Model Averaging in Predictive Regressions

Chu-An Liu and Biing-Shen Kuo

Academia Sinica and National Chengchi University

Mar 7, 2016

Introduction

- Model uncertainty: the challenge of empirical studies is that one does not know exactly what predictors should be included in the model.
- Two methods to deal with model uncertainty: model selection and model averaging.
- Model averaging: a weighted average of estimates from candidate models.
- Two model averaging approaches: Bayesian model averaging and Frequentist model averaging.

Introduction

- Since the seminal work of Bates and Granger (1969), forecast combination has been widely used in economics and statistics.
- How to form the forecast weights is still an open question.
  - Many methods have been proposed for forecast combination, including Granger and Ramanathan (1984), Min and Zellner (1993), Raftery, Madigan, and Hoeting (1997), Buckland, Burnham, and Augustin (1997), Yang (2004), Hansen (2008), Elliott, Gargano, and Timmermann (2013), and Cheng and Hansen (2015), among others.
- We propose a new method for weight selection for linear models.
  - We do not impose the i.i.d. normal assumption.
  - The set of candidate models could be nested or non-nested.
  - The proposed averaging criterion is an an asymptotically unbiased estimator of the mean squared forecast error.
  - We balance the trade-off between model biases and estimation variances.

Model and Estimation

- We deal with the one-step-ahead forecasting model:

$$y_{t+1} = \mathbf{x}_t'\boldsymbol{\beta} + \mathbf{z}_t'\boldsymbol{\gamma} + e_{t+1},$$
$$\mathrm{E}(\mathbf{h}_t e_{t+1}) = 0.$$

- $\mathbf{h}_t = (\mathbf{x}_t', \mathbf{z}_t')'$.
- $\mathbf{x}_t$ is a set of "must-have" predictors
- $\mathbf{z}_t$ is a set of "potentially relevant" predictors. It could be lags of y_t, deterministic terms, or the interaction terms between the predictors.
- The error term is allowed to be heteroskedastic.
- The goal is to construct a point forecast of y_{T+1} given $(\mathbf{x}_T, \mathbf{z}_T)$.

Approximating Models

- Submodel: 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 for $m = 1, ..., M$.
- The set of models could be nested or non-nested.
- The least-squares estimator of β in the mth model is

$$\widehat{\theta}_m = (\mathbf{H}_m'\mathbf{H}_m)^{-1}\mathbf{H}_m'\mathbf{y},$$

  where $\mathbf{H} = (\mathbf{X}, \mathbf{Z})$, $\mathbf{H}_m = (\mathbf{X}, \mathbf{Z}_m) = \mathbf{H}\mathbf{S}_m$, and $\mathbf{S}_m$ is a $(p + q) \times (p + q_m)$ selection matrix.
- The predicted value is $\widehat{\mathbf{y}}(m) = \mathbf{H}_m\widehat{\theta}_m = \mathbf{H}\mathbf{S}_m\widehat{\theta}_m$.
- The one-step-ahead forecast is $\widehat{y}_{T+1|T}(m) = \mathbf{h}_T'\mathbf{S}_m\widehat{\theta}_m$.

One-Step-Ahead Combination Forecast

- 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

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

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

Question and Contributions

- Question: How to assign the model weights?
- Contributions:
  - Optimal Weights: Show that the optimal model weights that minimize the mean squared forecast error (MSFE) depend on the local parameters and the covariance matrix of the predictive regression.
  - Data-Driven Weights: Propose a plug-in estimator of the infeasible optimal weights and use these estimated weights to construct the forecast combination.

Outline

Outline

MSE and MSFE

- Goal: Select weights 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.
- The in-sample mean squared error (MSE):

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

- The one-step-ahead mean squared forecast error:

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

- Therefore the optimal weight vector that minimizes the $MSE(\mathbf{w})$ is expected to minimizes the $MSFE(\mathbf{w})$.

Asymptotic Risk

- How to approximate the MSE?
- Use the asymptotic risk to approximate the MSE.
  - Let $\mathbf{Q} = \mathrm{E}(\mathbf{h}_t \mathbf{h}_t')$. Define the asymptotic trimmed risk or weighted MSE of an estimator $\tilde{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ as

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

  - The weighted MSE function plus σ^2 corresponds to one-step-ahead MSFE.
- Use the information from the sum of squared errors.
  - Define $\mathbf{P}(\mathbf{w}) = \sum_{m=1}^{M} w_m \mathbf{H}_m (\mathbf{H}_m' \mathbf{H}_m)^{-1} \mathbf{H}_m'$ and $\widehat{\mathbf{e}}(\mathbf{w}) = \mathbf{y} - \mathbf{H}\hat{\boldsymbol{\theta}}(\mathbf{w})$.
  - Then, we have $\mathrm{E}(\widehat{\mathbf{e}}(\mathbf{w})' \widehat{\mathbf{e}}(\mathbf{w})) = MSE(\mathbf{w}) + T\sigma^2 - 2\mathrm{E}(\mathbf{e}' \mathbf{P}(\mathbf{w}) \mathbf{e})$.
  - Mallows criterion (Hansen, 2007): $C_T(\mathbf{w}) = \widehat{\mathbf{e}}(\mathbf{w})' \widehat{\mathbf{e}}(\mathbf{w}) + 2\sigma^2 \mathbf{k}' \mathbf{w}$ where $\mathbf{k} = (k_1, ..., k_M)'$.

Outline

Local Asymptotic Framework

- We follow Hjort and Claeskens (2003, JASA) and use a local-to-zero asymptotic framework to approximate the MSE.
- Assumption 1. $\gamma = \gamma_T = \delta/\sqrt{T}$, where δ is an unknown constant vector.
- The local-to-zero framework is canonical in the sense that both squared model biases and estimator variances have the same order $O(T^{-1})$.
- We can decompose $\hat{\theta}_m$ as

$$\hat{\theta}_m = \theta_m + (\mathbf{H}_m'\mathbf{H}_m)^{-1}\mathbf{H}_m'\mathbf{Z}(\mathbf{I}_q - \mathbf{\Pi}_m'\mathbf{\Pi}_m)\gamma_T + (\mathbf{H}_m'\mathbf{H}_m)^{-1}\mathbf{H}_m'\mathbf{e}$$

  where $\theta_m = (\beta', \gamma_m')'$ and $\mathbf{\Pi}_m$ is a $q_m \times q$ selection matrix.

  - $(\mathbf{I}_q - \mathbf{\Pi}_m'\mathbf{\Pi}_m)$ is the selection matrix that chooses the omitted auxiliary regressors.
  - If γ_T converges to $\mathbf{0}$ slower than $T^{-1/2}$, the asymptotic bias goes to infinity.
  - If γ_T converges to $\mathbf{0}$ faster than $T^{-1/2}$, the asymptotic bias goes to zero.

Local Asymptotic Framework

- 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 $\mathcal{F}_t = \sigma(\mathbf{h}_t, \mathbf{h}_{t-1}, ...; e_t, e_{t-1}, ...)$.
- Assumption 2 states that data is strictly stationary.
- It implies that e_{t+1} is conditionally unpredictable at time t.
- It is sufficient to imply that

$$T^{-1}\mathbf{H}'\mathbf{H} \xrightarrow{p} \mathbf{Q}$$

$$T^{-1/2}\mathbf{H}'\mathbf{e} \xrightarrow{d} \mathbf{R} \sim \mathbf{N}(\mathbf{0}, \mathbf{\Omega})$$

  where $\mathbf{\Omega} = \mathrm{E}\left(\mathbf{h}_t\mathbf{h}_t' e_{t+1}^2\right)$.

- Also, 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.

Asymptotic Normality of the Averaging Estimator

Theorem 1.

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

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

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

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

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

Remark: $\mathbf{A}(\mathbf{w})\delta$ represents the bias term. The magnitude of the bias is determined by the covariance matrix $\mathbf{Q}$ and the local parameter δ.

Asymptotic Trimmed Risk of the Averaging Estimator

We derive the asymptotic trimmed risk of the model averaging estimator and characterize the optimal weights in a local asymptotic framework.

Theorem 2.

Suppose Assumptions 1-2 hold. We have

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

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

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

Note that $\mathbf{C}_m$ and $\mathbf{P}_m$ are functions of $\mathbf{Q}$ and the selection matrix.

