# Model Averaging in Predictive Regressions

Liu, Chu-An and Kuo, Biing-Shen

National University of Singapore, National Chengchi University

7 March 2014

# Model Averaging in Predictive Regressions*

Chu-An Liu† and Biing-Shen Kuo‡

March 7, 2014

**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 propose a frequentist model averaging criterion, an asymptotically unbiased estimator of the mean squared forecast error (MSFE), to select forecast weights. In contrast to the existing literature, we derive the MSFE in a local asymptotic framework without the i.i.d. normal assumption. This result allows us to decompose the MSFE into the bias and variance components and also to account for the correlations between candidate models. Monte Carlo simulations show that our averaging estimator has much lower MSFE than alternative methods such as weighted AIC, weighted BIC, Mallows model averaging, and jackknife model averaging. We apply the proposed method to stock return predictions.

JEL Classification: C52, C53

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

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

†Department of Economics, National University of Singapore. Email: ecslca@nus.edu.sg.

‡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.

This paper proposes a new frequentist model averaging criterion for forecast combination. For a given set of potentially relevant predictors, we construct the point forecast by combining predictions from all possible linear regression models. Building on the idea of the weighted focused information criterion (wFIC) proposed by Claeskens and Hjort (2008), we introduce a model averaging criterion to select the weights for candidate models and study its properties.[1] The proposed model averaging criterion is an estimate of the mean squared forecast error (MSFE). Therefore, the data-driven weights that minimize the model averaging criterion are expected to close to the optimal weights that minimize the MSFE. In contrast to the existing literature, we derive the MSFE of forecast combination in a local asymptotic framework without the i.i.d. normal assumption. This result allows us to decompose the MSFE 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. Furthermore, the criterion also accounts for the correlations between candidate models instead of assuming perfect correlation in most existing methods.

To yield a good approximation to the finite sample behavior, we investigate forecast combination in a local asymptotic framework where the regression coefficients of predictors are in a local $T^{-1/2}$ neighborhood of zero, which is similar to that used in weak instrument theory (Staiger and Stock, 1997). This local-to-zero framework ensures the consistency of the averaging estimator while in general presents an asymptotic bias. Since both squared model

[1]The idea of the focused information criterion proposed by Claeskens and Hjort (2003) 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, 2013), and the propensity score weighted estimation of the treatment effects (Kitagawa and Muris, 2013).

biases and estimator variances have the same stochastic order, the trade-off between omitted variable bias and estimation variance remains in the asymptotic theory. Thus, the forecast combination with optimal weights achieves the best trade-off between bias and variance in this context.

We show that the optimal weights can be characterized by the local parameters and the covariance matrix of the predictive regression. We then propose a plug-in estimator of the infeasible optimal weights and use these estimated weights to construct the forecast combination. Since the estimated weights depend on the covariance matrix, it is quite easy to model the heteroskedasticity and serial correlation by the plug-in method.

To illustrate the plug-in 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. In particular, the plug-in forecast combination approach achieves lower cumulative squared prediction error than those produced by other averaging methods. 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.

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 alternative forecast combination methods are proposed by 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 (2013), among others.

In a recent paper, Hansen (2008) proposes to construct forecast combinations using the weights by minimizing the Mallows model averaging (MMA) criterion introduced in Hansen (2007). Under the homoskedasticity assumption, the MMA criterion is an asymptotically unbiased estimator of the MSFE. The MMA criterion is based on the sum of squared errors and a penalty term that estimates the difference between MSFE and the expectation of the sum of squared errors. Hence, the MMA criterion addresses the trade-off between the model fit and model complexity. Like the MMA criterion, our model averaging criterion is also

an asymptotically unbiased estimator of the MSFE. We, however, employ a drifting asymptotic framework to approximate MSFE and address the trade-off between bias and variance. Compared to the MMA estimator, we do not restrict model errors to be homoskedastic and uncorrelated. Numerical comparisons show that our estimator achieves lower MSFE than the MMA estimator in most simulations.

One popular model averaging approach is the simple equal-weighted average. The simple equal-weighted average makes sense 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.[2] Instead of choosing the weights, the subset regression combinations have to choose the number of predictors $\kappa$, and the data-driven method for $\kappa$ still needs further investigation. Monte Carlo shows that the performance of complete subset regressions is sensitive to the choice of $\kappa$, 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 mean squared forecast error 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) demonstrates the asymptotic optimality of the Mallows model averaging estimator for nested and homoskedastic linear regression models. Wan, Zhang, and Zou (2010) extend the asymptotic optimality of the Mallows model averaging estimator for continuous weights and a non-nested setup. Hansen and Racine (2012) propose the jackknife model

[2]One limitation of subset regression combinations is that the approach is not suitable for the nested models. Suppose 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.

averaging (JMA) 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. These asymptotic theories, however, are limited to the random sample and hence are not directly applicable to forecast combination for dependent data.[3]

The outline of the paper is as follows. Section 2 introduces the model and forecast combination. Section 3 shows that the weight vector that minimizes the MSFE is equivalent to the weight vector that minimizes the MSE. Section 4 characterizes the optimal weights and presents the plug-in estimator for forecast combination. Section 5 evaluates the finite sample MSFE of the plug-in averaging estimator and other averaging estimators in two simulation experiments. Section 6 applies the plug-in forecast combination to the predictability of U.S. stock returns. Section 7 concludes. Proofs and figures are included in the Appendix.

# 2 Model and Forecast Combination

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

$$y_t = \mathbf{x}_{t-1}'\boldsymbol{\beta} + e_t, \tag{2.1}$$

$$\mathrm{E}(e_t|\mathbf{x}_{t-1}) = 0, \tag{2.2}$$

$$\mathrm{E}(e_t^2|\mathbf{x}_{t-1}) = \sigma^2(\mathbf{x}_{t-1}), \tag{2.3}$$

where $y_t$ is a scalar dependent variable, $\mathbf{x}_{t-1}$ is a $k\times 1$ vector of potentially relevant predictors, $\boldsymbol{\beta}$ is a $k \times 1$ vector of unknown parameters, and $e_t$ is an unobservable error term. The 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 serially correlated, and there is no further assumption on the distribution of the error term. We assume throughout that $1 \leq k \leq T - 1$, and we do not

[3]In a recent paper, Zhang, Wan, and Zou (2013) show the asymptotic optimality of the JMA estimator in the presence of lagged dependent variables. They assume that the dependent variable follows the stationary AR($\infty$) process. A more general theory needs to be developed in the future study.

let the number of predictors $k$ increase with the sample size $T$.

We now consider a set of $M$ approximating models indexed by $m = 1, ..., M$, where the $m$th model includes a subset of predictors $\mathbf{x}_{t-1}$. The $m$th model has $k_m$ 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 $M = k + 1$. If we consider all possible combinations of the predictor variables, then $M = 2^k$. Let $\mathbf{\Pi}_m$ be a $k_m \times k$ selection matrix that selects the included predictors in the $m$th model. For example, suppose that $k = 5$ and the $m$th model has three predictors, $x_{1t}$, $x_{2t}$, and $x_{4t}$. Then

$$\mathbf{\Pi}_m = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

In matrix notation, we write the model as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, where $\mathbf{y} = (y_1, y_2, ..., y_T)$ is $T \times 1$, $\mathbf{X} = (\mathbf{x}_0', \mathbf{x}_1', ..., \mathbf{x}_{T-1}')'$ is $T \times k$, and $\mathbf{e} = (e_1, e_2, ..., e_T)$ is $T \times 1$. The least squares estimator of $\boldsymbol{\beta}$ in the $m$th model is $\widehat{\boldsymbol{\beta}}_m = (\mathbf{X}_m'\mathbf{X}_m)^{-1}\mathbf{X}_m'\mathbf{y}$, where $\mathbf{X}_m = (\mathbf{X}\mathbf{\Pi}_m')$. The predicted value is $\widehat{\mathbf{y}}(m) = \mathbf{X}_m\widehat{\boldsymbol{\beta}}_m = \mathbf{X}\mathbf{\Pi}_m'\widehat{\boldsymbol{\beta}}_m$. Thus, the one-step-ahead forecast given information up to period $T$ from this $m$th model is

$$\widehat{y}_{T+1|T}(m) = \mathbf{x}_T'\mathbf{\Pi}_m'\widehat{\boldsymbol{\beta}}_m. \tag{2.4}$$

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, the weight vector lies in the unit simplex in $\mathbb{R}^M$: $\mathcal{H}^M = \left\{\mathbf{w} \in [0,1]^M : \sum_{m=1}^M w_m = 1\right\}$. The sum of the weight vector is required to be one. Otherwise, the averaging estimator of $\boldsymbol{\beta}$ is not consistent. The one-step-ahead combination forecast is

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

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

# 3 MSE and MSFE

The previous section defines the one-step-ahead combination forecast with fixed weights. Our goal is to select weights to minimize the one-step-ahead mean squared forecast error (MSFE) over the set of all possible forecast combinations. In this section, we show that the one-step-ahead MSFE is approximately the in-sample mean squared error (MSE) plus a constant term when the observations are strictly stationary.[4] As a result, the weight vector that minimizes the in-sample MSE is equivalent to the weight vector that minimizes the one-step-ahead MSFE.

We first write the conditional mean in (2.1) as $\mu_{t-1}$ so that the equation is $y_t = \mu_{t-1} + e_t$. Similarly for any weight vector, we write $\bar{\mu}_{t-1}(\mathbf{w}) = \mathbf{x}'_{t-1}\bar{\boldsymbol{\beta}}(\mathbf{w})$. We consider the quadratic loss function and define the in-sample mean squared error (risk) as

$$\begin{aligned} MSE(\mathbf{w}) &= \mathrm{E}\left(\frac{1}{T}\sum_{t=1}^{T}\left(\mu_{t-1} - \bar{\mu}_{t-1}(\mathbf{w})\right)^2\right) \\ &= \mathrm{E}\left(\left(\bar{\boldsymbol{\beta}}(\mathbf{w}) - \boldsymbol{\beta}\right)'\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_{t-1}\mathbf{x}'_{t-1}\right)\left(\bar{\boldsymbol{\beta}}(\mathbf{w}) - \boldsymbol{\beta}\right)\right). \end{aligned} \tag{3.1}$$

The in-sample MSE measures the global fit of the averaging estimator since it is constructed using the entire sample.

For any weight vector, the one-step-ahead mean squared forecast error is $MSFE(\mathbf{w}) = \mathrm{E}\left(y_{T+1} - \bar{y}_{T+1|T}(\mathbf{w})\right)^2$. Let $\sigma^2 = \mathrm{E}(e_t^2)$. Expanding the square, we find

$$\begin{aligned} MSFE(\mathbf{w}) &= \mathrm{E}\left(e_{T+1} + \mathbf{x}'_T\left(\bar{\boldsymbol{\beta}}(\mathbf{w}) - \boldsymbol{\beta}\right)\right)^2 \\ &= \sigma^2 + \mathrm{E}\left(\left(\bar{\boldsymbol{\beta}}(\mathbf{w}) - \boldsymbol{\beta}\right)'\mathbf{x}_T\mathbf{x}'_T\left(\bar{\boldsymbol{\beta}}(\mathbf{w}) - \boldsymbol{\beta}\right)\right) \\ &\simeq \sigma^2 + \mathrm{E}\left(\left(\bar{\boldsymbol{\beta}}(\mathbf{w}) - \boldsymbol{\beta}\right)'\mathbf{x}_{t-1}\mathbf{x}'_{t-1}\left(\bar{\boldsymbol{\beta}}(\mathbf{w}) - \boldsymbol{\beta}\right)\right) \\ &= \sigma^2 + MSE(\mathbf{w}). \end{aligned} \tag{3.2}$$

Note that $\mathbf{x}_{t-1}$ and $\bar{\boldsymbol{\beta}}(\mathbf{w})$ are independent in large samples when $(\mathbf{x}_{t-1}, e_t)$ are strictly

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

stationary and ergodic. A similar argument can apply to the independence of $\mathbf{x}_T$ and $\bar{\boldsymbol{\beta}}(\mathbf{w})$. As a result, the approximation in the third line is valid.

