How to Use Model Confidence Sets for Forecasting and Impulse Response Estimation and the Value of Model Averaging

Niels Aka*[a] and Rolf Tschernig[b]

[a]German Institute for Economic Research (DIW), D-10108 Berlin, Germany
[b]University of Regensburg, Department of Economics and Econometrics, D-93040 Regensburg, Germany

February 15, 2017

Abstract: Model confidence sets frequently include more than one model. We suggest to use all of them in the subsequent analysis by either averaging across the parameters of the estimated models included or the final quantities of interest. The weights for the included models can be chosen to be equal or by Jackknife model averaging. In Monte Carlo simulations these procedures are compared to classical model selection, shrinkage estimation and Jackknife model averaging by computing multi-step ahead forecasts and impulse responses. Applying Jackknife model averaging to model confidence sets turns out to be beneficial in small samples and robust in larger samples where using the Schwarz criterion has great merits.

Keywords: Forecast combination, impulse response analysis, Jackknife model averaging, Lasso, lag selection, model confidence sets, Monte Carlo, multi-step ahead forecasting, Postlasso, Ridge, univariate time series

*Corresponding author. Current address: DIW Berlin, D-10108 Berlin, Germany. Phone: +49 (0) 30 89789 581, Mail: naka@diw.de. Both authors thank the conference participants of the 2015 IMS China meeting, the Statistische Woche Hamburg 2015 and the Ausschuss für Ökonometrie 2016 for comments on earlier versions. We further thank the seminar participants at the DIW Berlin, Freie Universität Berlin, Humboldt Universität zu Berlin, Soochow University and Nanjing University of Aeronautics and Astronautics.

1 Introduction

We explore the merits of using model confidence sets to handle model uncertainty in two scenarios: forecasting exercises and impulse response analysis. We further compare this approach to several other statistical procedures in the framework of a Monte Carlo simulation study. In the simulation, we apply the different procedures to identify suitable autoregressive specifications within a larger pool of candidate specifications and then continue with that choice for forecasting and impulse response estimation. Thus, the simulation allows judging the performance of each procedure in terms of (forecast) errors. To include model uncertainty, most procedures follow one of two strategies. Either they average across model specifications or across the quantity of interest, i.e. forecasts or impulse responses. Common to both is that they rely on a weight vector for averaging. To create the weight vector, we employ the model confidence set (MCS), jackknife model averaging (JMA), standard information criteria, and combinations thereof. For comparison, we also specify models using the least absolute shrinkage and selection operator (Lasso), the Postlasso and ridge regression. Taking into account all the different approaches, we thus explore and compare the merits of 18 selection procedures.

In our simulations, the task is to compute h-step ahead forecasts up to horizon 15 and impulse responses up to 20 periods ahead. As data generating processes (DGPs) we take three different univariate autoregressive processes of order eight with some of the parameters set to zero. The AR(8) processes differ in terms of their signal-to-noise ratios, frequency characteristics and degrees of persistence. There are six sample sizes ranging from 40 to 500 observations. The initial set of models, from which suitable specifications will be selected, includes a total of 256 autoregressive processes based on all lag combinations up to lag eight. We therefore conduct a full subset specification search.

Based on mean squared error, the results suggest that using the Schwarz criterion works well for model selection in larger samples and across DGPs, but may perform poorly in small samples, in particular for impulse response estimation. In the latter case, applying JMA to the models inside the MCS turns out to be a robust strategy. This combination is found among the best strategies in small samples and performs comparable to the best competitors excluding Schwarz in larger samples. For computing impulse responses, model averaging is found to be superior to combining impulse responses of each model in the MCS.

Model selection, and its impact on estimation and inference, has been a long standing topic in econometric modelling (see Theil, 1957; Leamer, 1978). The reason being that the most parsimonious model containing the data generating process (DGP) is generally unknown in empirical work. This so-called true model therefore has to be selected from a collection of models that the econometrician has initially chosen. Frequently, the true model is too complex for reliable estimation in finite samples. In this case, a more parsimonious model has to be selected that approximates the DGP reasonably well.

This selection automatically implies a trade-off between the approximation error and estimation uncertainty.