Optimal Weights and Plug-In Weights

- The optimal weight vector is the value that minimizes the asymptotic risk over $\mathbf{w} \in \mathcal{H}^M$:

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

- The weight vector of the plug-in estimator is defined as

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

  where $\mathbf{w}' \widehat{\psi} \mathbf{w}$ is the sample analog of $\mathbf{w}' \psi \mathbf{w}$.

- The objection function is linear-quadratic in $\mathbf{w}$, which can be solved numerically via quadratic programming.

- The plug-in one-step-ahead combination forecast is

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

Construct the Data-Driven Weights

- We now discuss the plug-in estimator $\widehat{\psi}_{m,\ell}$.
- Recall that $\psi_{m,\ell} = \mathrm{tr}\left(\mathbf{Q}\mathbf{C}_m\boldsymbol{\delta}\boldsymbol{\delta}'\mathbf{C}_\ell'\right) + \mathrm{tr}\left(\mathbf{Q}\mathbf{P}_m\boldsymbol{\Omega}\mathbf{P}_\ell\right)$.
- We use the method of moments estimator for covariance matrices $\mathbf{Q}$ and $\boldsymbol{\Omega}$. Thus, it is quite easy to model the heteroskedasticity and serial correlation.
- We use the asymptotically unbiased estimator for the local parameter $\boldsymbol{\delta}$.
  - $\widehat{\boldsymbol{\delta}} = \sqrt{T}\widehat{\boldsymbol{\gamma}} \xrightarrow{d} \mathbf{R}_{\boldsymbol{\delta}} \sim N(\boldsymbol{\delta}, \mathbf{S}_0'\mathbf{Q}^{-1}\boldsymbol{\Omega}\mathbf{Q}^{-1}\mathbf{S}_0)$.
  - An alternative estimator is $\widehat{\boldsymbol{\delta}\boldsymbol{\delta}'} = \widehat{\boldsymbol{\delta}}\widehat{\boldsymbol{\delta}}' - \mathbf{S}_0'\widehat{\mathbf{Q}}^{-1}\widehat{\boldsymbol{\Omega}}\widehat{\mathbf{Q}}^{-1}\mathbf{S}_0$.
- The asymptotic trimmed risk of the plug-in averaging estimator:
  - $\widehat{\mathbf{w}} \xrightarrow{d} \mathbf{w}^* = \underset{\mathbf{w}\in\mathcal{H}^M}{\mathrm{argmin}}\, \mathbf{w}'\boldsymbol{\psi}^*\mathbf{w}$ where $\psi^*_{m,\ell} = \mathrm{tr}\left(\mathbf{Q}\mathbf{C}_m\mathbf{R}_{\boldsymbol{\delta}}\mathbf{R}_{\boldsymbol{\delta}}'\mathbf{C}_\ell'\right) + \mathrm{tr}\left(\mathbf{Q}\mathbf{P}_m\boldsymbol{\Omega}\mathbf{P}_\ell\right)$.
  - $R(\widehat{\boldsymbol{\theta}}(\widehat{\mathbf{w}}), \boldsymbol{\theta}) = \mathrm{E}\left((\mathbf{A}(\mathbf{w}^*)\boldsymbol{\delta} + \mathbf{P}(\mathbf{w}^*)\mathbf{R})'\mathbf{Q}(\mathbf{A}(\mathbf{w}^*)\boldsymbol{\delta} + \mathbf{P}(\mathbf{w}^*)\mathbf{R})\right)$.
  - $\mathbf{A}(\mathbf{w}^*) = \sum_{m=1}^M w_m^*\left(\mathbf{P}_m\mathbf{Q} - \mathbf{I}_{p+q}\right)\mathbf{S}_0$ and $\mathbf{P}(\mathbf{w}^*) = \sum_{m=1}^M w_m^*\mathbf{P}_m$.

Outline

Finite Sample Investigation

We consider two simulation setups.

- The first design is the 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.

We consider the following estimators:

(1) Smoothed AIC (S-AIC; Buckland, Burnham, and Augustin (1997))
(2) Smoothed BIC (S-BIC)
(3) Mallows model averaging (MMA; Hansen (2007))
(4) Jackknife model averaging (JMA; Hansen and Racine (2012))
(5) Complete subset regression (Elliott, Gargano, and Timmermann (2013))
(6) Plug-In averaging estimators

Six Forecast Combination Methods

Weight choice:

- S-AIC: $\widehat{w}_m = \exp(-\frac{1}{2}\mathrm{AIC}_m)/\sum_{j=1}^{M}\exp(-\frac{1}{2}\mathrm{AIC}_j)$.
- S-BIC: $\widehat{w}_m = \exp(-\frac{1}{2}\mathrm{BIC}_m)/\sum_{j=1}^{M}\exp(-\frac{1}{2}\mathrm{BIC}_j)$.
- MMA: $C_T(\mathbf{w}) = \widehat{\mathbf{e}}(\mathbf{w})'\widehat{\mathbf{e}}(\mathbf{w}) + 2\sigma^2\mathbf{k}'\mathbf{w}$
- JMA: $CV_n(\mathbf{w}) = \mathbf{w}'\tilde{\mathbf{e}}'\tilde{\mathbf{e}}\mathbf{w}$ 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.

Remark:

- MMA is limited to the homoskedastic model.
- Both MMA and JMA are limited to the random sample.

Complete Subset Regressions

- Elliott, Gargano, and Timmermann (2013, JoE) propose a new forecast combination method based on complete subset regressions.
- For a given set of potential predictors, they construct the forecast combination by using equal-weighted combination based on all possible models that include κ predictors.
- The one-step-ahead combination forecast is

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

  where $n_{\kappa,k} = k!/(\kappa!(k-\kappa)!)$ is the number of models considered based on κ subset regressions.
- Remark: Complete subset regressions is not suitable for the nested models.

DGP 1: Regression Model

- The data generation process for the first design is

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

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

- x_{1t} is the intercept. x_{jt} for $j \geq 2$ are AR(1) processes with $\rho_x = 0.5$, and 0.9.
- The predictors x_{jt} are correlated. $(u_{2t}, ..., u_{kt})' \sim N(\mathbf{0}, \mathbf{Q}_u)$ 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.
- $\boldsymbol{\beta} = \left(1, \frac{k-1}{k}, ..., \frac{1}{k}\right)' c/\sqrt{T}$ and $\delta_j = \sqrt{T}\beta_j = c(k-j+1)/k$.
- Homoskedastic simulation: $e_t \sim N(0,1)$.
  Heteroskedastic simulation: $e_t = 3^{-1/2}(1-\rho_x^2)x_{kt}^2\epsilon_t$ and ϵ_t follows AR(1).