Let the optimal weight vector be the value that minimizes $MSFE(\mathbf{w})$ over $\mathbf{w} \in \mathcal{H}^M$. Since $\sigma^2$ is a constant and not related to the weight vector $\mathbf{w}$, we have

$$\underset{\mathbf{w}\in\mathcal{H}^M}{\operatorname{argmin}} MSE(\mathbf{w}) = \underset{\mathbf{w}\in\mathcal{H}^M}{\operatorname{argmin}} MSFE(\mathbf{w}). \tag{3.3}$$

Equation (3.3) means the optimal weight vector that minimizes the MSE also minimizes the MSFE.

One straightforward way to compute the MSE defined in (3.1) is to use the limiting distribution of $\left(\bar{\boldsymbol{\beta}}(\mathbf{w}) - \boldsymbol{\beta}\right)$ to approximate the MSE. In order to obtain a good approximation to the finite sample behavior, we study the MSE in a local asymptotic framework, which we will describe in the following section.

Another method to approximate the MSE is to use the information from the sum of squared errors, which is the idea behind the Mallows criterion. Let $\widehat{\mathbf{e}}(\mathbf{w}) = \mathbf{y} - \mathbf{X}\bar{\boldsymbol{\beta}}(\mathbf{w})$ be the averaging residual vector. Define $\mathbf{P}(\mathbf{w}) = \sum_{m=1}^{M} w_m \mathbf{X}_m (\mathbf{X}_m'\mathbf{X}_m)^{-1}\mathbf{X}_m'$. Expanding the sum of squared errors we have

$$\widehat{\mathbf{e}}(\mathbf{w})'\widehat{\mathbf{e}}(\mathbf{w}) = \left(\bar{\boldsymbol{\beta}}(\mathbf{w}) - \boldsymbol{\beta}\right)' \mathbf{X}'\mathbf{X} \left(\bar{\boldsymbol{\beta}}(\mathbf{w}) - \boldsymbol{\beta}\right) + \mathbf{e}'\mathbf{e} + 2\mathbf{e}'(\mathbf{I} - \mathbf{P}(\mathbf{w}))\mathbf{X}\boldsymbol{\beta} - 2\mathbf{e}'\mathbf{P}(\mathbf{w})\mathbf{e}. \tag{3.4}$$

Under the homoskedasticity assumption, we take expectation on both sides and obtain

$$\mathrm{E}(\widehat{\mathbf{e}}(\mathbf{w})'\widehat{\mathbf{e}}(\mathbf{w})) = MSE(\mathbf{w}) + T\sigma^2 - 2\mathrm{E}(\mathbf{e}'\mathbf{P}(\mathbf{w})\mathbf{e}). \tag{3.5}$$

The Mallows model averaging (MMA) criterion proposed by Hansen (2007) is

$$C_T(\mathbf{w}) = \widehat{\mathbf{e}}(\mathbf{w})'\widehat{\mathbf{e}}(\mathbf{w}) + 2\sigma^2\mathbf{k}'\mathbf{w}, \tag{3.6}$$

where $\mathbf{k} = (k_1, ..., k_M)'$ and $2\sigma^2\mathbf{k}'\mathbf{w}$ is an estimate of the final term in (3.5). The second term of (3.6) serves as a penalty term of the criterion function since $\mathbf{k}'\mathbf{w}$ measures the effective number of parameters. Therefore, we can interpret the MMA criterion as a measure of

model fit and model complexity. Hansen (2008) shows that the MMA criterion is an unbiased estimate of the in-sample mean squared error plus a constant for stationary dependent observations. Our approach uses the asymptotic mean squared error (AMSE) to approximate the MSE, which is different from the MMA estimator.

# 4 Weight Selection

This section characterizes the optimal weights of forecast combinations and presents a plug-in method to estimate the infeasible optimal weights.

## 4.1 Optimal Weights

We first investigate the in-sample MSE of the averaging estimator. In finite samples, the least squares estimator for all models except the model including all predictors has omitted variable bias. For nonzero and fixed values of $\beta$, the risk of these models tends to infinity with the sample size, and hence the asymptotic approximations break down. We therefore follow Hjort and Claeskens (2003) and Claeskens and Hjort (2003), and use a local-to-zero asymptotic framework similar to weak instrument theory to approximate the in-sample MSE. More precisely, the parameters $\boldsymbol{\beta}$ are modeled as being in a local $T^{-1/2}$ neighborhood of zero. We first establish the asymptotic distribution of the averaging estimator with fixed weights. Define $\mathbf{Q} = \mathrm{E}\left(\mathbf{x}_{t-1}\mathbf{x}_{t-1}'\right)$ and $\boldsymbol{\Omega} = \lim_{T\to\infty} T^{-1}\sum_{s=1}^{T}\sum_{t=1}^{T}\mathrm{E}\left(\mathbf{x}_{s-1}\mathbf{x}_{t-1}'e_s e_t\right)$.[5]

**Assumption 1.** $\boldsymbol{\beta} = \boldsymbol{\beta}_T = \boldsymbol{\delta}/\sqrt{T}$, where $\boldsymbol{\delta}$ ia a fixed vector.

**Assumption 2.** As $T \to \infty$, $T^{-1}\mathbf{X}'\mathbf{X} \xrightarrow{p} \mathbf{Q}$ and $T^{-1/2}\mathbf{X}'\mathbf{e} \xrightarrow{d} \mathbf{Z} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\Omega})$.

Assumption 1 assumes that $\boldsymbol{\beta}$ is local to zero. This assumption ensures that the asymptotic mean squared error of the averaging estimator remains finite. It is a common technique to analyze the finite sample properties of the model selection and averaging estimator, for example, Leeb and Pötscher (2005), Pötscher (2006), Elliott, Gargano, and Timmermann

[5]If the error term $e_t$ is serially uncorrelated and identically distributed, $\boldsymbol{\Omega}$ can be simplified as $\boldsymbol{\Omega} = \mathrm{E}\left(\mathbf{x}_{t-1}\mathbf{x}_{t-1}'e_t^2\right)$, and if the error term is i.i.d. and homoskedastic, then $\boldsymbol{\Omega} = \sigma^2\mathbf{Q}$.

(2013), and Hansen (2013). This assumption implies that as the sample size increases, all of the candidate models are close to each other. Under this framework, it is informative to know if we can improve by forecast combinations instead of relying on one single prediction model. Also 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.

Assumption 2 is a high-level condition that permits the application of cross-section, panel, and time series data. This condition holds under appropriate primitive assumptions. For example, if $y_t$ is a stationary and ergodic martingale difference sequence with finite fourth moments, then the condition follows from the weak law of large numbers and the central limit theorem for martingale difference sequences. Since the selection matrix $\mathbf{\Pi}_m$ is non-random with elements either 0 or 1, for the $m$th model we have $T^{-1}\mathbf{X}_m'\mathbf{X}_m \xrightarrow{p} \mathbf{Q}_m$ where $\mathbf{Q}_m = \mathbf{\Pi}_m \mathbf{Q} \mathbf{\Pi}_m'$ is nonsingular. Let $\mathbf{I}_k$ be a $k \times k$ identity matrix.

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

$$\begin{aligned}
&\sqrt{T}\left(\bar{\boldsymbol{\beta}}(\mathbf{w}) - \boldsymbol{\beta}\right) \xrightarrow{d} \mathbf{N}\left(\mathbf{A}(\mathbf{w})\boldsymbol{\delta}, \mathbf{V}(\mathbf{w})\right) \\
&\mathbf{A}(\mathbf{w}) = \sum_{m=1}^{M} w_m \mathbf{A}_m \\
&\mathbf{V}(\mathbf{w}) = \sum_{m=1}^{M} w_m^2 \mathbf{B}_m \mathbf{\Omega} \mathbf{B}_m + 2 \sum \sum_{m \neq \ell} w_m w_\ell \mathbf{B}_m \mathbf{\Omega} \mathbf{B}_\ell
\end{aligned}$$

*where* $\mathbf{A}_m = \mathbf{\Pi}_m' \mathbf{Q}_m^{-1} \mathbf{\Pi}_m \mathbf{Q} - \mathbf{I}_k$ *and* $\mathbf{B}_m = \mathbf{\Pi}_m' \mathbf{Q}_m^{-1} \mathbf{\Pi}_m$.

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 $\mathbf{Q}^{-1}\mathbf{\Omega}\mathbf{Q}^{-1}$. Note that $\mathbf{A}(\mathbf{w})\boldsymbol{\delta}$ represents the asymptotic bias term of the averaging estimator $\bar{\boldsymbol{\beta}}(\mathbf{w})$. 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.

Theorem 1 shows the asymptotic normality of the averaging estimator with non-random weights. We use this result to compute the in-sample mean squared error of the averaging estimator in the next theorem. The distribution result is also useful for inference.

**Theorem 2.** *Suppose Assumptions 1-2 hold. We have*

$$MSE(\mathbf{w}) = \frac{1}{T}\mathbf{w}'\mathbf{C}_M\mathbf{w} + O(T^{-3/2})$$

*where* $\mathbf{C}_M$ *is an* $M \times M$ *matrix with the* $(m,\ell)$*th element*

$$c_{m,\ell} = \operatorname{tr}(\mathbf{Q}\mathbf{A}_m\boldsymbol{\delta}\boldsymbol{\delta}'\mathbf{A}_\ell') + \operatorname{tr}(\mathbf{Q}\mathbf{B}_m\boldsymbol{\Omega}\mathbf{B}_\ell)$$

*where* $\mathbf{A}_m$ *and* $\mathbf{B}_m$ *are defined in Theorem 1.*

Theorem 2 presents the risk of the averaging estimator in the local asymptotic framework. The $m$th diagonal element of $\mathbf{C}_M$ characterizes the bias and variance of the $m$th model while the off-diagonal elements measure the product of biases and covariance between different models. Let $\mathbf{w}_m^0$ be an $M \times 1$ vector in which the $m$th element is one and the others are zeros. Then we obtain the risk of the $m$th model, i.e., $MSE(\mathbf{w}_m^0) = T^{-1}\operatorname{tr}(\mathbf{Q}\mathbf{A}_m\boldsymbol{\delta}\boldsymbol{\delta}'\mathbf{A}_m') + T^{-1}\operatorname{tr}(\mathbf{Q}\mathbf{B}_m\boldsymbol{\Omega}\mathbf{B}_m) + O(T^{-3/2})$. Claeskens and Hjort (2008) propose to use the estimate of $MSE(\mathbf{w}_m^0)$ for model selection. Here we generalize their results from model selection to model averaging.[6]

Theorem 2 is also 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.