In applied work, the current standard approach to address this trade-off is to select a single model that minimizes a selection criterion such as the AIC, Hannan-Quinn (HQC) or Schwarz information criterion (SIC). While popular and straightforward to apply, this simple approach has its drawbacks. First, different model selection criteria may select differently. Second, by selecting a *single* model, the practitioner ignores models that are ranked close to the preferred one. Such models may be equally good or even better but were not selected due to noise. In such a case they should also be considered in any further analysis. Third, classical frequentist inference conditions on a given model and will therefore suffer from size distortions if models were actually selected beforehand. This result has been established theoretically (cf. Leeb and Pötscher, 2005) and empirically (cf. Demetrescu, Hassler, and Kuzin, 2011).

While we do not focus on inference, we do address the first two concerns by using the MCS and JMA procedures. The MCS was suggested by Hansen, Lunde, and Nason (2011) to estimate a set of superior models from an initially chosen set, where superiority is defined by a user-specified loss function. The authors device a stepwise procedure in which they repeatedly test the null hypothesis that all currently selected models are identical with respect to expected loss. Upon rejection, they remove a model from the current set and repeat until rejection fails. This approach has an important advantage over standard practice. It allows for an (asymptotic) control of the family-wise error rate within the process of model selection. When selecting a model by comparing model selection criteria, one implicitly conducts a sequence of pairwise tests and thus a multiple testing issue arises. Controlling the overall significance level then requires controlling the family-wise error rate. Only recently, methods for controlling the family-wise error rate have become available that can effectively take into account dependence among the tests by using bootstrap methods (e. g. Romano and Wolf, 2005). Hansen, Lunde, and Nason (2011) succeeded in adapting this framework to the model selection problem.

If, however, the estimated MCS contains more than one model, it is uncertain how best to continue with the empirical analysis. The most natural approach for computing the quantities of interest, such as h-step ahead forecasts or impulse responses, is to take the underlying null hypothesis seriously and to treat all models in the MCS as equal. Therefore one assigns to all models in the MCS the same weight and to all other models weight zero. These weights can then be used to average across parameters of different models or across forecasts or other quantities of interest. The issue of averaging across parameters has already been intensively studied in the literature on (frequentist) model averaging. Such methods require choosing weights, which are in some specified sense optimal, for averaging the parameters across models. Claeskens and Hjort (2008) provide the first comprehensive book on frequentist model averaging and Moral-Benito (2015) provides a very recent survey. Based on the results of this literature, we suggest, as an

alternative to using equal weights, to apply Jackknife model averaging (JMA) to the models contained in the MCS. Hansen and Racine (2012) and Zhang, Wan, and Zou (2013) showed that JMA is asymptotically optimal in various situations. It may be noted that in Bayesian econometrics, model averaging has a much longer tradition.

Instead of weighting the parameters of different models, one may compute a weighted average of the quantity of interest computed from each model. If this quantity is an out-of-sample forecast, one obtains forecast combinations. This is a very active field of research by itself and recent surveys on forecast combinations are provided byAiolfi, Capistrán, and Timmermann (2011) or Timmermann (2006). A very recent comprehensive treatment is found in the book by Elliot and Timmermann (2016). We therefore also consider using the models included in the MCS for combining forecasts or impulse response estimates.

The paper is organized as follows. Section 2 describes all relevant methods with a particular emphasis on model averaging and model confidence sets. The setup of the Monte Carlo simulation is laid out in Section 3. Section 4 reports the results and Section 5 briefly summarizes.

2 Methods for Model Selection and Model Averaging

In this section we briefly describe all methods used in our simulation study. We first outline the general setup of model averaging and forecast combination. Then we describe Jackknife model averaging in more detail and show how to use the MCS for model averaging. Next we sketch shrinkage methods, in particular Ridge and Lasso estimation. Finally we propose two new combinations of existing methods.

2.1 Setup

We only consider dynamic regression models for a scalar dependent variable y_t. All approaches considered in this paper aim at estimating one or several quantities of interest such as the conditional mean, marginal effects, h-step ahead predictions or impulse response functions. They all require the user to specify an initial collection of models for further consideration. Denote this set of models by $\mathcal{M}^0$ and index all models in the set by $i = 1, \ldots, m_0$. To some extent we follow the notation of Hansen, Lunde, and Nason (2011), hereafter HLN. To facilitate the presentation, consider the estimation of the conditional mean $\mu_t \equiv E[y_t|\boldsymbol{x}_t]$ based on an observed sample $(y_t, \boldsymbol{x}_t)$, $t = 1, 2, \ldots, n$, where the $(1 \times k_{max})$ vector $\boldsymbol{x}_t$ denotes all explanatory variables available in the sample and may include lagged y_t's. We assume that $\boldsymbol{x}_t$ belongs to the information set Ω_t which includes all potential explanatory variables that are predetermined w.r.t. the error term of the DGP. Note that at this point we allow for the possibility that μ_t is misspecified. In this case, there exists a vector of further explanatory variables $\boldsymbol{x}_t^+ \in \Omega_t$ such that $\mu_t = E[y_t|\boldsymbol{x}_t] \neq E[y_t|\boldsymbol{x}_t, \boldsymbol{x}_t^+]$. Then μ_t exhibits an approximation error.

