Elliott, Gargano, and Timmermann (2013) propose an equal-weighted forecast combination approach to the subset predictive regression. They find that forecast combinations achieve significant gains on out-of-sample predictions relative to the historical average. We apply the forecast combination with data-driven weights instead of equal weights to the U.S. stock market.

[10] We compute the average model size by computing averages across 5000 simulation draws, that is, $\frac{1}{S}\sum_{s=1}^{S}\sum_{m=1}^{M}\widehat{w}_{s,m}k_m$.

5.1 Data

We estimate the following predictive regression $r_{t+1} = \beta + \mathbf{z}_t'\boldsymbol{\gamma} + e_{t+1}$, where r_{t+1} is the equity premium, $\mathbf{z}_t$ are the economic variables, and e_{t+1} is an unobservable disturbance term. The goal is to select weights to achieve the lowest cumulative squared prediction error.

The quarterly data are taken from Welch and Goyal (2008) and are up to date through 2011.[11] The total sample size is 260 over the period 1947–2011. The stock returns are measured as the difference between the continuously compounded return on the S&P 500 index including dividends and the Treasury bill rate. We consider 10 economic variables and a total of 1025 possible models, including a null model.[12] The 10 economic variables are as follows: dividend price ratio, dividend yield, earnings price ratio, book-to-market ratio, net equity expansion, Treasury bill, long-term return, default yield spread, default return spread, and inflation; see Welch and Goyal (2008) for a detailed description of the data and their source.

We follow Welch and Goyal (2008) and calculate the out-of-sample forecast of the equity premium using a recursively expanding estimation window. We first divide the total sample into an in-sample period (1947:1–1964:4) and an out-of-sample evaluation period (1965:1–2011:4). The first out-of-sample forecast is for 1965:1, while the last out-of-sample forecast is for 2011:4. For each out-of-sample forecast, we estimate the predictive regression based on all available samples up to that point. That is, for the first out-of-sample forecast, we calculate different combination forecast methods based on the sample 1947:1–1964:4. We then expand the estimation window to the sample 1947:1–1965:1 and construct the combination forecast for 1965:2 and so on. Note that we re-estimate the data-driven weights for each scheme.

5.2 Out-Of-Sample Forecasting Results

We follow Welch and Goyal (2008) and use the historical average of the equity premium as a benchmark. As shown in Welch and Goyal (2008) and Rapach, Strauss, and Zhou (2010), none of the forecasts based on the individual economics variable consistently outperforms the forecast based on the historical average.

[11] The data are available at http://www.hec.unil.ch/agoyal/.

[12] Elliott, Gargano, and Timmermann (2013) consider 12 variables, which are slightly different from the variables used in Rapach, Strauss, and Zhou (2010). We use the variables that are both considered in two articles. All the models except the null model include the constant term. The null model does not include any predictor.

Figure 7 presents the time series plots of the differences between the cumulative squared prediction error of the historical average benchmark forecast and the cumulative squared prediction error of the forecast combinations based on different model averaging approaches. When the curve in each panel is greater than zero, the forecast combination method outperforms the historical average.

The upper panel of Figure 7 shows that MMA, JMA, PIA(1), and PIA(2) consistently beat the historical average in terms of MSFE. S-AIC and S-BIC have a smaller cumulative squared prediction error than the historical average before 1997, but neither estimator outperforms the historical average after 1997. It is clear to see that PIA(2) and MMA perform similarly and both estimators achieve smaller cumulative squared prediction errors as compared to other estimators. Note that PIA(2) performs better than PIA(1), which is consistent with the finding in our simulations. One interesting observation is that the ranking of estimators almost remains the same in the out-of-sample evaluation period.

The two lower panels of Figure 7 compare the cumulative squared prediction error of PIA(2) to those of the complete subset regressions. As we can see from these two panels, the complete subset regressions that use $\kappa = 4$ or 5 predictors produce the lowest cumulative squared prediction error. Our finding of the optimal value of κ is slightly larger than that in Elliott, Gargano, and Timmermann (2013). PIA(2) has similar performance to the complete subset regressions with $\kappa = 4$ or 5, and PIA(2) outperforms the complete subset regressions when $\kappa < 4$ or $\kappa > 5$. It is clear to see that the choice of κ has a great influence on the performance of the complete subset regressions, and in practice the optimal choice of κ is unknown. Examining these three panels in Figure 7, there are four undominated methods (PIA(2), MMA, $\kappa = 4$, and $\kappa = 5$), and there is no one forecast combination method that uniformly dominates the others.

For a formal comparison between the plug-in forecast combination and the historical average benchmark, we compute the out-of-sample R^2 statistics.[13] The out-of-sample R^2 value of PIA(2) is 2.7257 with the associated p-value 0.0173, which means PIA(2) has a significantly lower MSFE than the historical average benchmark forecast. Therefore, our results support the findings of Rapach, Strauss, and Zhou (2010) and Elliott, Gargano, and Timmermann (2013) that forecast combinations provide significant gains

[13]Let $\bar{r}_{\tau+1|\tau} = \sum_{t=1}^{\tau} r_t$ be the historical average and $\widehat{r}_{T+1|T}(\widehat{\mathbf{w}})$ the equity premium forecast based on forecast combination. The out-of-sample R^2 value is computed as $R^2_{OOS} = 1 - \sum_{\tau=\tau_0}^{T-1} \left(r_{\tau+1} - \widehat{r}_{\tau+1|\tau}(\widehat{\mathbf{w}})\right)^2 / \sum_{\tau=\tau_0}^{T-1} \left(r_{\tau+1} - \bar{r}_{\tau+1|\tau}\right)^2$. The associated p-value is based on Clark and West (2007) to test the null hypothesis that $R^2_{OOS} \leq 0$.

on equity premium predictions relative to the historical average.

6 Conclusion

This paper studies the weight selection for forecast combination in a predictive regression when the goal is minimizing the asymptotic risk. We derive the asymptotic distribution and asymptotic risk of the averaging estimator in a local asymptotic framework without the i.i.d. normal assumption. We then develop a frequentist model averaging criterion, an asymptotically unbiased estimator of the asymptotic risk, to select forecast weights. While this paper has focused on the one-step-ahead forecasting model, the proposed plug-in averaging method can be easily extended to the h-step-ahead forecasting model.[14] Simulations show that the proposed estimator achieves lower MSFE relative risk than other existing model averaging methods in most cases.

Appendix

A Proofs

The following Lemma (Lemma 1 in Liu (2015)) shows the asymptotic distributions of the least squares estimators in the mth model. Let $\boldsymbol{\theta}_m = (\boldsymbol{\beta}', \boldsymbol{\gamma}_m')' = (\boldsymbol{\beta}', \boldsymbol{\gamma}'\boldsymbol{\Pi}_m')' = \mathbf{S}_m'\boldsymbol{\theta}$.

Lemma 1. *[Liu (2015)] Suppose that Assumptions 1–2 hold. As $T \to \infty$, we have*

$$\sqrt{T}\left(\widehat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m\right) \xrightarrow{d} \mathbf{A}_m\boldsymbol{\delta} + \mathbf{B}_m\mathbf{R} \sim \mathbf{N}\left(\mathbf{A}_m\boldsymbol{\delta},\ \mathbf{Q}_m^{-1}\boldsymbol{\Omega}_m\mathbf{Q}_m^{-1}\right),$$

where $\mathbf{A}_m = \mathbf{Q}_m^{-1}\mathbf{S}_m'\mathbf{Q}\mathbf{S}_0\left(\mathbf{I}_q - \boldsymbol{\Pi}_m'\boldsymbol{\Pi}_m\right)$ *and* $\mathbf{B}_m = \mathbf{Q}_m^{-1}\mathbf{S}_m'$.