Following Theorem 2, we define the optimal weight vector as the value that minimizes the leading term of $MSE(\mathbf{w})$ over $\mathbf{w} \in \mathcal{H}^M$:

$$\mathbf{w}^o = \underset{\mathbf{w}\in\mathcal{H}^M}{\operatorname{argmin}}\, \mathbf{w}'\mathbf{C}_M\mathbf{w}. \tag{4.1}$$

[6]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.

Combining Theorem 2 with (3.3), we deduce that $\mathbf{w}^o$ is also 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.

## 4.2 Plug-In Weights

The optimal weights, however, are infeasible, since they depend on the unknown parameters, $\boldsymbol{\delta}$, $\mathbf{Q}$, $\boldsymbol{\Omega}$, $\mathbf{A}_m$, and $\mathbf{B}_m$. Similar to Liu (2013), we propose a plug-in estimator to estimate the optimal weights for the forecasting model. We estimate the leading term of the $MSE(\mathbf{w})$ given in Theorem 2 by plugging in an asymptotically unbiased estimator and choose the data-driven weights by minimizing the sample analog of the MSE. We then use these estimated weights to construct the one-step-ahead forecast combination.

We first consider the estimate of the second term of $c_{m,\ell}$. It is not problematic since the unknown parameters $\mathbf{Q}$, $\boldsymbol{\Omega}$, and $\mathbf{B}_m$ can be consistently estimated by the sample analogue. Let $\widehat{\boldsymbol{\beta}}_f = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ and $\widehat{e}_t = y_t - \mathbf{x}_{t-1}'\widehat{\boldsymbol{\beta}}_f$ be the least squares estimator and the residuals for the full model. The covariance matrix $\mathbf{Q}$ and $\boldsymbol{\Omega}$ can be consistently estimated by the method of moments estimator and the heteroskedasticity and autocorrelation consistent covariance matrix estimator, i.e.,

$$\widehat{\mathbf{Q}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_{t-1}\mathbf{x}_{t-1}' \text{ and } \widehat{\boldsymbol{\Omega}} = \sum_{j=-T}^{T} K\left(\frac{j}{S_T}\right)\widehat{\boldsymbol{\Gamma}}(j),$$

where $K(\cdot)$ is a kernel function, $S_T$ is the bandwidth, $\widehat{\boldsymbol{\Gamma}}(j) = T^{-1}\sum_{t=1}^{T-j}\mathbf{x}_{t-1}\mathbf{x}_{t-1+j}'\widehat{e}_t\widehat{e}_{t+j}$ for $j \geq 0$, and $\widehat{\boldsymbol{\Gamma}}(j) = \widehat{\boldsymbol{\Gamma}}(-j)'$ for $j < 0$.[7] Note that both $\widehat{\mathbf{A}}_m$ and $\widehat{\mathbf{B}}_m$ are functions of $\mathbf{Q}$ and selection matrix $\boldsymbol{\Pi}_m$, which can also be consistently estimated under Assumption 2. Therefore, we have $\operatorname{tr}\big(\widehat{\mathbf{B}}_m\widehat{\mathbf{Q}}\widehat{\mathbf{B}}_\ell\widehat{\boldsymbol{\Omega}}\big) \xrightarrow{p} \operatorname{tr}\left(\mathbf{Q}\mathbf{B}_m\boldsymbol{\Omega}\mathbf{B}_\ell\right)$.

We next consider the estimate of the first term of $c_{m,\ell}$. 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

[7] Note that $\widehat{\mathbf{Q}} \xrightarrow{p} \mathbf{Q}$ under Assumption 2. Also, $\widehat{\boldsymbol{\Omega}} \xrightarrow{p} \boldsymbol{\Omega}$ under some regularity conditions, 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 $\widehat{\boldsymbol{\Omega}} = T^{-1}\sum_{t=1}^{T}\mathbf{x}_{t-1}\mathbf{x}_{t-1}'\widehat{e}_t^2$, the heteroskedasticity-consistent covariance matrix estimator proposed by White (1980).

the estimator from the full model. That is, $\widehat{\boldsymbol{\delta}} = \sqrt{T}\widehat{\boldsymbol{\beta}}_f$. In the proof of Theorem 1, we show that

$$\widehat{\boldsymbol{\delta}} = \sqrt{T}\widehat{\boldsymbol{\beta}}_f \xrightarrow{d} \mathbf{Z}_{\boldsymbol{\delta}} = \boldsymbol{\delta} + \mathbf{Q}^{-1}\mathbf{Z} \sim N(\boldsymbol{\delta}, \mathbf{Q}^{-1}\boldsymbol{\Omega}\mathbf{Q}^{-1}).$$

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

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

Thus, the weight vector of the plug-in estimator is defined as

$$\widehat{\mathbf{w}} = \underset{\mathbf{w}\in\mathcal{H}^M}{\operatorname{argmin}}\, \mathbf{w}'\widehat{\mathbf{C}}_M\mathbf{w}, \tag{4.2}$$

where $\widehat{\mathbf{C}}_M$ is a sample analog of $\mathbf{C}_M$ with the $(m,\ell)$th element

$$\widehat{c}_{m,\ell} = \operatorname{tr}\left(\widehat{\mathbf{Q}}\widehat{\mathbf{A}}_m\widehat{\boldsymbol{\delta}\boldsymbol{\delta}'}\widehat{\mathbf{A}}_\ell'\right) + \operatorname{tr}\left(\widehat{\mathbf{B}}_m\widehat{\mathbf{Q}}\widehat{\mathbf{B}}_\ell\widehat{\boldsymbol{\Omega}}\right), \tag{4.3}$$

and $T^{-1}\mathbf{w}'\widehat{\mathbf{C}}_M\mathbf{w}$ is an asymptotically unbiased estimator of $MSE(\mathbf{w})$.[8] The plug-in one-step-ahead combination forecast is

$$\bar{y}_{T+1|T}(\widehat{\mathbf{w}}) = \mathbf{x}_T'\bar{\boldsymbol{\beta}}(\widehat{\mathbf{w}}). \tag{4.4}$$

As mentioned by Hjort and Claeskens (2003), we can also estimate $MSE(\mathbf{w})$ by inserting $\hat{\boldsymbol{\delta}}$ for $\boldsymbol{\delta}$. The alternative estimator of $c_{m,\ell}$ is

$$\tilde{c}_{m,\ell} = \operatorname{tr}\left(\widehat{\mathbf{Q}}\widehat{\mathbf{A}}_m\widehat{\boldsymbol{\delta}}\widehat{\boldsymbol{\delta}}'\widehat{\mathbf{A}}_\ell'\right) + \operatorname{tr}\left(\widehat{\mathbf{B}}_m\widehat{\mathbf{Q}}\widehat{\mathbf{B}}_\ell\widehat{\boldsymbol{\Omega}}\right). \tag{4.5}$$

Although $\tilde{c}_{m,\ell}$ is not an asymptotically unbiased estimator, the simulation shows that the

[8] Claeskens and Hjort (2008) suggest estimating the first term of $c_{m,\ell}$ by $\max\left\{0, \operatorname{tr}\left(\widehat{\mathbf{Q}}\widehat{\mathbf{A}}_m\widehat{\boldsymbol{\delta}\boldsymbol{\delta}'}\widehat{\mathbf{A}}_\ell'\right)\right\}$ to avoid the negative estimate for the squared bias term. However, our simulations show that this modified estimator has less performance than the estimator (4.3). Therefore, we focus on the estimator (4.3) in this paper.

estimator (4.5) has better finite sample performance than the estimator (4.3) in most ranges of the parameter space.

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

# 5 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; that is, the candidate models are nonnested. The second design is a moving average model with exogenous inputs, and we consider a sequence of nested candidate models.

## 5.1 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 (4.3) (labeled PIA(1)), and the plug-in averaging estimator based on (4.5) (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 $\widehat{w}_m = \exp(-\frac{1}{2}\mathrm{AIC}_m)/\sum_{j=1}^{M}\exp(-\frac{1}{2}\mathrm{AIC}_j)$ where $\mathrm{AIC}_m = T\log(\widehat{\sigma}_m^2) + 2k_m$, $\widehat{\sigma}_m^2 = T^{-1}\sum_{t=1}^{T}\widehat{e}_{m,t}^2$, 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 $\widehat{w}_m = \exp(-\frac{1}{2}\mathrm{BIC}_m)/\sum_{j=1}^{M}\exp(-\frac{1}{2}\mathrm{BIC}_j)$ where

$\text{BIC}_m = T\log(\widehat{\sigma}_m^2) + \log(T)k_m$.

The Mallows model averaging estimator is proposed by Hansen (2007), and the weight selection criterion is defined in (3.6). One restriction of the MMA estimator is that it is limited to the homoskedastic model. The homoskedasticity restriction is relaxed by the jackknife model averaging estimator proposed by Hansen and Racine (2012). The weights of the JMA estimator are chosen by minimizing a leave-one-out cross-validation criterion

$$\text{CV}_T(\mathbf{w}) = \frac{1}{T}\mathbf{w}'\tilde{\mathbf{e}}'\tilde{\mathbf{e}}\mathbf{w}, \tag{5.1}$$

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 $t$th 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.

The one-step-ahead combination forecast based on the above four estimators and the plug-in forecast combination is $\sum_{m=1}^{M} \widehat{w}_m \widehat{y}_{T+1|T}(m)$ 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 $\kappa$ be 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 $\kappa$ predictors. Let $n_{\kappa,k} = k!/((k-\kappa)!\kappa!)$ be the number of models considered based on $\kappa$ subset regressions. The one-step-ahead combination forecast based on complete subset regression method is

$$\bar{y}_{T+1|T}(\kappa) = \frac{1}{n_{\kappa,k}} \sum_{m=1}^{n_{\kappa,k}} \mathbf{x}_T' \mathbf{\Pi}_m' \widehat{\boldsymbol{\beta}}_m \quad \text{s.t.} \quad \text{tr}(\mathbf{\Pi}_m' \mathbf{\Pi}_m) = \kappa. \tag{5.2}$$

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

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 $\bar{y}_{T+1|T}(\widehat{\mathbf{w}})$ 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:

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

where $S$ is the number of simulations. We set $S = 5000$ for all experiments. The lower relative risk means better performance on predictions.

## 5.2 Linear Regression Models

The data generation process for the first design is

$$y_t = \sum_{j=1}^{k}\beta_j x_{jt-1} + e_t, \tag{5.3}$$

$$x_{jt} = \rho_x x_{jt-1} + u_{jt}, \text{ for } j \geq 2. \tag{5.4}$$

We set $x_{1t} = 1$ to be the intercept and remaining $x_{jt}$ are AR(1) processes with $\rho_x = 0.5$ and 0.9. The predictors $x_{jt}$ are correlated. 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 $\rho_u$. 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 control 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 $\mathbf{Q}_x = (1-\rho_x^2)^{-1}\mathbf{Q}_u$. The population $R^2$ varies on a grid between 0 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 $M = 32$.

Figures 1–8 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 for linear regression models, 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 population $R^2$. 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. One interesting observation is that all estimators have decreasing relative risk as $R^2$ increases or $\rho_u$ increases.