We have by definition

$$y_t = \mu_t + u_t, \tag{1}$$

$$E[u_t|\boldsymbol{x}_t] = 0. \tag{2}$$

The conditional variance is denoted by $\sigma_t^2 \equiv E[u_t^2|\boldsymbol{x}_t]$.

If the set of available regressors $\mathcal{M}^0$ contains irrelevant regressors, then estimation efficiency can be increased by only using all relevant regressors. The underlying subset of regressors is one dimension in which the m_0 models in $\mathcal{M}^0$ may vary and on which we focus in this paper. In case of linear models one may write a subset model as

$$y_t = \boldsymbol{x}_{t,i}\boldsymbol{\beta}_i + u_{t,i}, \tag{3}$$

where the $(1 \times k_i)$ vector $\boldsymbol{x}_{t,i}$ and the $(k_i \times 1)$ vector $\boldsymbol{\beta}_i$ denote the regressors and the parameters of model i. Note that we index the error term also by the model to explicitly indicate that its properties depend on the selected model. For example, if relevant regressors are omitted from $\boldsymbol{x}_{t,i}$, then $u_{t,i}$ contains u_t and the omitted regressor. In matrix notation we have

$$\boldsymbol{y} = \boldsymbol{X}_i\boldsymbol{\beta}_i + \boldsymbol{u}_i, \quad i = 1, 2, \ldots, m_0, \tag{4}$$

where $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$ and $\boldsymbol{X}_i$ denote the sample vector of dependent variables and the $(n \times k_i)$ sample matrix of explanatory observations used in model i, respectively. The $(n \times 1)$ sample error vector is denoted by $\boldsymbol{u}_i$.

The set of models $\mathcal{M}^0$ may not contain all possible combinations of regressors based on $\boldsymbol{x}_t$. We denote this complete set of possible models by $\mathcal{M}^{all} = \{1, 2, \ldots, m_{all}\}$ with $m_{all} = 2^{k_{max}}$. Note that in this setup $\mathcal{M}^0 \subseteq \mathcal{M}^{all}$ holds. When referring to the index of the complete model set $\mathcal{M}^{all}$, we use the index s. For a generic set of models we simply write $\mathcal{M}$.

In the sequel we will assign $s = m_{all}$ to the model which includes all available k_{max} regressors $\boldsymbol{X} = (\boldsymbol{x}_1', \ldots, \boldsymbol{x}_t', \ldots, \boldsymbol{x}_n')'$ and

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_{m_{all}} + \boldsymbol{u}. \tag{5}$$

can therefore be called the encompassing model. Note that the encompassing model may not be included in the initial model set $\mathcal{M}^0$. The model with the largest number of regressors in $\mathcal{M}^0$ is indexed by m_0.

For each of the m_0 models (4) in the initial model set $\mathcal{M}^0$ one can produce an estimate $\hat{\mu}_{t,i}$ of the conditional expectation μ_t, e. g. by OLS, and therefore also of the observation y_t. In order to have a common notation we will use $\hat{y}_{t,i}$ in place of $\hat{\mu}_{t,i}$ and h-step ahead forecasts, if available, will be denoted by $\hat{y}_{t+h|t,i}$, $h = 1, \ldots, H$, where (sample) information up to time t is used.

Model selection corresponds to making some choice out of the m_0 available estimated models for μ_t. If the models allow for computing h-step ahead predictions, as for example in case of autoregressive models of finite order, then model selection also implies a choice from the h-step ahead predictions for y_{t+h}, $h = 1, \ldots, H$, given by the set $\{\hat{y}_{t+h|t,1}, \hat{y}_{t+h|t,2}, \ldots, \hat{y}_{t+h|t,m_0}\}$.