[14] Consider an h-step-ahead forecasting model: $y_{t+h} = \mathbf{x}_t'\boldsymbol{\beta} + \mathbf{z}_t'\boldsymbol{\gamma} + e_{t+h}$ and $\mathrm{E}(\mathbf{h}_t e_{t+h}) = 0$. The h-step-ahead forecast from the mth model is $\widehat{y}_{T+h|T}(m) = \mathbf{h}_T'\mathbf{S}_m\widehat{\boldsymbol{\theta}}_m$ where $\widehat{\boldsymbol{\theta}} = (\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\mathbf{y}$, and the h-step-ahead combination forecast is $\widehat{y}_{T+h|T}(\mathbf{w}) = \mathbf{h}_T'\widehat{\boldsymbol{\theta}}(\mathbf{w})$, where $\widehat{\boldsymbol{\theta}}(\mathbf{w}) = \sum_{m=1}^{M} w_m\mathbf{S}_m\widehat{\boldsymbol{\theta}}_m$. We now modify Assumption 2 as follows: **Assumption 2'**. $\{y_{t+h}, \mathbf{h}_t\}$ is a strictly stationary and ergodic time series with finite $r > 4$ moments and $\mathrm{E}(e_{t+h}|\mathcal{F}_t) = 0$, where $\mathcal{F}_t = \sigma(\mathbf{h}_t, \mathbf{h}_{t-1}, ...; e_t, e_{t-1}, ...)$. Suppose that Assumptions 1 and 2' hold. Then the results in Theorems 1–3 still hold except the definition of $\boldsymbol{\Omega}$ is replaced by $\boldsymbol{\Omega} = \lim_{T\to\infty}\frac{1}{T}\sum_{s=1}^{T}\sum_{t=1}^{T}\mathrm{E}\left(\mathbf{h}_s\mathbf{h}_t'e_{s+h}e_{t+h}\right)$. Therefore, we can construct the plug-in combination forecast in the same way as (3.13).

Proof of Theorem 1: Recall that $\widehat{\boldsymbol{\theta}}(\mathbf{w}) = \sum_{m=1}^{M} w_m \mathbf{S}_m \widehat{\boldsymbol{\theta}}_m$. By Lemma 1, we have

$$\begin{aligned}\sqrt{T}\mathbf{S}_m\left(\widehat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m\right) &\xrightarrow{d} \mathbf{S}_m\left(\mathbf{Q}_m^{-1}\mathbf{S}_m'\mathbf{Q}\mathbf{S}_0\left(\mathbf{I}_q - \boldsymbol{\Pi}_m'\boldsymbol{\Pi}_m\right)\boldsymbol{\delta} + \mathbf{Q}_m^{-1}\mathbf{S}_m'\mathbf{R}\right) \\ &= \mathbf{P}_m\mathbf{Q}\mathbf{S}_0\left(\mathbf{I}_q - \boldsymbol{\Pi}_m'\boldsymbol{\Pi}_m\right)\boldsymbol{\delta} + \mathbf{P}_m\mathbf{R},\end{aligned}$$

where $\mathbf{P}_m = \mathbf{S}_m\mathbf{Q}_m^{-1}\mathbf{S}_m'$. Also, by Assumption 1, it follows that $\sqrt{T}(\mathbf{S}_m\boldsymbol{\theta}_m - \boldsymbol{\theta}) = \mathbf{S}_0(\boldsymbol{\Pi}_m'\boldsymbol{\Pi}_m - \mathbf{I}_q)\boldsymbol{\delta}$, where $\mathbf{S}_0 = (\mathbf{0}_{q\times p}, \mathbf{I}_q)'$. Therefore, by Assumptions 1–2 and the application of Slutsky's theorem, we have

$$\begin{aligned}\sqrt{T}\left(\mathbf{S}_m\widehat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}\right) &= \sqrt{T}\mathbf{S}_m\left(\widehat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m\right) + \sqrt{T}(\mathbf{S}_m\boldsymbol{\theta}_m - \boldsymbol{\theta}) \\ &\xrightarrow{d} \mathbf{P}_m\mathbf{Q}\mathbf{S}_0\left(\mathbf{I}_q - \boldsymbol{\Pi}_m'\boldsymbol{\Pi}_m\right)\boldsymbol{\delta} + \mathbf{P}_m\mathbf{R} - \mathbf{S}_0(\mathbf{I}_q - \boldsymbol{\Pi}_m'\boldsymbol{\Pi}_m)\boldsymbol{\delta} \\ &= \left(\mathbf{S}_m\mathbf{Q}_m^{-1}\mathbf{S}_m'\mathbf{Q}\mathbf{S}_0 - \mathbf{S}_0\right)\left(\mathbf{I}_q - \boldsymbol{\Pi}_m'\boldsymbol{\Pi}_m\right)\boldsymbol{\delta} + \mathbf{P}_m\mathbf{R} \\ &= \left(\mathbf{S}_m\mathbf{Q}_m^{-1}\mathbf{S}_m'\mathbf{Q}\mathbf{S}_0 - \mathbf{S}_0\right)\boldsymbol{\delta} + \mathbf{P}_m\mathbf{R} \\ &= \left(\mathbf{P}_m\mathbf{Q} - \mathbf{I}_{p+q}\right)\mathbf{S}_0\boldsymbol{\delta} + \mathbf{P}_m\mathbf{R} \equiv \boldsymbol{\Lambda}_m, \qquad \text{(A.1)}\end{aligned}$$

where the third equality holds by the fact that $\mathbf{S}_0\boldsymbol{\Pi}_m' = \mathbf{S}_m\left(\mathbf{0}_{p\times q_m}', \mathbf{I}_{q_m}\right)'$.

From (A.1), there is joint convergence in distribution of all $\sqrt{T}\left(\mathbf{S}_m\widehat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}\right)$ to $\boldsymbol{\Lambda}_m$, since all of $\boldsymbol{\Lambda}_m$ can be expressed in terms of the same normal vector $\mathbf{R}$. Because the weights are nonrandom, it follows that

$$\sqrt{T}\left(\widehat{\boldsymbol{\theta}}(\mathbf{w}) - \boldsymbol{\theta}\right) = \sum_{m=1}^{M} w_m\sqrt{T}\left(\mathbf{S}_m\widehat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}\right) \xrightarrow{d} \sum_{m=1}^{M} w_m\boldsymbol{\Lambda}_m \equiv \boldsymbol{\Lambda}. \qquad \text{(A.2)}$$

By standard algebra, we can show the mean vector of $\boldsymbol{\Lambda}$ as

$$\mathrm{E}\left(\sum_{m=1}^{M} w_m\boldsymbol{\Lambda}_m\right) = \sum_{m=1}^{M} w_m\mathrm{E}\left(\boldsymbol{\Lambda}_m\right) = \sum_{m=1}^{M} w_m\left(\mathbf{P}_m\mathbf{Q} - \mathbf{I}_{p+q}\right)\mathbf{S}_0\boldsymbol{\delta} = \mathbf{A}(\mathbf{w})\boldsymbol{\delta},$$

where $\mathbf{A}(\mathbf{w}) = \sum_{m=1}^{M} w_m\left(\mathbf{P}_m\mathbf{Q} - \mathbf{I}_{p+q}\right)\mathbf{S}_0$.

Next we want to show the covariance matrix of $\boldsymbol{\Lambda}$. Let $\mathbf{C}_m = \left(\mathbf{P}_m\mathbf{Q} - \mathbf{I}_{p+q}\right)\mathbf{S}_0$. For any two models, we have

$$\begin{aligned}Cov(\boldsymbol{\Lambda}_m, \boldsymbol{\Lambda}_\ell) &= \mathrm{E}\left(\left(\mathbf{C}_m\boldsymbol{\delta} + \mathbf{P}_m\mathbf{R} - \mathrm{E}(\mathbf{C}_m\boldsymbol{\delta} + \mathbf{P}_m\mathbf{R})\right)\left(\mathbf{C}_\ell\boldsymbol{\delta} + \mathbf{P}_\ell\mathbf{R} - \mathrm{E}(\mathbf{C}_\ell\boldsymbol{\delta} + \mathbf{P}_\ell\mathbf{R})\right)'\right) \\ &= \mathrm{E}\left(\mathbf{P}_m\mathbf{R}\mathbf{R}'\mathbf{P}_\ell'\right) = \mathbf{P}_m\mathrm{E}\left(\mathbf{R}\mathbf{R}'\right)\mathbf{P}_\ell' = \mathbf{P}_m\boldsymbol{\Omega}\mathbf{P}_\ell,\end{aligned}$$

where the second equality holds by the fact that $\mathbf{C}_m$, $\mathbf{P}_m$, and $\boldsymbol{\delta}$ are constant vectors and $\mathbf{R} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\Omega})$. Therefore, the covariance matrix of $\boldsymbol{\Lambda}$ is

$$\begin{aligned} Var\left(\sum_{m=1}^{M} w_m \boldsymbol{\Lambda}_m\right) &= \sum_{m=1}^{M} w_m^2 Var(\boldsymbol{\Lambda}_m) + 2\sum\sum_{m\neq\ell} w_m w_\ell Cov(\boldsymbol{\Lambda}_m, \boldsymbol{\Lambda}_\ell) \\ &= \sum_{m=1}^{M} w_m^2 \mathbf{P}_m \boldsymbol{\Omega} \mathbf{P}_m + 2\sum\sum_{m\neq\ell} w_m w_\ell \mathbf{P}_m \boldsymbol{\Omega} \mathbf{P}_\ell \equiv \mathbf{V}_{\mathbf{w}}. \end{aligned}$$