Figures 3 and 4 display the relative risk for the large AR(1) coefficient. The relative performance of six estimators depends strongly on $R^2$ and $\rho_u$. 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.

Figures 5 and 6 show the relative risk when $R^2$ varies between 0 and 0.1. It is clear that PIA(1) achieves lower relative risk than PIA(2) when $R^2$ is small in both homoskedastic and heteroskedastic simulations. For the homoskedastic simulation, after passing the transition point where S-BIC and PIA(2) have equal risk, the relative risk of S-BIC is decreasing while the relative risk of other estimators is increasing sharply. The transition point is getting close to zero as $\rho_u$ decreases. Similar results for the AIC and BIC model selection estimators are also found in Yang (2007) and Ng (2013). However, the advantage of S-BIC for small $R^2$ value does not exist in the heteroskedastic simulation.

Figures 7 and 8 compare the relative risk between the plug-in averaging estimator and the

complete subset regressions. The performance of the subset regression approach is sensitive to the choice of $\kappa$, the number of the predictors included in the model. As $R^2$ increases, the optimal value of $\kappa$ 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 $\kappa$.

## 5.3 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

$$y_t = x_t + 0.5x_{t-1} + e_t + \beta e_{t-1}, \tag{5.5}$$

$$x_t = 0.5x_{t-1} + u_t. \tag{5.6}$$

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 $\sigma_t^2 = 0.5$ for the homoskedastic simulation and $\sigma_t^2 = 1+x_t^2$ for the heteroskedastic simulation. The parameter $\beta$ 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

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

The number of models is $M = 7$. 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.

Figures 9–12 show the results for the second simulation setup. In each figure, the relative risk and model size is displayed for $T = 100$, 200, 500, and 1000 for MAX$(1, 1)$ models,

respectively.

Figures 9 and 10 display the relative risk when the moving average coefficient $\beta$ varies between $-0.5$ and 0.5. We analyze the behavior of the estimators in two regions, small $|\beta|$ and large $|\beta|$. The S-BIC estimator has the lowest relative risk when $|\beta|$ is small, and PIA(2) performs better than other estimators when $|\beta|$ is large. However, when the sample size is small, S-BIC has poor performance in both regions. These findings from MAX$(1,1)$ models are consistent with those of regression models with small $R^2$ values in Figures 5 and 6.

Figures 11 and 12 compare the model size of six estimators. The model size is defined as the average number of predictors selected by each combination method across 5000 simulation draws. As we expected, the model size of S-BIC is smaller than those of other estimators. S-AIC and PIA(2) have similar model sizes, and they tend to select the larger models compared to MMA, JMA, and PIA(1). An interesting observation is that all estimators have smaller model sizes when $\beta$ is large, but the model size is not monotone in $\beta$.

# 6 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. In a recent article, Welch and Goyal (2008) argue that numerous economic variables have poor out-of-sample predictions and these forecasting models are unstable to consistently provide forecasting gain relative to the historical average. 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 U.S. stock market.

## 6.1 Data

We estimate the following predictive regression $r_t = \alpha + \mathbf{x}_{t-1}'\boldsymbol{\beta} + e_t$ where $r_t$ is the equity premium, $\mathbf{x}_{t-1}$ are the economic variables, and $e_t$ 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.[9] 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.[10] The 10 economic variables are as follows: dividend price ratio, dividend yield, earnings price ratio, book-to-market ratio, net equity expansion, Treasure 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. For example, the out-of-sample forecast for 1965:2 is generated by the sample from 1947:1–1965:1.

## 6.2 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 forecast based on the individual economics variable consistently outperforms the forecast based on the historical average.

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

[10]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 13 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 13 shows that MMA, JMA, PIA(1), and PIA(2) consistently beat the historical average in terms of MSFE, while S-AIC and S-BIC have worse performance than the historical average after 1997. It is clear to see that both PIA(2) and MMA have smaller cumulative squared prediction error than other estimators. 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.[11] Therefore, our results support the findings of Rapach, Strauss, and Zhou (2010) and Elliott, Gargano, and Timmermann (2013) that forecast combinations provide significant gains on equity premium predictions relative to the historical average.

The two lower panels of Figure 13 compare the cumulative squared prediction error of PIA(2) to those of the complete subset regressions. As we can see from the results, complete subset regressions that use $\kappa = 4$ or 5 predictors produce the lowest cumulative squared prediction error, whihc is similar to that of PIA(2). However, the choice of $\kappa$ has a great influence on the performance of the complete subset regressions, and in practice the optimal choice of $\kappa$ is unknown. Examining these three panels in Figure 13, there is no one forecast combination method that uniformly dominates the others.

# 7 Conclusion

This paper studies the weight selection for forecast combination in a predictive regression when the goal is minimizing the MSFE. In contrast to the existing literature, we derive the MSFE in a local asymptotic framework without the i.i.d. normal assumption. We show that

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

the optimal model weights that minimize the MSFE depend on the local parameters and the covariance matrix of the predictive regression. We then propose a frequentist model averaging criterion, an asymptotically unbiased estimator of MSFE, to select forecast weights. Simulations show that the proposed estimator achieves much lower MSFE than other existing model averaging methods.

# Appendix

## A Proofs

**Proof of Theorem 1:** Note that $\bar{\boldsymbol{\beta}}(\mathbf{w}) = \sum_{m=1}^{M} w_m \boldsymbol{\Pi}_m' \widehat{\boldsymbol{\beta}}_m$. In order to derive the asymptotic distribution of the averaging estimator $\bar{\boldsymbol{\beta}}(\mathbf{w})$, we first derive the asymptotic distribution of $\sqrt{T}\left(\boldsymbol{\Pi}_m' \widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}\right)$ and then show that there is joint convergence in distribution of all $\widehat{\boldsymbol{\beta}}_m$.

Let $\widehat{\boldsymbol{\beta}}_f$ be the least squares estimator of $\boldsymbol{\beta}$ for the full model, i.e., $\widehat{\boldsymbol{\beta}}_f = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. By Assumptions 1 and 2 and the application of the continuous mapping theorem, it follows that

$$\sqrt{T}\widehat{\boldsymbol{\beta}}_f = \sqrt{T}\boldsymbol{\beta} + \sqrt{T}\left(\widehat{\boldsymbol{\beta}}_f - \boldsymbol{\beta}\right) = \boldsymbol{\delta} + \left(\frac{1}{T}\mathbf{X}'\mathbf{X}\right)^{-1}\left(\frac{1}{\sqrt{T}}\mathbf{X}'\mathbf{e}\right) \xrightarrow{d} \boldsymbol{\delta} + \mathbf{Q}^{-1}\mathbf{Z}. \tag{A.1}$$

Note that

$$\widehat{\boldsymbol{\beta}}_m = (\mathbf{X}_m'\mathbf{X}_m)^{-1}\mathbf{X}_m'\mathbf{y} = (\mathbf{X}_m'\mathbf{X}_m)^{-1}\boldsymbol{\Pi}_m\mathbf{X}'\mathbf{y} = (\mathbf{X}_m'\mathbf{X}_m)^{-1}\boldsymbol{\Pi}_m\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}}_f.$$

Therefore, we have

$$\begin{aligned}
\sqrt{T}\left(\boldsymbol{\Pi}_m'\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}\right) &= \boldsymbol{\Pi}_m'\left(\frac{1}{T}\mathbf{X}_m'\mathbf{X}_m\right)^{-1}\boldsymbol{\Pi}_m\left(\frac{1}{T}\mathbf{X}'\mathbf{X}\right)\sqrt{T}\widehat{\boldsymbol{\beta}}_f - \sqrt{T}\boldsymbol{\beta} \\
&\xrightarrow{d} \boldsymbol{\Pi}_m'\mathbf{Q}_m^{-1}\boldsymbol{\Pi}_m\mathbf{Q}\left(\boldsymbol{\delta} + \mathbf{Q}^{-1}\mathbf{Z}\right) - \boldsymbol{\delta} \\
&= \left(\boldsymbol{\Pi}_m'\mathbf{Q}_m^{-1}\boldsymbol{\Pi}_m\mathbf{Q} - \mathbf{I}_k\right)\boldsymbol{\delta} + \boldsymbol{\Pi}_m'\mathbf{Q}_m^{-1}\boldsymbol{\Pi}_m\mathbf{Z} \\
&= \mathbf{A}_m\boldsymbol{\delta} + \mathbf{B}_m\mathbf{Z} \equiv \boldsymbol{\Lambda}_m
\end{aligned} \tag{A.2}$$

where $\mathbf{A}_m = \mathbf{\Pi}_m'\mathbf{Q}_m^{-1}\mathbf{\Pi}_m\mathbf{Q} - \mathbf{I}_k$ and $\mathbf{B}_m = \mathbf{\Pi}_m'\mathbf{Q}_m^{-1}\mathbf{\Pi}_m$. Note that (A.2) implies joint convergence in distribution of all $\sqrt{T}\left(\mathbf{\Pi}_m'\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}\right)$ to $\mathbf{\Lambda}_m$, since all of $\mathbf{\Lambda}_m$ can be expressed in terms of the same normal vector $\mathbf{Z}$.

Because the weights are non-random, it follows that

$$\sqrt{T}\left(\bar{\boldsymbol{\beta}}(\mathbf{w}) - \boldsymbol{\beta}\right) = \sum_{m=1}^{M} w_m\sqrt{T}\left(\mathbf{\Pi}_m'\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}\right) \xrightarrow{d} \sum_{m=1}^{M} w_m\mathbf{\Lambda}_m \equiv \mathbf{\Lambda}.$$

By (A.2) and standard algebra, we can show the mean vector of $\mathbf{\Lambda}$ as

$$\mathrm{E}\left(\sum_{m=1}^{M} w_m\mathbf{\Lambda}_m\right) = \sum_{m=1}^{M} w_m\mathrm{E}\left(\mathbf{\Lambda}_m\right) = \sum_{m=1}^{M} w_m\left(\mathbf{\Pi}_m'\mathbf{Q}_m^{-1}\mathbf{\Pi}_m\mathbf{Q} - \mathbf{I}_k\right)\boldsymbol{\delta} = \mathbf{A}(\mathbf{w})\boldsymbol{\delta}$$

where $\mathbf{A}(\mathbf{w}) = \sum_{m=1}^{M} w_m\left(\mathbf{\Pi}_m'\mathbf{Q}_m^{-1}\mathbf{\Pi}_m\mathbf{Q} - \mathbf{I}_k\right) = \sum_{m=1}^{M} w_m\mathbf{A}_m$.