All methods for model selection are in one way or the other based on an expected loss. The loss caused by the deviation of the forecast $\hat{y}_{t+h|t,i}$ of model i from the observation y_{t+h} is measured by the user-specified loss function $L(y_{t+h}, \hat{y}_{t+h|t,i})$. The expected loss or risk for model i is defined as $E[L(y_{t+h}, \hat{y}_{t+h|t,i})]$ where the expectation is taken over the estimation and the evaluation sample both w.r.t. the DGP. Mostly used is the mean squared error (of prediction) (MSEP) based on the quadratic loss $L(y_{t+h}, \hat{y}_{t+h|t,i}) = (y_{t+h} - \hat{y}_{t+h|t,i})^2$

$$MSEP(y_{t+h}, i) \equiv E\left[(y_{t+h} - \hat{y}_{t+h|t,i})^2\right]. \tag{6}$$

Optimally, one would like to choose the model(s) with the lowest risk. For an arbitrary loss function $L_{t,i}$ for predicting y_t using model i Hansen, Lunde, and Nason (2011) suggest to evaluate differences between models in terms of their expected loss differences $\mu_{ij} \equiv E[L_{t,i} - L_{t,j}]$ for all $i, j \in \mathcal{M}^0$ where it is assumed that all expected loss differences are finite and independent of t such that a ranking of models is possible. Note that this allows for the possibility that the expected values of each loss function may be nonstationary. Hansen, Lunde, and Nason (2011, Definition 1) define a set of superior models $\mathcal{M}^*$ as

$$\mathcal{M}^* \equiv \{i \in \mathcal{M}^0 : \mu_{ij} \leq 0 \text{ for all } j \in \mathcal{M}^0\}. \tag{7}$$

Inserting $L_{t,i} = L(y_t - \hat{y}_{t|t-h,i})^2$ into (7), one obtains the set of superior models for an h-step ahead forecast by

$$\mathcal{M}_h^* := \{i \in \mathcal{M}^0 : \operatorname*{argmin}_i E[L(y_t, \hat{y}_{t|t-h,i})]\}. \tag{8}$$

Note that $\mathcal{M}^*$ and thus $\mathcal{M}_h^*$ may depend on sample size. For example, when using the MSEP, the best model(s) show the optimal trade-off between squared bias and estimation variance.

In practice, the MSEPs of each model are unknown and have to be estimated. Due to the estimation error the model exhibiting minimal MSEP may not be chosen. In the Monte Carlo simulations in Section 4 where we consider h-step ahead forecasts, we compare procedures based on model selection, model averaging, forecast combinations, and shrinkage estimation. In addition, we propose combinations of those. This holds in particular for the MCS-based model selection procedure that allows to select more than a single MSEP-optimal model. In the remainder of this section we present each of the procedures.

2.2 Model Selection

After estimating the MSEP it is common in practice to select a single model by choosing the one with the lowest estimated MSEP. The MSEP for $h = 1$ can be estimated by AIC or cross-validation. While not exactly estimating the MSEP, the Hannan-Quinn (HQ) or Schwarz criterion (SC) are equally common. In Section 4 we use AIC, HQ, and SC to estimate the conditional mean

$$\hat{y}_{t,\hat{i}_q} = \boldsymbol{X}_{\hat{i}_q}\hat{\boldsymbol{\beta}}_{\hat{i}_q}, \quad q = AIC, HQ, SC, \quad \hat{i}_q \in \mathcal{M}^0, \tag{9}$$

and the h-step ahead forecasts $\hat{y}_{t+h|t,\hat{i}_q}$ if the models in $\mathcal{M}^0$ allow for this.

Instead of explicitly selecting a model in $\mathcal{M}^0$ one may use shrinkage estimation. Depending on the type of the regularization term, shrinkage implicitly does model selection (Lasso, Postlasso) or does not (Ridge). In the latter case, the largest model m_0 is always used albeit with reduced flexibility due to the shrinkage parameter. Shrinkage methods require a proper choice of the shrinkage parameter. All shrinkage methods used in Section 4 are briefly presented in Section 2.7.

If the set of superior models $\mathcal{M}^*$ given by (8) contains more than one model, all procedures selecting a single model fail to estimate $\mathcal{M}^*$. This more general case is included by the MCS model selection procedure described in Section 2.6. The MCS procedure also allows to control the size in the underlying sequence of tests.