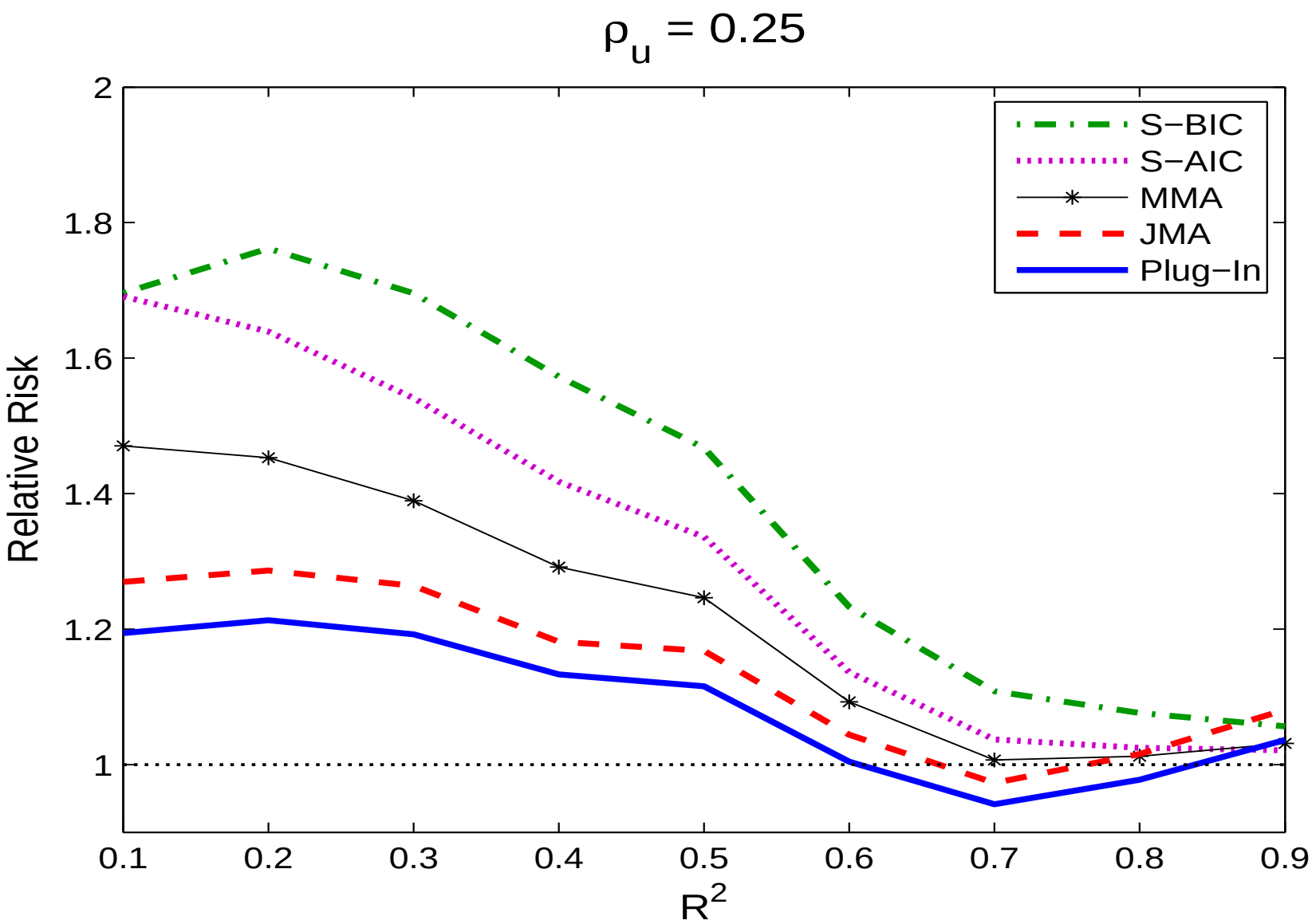
DGP 1: Regression Model

- The sample size is $T = 200$.
- The number of predictors is $k = 5$.
- The number of models is $M = 32$.
- We report the relative risk:

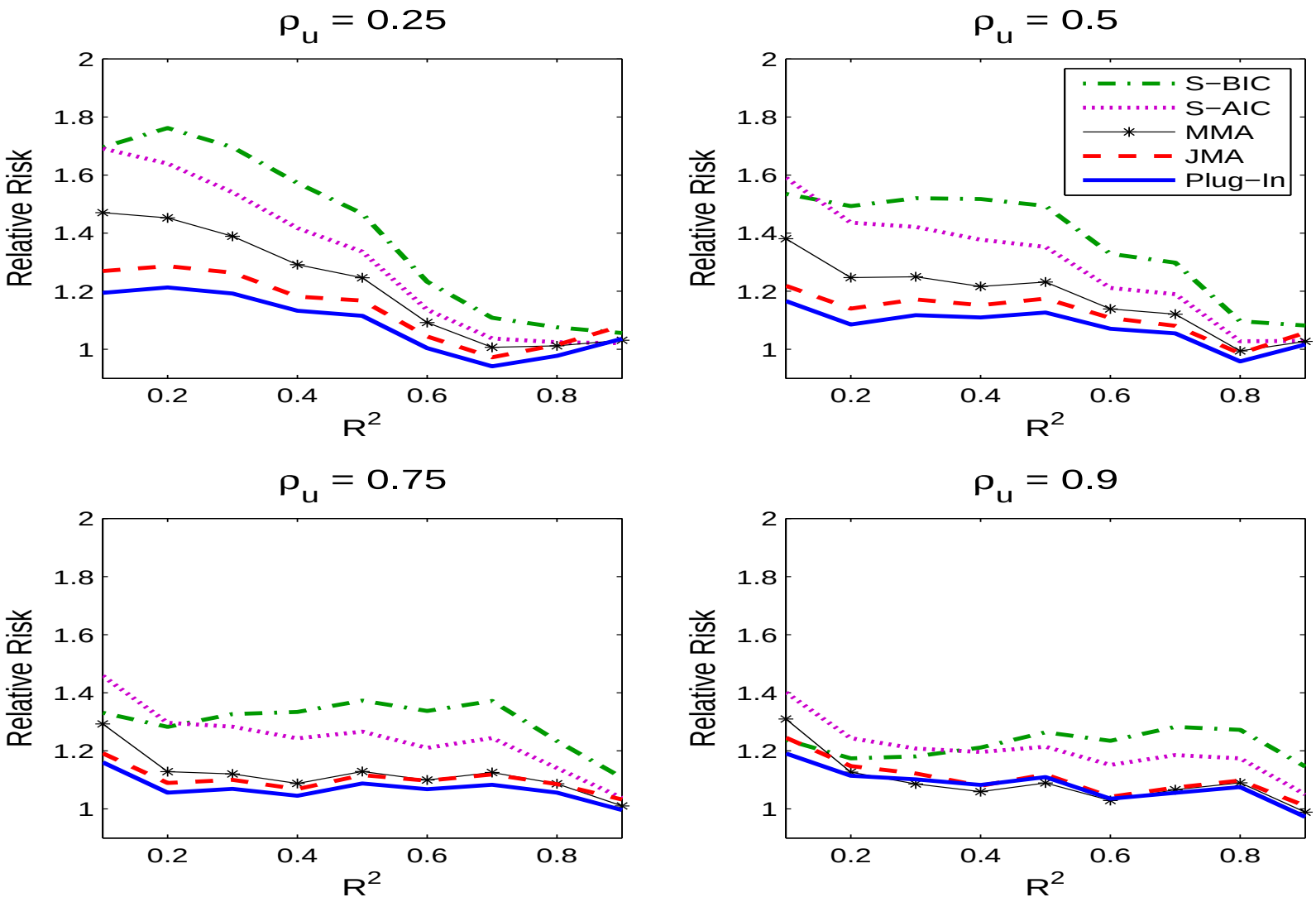
$$\frac{\frac{1}{S}\sum_{s=1}^{S}\left(y_{s,T+1|T} - \widehat{y}_{s,T+1|T}(\widehat{\mathbf{w}})\right)^2}{\min\limits_{m\in\{1,\ldots,M\}}\frac{1}{S}\sum_{s=1}^{S}\left(y_{s,T+1|T} - \widehat{y}_{s,T+1|T}(m)\right)^2},$$

where $\widehat{y}_{T+1|T}(m)$ is the prediction based on the model m and $\widehat{y}_{T+1|T}(\widehat{\mathbf{w}})$ is the prediction based on the averaging estimator. $(S = 5000)$