Next we want to show the covariance matrix of $\mathbf{\Lambda}$. For any two models, we have

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

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

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

This completes the proof. ■

**Proof of Theorem 2:** Note that

$$MSE(\mathbf{w}) = \mathrm{E}\left(\left(\bar{\boldsymbol{\beta}}(\mathbf{w}) - \boldsymbol{\beta}\right)' \left(\frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_{t-1}\mathbf{x}_{t-1}'\right)\left(\bar{\boldsymbol{\beta}}(\mathbf{w}) - \boldsymbol{\beta}\right)\right) = \frac{1}{T}\mathrm{E}\left(\xi_T\right)$$

where

$$\xi_T = \left(\sqrt{T}\left(\bar{\boldsymbol{\beta}}(\mathbf{w}) - \boldsymbol{\beta}\right)' \left(\frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_{t-1}\mathbf{x}_{t-1}'\right)\sqrt{T}\left(\bar{\boldsymbol{\beta}}(\mathbf{w}) - \boldsymbol{\beta}\right)\right).$$

Define $\mathbf{B}(\mathbf{w}) = \sum_{m=1}^{M} w_m \mathbf{B}_m$. We can rewrite the asymptotic distribution of the averaging estimator $\bar{\boldsymbol{\beta}}(\mathbf{w})$ as

$$\begin{aligned}\sqrt{T}\left(\bar{\boldsymbol{\beta}}(\mathbf{w}) - \boldsymbol{\beta}\right) &= \sum_{m=1}^{M} w_m \sqrt{T}\left(\boldsymbol{\Pi}_m' \widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}\right) \\ &\xrightarrow{d} \sum_{m=1}^{M} w_m \left(\mathbf{A}_m \boldsymbol{\delta} + \mathbf{B}_m \mathbf{Z}\right) = \mathbf{A}(\mathbf{w})\boldsymbol{\delta} + \mathbf{B}(\mathbf{w})\mathbf{Z}. \qquad \text{(A.3)}\end{aligned}$$

By (A.3) and the application of the continuous mapping theorem, it follows that

$$\xi_T \xrightarrow{d} \left(\mathbf{A}(\mathbf{w})\boldsymbol{\delta} + \mathbf{B}(\mathbf{w})\mathbf{Z}\right)' \mathbf{Q}\left(\mathbf{A}(\mathbf{w})\boldsymbol{\delta} + \mathbf{B}(\mathbf{w})\mathbf{Z}\right).$$

Suppose $\xi_T$ is uniformly integrable, then we have

$$\begin{aligned}\mathrm{E}\left(\xi_T\right) &\xrightarrow{d} \mathrm{E}\left(\left(\mathbf{A}(\mathbf{w})\boldsymbol{\delta} + \mathbf{B}(\mathbf{w})\mathbf{Z}\right)' \mathbf{Q}\left(\mathbf{A}(\mathbf{w})\boldsymbol{\delta} + \mathbf{B}(\mathbf{w})\mathbf{Z}\right)\right) \\ &= \mathrm{E}\left(\boldsymbol{\delta}'\mathbf{A}(\mathbf{w})'\mathbf{Q}\mathbf{A}(\mathbf{w})\boldsymbol{\delta} + \mathbf{Z}'\mathbf{B}(\mathbf{w})'\mathbf{Q}\mathbf{B}(\mathbf{w})\mathbf{Z} + 2\boldsymbol{\delta}'\mathbf{A}(\mathbf{w})'\mathbf{Q}\mathbf{B}(\mathbf{w})\mathbf{Z}\right) \\ &= \boldsymbol{\delta}'\mathbf{A}(\mathbf{w})'\mathbf{Q}\mathbf{A}(\mathbf{w})\boldsymbol{\delta} + \mathrm{E}\left(\mathbf{Z}'\mathbf{B}(\mathbf{w})'\mathbf{Q}\mathbf{B}(\mathbf{w})\mathbf{Z}\right) \\ &= \mathrm{tr}\left(\mathbf{Q}\mathbf{A}(\mathbf{w})\boldsymbol{\delta}\boldsymbol{\delta}'\mathbf{A}(\mathbf{w})'\right) + \mathrm{tr}\left(\mathbf{B}(\mathbf{w})'\mathbf{Q}\mathbf{B}(\mathbf{w})\boldsymbol{\Omega}\right) \\ &= \mathbf{w}'\mathbf{C}_M\mathbf{w}\end{aligned}$$

where $\mathbf{C}_M$ is an $M \times M$ matrix with the $(m,\ell)$th element $c_{m,\ell} = \mathrm{tr}\left(\mathbf{Q}\mathbf{A}_m\boldsymbol{\delta}\boldsymbol{\delta}'\mathbf{A}_\ell'\right) + \mathrm{tr}\left(\mathbf{B}_m\mathbf{Q}\mathbf{B}_\ell\boldsymbol{\Omega}\right)$. Therefore, we have $\mathrm{E}\left(\xi_T\right) = \mathbf{w}'\mathbf{C}_M\mathbf{w} + O(T^{-1/2})$ and $MSE(\mathbf{w}) = T^{-1}\mathbf{w}'\mathbf{C}_M\mathbf{w} + O(T^{-3/2})$. This completes the proof. ■

# B Figures

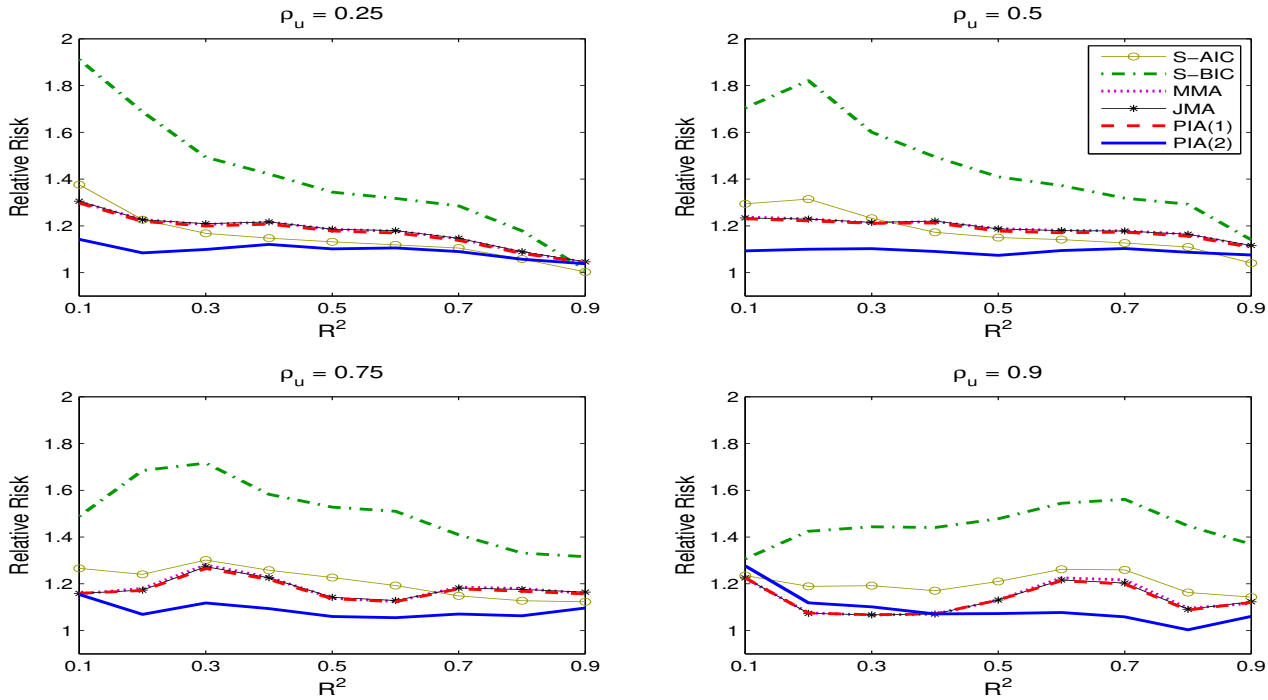


Figure 1: Relative risk for linear regression models with homoskedastic errors when $\rho_x = 0.5$ and $R^2$ varies between 0.1 and 0.9.

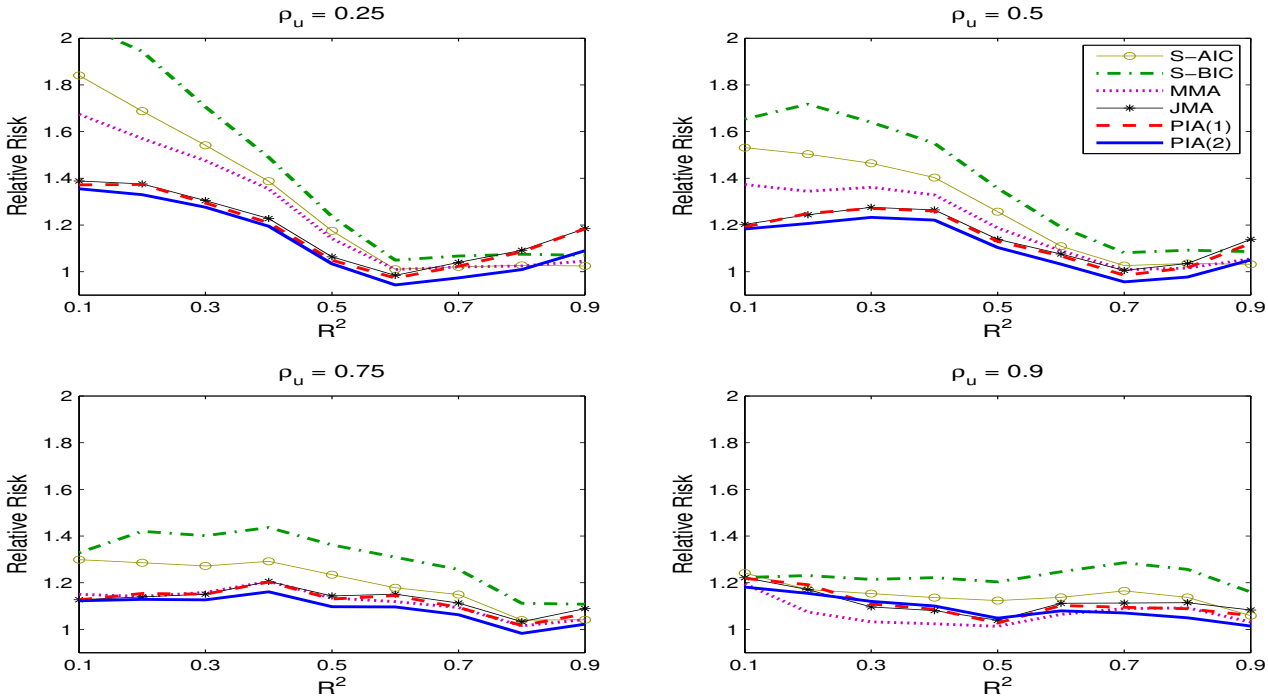


Figure 2: Relative risk for linear regression models with heteroskedastic errors when $\rho_x = 0.5$ and $R^2$ varies between 0.1 and 0.9.