2.3 Model Averaging

Selecting models always implies a discrete choice which can be avoided if all models of $\mathcal{M}^0$ are considered by properly averaging across all of them. By using continuous weights, a continuous choice is available. Let $\boldsymbol{b}_i$ denote the $(k_{max} \times 1)$ vector which contains the entries of $\boldsymbol{\beta}_i$ at those rows where the explanatory variables in $\boldsymbol{X}_i$ correspond to the columns in $\boldsymbol{X}$. All other entries in $\boldsymbol{b}_i$ are zero. For the encompassing model (5) one has $\boldsymbol{b}_{m_{all}} = \boldsymbol{\beta}_{m_{all}}$.

Since in Section 2.8 we will use model averaging not only for all models contained in the initial model set $\mathcal{M}^0$, we describe the general setup for a set of models $\mathcal{M} = \{1, 2, \ldots, m\}$ to be specified later. Model averaging computes the weighted parameter average across all models in $\mathcal{M}$

$$\hat{\boldsymbol{b}}_{ma}(\boldsymbol{w}) \equiv \sum_{i=1}^{m} \hat{\boldsymbol{b}}_i w_i = \sum_{i \in \mathcal{M}} \hat{\boldsymbol{b}}_i w_i \tag{10}$$

where the index ma indicates model averaging and $\boldsymbol{w} = (w_1, w_2, \ldots, w_m)'$ denotes the vector of weights which sum to unity, $\sum_{i=1}^{m} w_i = 1$ (Claeskens and Hjort, 2008, Section 7). In this paper we follow Hansen and Racine (2012) and impose the stronger condition of non-negative weights bounded by one, $w_i \in [0, 1]$. We denote the set of possible

weight vectors by

$$\mathcal{H}_n = \{\boldsymbol{w} \in [0,1]^m : \sum_{i=1}^{m} w_i = 1\}. \tag{11}$$

When applied to $\mathcal{M}^0$, the averaging estimator (10) can be viewed as a restricted estimator of $\boldsymbol{\beta}_{m_0}$ of the largest model in $\mathcal{M}^0$. In this sense, model averaging can be viewed as an estimator of the single model m_0 with a particular way of regularization.

Due to the linearity in parameters of the regression setup, averaging across parameters is equivalent to averaging across estimated conditional means

$$\hat{\boldsymbol{y}}_{ma}(\boldsymbol{w}) \equiv \sum_{i=1}^{m} \boldsymbol{X}\hat{\boldsymbol{b}}_i w_i = \boldsymbol{X}\sum_{i=1}^{m} \hat{\boldsymbol{b}}_i w_i = \boldsymbol{X}\hat{\boldsymbol{b}}_{ma}(\boldsymbol{w}). \tag{12}$$

The equivalence between averaging across parameters first and then computing the quantity of interest and the other way round no longer holds if the quantity of interest is a nonlinear function of parameters of which a prominent case are h-step ahead predictions. If the models in $\mathcal{M}^0$ allow for computing h-step ahead forecasts, applying model averaging leads to computing a single h-step ahead forecast using the averaged parameter estimate $\hat{\boldsymbol{b}}_{ma}(\boldsymbol{w})$ given by (10). We denote this forecast by $\hat{y}_{t+h|t,ma}(\hat{\boldsymbol{b}}_{ma}(\boldsymbol{w}))$. An alternative is discussed in the next section.

Note that the model selection procedures mentioned in Section 2.2 are a special form of model averaging with weight 1 assigned to that model which is selected and weight 0 to all other models. We denote the corresponding weight vectors as $\hat{\boldsymbol{w}}_{AIC}$, $\hat{\boldsymbol{w}}_{HQ}$, and $\hat{\boldsymbol{w}}_{SC}$, etc.

2.4 Forecast Combinations and Combinations of Impulse Responses

An alternative to computing h-step ahead forecasts $\hat{y}_{t+h|t,ma}(\hat{\boldsymbol{b}}_{ma}(\boldsymbol{w}))$ by model averaging is to compute the h-step ahead forecasts for each of the m models in $\mathcal{M}$ using $\hat{\boldsymbol{\beta}}_i$ and then average across all m individual h-step forecasts

$$\hat{y}_{t+h|t,fc}(\boldsymbol{w}) \equiv \sum_{i=1}^{m} w_i \hat{y}_{t+h|t,i} = \sum_{i\in\mathcal{M}} w_i \hat{y}_{t+h|t,i} \tag{13}$$

The procedure (13) is called forecast averaging which is indicated by the index fc. This approach can also be used in more general settings where various forecasts are available but not the data underlying some of the forecasts (c.f., e.g. Aiolfi, Capistrán, and Timmermann, 2011).