This completes the proof. ■

Proof of Theorem 2: The argument is similar to the proof of Theorem 3 and we omit it for brevity. ■

Proof of Theorem 3: The asymptotic trimmed risk is easy to calculate when the estimator $\tilde{\boldsymbol{\theta}}$ has an asymptotic distribution. Suppose that $\sqrt{T}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \mathbf{Z} \sim N(0, \mathbf{V})$. Then by Lemma 1 of Hansen (2016), the asymptotic trimmed risk equals $\mathrm{E}(\mathbf{Z}'\mathbf{Q}\mathbf{Z})$. We first rewrite the asymptotic distribution of the averaging estimator in (A.2) as

$$\begin{aligned} \sqrt{T}\left(\widehat{\boldsymbol{\theta}}(\mathbf{w}) - \boldsymbol{\theta}\right) \xrightarrow{d} \sum_{m=1}^{M} w_m \boldsymbol{\Lambda}_m &= \sum_{m=1}^{M} w_m \left(\left(\mathbf{P}_m \mathbf{Q} - \mathbf{I}_{p+q}\right) \mathbf{S}_0 \boldsymbol{\delta} + \mathbf{P}_m \mathbf{R}\right) \\ &= \mathbf{A}(\mathbf{w})\boldsymbol{\delta} + \mathbf{P}_{\mathbf{w}}\mathbf{R}, \end{aligned} \tag{A.3}$$

where $\mathbf{P}_{\mathbf{w}} = \sum_{m=1}^{M} w_m \mathbf{P}_m$. Note that $\mathbf{A}(\mathbf{w}) = \sum_{m=1}^{M} w_m \left(\mathbf{P}_m \mathbf{Q} - \mathbf{I}_{p+q}\right) \mathbf{S}_0 = \sum_{m=1}^{M} w_m \mathbf{C}_m$, where $\mathbf{C}_m = \left(\mathbf{P}_m \mathbf{Q} - \mathbf{I}_{p+q}\right) \mathbf{S}_0$. Thus, the asymptotic trimmed risk of $\widehat{\boldsymbol{\theta}}(\mathbf{w})$ is

$$\begin{aligned} R(\widehat{\boldsymbol{\theta}}(\mathbf{w}), \boldsymbol{\theta}) &= \mathrm{E}\left(\left(\mathbf{A}(\mathbf{w})\boldsymbol{\delta} + \mathbf{P}_{\mathbf{w}}\mathbf{R}\right)'\mathbf{Q}\left(\mathbf{A}(\mathbf{w})\boldsymbol{\delta} + \mathbf{P}_{\mathbf{w}}\mathbf{R}\right)\right) \\ &= \mathrm{E}\left(\boldsymbol{\delta}'\mathbf{A}(\mathbf{w})'\mathbf{Q}\mathbf{A}(\mathbf{w})\boldsymbol{\delta}\right) + 2\mathrm{E}\left(\mathbf{R}'\mathbf{P}_{\mathbf{w}}'\mathbf{Q}\mathbf{A}(\mathbf{w})\boldsymbol{\delta}\right) + \mathrm{E}\left(\mathbf{R}'\mathbf{P}_{\mathbf{w}}'\mathbf{Q}\mathbf{P}_{\mathbf{w}}\mathbf{R}\right) \\ &= \boldsymbol{\delta}'\mathbf{A}(\mathbf{w})'\mathbf{Q}\mathbf{A}(\mathbf{w})\boldsymbol{\delta} + \mathrm{E}(\mathbf{R}'\mathbf{P}_{\mathbf{w}}'\mathbf{Q}\mathbf{P}_{\mathbf{w}}\mathbf{R}) \\ &= \mathrm{tr}\left(\mathbf{Q}\mathbf{A}(\mathbf{w})\boldsymbol{\delta}\boldsymbol{\delta}'\mathbf{A}(\mathbf{w})'\right) + \mathrm{tr}\left(\mathbf{Q}\mathbf{P}_{\mathbf{w}}'\boldsymbol{\Omega}\mathbf{P}_{\mathbf{w}}\right) \\ &= \mathbf{w}'\boldsymbol{\psi}\mathbf{w}, \end{aligned}$$

where $\boldsymbol{\psi}$ is an $M \times M$ matrix with the (m,ℓ)th element $\psi_{m,\ell} = \mathrm{tr}\left(\mathbf{Q}\mathbf{C}_m\boldsymbol{\delta}\boldsymbol{\delta}'\mathbf{C}_\ell'\right) + \mathrm{tr}\left(\mathbf{Q}\mathbf{P}_m\boldsymbol{\Omega}\mathbf{P}_\ell\right)$. This completes the proof. ■

Proof of Theorem 4: Recall that $\widehat{\psi}_{m,\ell} = \mathrm{tr}(\widehat{\mathbf{Q}}\widehat{\mathbf{C}}_m\widehat{\boldsymbol{\delta}\boldsymbol{\delta}'}\widehat{\mathbf{C}}_\ell') + \mathrm{tr}(\widehat{\mathbf{Q}}\widehat{\mathbf{P}}_m\widehat{\boldsymbol{\Omega}}\widehat{\mathbf{P}}_\ell)$. We first show that $\mathbf{w}'\widehat{\boldsymbol{\psi}}\mathbf{w} \xrightarrow{p} \mathbf{w}'\boldsymbol{\psi}^*\mathbf{w}$ and $\widehat{\mathbf{w}} \xrightarrow{d} \mathbf{w}^* = \mathrm{argmin}_{\mathbf{w}\in\mathcal{H}^M} \mathbf{w}'\boldsymbol{\psi}^*\mathbf{w}$. Since $\widehat{\mathbf{Q}}$ and $\widehat{\boldsymbol{\Omega}}$

are consistent estimators for $\mathbf{Q}$ and $\mathbf{\Omega}$, we have $\text{tr}(\widehat{\mathbf{Q}}\widehat{\mathbf{P}}_m\widehat{\mathbf{\Omega}}\widehat{\mathbf{P}}_\ell) \xrightarrow{p} \text{tr}(\mathbf{Q}\mathbf{P}_m\mathbf{\Omega}\mathbf{P}_\ell)$ by the continuous mapping theorem. Then by (3.5), (3.6), and the application of Slutsky's theorem, we have

$$\widehat{\psi}_{m,\ell} \xrightarrow{d} \text{tr}\left(\mathbf{Q}\mathbf{C}_m(\mathbf{R}_{\boldsymbol{\delta}}\mathbf{R}_{\boldsymbol{\delta}}' - \mathbf{S}_0'\mathbf{Q}^{-1}\mathbf{\Omega}\mathbf{Q}^{-1}\mathbf{S}_0)\mathbf{C}_\ell'\right) + \text{tr}\left(\mathbf{Q}\mathbf{P}_m\mathbf{\Omega}\mathbf{P}_\ell\right) = \psi_{m,\ell}^*.$$

Note that all of $\psi_{m,\ell}^*$ can be expressed in terms of the normal random vector $\mathbf{R}$. Therefore, there is joint convergence in distribution of all $\widehat{\psi}_{m,\ell}$ to $\psi_{m,\ell}^*$. Thus we have $\mathbf{w}'\widehat{\boldsymbol{\psi}}\mathbf{w} \xrightarrow{p} \mathbf{w}'\boldsymbol{\psi}^*\mathbf{w}$. Next observe that $\mathbf{w}'\boldsymbol{\psi}^*\mathbf{w}$ is a convex minimization problem because $\mathbf{w}'\boldsymbol{\psi}^*\mathbf{w}$ is quadratic and $\boldsymbol{\psi}^*$ is positive definite. Hence, the limiting process $\mathbf{w}'\boldsymbol{\psi}^*\mathbf{w}$ is continuous in $\mathbf{w}$ and has a unique minimum. Also note that $\widehat{\mathbf{w}} = O_p(1)$ by the fact that $\mathcal{H}^M$ is convex. Therefore, by Theorem 3.2.2 of Van der Vaart and Wellner (1996) or Theorem 2.7 of Kim and Pollard (1990), the minimizer $\widehat{\mathbf{w}}$ converges in distribution to the minimizer of $\mathbf{w}'\boldsymbol{\psi}^*\mathbf{w}$, which is $\mathbf{w}^*$.