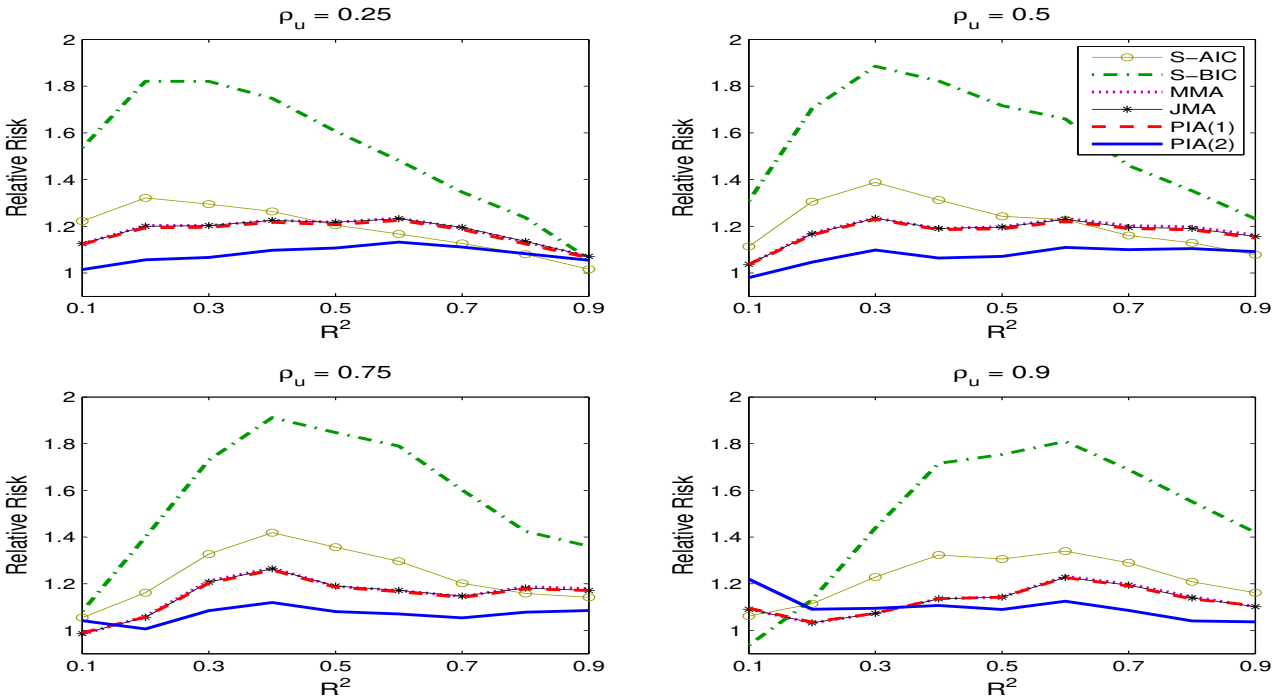


Figure 3: Relative risk for linear regression models with homoskedastic errors when $\rho_x = 0.9$ and $R^2$ varies between 0.1 and 0.9.

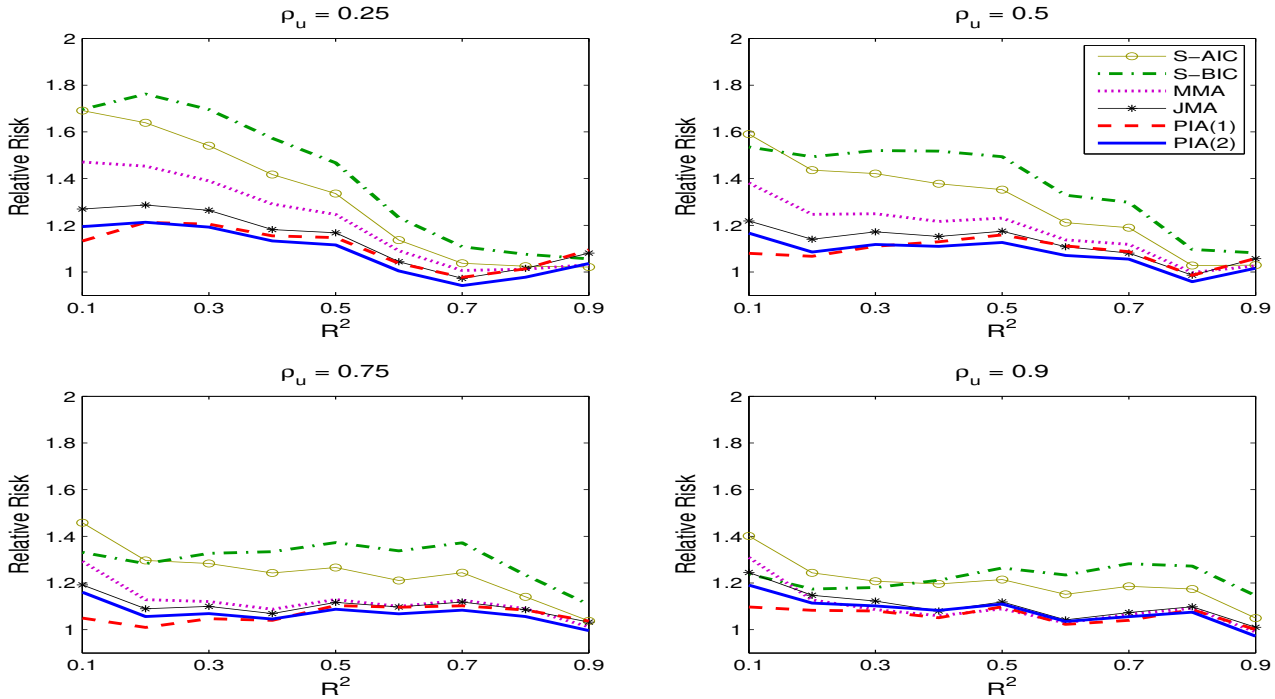


Figure 4: Relative risk for linear regression models with heteroskedastic errors when $\rho_x = 0.9$ and $R^2$ varies between 0.1 and 0.9.

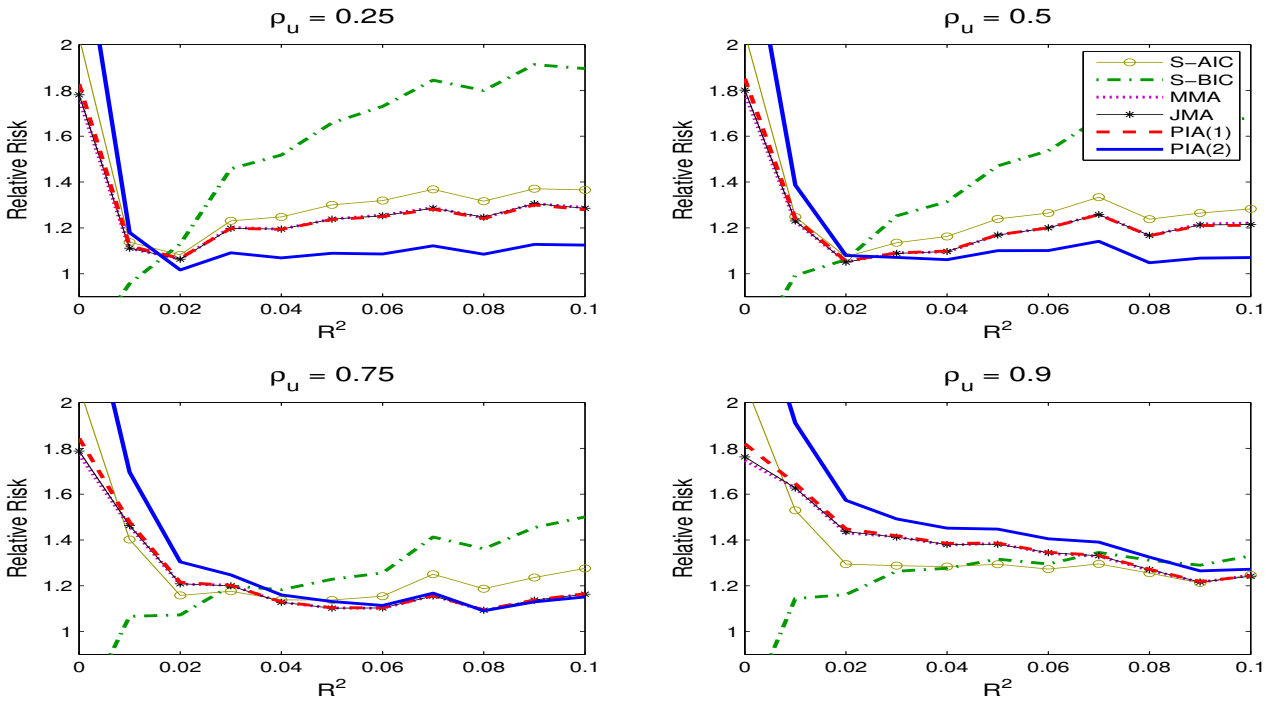


Figure 5: Relative risk for linear regression models with homoskedastic errors when $\rho_x = 0.5$ and $R^2$ varies between 0 and 0.1.

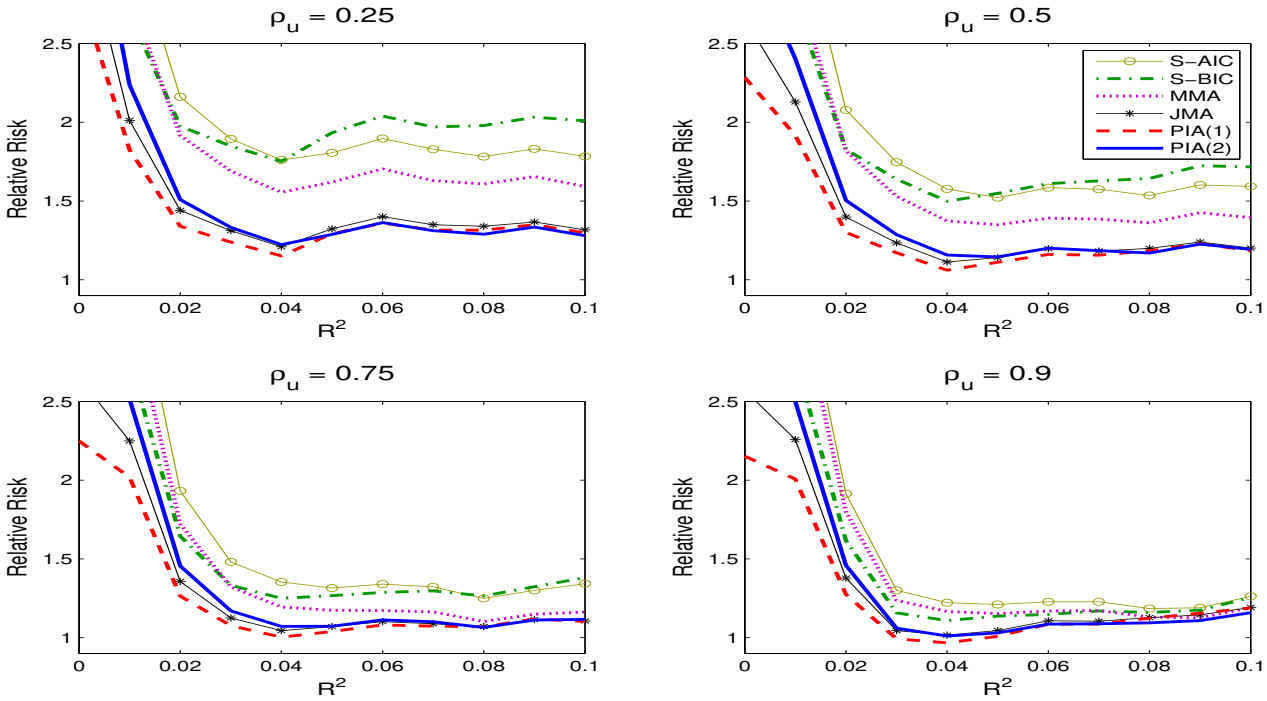


Figure 6: Relative risk for linear regression models with heteroskedastic errors when $\rho_x = 0.5$ and $R^2$ varies between 0 and 0.1.