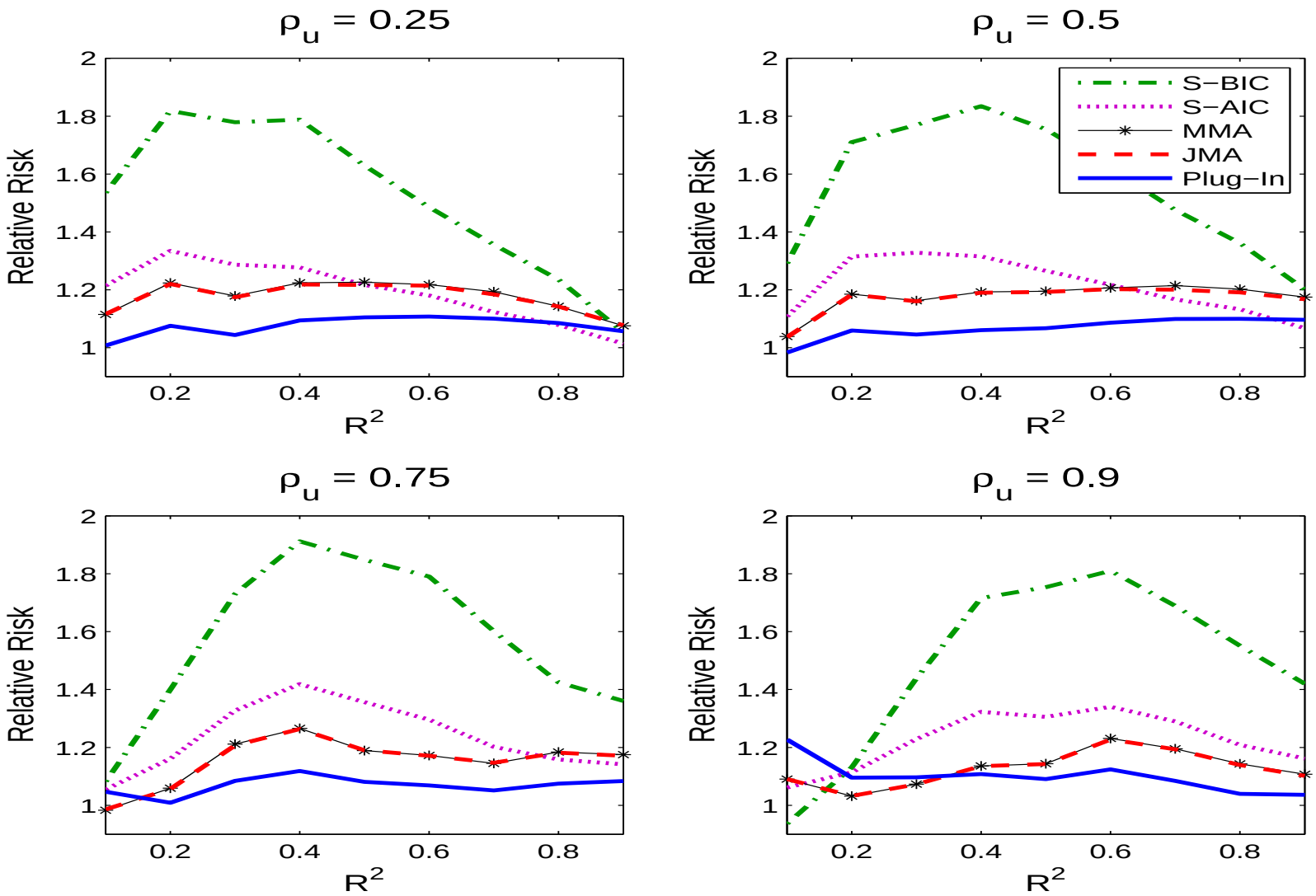
Heteroskedastic simulation, $\rho_x = 0.9$.



Heteroskedastic simulation, $\rho_x = 0.9$.



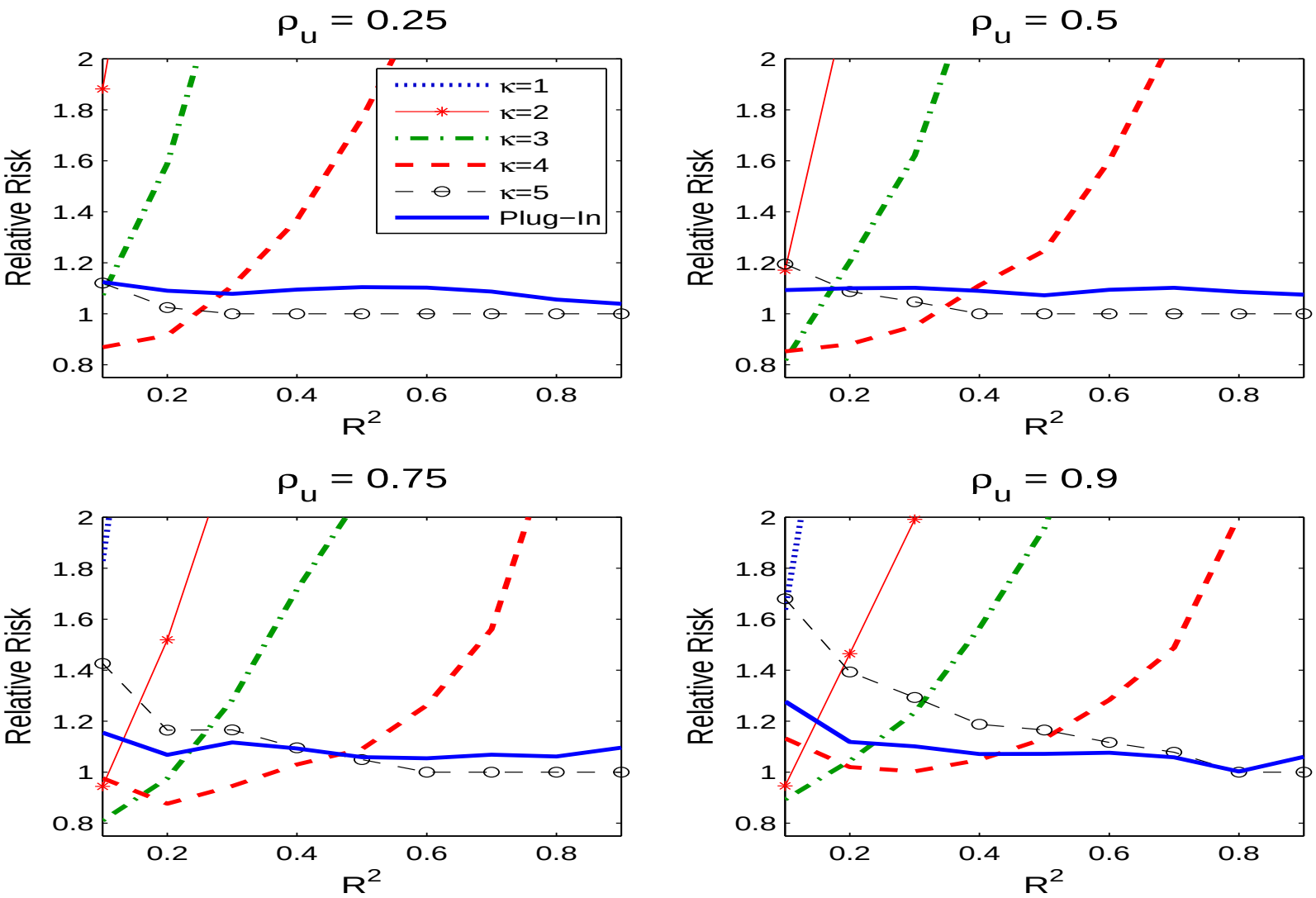
Homoskedastic simulation, $\rho_x = 0.9$.