We now derive the asymptotic distribution and asymptotic trimmed risk of the plug-in averaging estimator. Since both $\mathbf{\Lambda}_m$ and w_m^* can be expressed in terms of the same normal random vector $\mathbf{R}$, there is joint convergence in distribution of all $\widehat{\boldsymbol{\theta}}(m)$ and $\hat{w}_m$. By (A.2), (A.3), and $\widehat{\mathbf{w}} \xrightarrow{d} \mathbf{w}^*$, it follows that

$$\sqrt{T}\left(\widehat{\boldsymbol{\theta}}(\widehat{\mathbf{w}}) - \boldsymbol{\theta}\right) = \sum_{m=1}^{M} \widehat{w}_m\sqrt{T}\left(\mathbf{S}_m\widehat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}\right) \xrightarrow{d} \sum_{m=1}^{M} w_m^*\mathbf{\Lambda}_m = \mathbf{A}(\mathbf{w}^*)\boldsymbol{\delta} + \mathbf{P}(\mathbf{w}^*)\mathbf{R},$$

where $\mathbf{A}(\mathbf{w}^*) = \sum_{m=1}^{M} w_m^*\left(\mathbf{P}_m\mathbf{Q} - \mathbf{I}_{p+q}\right)\mathbf{S}_0$ and $\mathbf{P}(\mathbf{w}^*) = \sum_{m=1}^{M} w_m^*\mathbf{P}_m$. Therefore, the asymptotic trimmed risk of $\widehat{\boldsymbol{\theta}}(\widehat{\mathbf{w}})$ is

$$R(\widehat{\boldsymbol{\theta}}(\widehat{\mathbf{w}}), \boldsymbol{\theta}) = \text{E}\left((\mathbf{A}(\mathbf{w}^*)\boldsymbol{\delta} + \mathbf{P}(\mathbf{w}^*)\mathbf{R})'\mathbf{Q}(\mathbf{A}(\mathbf{w}^*)\boldsymbol{\delta} + \mathbf{P}(\mathbf{w}^*)\mathbf{R})\right).$$

This completes the proof. ■

Derivation of Equation (3.20): We first show the first term of (3.20). Note that

$$\begin{aligned}\text{tr}\Big(\widehat{\mathbf{Q}}(\widehat{\mathbf{P}}_m\widehat{\mathbf{Q}} - \mathbf{I}_q)\widehat{\boldsymbol{\delta}}\widehat{\boldsymbol{\delta}}'(\widehat{\mathbf{Q}}\widehat{\mathbf{P}}_\ell - \mathbf{I}_q)\Big) &= \text{tr}(\widehat{\mathbf{Q}}\widehat{\mathbf{P}}_m\widehat{\mathbf{Q}}\widehat{\boldsymbol{\delta}}\widehat{\boldsymbol{\delta}}'\widehat{\mathbf{Q}}\widehat{\mathbf{P}}_\ell) - \text{tr}(\widehat{\mathbf{Q}}\widehat{\mathbf{P}}_m\widehat{\mathbf{Q}}\widehat{\boldsymbol{\delta}}\widehat{\boldsymbol{\delta}}') \\ &\quad - \text{tr}(\widehat{\mathbf{Q}}\widehat{\boldsymbol{\delta}}\widehat{\boldsymbol{\delta}}'\widehat{\mathbf{Q}}\widehat{\mathbf{P}}_\ell) + \text{tr}(\widehat{\mathbf{Q}}\widehat{\boldsymbol{\delta}}\widehat{\boldsymbol{\delta}}'). \qquad \text{(A.4)}\end{aligned}$$

Recall that $\widehat{\boldsymbol{\delta}} = \sqrt{T}\widehat{\boldsymbol{\gamma}} = \sqrt{T}(\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\mathbf{y}$. Thus, we have $\widehat{\boldsymbol{\delta}}\widehat{\boldsymbol{\delta}}' = T^{-1}\widehat{\mathbf{Q}}^{-1}\mathbf{H}'\mathbf{y}\mathbf{y}'\mathbf{H}\widehat{\mathbf{Q}}^{-1}$. Note that $\mathbf{H}_m = \mathbf{H}\mathbf{S}_m$ and $\widehat{\mathbf{P}}_m = \mathbf{S}_m\widehat{\mathbf{Q}}_m^{-1}\mathbf{S}_m'$. Then the first term of (A.4) can be

rewritten as

$$\begin{aligned}\mathrm{tr}(\widehat{\mathbf{Q}}\widehat{\mathbf{P}}_m\widehat{\mathbf{Q}}\widehat{\boldsymbol{\delta}}\widehat{\boldsymbol{\delta}}'\widehat{\mathbf{Q}}\widehat{\mathbf{P}}_\ell) &= \mathrm{tr}(T^{-1}\widehat{\mathbf{Q}}\mathbf{S}_m\widehat{\mathbf{Q}}_m^{-1}\mathbf{S}_m'\widehat{\mathbf{Q}}\widehat{\mathbf{Q}}^{-1}\mathbf{H}'\mathbf{y}\mathbf{y}'\mathbf{H}\widehat{\mathbf{Q}}^{-1}\widehat{\mathbf{Q}}\mathbf{S}_\ell\widehat{\mathbf{Q}}_\ell^{-1}\mathbf{S}_\ell') \\ &= \mathrm{tr}(\mathbf{H}_m(\mathbf{H}_m'\mathbf{H}_m)^{-1}\mathbf{H}_m'\mathbf{y}\mathbf{y}'\mathbf{H}_\ell(\mathbf{H}_\ell'\mathbf{H}_\ell)^{-1}\mathbf{H}_\ell') \\ &= \mathbf{y}'\mathcal{P}_m\mathcal{P}_\ell\mathbf{y},\end{aligned}$$

where $\mathcal{P}_m = \mathbf{H}_m(\mathbf{H}_m'\mathbf{H}_m)^{-1}\mathbf{H}_m'$. Define $\mathcal{P} = \mathbf{H}(\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'$. Thus, the equation (A.4) can be rewritten as

$$\begin{aligned}\mathrm{tr}(\widehat{\mathbf{Q}}(\widehat{\mathbf{P}}_m\widehat{\mathbf{Q}} - \mathbf{I}_q)\widehat{\boldsymbol{\delta}}\widehat{\boldsymbol{\delta}}'(\widehat{\mathbf{Q}}\widehat{\mathbf{P}}_\ell - \mathbf{I}_q)) &= \mathbf{y}'\mathcal{P}_m\mathcal{P}_\ell\mathbf{y} - \mathbf{y}'\mathcal{P}_m\mathbf{y} - \mathbf{y}'\mathcal{P}_\ell\mathbf{y} + \mathbf{y}'\mathcal{P}\mathbf{y} \\ &= \mathbf{y}'(\mathbf{I} - \mathcal{P}_m)(\mathbf{I} - \mathcal{P}_\ell)\mathbf{y} - \mathbf{y}'(\mathbf{I} - \mathcal{P})\mathbf{y} \\ &= \widehat{\mathbf{e}}_m'\widehat{\mathbf{e}}_\ell - \widehat{\mathbf{e}}'\widehat{\mathbf{e}}, \qquad \text{(A.5)}\end{aligned}$$

where $\widehat{\mathbf{e}} = \mathbf{y} - \mathbf{H}\widehat{\boldsymbol{\theta}}$ and $\widehat{\mathbf{e}}_m = \mathbf{y} - \mathbf{H}_m\widehat{\boldsymbol{\theta}}_m$.