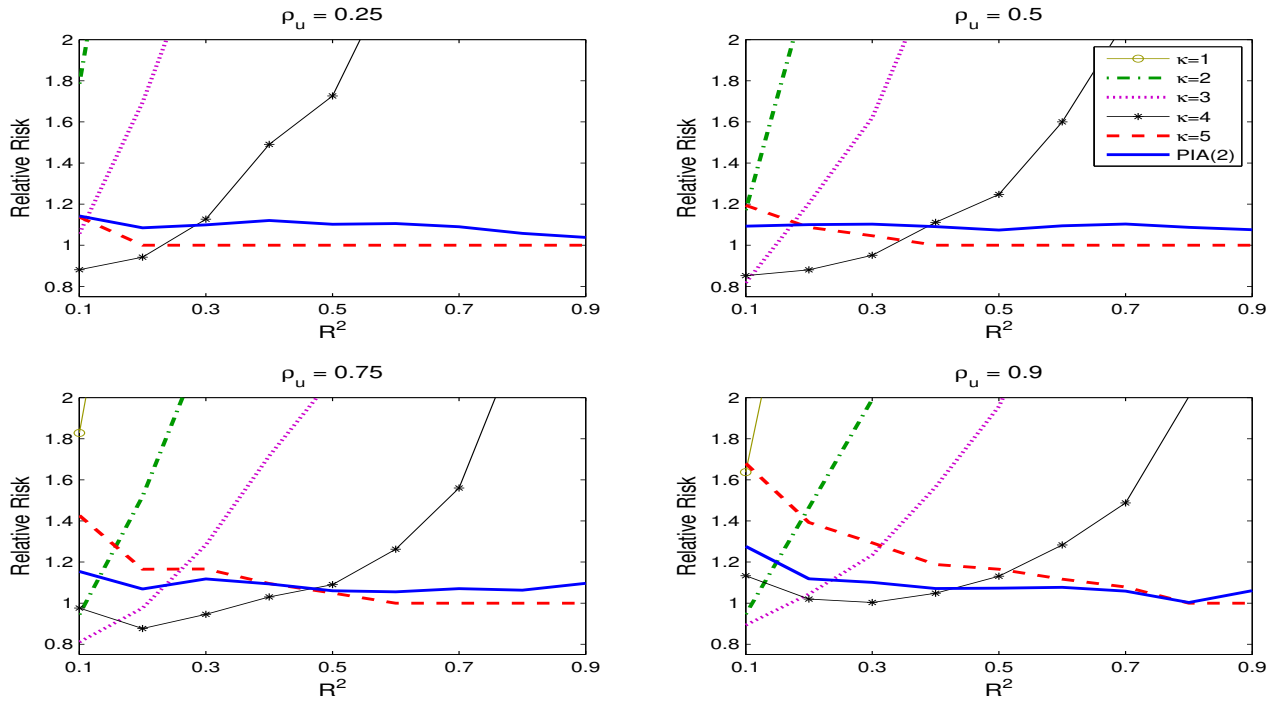


Figure 7: Relative risk for linear regression models with homoskedastic errors when $\rho_x = 0.5$ and $R^2$ varies between 0.1 and 0.9.

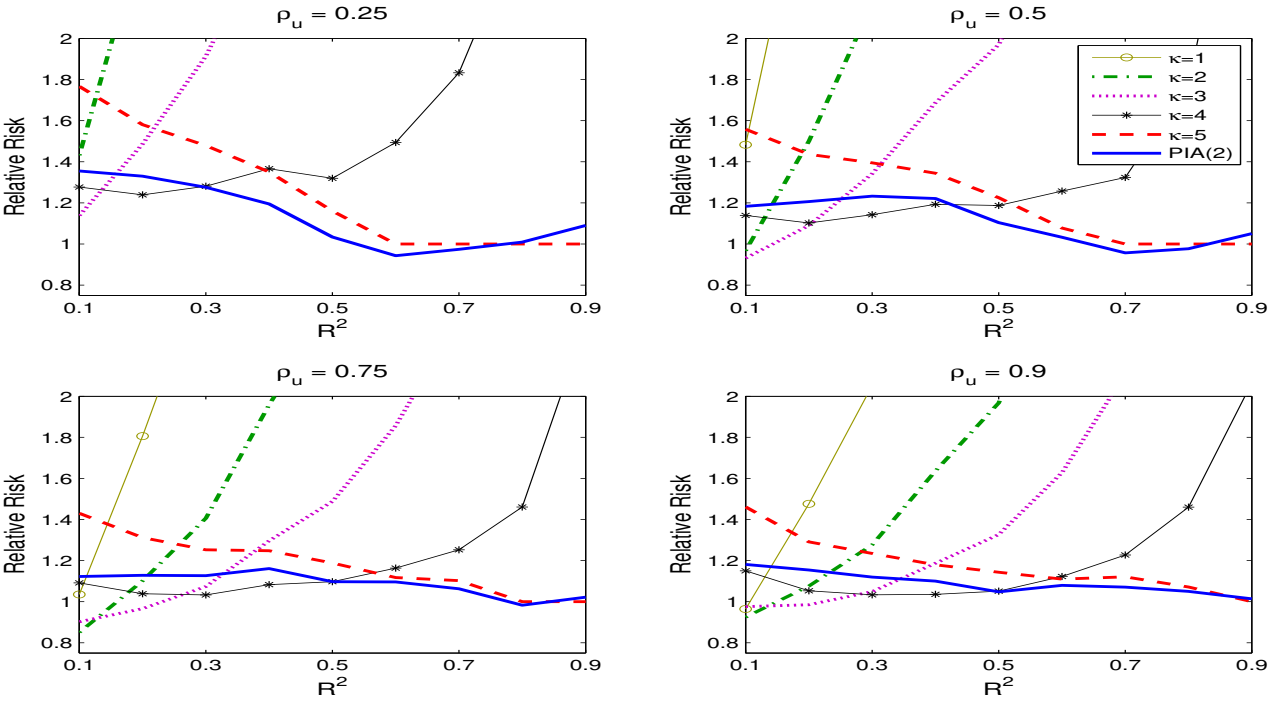


Figure 8: Relative risk for linear regression models with heteroskedastic errors when $\rho_x = 0.5$ and $R^2$ varies between 0.1 and 0.9.

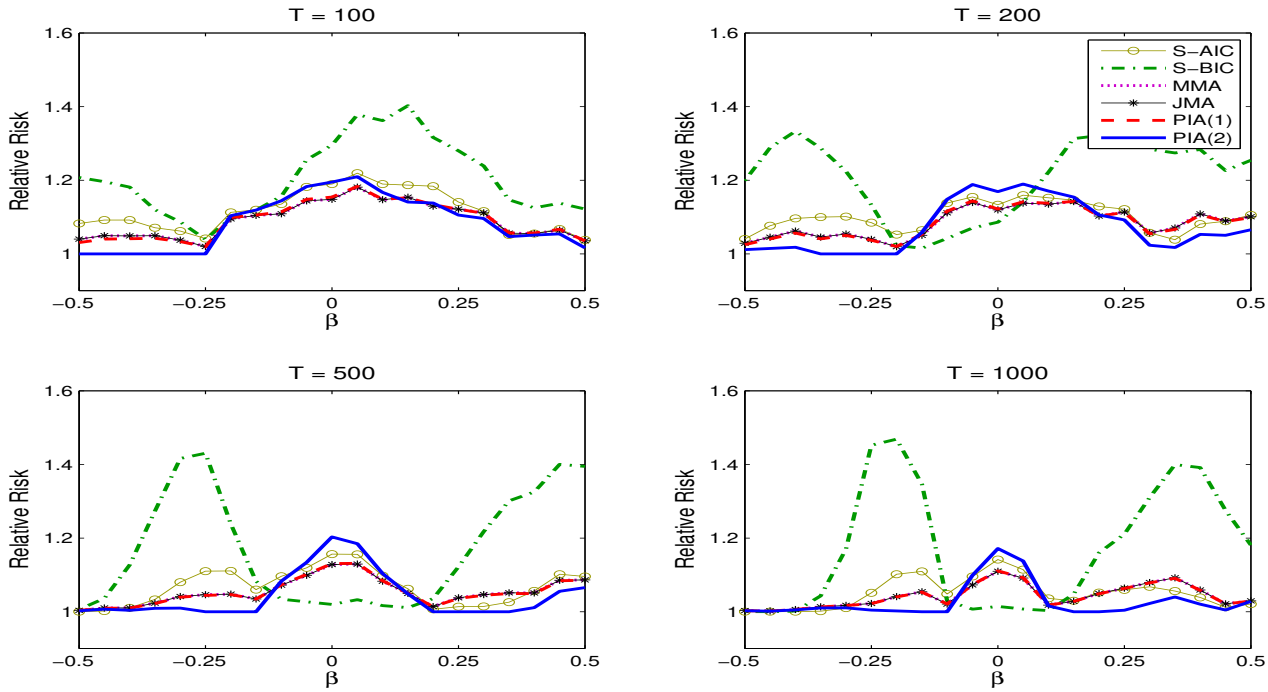


Figure 9: Relative risk for MAX(1, 1) models with homoskedastic errors when $\beta$ varies between -0.5 and 0.5.

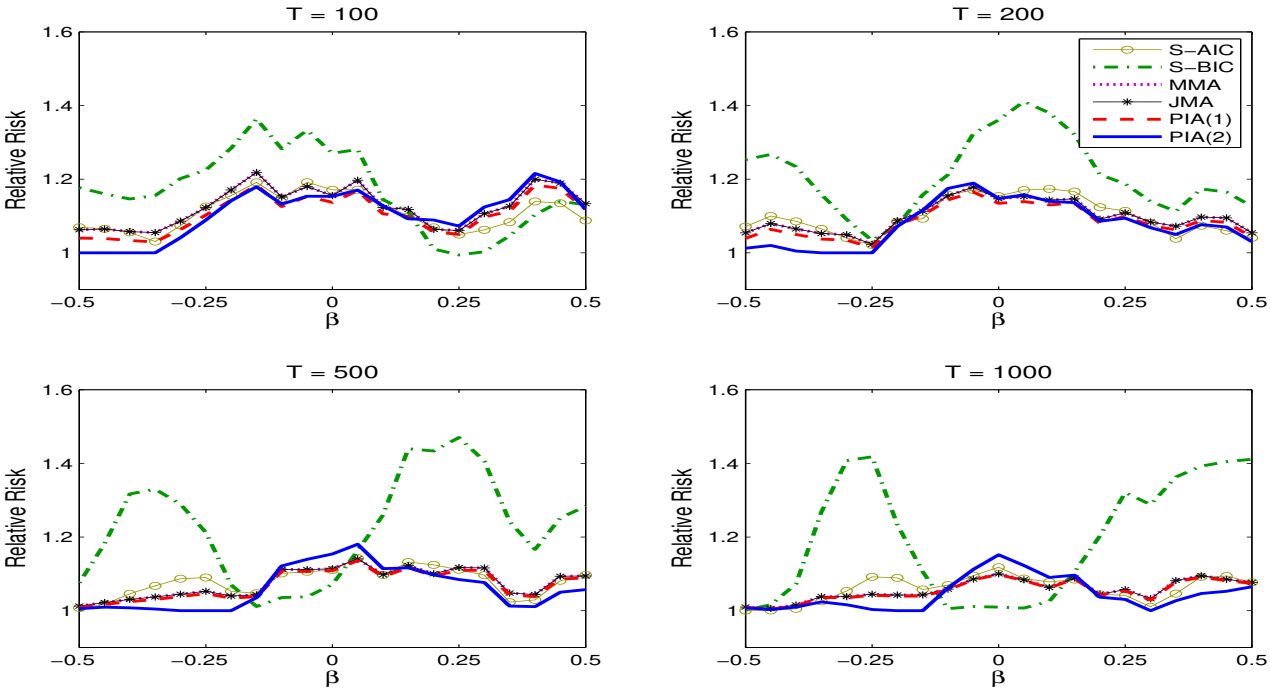


Figure 10: Relative risk for MAX(1, 1) models with heteroskedastic errors when $\beta$ varies between -0.5 and 0.5.