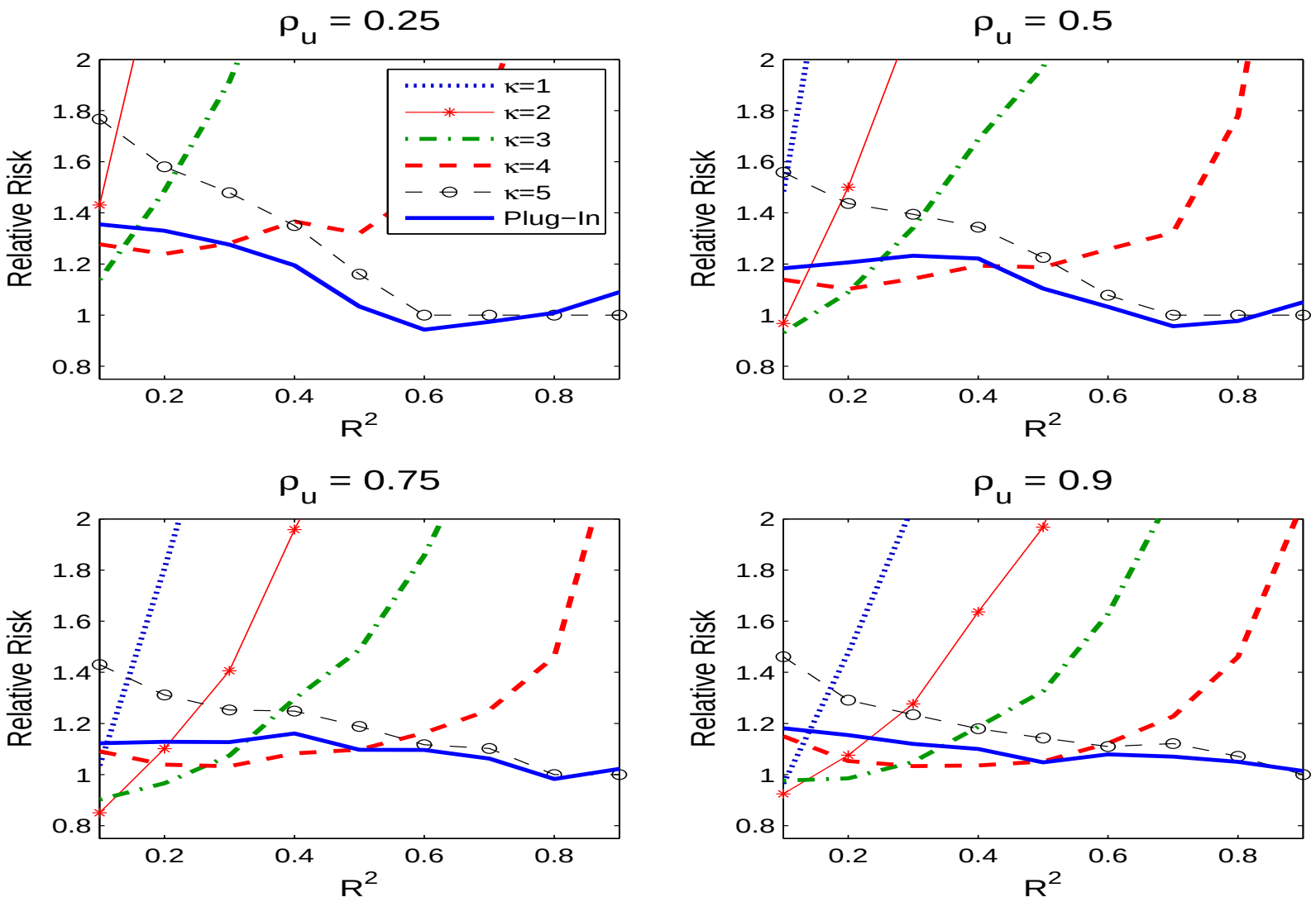
Simulation Results

- Monte Carlo simulations show that the plug-in averaging estimtor has much lower MSFE than other model averaging estimators in both homoskedastic and heteroskedastic settings.
- JMA has lower reative risk than MMA and S-AIC in the heteroskedastic simulation.
- S-BIC has poor performance in both homoskedastic and heteroskedastic settings.
- We now compare the plug-in averaging estimator with complete subset regressions.

Homoskedastic simulation, $\rho_x = 0.5$.



Heteroskedastic simulation, $\rho_x = 0.5$.



DGP 2: MAX(1, 1)

- The data generation process for the second design is

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

  - x_t is an AR(1) process and $u_t \sim N(0, 1)$.
  - $e_t \sim N(0, \sigma_t^2)$, where $\sigma_t^2 = 0.5$ for the homoskedastic simulation and $\sigma_t^2 = 1 + x_t^2$ for the heteroskedastic simulation.
  - The parameter β is varied on a grid from -0.5 to 0.5.

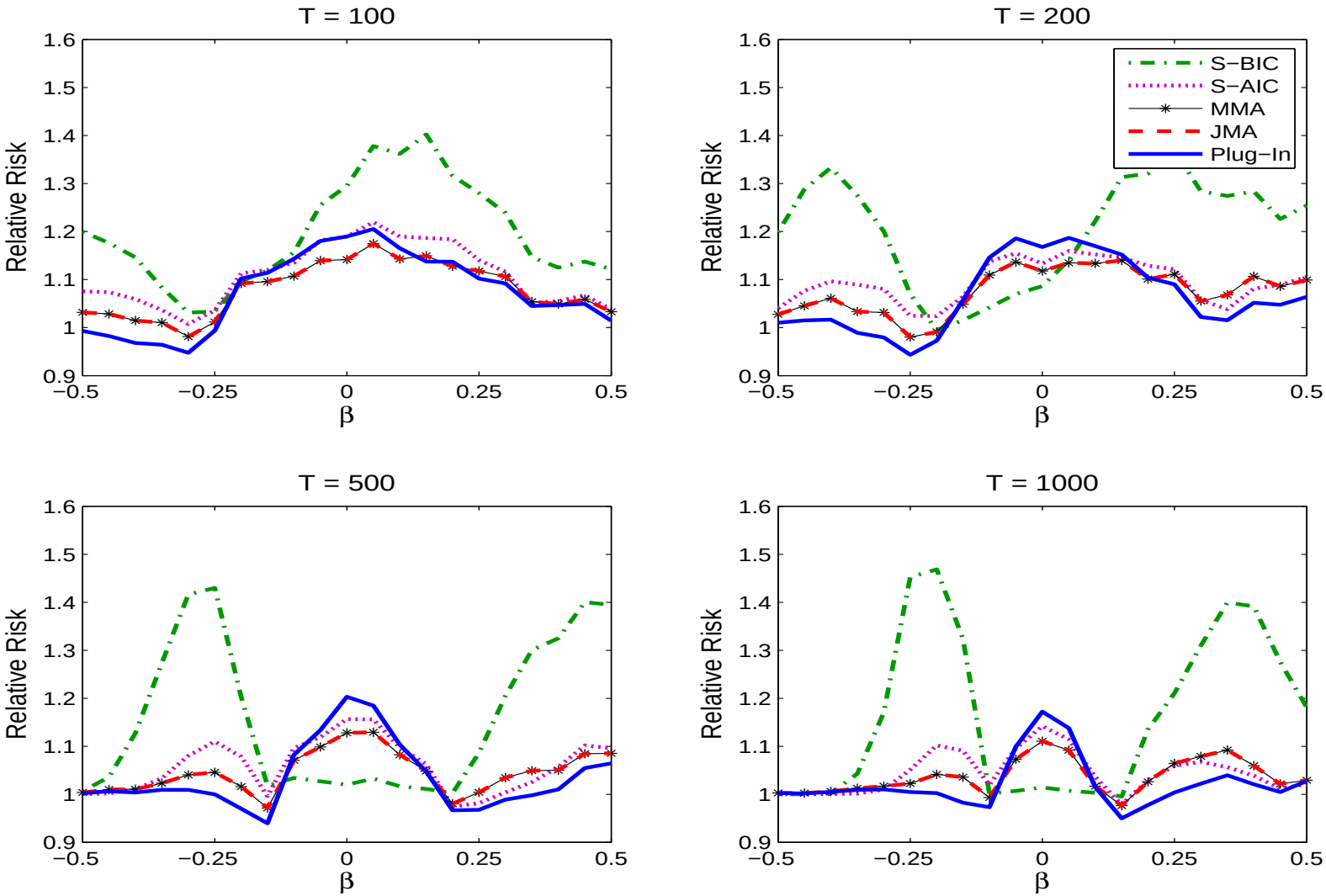
- We consider a sequence of nested models based on regressors

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

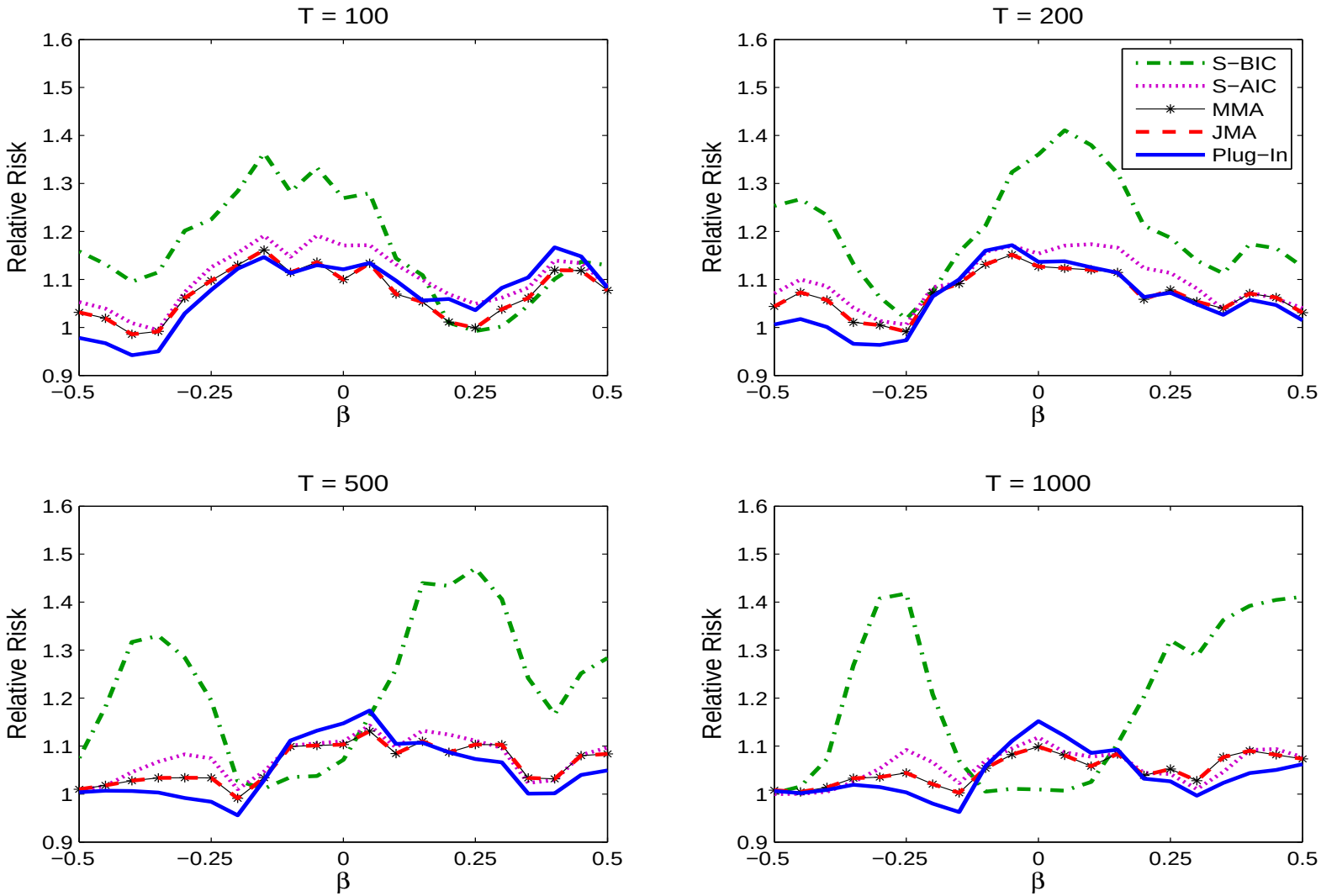
  - For $\beta \neq 0$, the true model is infinite dimensional.
  - For $\beta = 0$, all seven models are wrong.