We now show the second term of (3.20). By some algebra, it follows that

$$\begin{aligned}&\mathrm{tr}\Big(\widehat{\mathbf{Q}}(\widehat{\mathbf{P}}_m\widehat{\mathbf{Q}} - \mathbf{I}_q)\widehat{\mathbf{Q}}^{-1}\widehat{\boldsymbol{\Omega}}\widehat{\mathbf{Q}}^{-1}(\widehat{\mathbf{Q}}\widehat{\mathbf{P}}_\ell - \mathbf{I}_q) - \widehat{\mathbf{Q}}\widehat{\mathbf{P}}_m\widehat{\boldsymbol{\Omega}}\widehat{\mathbf{P}}_\ell\Big) \qquad \text{(A.6)} \\ &= \mathrm{tr}(\widehat{\mathbf{Q}}\widehat{\mathbf{P}}_m\widehat{\mathbf{Q}}\widehat{\mathbf{Q}}^{-1}\widehat{\boldsymbol{\Omega}}\widehat{\mathbf{Q}}^{-1}\widehat{\mathbf{Q}}\widehat{\mathbf{P}}_\ell - \widehat{\mathbf{Q}}\widehat{\mathbf{P}}_m\widehat{\mathbf{Q}}\widehat{\mathbf{Q}}^{-1}\widehat{\boldsymbol{\Omega}}\widehat{\mathbf{Q}}^{-1} \\ &\quad - \widehat{\mathbf{Q}}\widehat{\mathbf{Q}}^{-1}\widehat{\boldsymbol{\Omega}}\widehat{\mathbf{Q}}^{-1}\widehat{\mathbf{Q}}\widehat{\mathbf{P}}_\ell + \widehat{\mathbf{Q}}\widehat{\mathbf{Q}}^{-1}\widehat{\boldsymbol{\Omega}}\widehat{\mathbf{Q}}^{-1} - \widehat{\mathbf{Q}}\widehat{\mathbf{P}}_m\widehat{\boldsymbol{\Omega}}\widehat{\mathbf{P}}_\ell) \\ &= \mathrm{tr}(\widehat{\mathbf{Q}}\widehat{\mathbf{P}}_m\widehat{\boldsymbol{\Omega}}\widehat{\mathbf{P}}_\ell) - \mathrm{tr}(\widehat{\boldsymbol{\Omega}}\widehat{\mathbf{P}}_m) - \mathrm{tr}(\widehat{\boldsymbol{\Omega}}\widehat{\mathbf{P}}_\ell) + \mathrm{tr}(\widehat{\mathbf{Q}}^{-1}\widehat{\boldsymbol{\Omega}}) - \mathrm{tr}(\widehat{\mathbf{Q}}\widehat{\mathbf{P}}_m\widehat{\boldsymbol{\Omega}}\widehat{\mathbf{P}}_\ell) \\ &= \mathrm{tr}(\widehat{\mathbf{Q}}^{-1}\widehat{\boldsymbol{\Omega}}) - \mathrm{tr}(\widehat{\mathbf{Q}}_m^{-1}\widehat{\boldsymbol{\Omega}}_m) - \mathrm{tr}(\widehat{\mathbf{Q}}_\ell^{-1}\widehat{\boldsymbol{\Omega}}_\ell), \qquad \text{(A.7)}\end{aligned}$$

where $\widehat{\boldsymbol{\Omega}}_m = \mathbf{S}_m'\widehat{\boldsymbol{\Omega}}\mathbf{S}_m$. Combining (A.5) and (A.7), we have (3.20). ■

B Figures

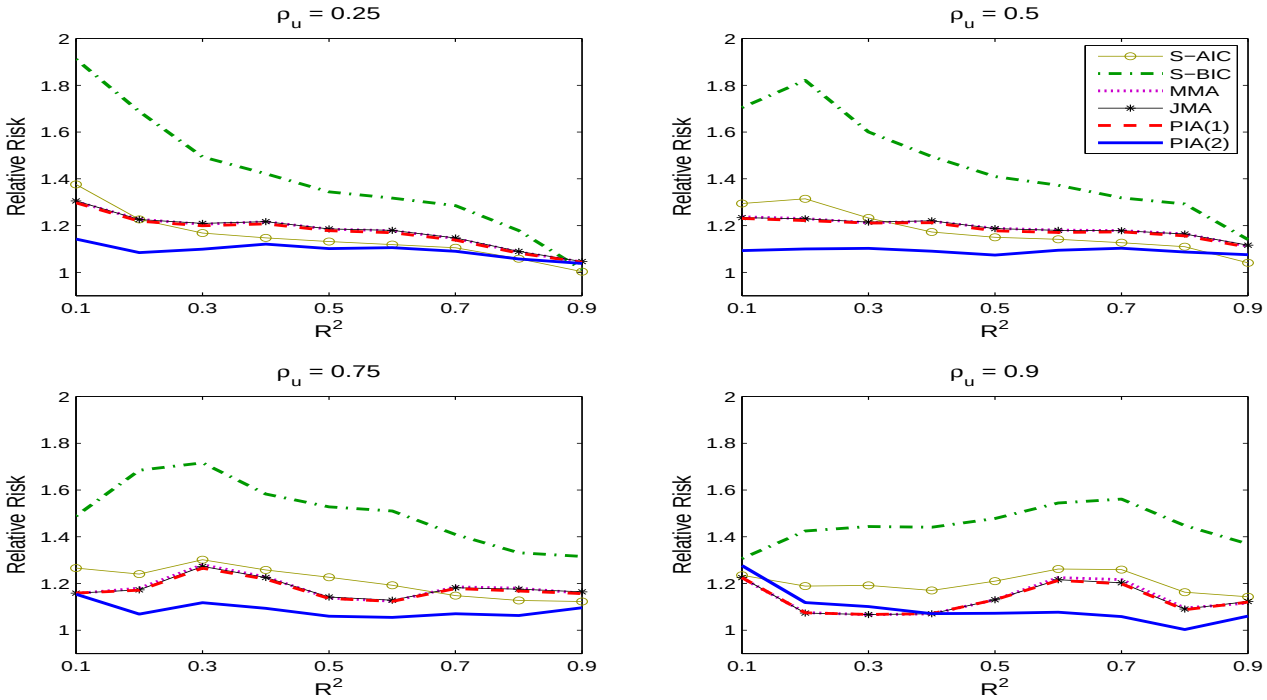

Figure 1: Relative risk for linear regression models with homoskedastic errors and $\rho_x = 0.5$.

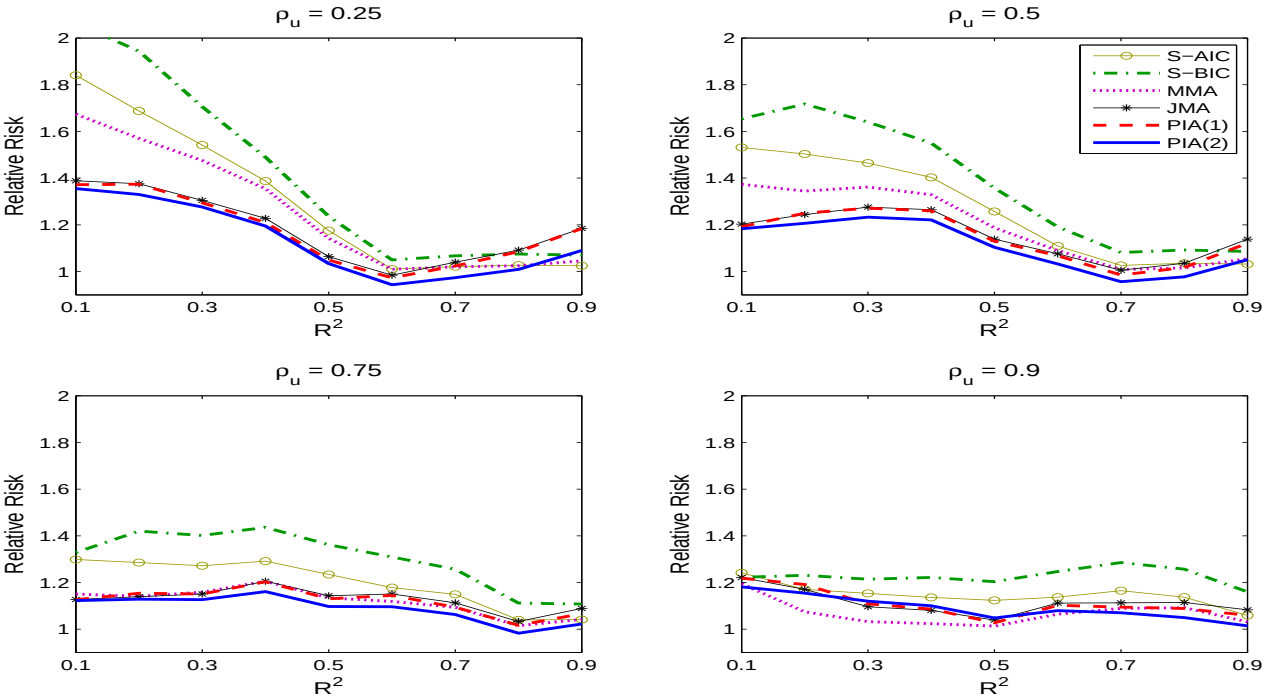

Figure 2: Relative risk for linear regression models with heteroskedastic errors and $\rho_x = 0.5$.