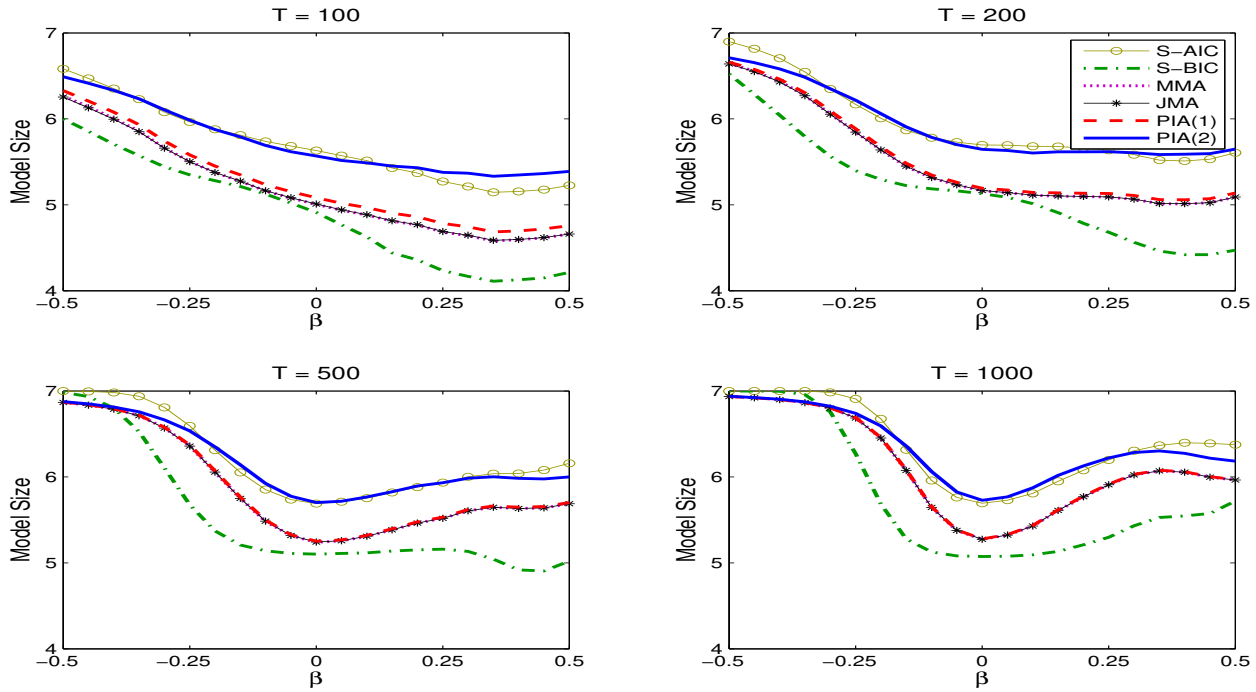


Figure 11: Model size for MAX$(1,1)$ models with homoskedastic errors when $\beta$ varies between -0.5 and 0.5.

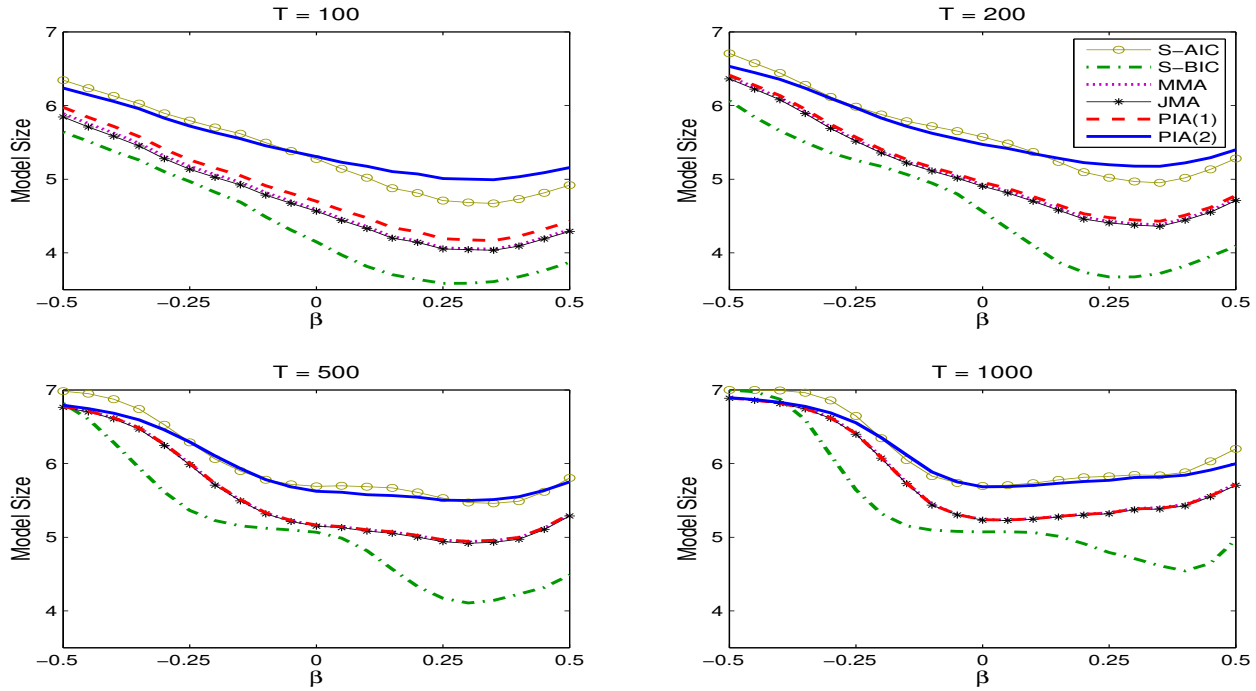


Figure 12: Model size for MAX$(1,1)$ models with heteroskedastic errors when $\beta$ varies between -0.5 and 0.5.

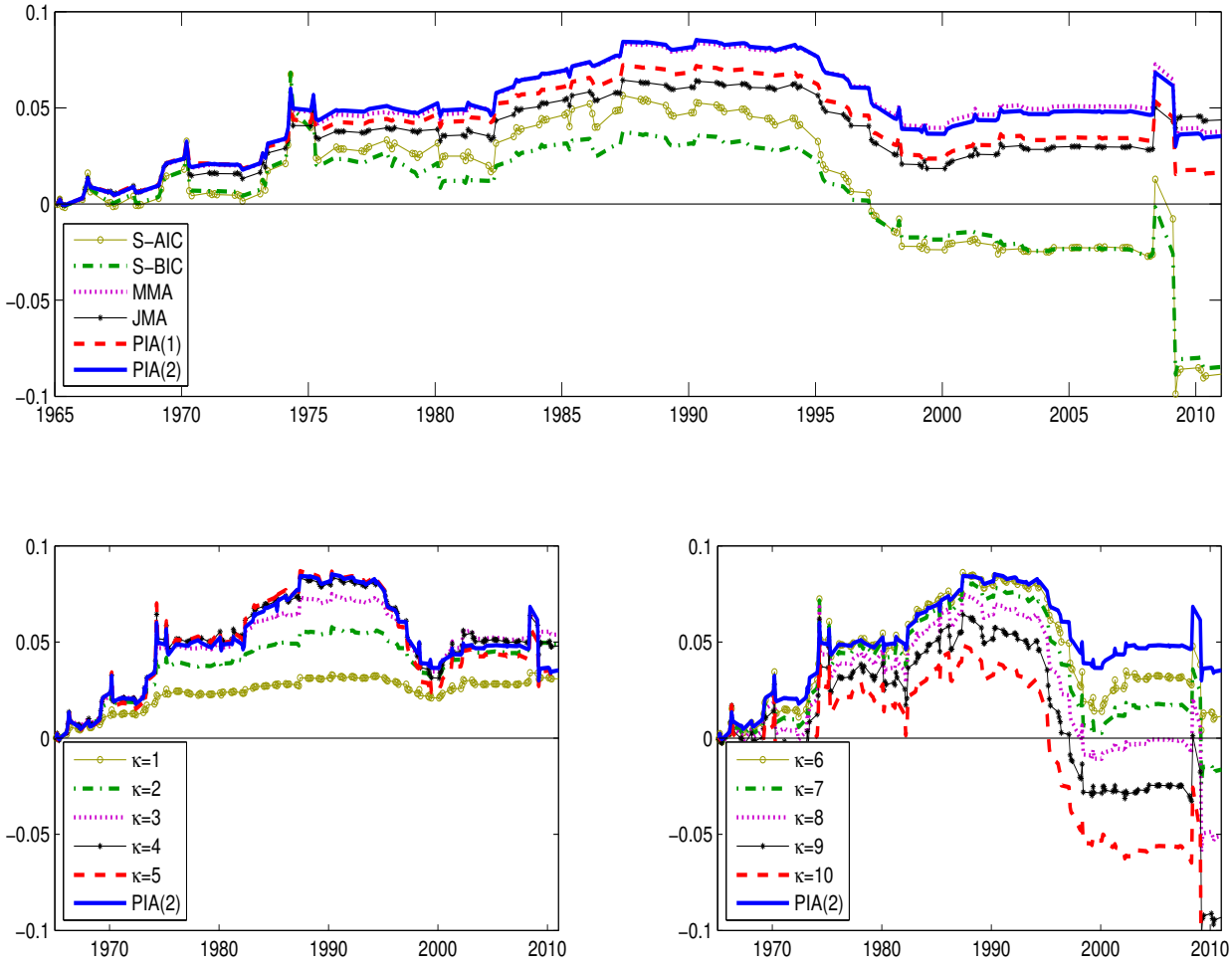


Figure 13: The differences between the cumulative squared prediction error of the historical average forecasting model and the cumulative squared prediction error of the forecast combination model for 1965:1–2011:4.

# 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.

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 (2013): "Forecasting with Factor-Augmented Regression: A Frequentist Model Averaging Approach," Forthcoming. *Journal of Econometrics*.

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. (2013): "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.

——— (2013): "Model Averaging, Asymptotic Risk, and Regressor Groups," Forthcoming. *Quantitative Economics*.

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

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.

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. (2013): "Distribution Theory of the Least Squares Averaging Estimator," Working Paper, National University of Singapore.

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

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.

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

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.

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