- Sample size: $T = 100$, 200, 500, and 1000.
- We report the relative risk and the average model size.

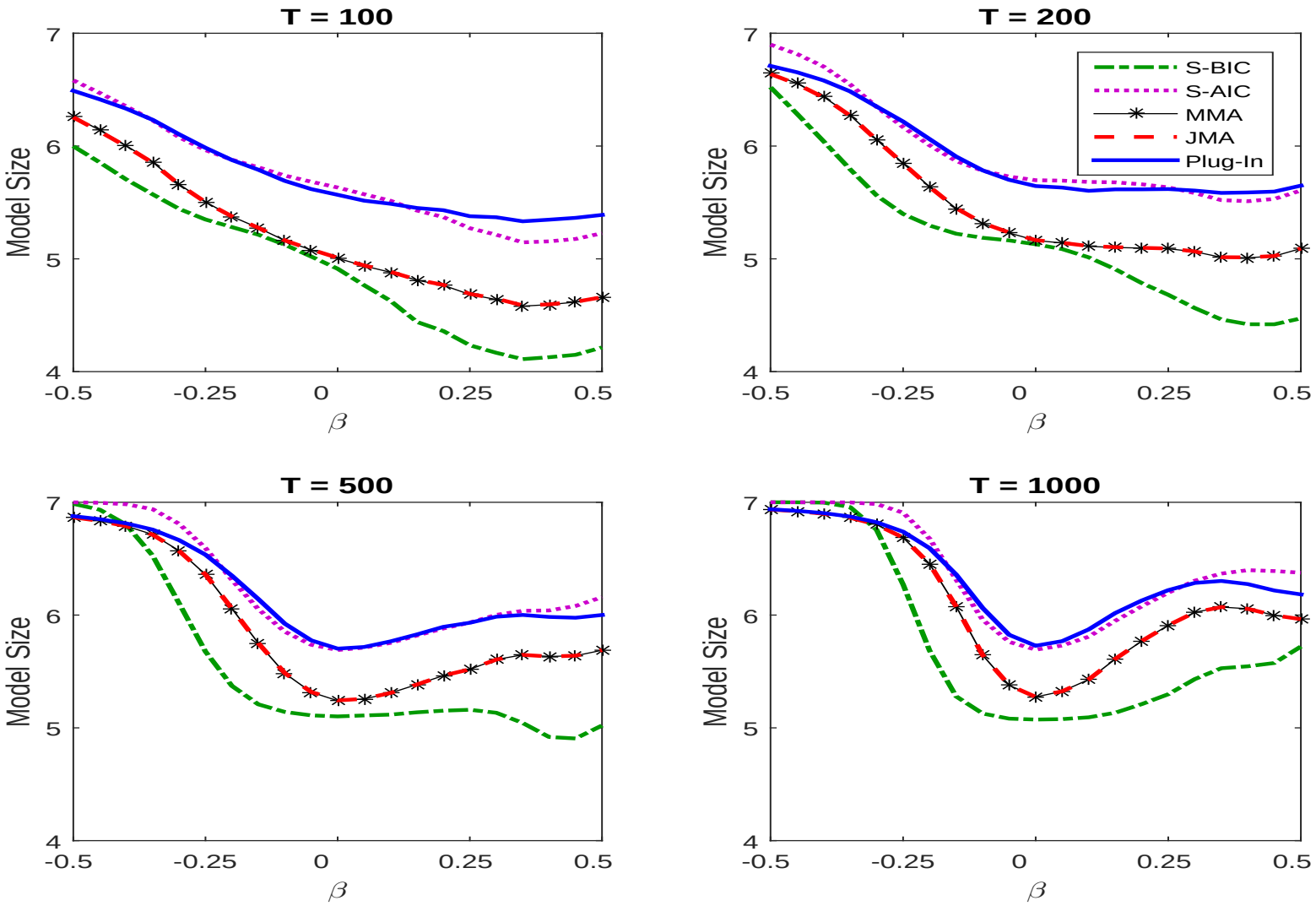
Relative risk, homoskedastic errors.



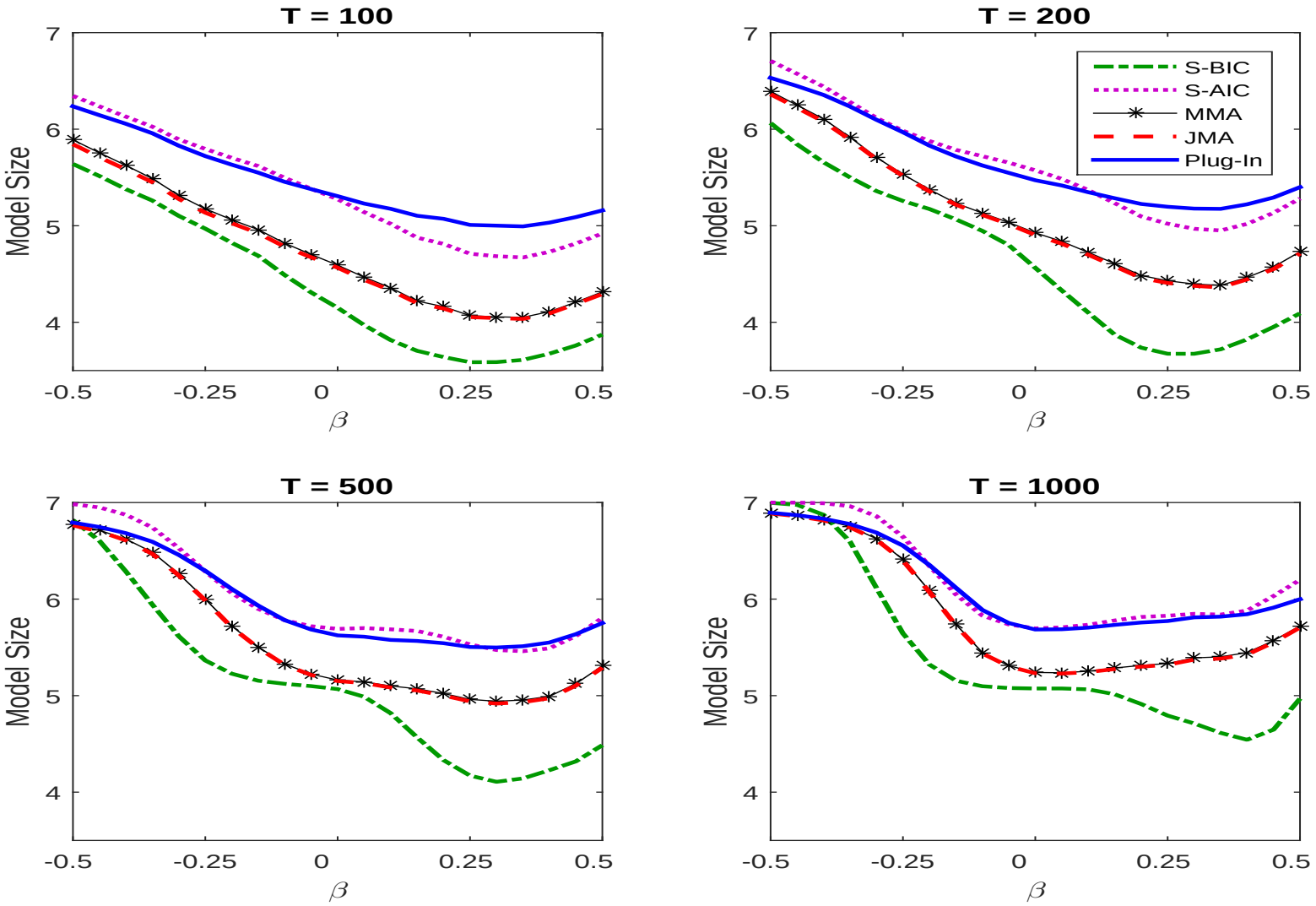
Relative risk, heteroskedastic errors.