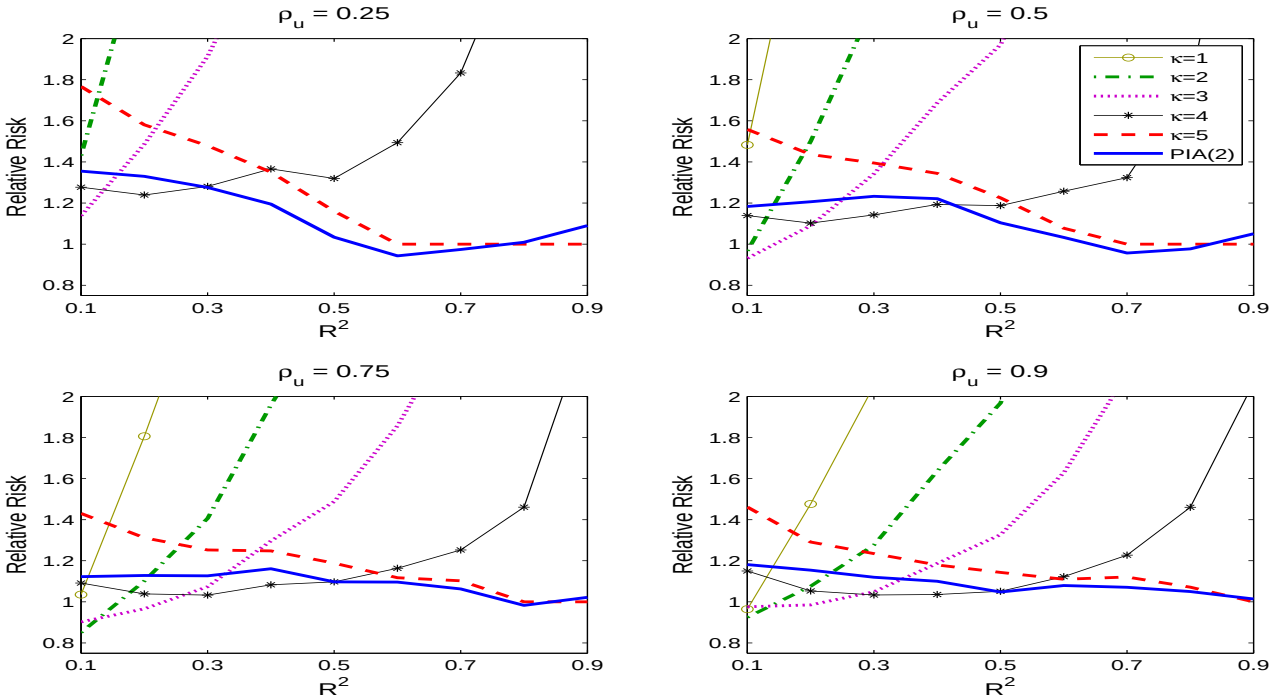

Figure 3: Relative risk for linear regression models with heteroskedastic errors and $\rho_x = 0.5$.

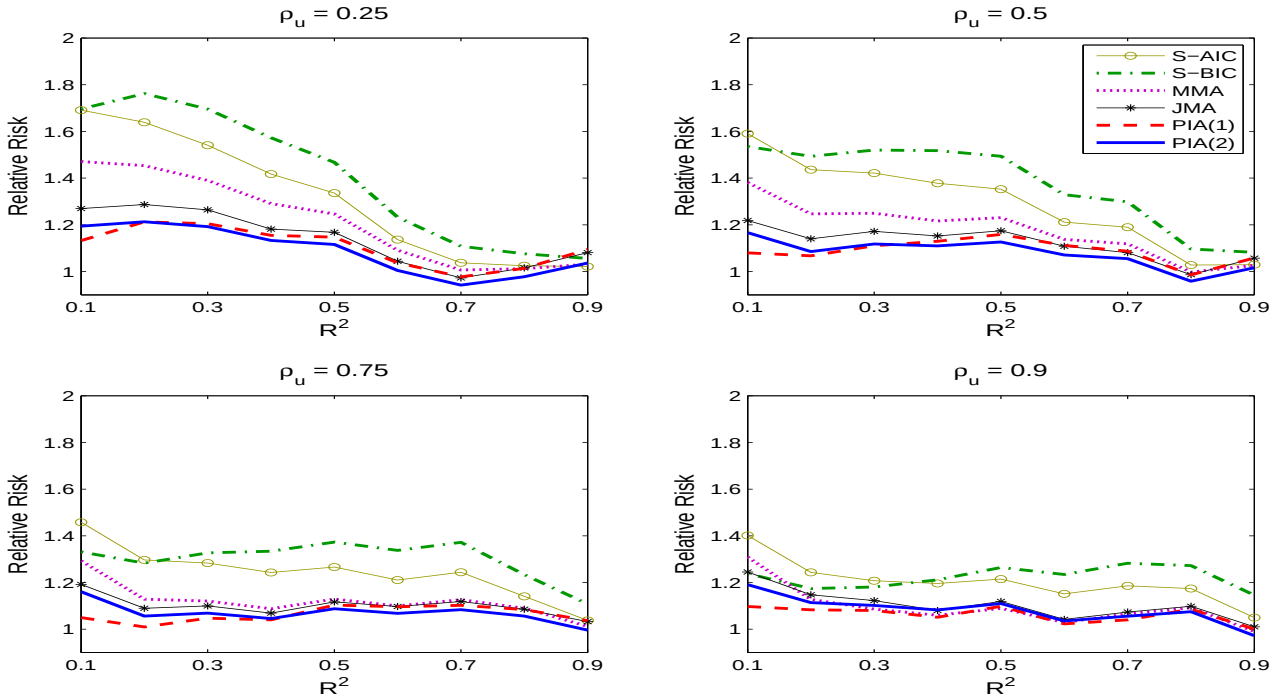

Figure 4: Relative risk for linear regression models with heteroskedastic errors and $\rho_x = 0.9$.

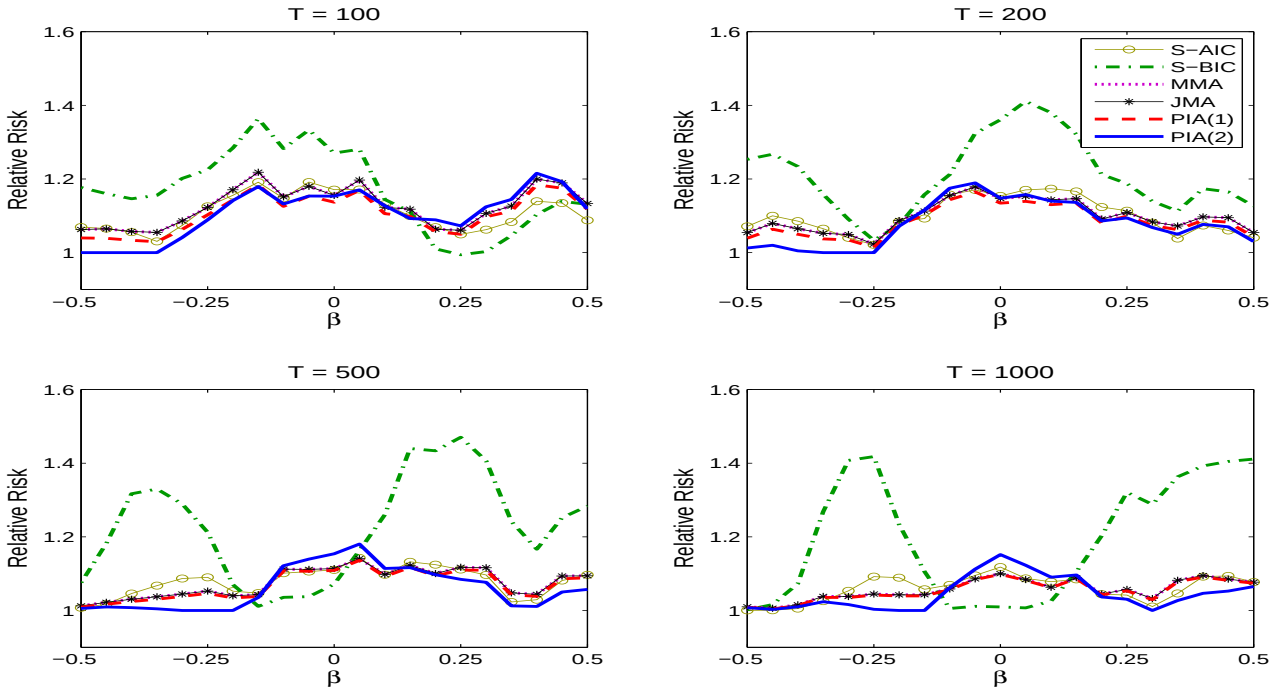

Figure 5: Relative risk for MAX(1, 1) models.

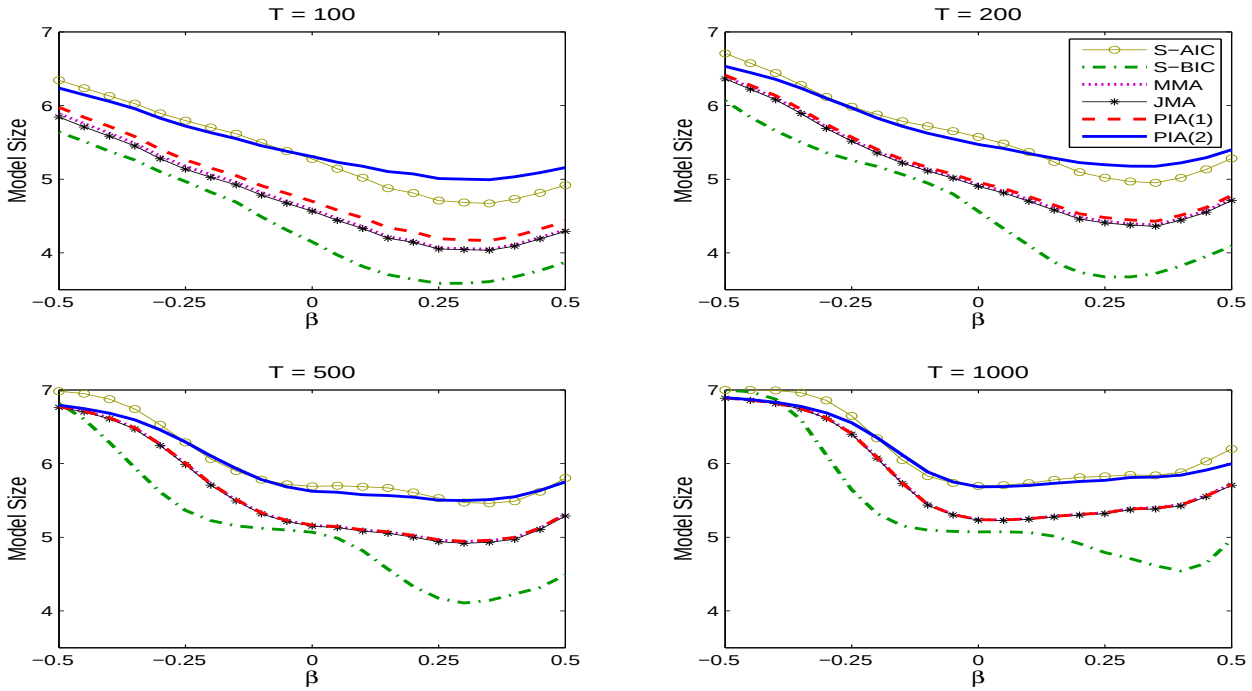

Figure 6: Average model size for MAX(1, 1) models.

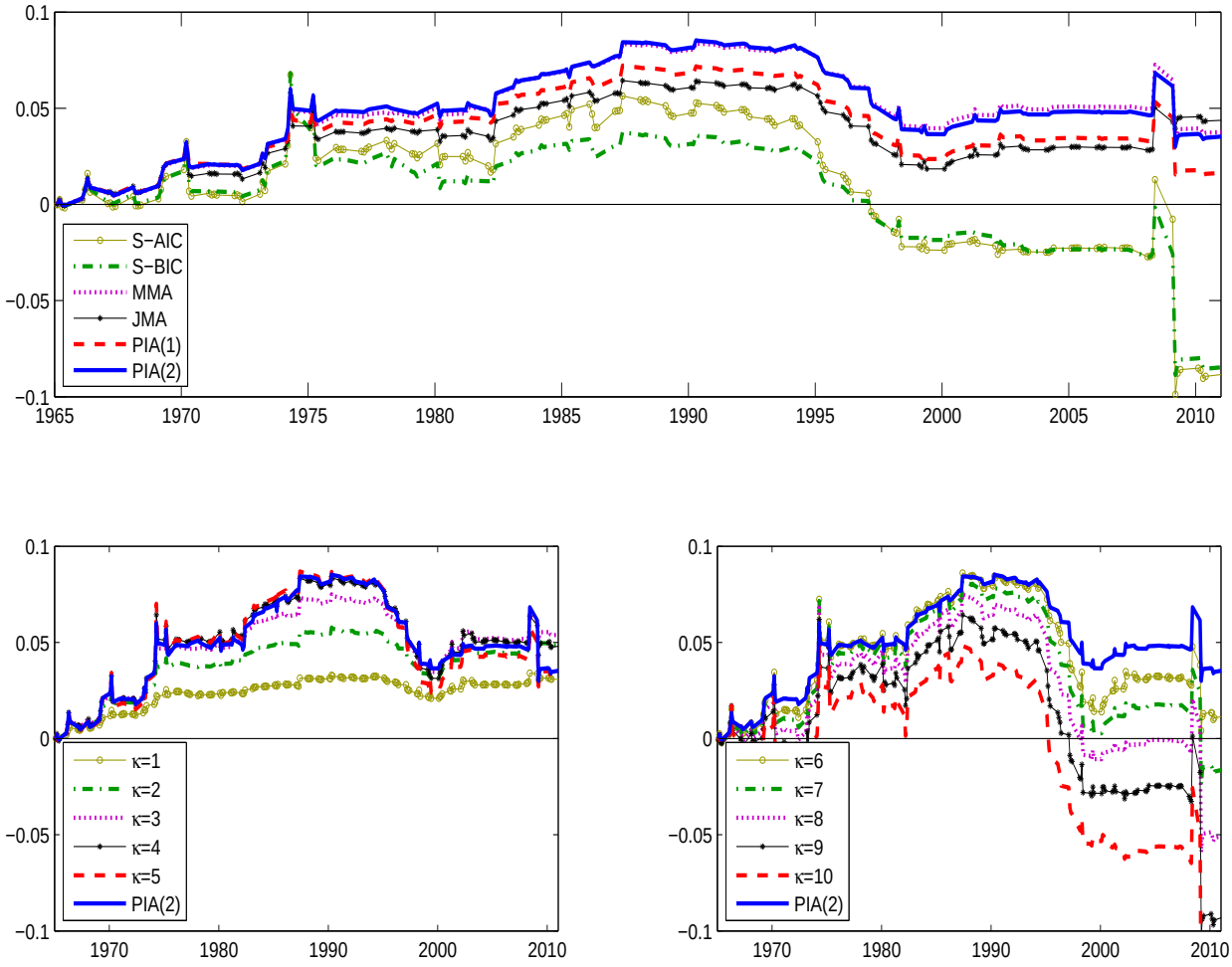

Figure 7: The differences between the cumulative squared prediction error of the historical average forecasting model and the cumulative squared prediction error of the forecast combination model for 1965:1–2011:4.