Model size, homoskedastic errors.



Model size, heteroskedastic errors.



Outline

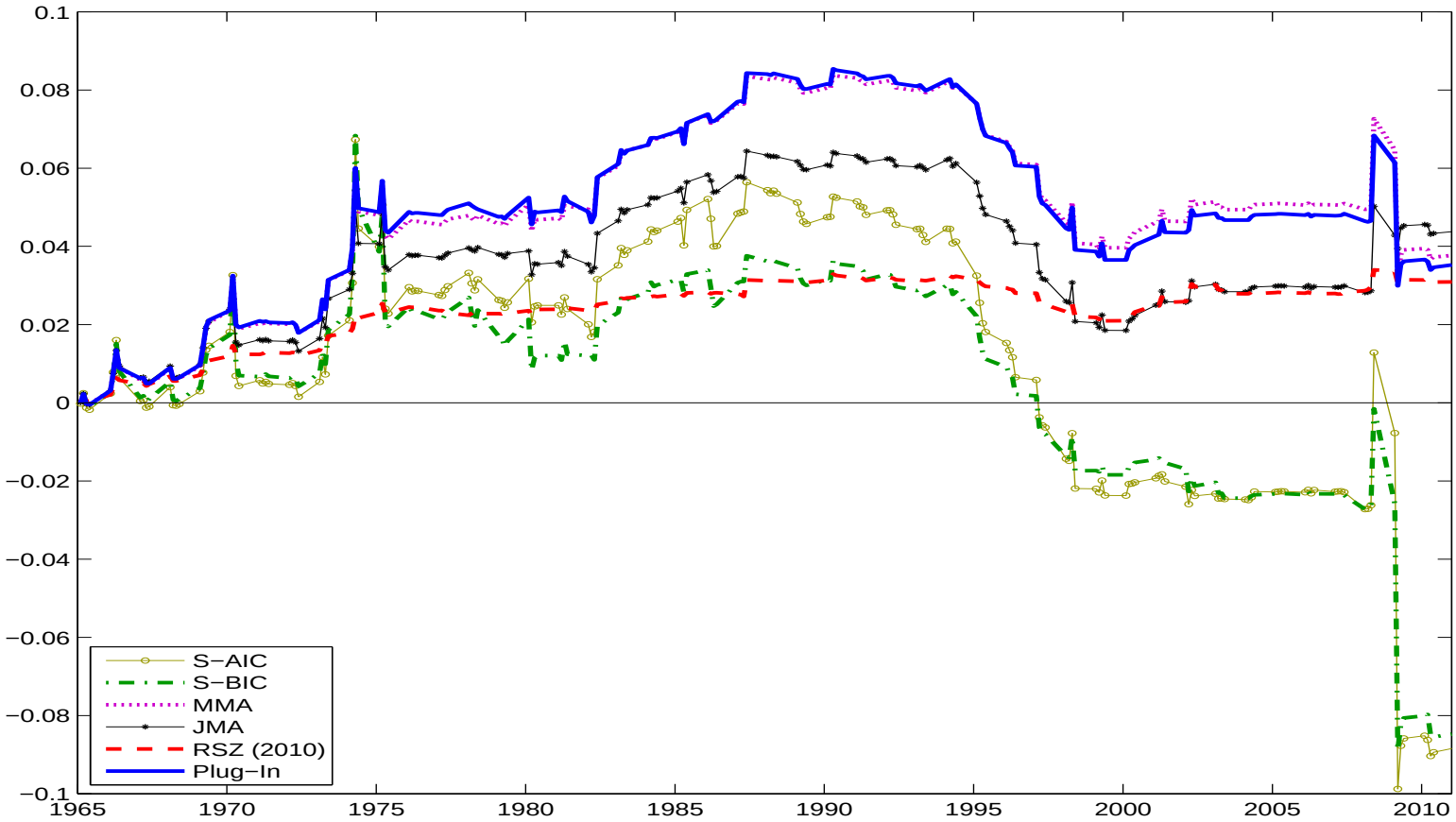
Empirical Application: Stock Returns Prediction

- We apply the plug-in forecast combination method to stock returns prediction.
- Different studies suggest different economic variables and models.
- Welch and Goyal (2008) argue that numerous economic variables have poor out-of-sample predictions.
- Rapach, Strauss, and Zhou (2010) propose a equal-weighted forecast combination approach to the subset predictive regression.
- We apply the forecast combination with data-driven weights instead of equal weights to U.S. stock market.