References

Andrews, D. W. K. (1991a): "Asymptotic Optimality of Generalized C_L, Cross-Validation, and Generalized Cross-Validation in Regression with Heteroskedastic Errors," *Journal of Econometrics*, 47, 359–377.

——— (1991b): "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59, 817–858.

Bates, J. and C. Granger (1969): "The Combination of Forecasts," *Operational Research Quarterly*, 20, 451–468.

Breiman, L. (1996): "Bagging Predictors," *Machine learning*, 24, 123–140.

Buckland, S., K. Burnham, and N. Augustin (1997): "Model Selection: An Integral Part of Inference," *Biometrics*, 53, 603–618.

Cheng, X. and B. E. Hansen (2015): "Forecasting with Factor-Augmented Regression: A Frequentist Model Averaging Approach," *Journal of Econometrics*, 186, 280–293.

Claeskens, G. and R. J. Carroll (2007): "An Asymptotic Theory for Model Selection Inference in General Semiparametric Problems," *Biometrika*, 94, 249–265.

Claeskens, G. and N. L. Hjort (2003): "The Focused Information Criterion," *Journal of the American Statistical Association*, 98, 900–916.

——— (2008): "Minimizing Average Risk in Regression Models," *Econometric Theory*, 24, 493–527.

Clark, T. and K. West (2007): "Approximately Normal Tests for Equal Predictive Accuracy in Nested Models," *Journal of Econometrics*, 138, 291–311.

Clemen, R. (1989): "Combining Forecasts: A Review and Annotated Bibliography," *International Journal of Forecasting*, 5, 559–583.

DiTraglia, F. (2014): "Using Invalid Instruments on Purpose: Focused Moment Selection and Averaging for GMM," Working Paper, University of Pennsylvania.

Elliott, G., A. Gargano, and A. Timmermann (2013): "Complete Subset Regressions," *Journal of Econometrics*, 177, 357–373.

Granger, C. (1989): "Combining Forecasts–Twenty Years Later," *Journal of Forecasting*, 8, 167–173.

Granger, C. and R. Ramanathan (1984): "Improved Methods of Combining Forecasts," *Journal of Forecasting*, 3, 197–204.

Hansen, B. E. (2007): "Least Squares Model Averaging," *Econometrica*, 75, 1175–1189.

——— (2008): "Least-Squares Forecast Averaging," *Journal of Econometrics*, 146, 342–350.

——— (2010): "Multi-Step Forecast Model Selection," Working Paper, University of Wisconsin.

——— (2014): "Model Averaging, Asymptotic Risk, and Regressor Groups," *Quantitative Economics*, 5, 495–530.

——— (2016): "Efficient Shrinkage in Parametric Models," *Journal of Econometrics*, 190, 115–132.

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

Hansen, P., A. Lunde, and J. Nason (2011): "The Model Confidence Set," *Econometrica*, 79, 453–497.

Hjort, N. L. and G. Claeskens (2003): "Frequentist Model Average Estimators," *Journal of the American Statistical Association*, 98, 879–899.

Ing, C.-K. and C.-Z. Wei (2005): "Order Selection for Same-Realization Predictions in Autoregressive Processes," *The Annals of Statistics*, 33, 2423–2474.

Inoue, A. and L. Kilian (2008): "How Useful is Bagging in Forecasting Economic Time Series? A Case Study of US Consumer Price Inflation," *Journal of the American Statistical Association*, 103, 511–522.

Kim, J. and D. Pollard (1990): "Cube Root Asymptotics," *The Annals of Statistics*, 18, 191–219.

Kitagawa, T. and C. Muris (2013): "Covariate Selection and Model Averaging in Semiparametric Estimation of Treatment Effects," Cemmap Working Paper.

Leeb, H. and B. Pötscher (2005): "Model Selection and Inference: Facts and Fiction," *Econometric Theory*, 21, 21–59.

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.

Liu, C.-A. (2015): "Distribution Theory of the Least Squares Averaging Estimator," *Journal of Econometrics*, 186, 142–159.

Liu, Q. and R. Okui (2013): "Heteroscedasticity-Robust C_p Model Averaging," *The Econometrics Journal*, 16, 463–472.

Lu, X. (2015): "A Covariate Selection Criterion for Estimation of Treatment Effects," *Journal of Business & Economic Statistics*, 33, 506–522.

Min, C.-K. and A. Zellner (1993): "Bayesian and Non-Bayesian Methods for Combining Models and Forecasts with Applications to Forecasting International Growth Rates," *Journal of Econometrics*, 56, 89–118.

Newey, W. and K. West (1987): "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–708.

Ng, S. (2013): "Variable Selection in Predictive Regressions," in *Handbook of Economic Forecasting*, ed. by G. Elliott and A. Timmermann, Elsevier, vol. 2, chap. 14, 752–789.

Pötscher, B. (2006): "The Distribution of Model Averaging Estimators and an Impossibility Result Regarding its Estimation," *Lecture Notes-Monograph Series*, 52, 113–129.

Raftery, A., D. Madigan, and J. Hoeting (1997): "Bayesian Model Averaging for Linear Regression Models," *Journal of the American Statistical Association*, 92, 179–191.

Rapach, D., J. Strauss, and G. Zhou (2010): "Out-of-Sample Equity Premium Prediction: Combination Forecasts and Links to the Real Economy," *Review of Financial Studies*, 23, 821–862.

Rapach, D. and G. Zhou (2012): "Forecasting Stock Returns," in *Handbook of Economic Forecasting*, Elsevier, vol. 2.

Shao, J. (1997): "An Asymptotic Theory for Linear Model Selection," *Statistica Sinica*, 7, 221–242.

Shibata, R. (1980): "Asymptotically Efficient Selection of the Order of the Model for Estimating Parameters of a Linear Process," *The Annals of Statistics*, 8, 147–164.

Staiger, D. and J. Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65, 557–586.

Stock, J. H. and M. W. Watson (2006): "Forecasting with Many Predictors," in *Handbook of Economic Forecasting*, ed. by G. Elliott, C. Granger, and A. Timmermann, Elsevier, vol. 1, 515–554.

Sueishi, N. (2013): "Generalized Empirical Likelihood-Based Focused Information Criterion and Model Averaging," *Econometrics*, 1, 141–156.

Tibshirani, R. (1996): "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, 58, 267–288.

Timmermann, A. (2006): "Forecast Combinations," in *Handbook of Economic Forecasting*, ed. by G. Elliott, C. Granger, and A. Timmermann, Elsevier, vol. 1, 135–196.

Van der Vaart, A. and J. Wellner (1996): *Weak Convergence and Empirical Processes*, Springer Verlag.

Wan, A., X. Zhang, and G. Zou (2010): "Least Squares Model Averaging by Mallows Criterion," *Journal of Econometrics*, 156, 277–283.

Welch, I. and A. Goyal (2008): "A Comprehensive Look at the Empirical Performance of Equity Premium Prediction," *Review of Financial Studies*, 21, 1455–1508.

White, H. (1980): "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817–838.

Yang, Y. (2004): "Combining Forecasting Procedures: Some Theoretical Results," *Econometric Theory*, 20, 176–222.

——— (2007): "Prediction/Estimation with Simple Linear Models: Is it Really that Simple?" *Econometric Theory*, 23, 1–36.

Zhang, X. and H. Liang (2011): "Focused Information Criterion and Model Averaging for Generalized Additive Partial Linear Models," *The Annals of Statistics*, 39, 174–200.

Zhang, X., A. T. Wan, and S. Z. Zhou (2012): "Focused Information Criteria, Model Selection, and Model Averaging in a Tobit Model with a Nonzero Threshold," *Journal of Business & Economic Statistics*, 30, 132–142.

Zhang, X., A. T. Wan, and G. Zou (2013): "Model Averaging by Jackknife Criterion in Models with Dependent Data," *Journal of Econometrics*, 174, 82–94.

Zhang, X., G. Zou, and H. Liang (2014): "Model Averaging and Weight Choice in Linear Mixed-Effects Models," *Biometrika*, 101, 205–218.

Zou, H. (2006): "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429.

Zou, H. and Y. Yang (2004): "Combining Time Series Models for Forecasting," *International Journal of Forecasting*, 20, 69–84.