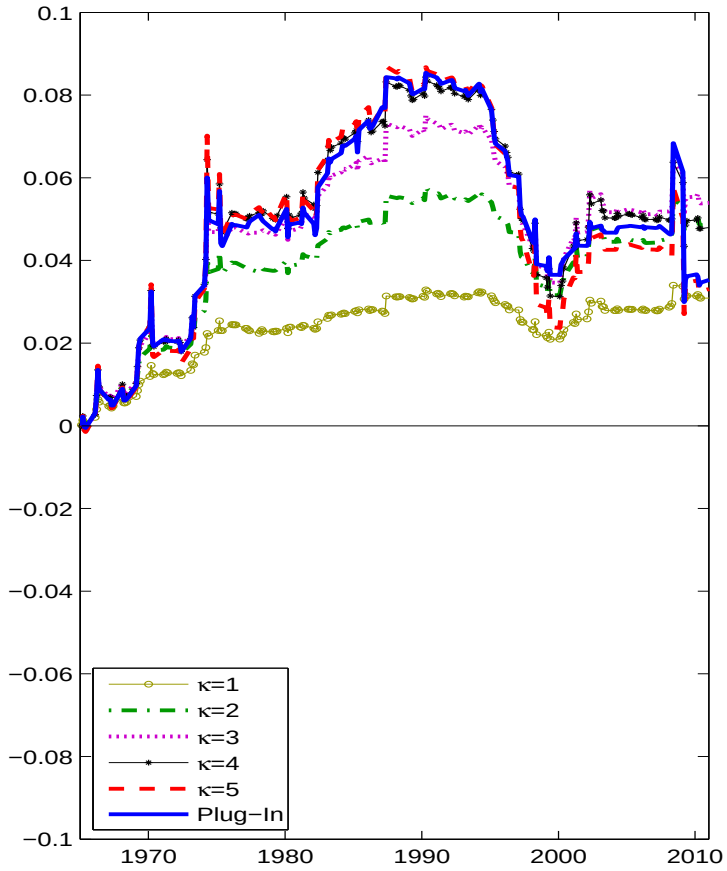
Model and Data

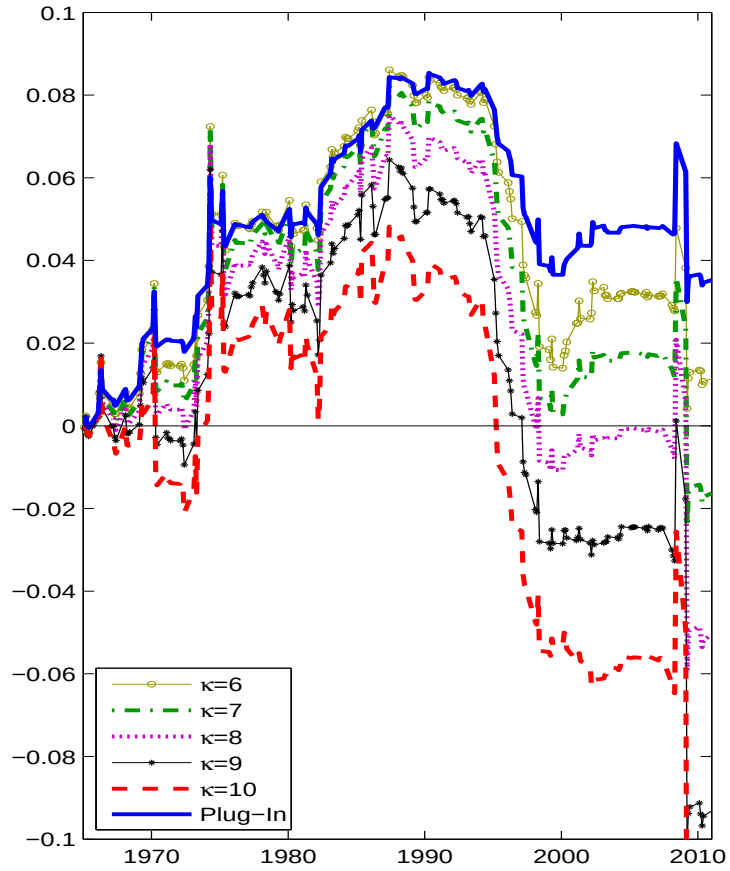
- The model: $r_{t+1} = \beta + \mathbf{z}_t'\gamma + e_{t+1}$
- r_{t+1} is the equity premium and $\mathbf{z}_t$ are the economic variables.
- We consider 10 economic variables and all possible models.
  - The 10 economic variables are Dividend Price Ratio, Dividend Yield, Earnings Price Ratio, Book-to-Market Ratio, Net Equity Expansion, Treasure Bill, Long Term Return, Default Yield Spread, Default Return Spread, and Inflation.
- The quarterly data are from Welch and Goyal (2008) for 1947-2011 (T=260).
- We follow Welch and Goyal (2008) and calculate the out-of-sample forecast of the equity premium using a recursively expanding estimation window.
  - In-sample period: 1947:1-1964:4
  - Out-of-sample evaluation period: 1965:1-2011:4
- We use the historical average of the equity premium as a benchmark.

The differences between the cumulative square prediction errors of the historical average forecasting model and the cumulative square prediction errors of the forecast combination model for 1965:1-2011:4.



The differences between the cumulative square prediction errors of the historical average forecasting model and the cumulative square prediction errors of the forecast combination model for 1965:1-2011:4.





Out-Of-Sample Forecasting Results

- The out-of-sample R^2 value of the plug-in averaging estimator is 2.7257 with the p-value 0.0173.
  - 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 $\bar{r}_{\tau+1|\tau} = \sum_{t=1}^{\tau} r_t$ 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 $R^2_{OOS} \leq 0$.
- Our results support that forecast combinations provide significant gains on equity premium predictions relative to the historical average.

Outline

Multi-Step Forecast: Model and Estimation

- We now consider the h-step-ahead forecasting model:

$$y_{t+h} = \mathbf{x}_t'\boldsymbol{\beta} + \mathbf{z}_t'\boldsymbol{\gamma} + e_{t+h}$$
$$\mathrm{E}(\mathbf{h}_t e_{t+h}) = 0.$$

- The goal is to construct a point forecast of y_{T+h} given $(\mathbf{x}_t', \mathbf{z}_t')$.
- 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 $\widehat{\boldsymbol{\theta}} = (\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\mathbf{y}$.

- The h-step-ahead combination forecast is

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

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

Optimal Weights and Plug-In Weights

- 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 and the optimal weight vector has the same form

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

  where the (m, ℓ)th element of ψ is $\psi_{m,\ell} = \mathrm{tr}\left(\mathbf{Q}\mathbf{C}_m \boldsymbol{\delta}\boldsymbol{\delta}' \mathbf{C}_\ell'\right) + \mathrm{tr}\left(\mathbf{Q}\mathbf{P}_m \boldsymbol{\Omega} \mathbf{P}_\ell\right)$ and $\boldsymbol{\Omega} = \lim_{T\to\infty} \frac{1}{T} \sum_{s=1}^{T} \sum_{t=1}^{T} \mathrm{E}\left(\mathbf{h}_s \mathbf{h}_t' e_{s+h} e_{t+h}\right)$.

- We also can construct the plug-in estimator by replacing the unknown paramters $\mathbf{Q}$, $\boldsymbol{\Omega}$, and $\boldsymbol{\delta}$ by the sample analogue.

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 $\widehat{\mathbf{C}}_m = \widehat{\mathbf{P}}_m\widehat{\mathbf{Q}} - \mathbf{I}_q$.
- The plug-in estimator can be rewritten as

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

where $\widehat{\mathbf{e}} = \mathbf{y} - \mathbf{H}\widehat{\theta}$, $\widehat{\mathbf{e}}_m = \mathbf{y} - \mathbf{H}_m\widehat{\theta}_m$, $\widehat{\mathbf{Q}}_m = \mathbf{S}_m'\widehat{\mathbf{Q}}\mathbf{S}_m$, and $\widehat{\Omega}_m = \mathbf{S}_m'\widehat{\Omega}\mathbf{S}_m$.

The Equivalent Result 1

- The criterion function for the plug-in averaging estimator is

$$\mathbf{w}'\widehat{\psi}\mathbf{w} = \mathbf{w}'\widetilde{\psi}\mathbf{w} - \widehat{\mathbf{e}}'\widehat{\mathbf{e}} - \mathrm{tr}(\widehat{\mathbf{Q}}^{-1}\widehat{\Omega}).$$

- The (m,ℓ)th element of $\widetilde{\psi}$ is

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

- Minimizing $\mathbf{w}'\widehat{\psi}\mathbf{w}$ over $\mathbf{w} = (w_1, ..., w_M)$ is equivalent to minimizing $\mathbf{w}'\widetilde{\psi}\mathbf{w}$.
- If the error term is i.i.d. and homoskedastic, then the covariance matrix $\mathbf{\Omega}$ can be consistently estimated by $\widehat{\Omega} = \widehat{\sigma}^2\widehat{\mathbf{Q}}$. Thus, $\mathrm{tr}(\widehat{\mathbf{Q}}_m^{-1}\widehat{\Omega}_m) = \widehat{\sigma}^2 k_m$.
- Define $\mathbf{\Sigma}$ as an $M \times M$ matrix whose (m,ℓ)th element is $k_m + k_\ell$.
- The criterion function for the plug-in averaging estimator is

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

which is the Mallows criterion proposed by Hansen (2007).

The Equivalent Result 2

- If the error term is serially uncorrelated and identically distributed, then $\mathbf{\Omega}$ can be consistently estimated by $\widehat{\mathbf{\Omega}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{h}_t \mathbf{h}_t' \widehat{e}_{t+1}^2$.
- Thus, $\mathrm{tr}(\widehat{\mathbf{Q}}_m^{-1}\widehat{\mathbf{\Omega}}_m) = \mathrm{tr}\left(\left(\sum_{t=1}^{T} \mathbf{h}_{m,t}\mathbf{h}_{m,t}'\right)^{-1}\left(\sum_{t=1}^{T} \mathbf{h}_{m,t}\mathbf{h}_{m,t}'\widehat{e}_{t+1}^2\right)\right) \equiv \widetilde{k}_m$.
- Define $\widetilde{\mathbf{\Sigma}}$ as an $M \times M$ matrix whose (m,ℓ)th element is $\widetilde{k}_m + \widetilde{k}_\ell$.
- The criterion function for the plug-in averaging estimator is

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

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

Conclusion

- We study the weight selection for forecast combination in a predictive regression when the goal is minimizing the MSFE.
- 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 propose a frequentist model averaging criterion, an asymptotically unbiased estimator of the asymptotic risk, to select forecast weights.
- Simulations show that the proposed estimator achieves lower MSFE relative risk than other existing model averaging methods in most cases.
- The proposed method can be easily extended to the multi-step forecast combination.