Analogously to combining forecasts (13) one may combine estimated impulse responses $\hat{\phi}_{h,i}$ computed for each of the m models delivering

$$\hat{\phi}_{h,fc}(\boldsymbol{w}) \equiv \sum_{i\in\mathcal{M}} w_i \hat{\phi}_{h,i} \tag{14}$$

One important question is whether forecast combinations are superior to forecasts that are computed with averaged parameters and similarly for impulse response estimation. We will investigate these issues in our Monte Carlo study in Section 4. For both, model averaging and forecast combinations it is crucial to select the weights $\boldsymbol{w}$ in some sense optimally. One approach, also applicable to time series, is presented next.

2.5 Jackknife Model Averaging

Initially, as Hansen and Racine (2012) mention, Wolpert (1992) and Breiman (1996) introduced the idea of Jackknife model averaging (JMA) which is to use leave-one-out cross-validation to choose the weights $\boldsymbol{w}$. Hastie, Tibshirani, and Friedman (2009, Section 8.8) call this procedure stacking. However, only recently Hansen and Racine (2012) showed the asymptotic optimality of JMA "in the sense of achieving the lowest possible expected squared error over the class of linear estimators constructed from a countable set of weigths (Hansen and Racine, 2012, p. 36)". Their procedure requires independent observations but in contrast to alternative procedures mentioned by the authors allows for "bounded heteroscedasticity of unknown form" and an unbounded number of models. Zhang, Wan, and Zou (2013) showed the asymptotic optimality for a wider class of data generating processes including stochastic processes. It is for these reasons that we have chosen JMA for representing model averaging. As a side remark Zhang, Wan, and Zou (2013) do no longer require the weights to be taken from a discrete grid of points.

Next, we briefly describe the algorighm of JMA. Hansen and Racine (2012), hereafter HR, consider linear estimators for which $\hat{\boldsymbol{\mu}}_i = \boldsymbol{P}_i\boldsymbol{y}$ holds and where the $n \times n$ matrix $\boldsymbol{P}_i$ does not depend on $\boldsymbol{y}$. For least-squares estimation $\boldsymbol{P}_i = \boldsymbol{X}_i \left(\boldsymbol{X}_i'\boldsymbol{X}_i\right)^{-1} \boldsymbol{X}_i'$ is a projection matrix. Using this notation, the model averaging estimator for the conditional mean given by (12) for a given weight vector $\boldsymbol{w}$ is

$$\hat{\boldsymbol{y}}_{ma}(\boldsymbol{w}) = \sum_{i=1}^{m} w_i \boldsymbol{P}_i \boldsymbol{y} = \boldsymbol{P}(\boldsymbol{w})\boldsymbol{y}, \quad \boldsymbol{P}(\boldsymbol{w}) \equiv \sum_{i=1}^{m} w_i \boldsymbol{P}_i. \tag{15}$$

In order to estimate the weight vector $\boldsymbol{w}$ HR use Jackknife estimation by applying leave-one-out cross-validation or n-fold cross-validation (Hastie, Tibshirani, and Friedman, 2009, Section 7.10.1). This estimator, denoted by $\widetilde{y}_{t,i}$, estimates the conditional mean μ_t without using the observation $(y_t, \boldsymbol{x}_t)$. The corresponding projection matrix is denoted by $\widetilde{\boldsymbol{P}}_i$ which is identical to $\boldsymbol{P}_i$ except for zeros on the diagonal. For a given weight vector $\boldsymbol{w}$ the Jackknife estimator is then given by

$$\widetilde{\boldsymbol{y}}_{ma}(\boldsymbol{w}) = \sum_{i=1}^{m} w_i \widetilde{\boldsymbol{P}}_i \boldsymbol{y} = \widetilde{\boldsymbol{P}}(\boldsymbol{w})\boldsymbol{y}, \quad \widetilde{\boldsymbol{P}}(\boldsymbol{w}) \equiv \sum_{i=1}^{m_0} w_i \widetilde{\boldsymbol{P}}_i. \tag{16}$$

In order to determine $\boldsymbol{w}$ HR estimate the MSEP by

$$CV_n(\boldsymbol{w}) = ||\boldsymbol{y} - \widetilde{y}_{ma}(\boldsymbol{w})||^2/n \tag{17}$$

and minimize it w.r.t. the weight vector $\boldsymbol{w}$

$$\widehat{\boldsymbol{w}}_{jma} = \underset{\boldsymbol{w}\in\mathcal{H}_n}{\operatorname{argmin}}\ CV_n(\boldsymbol{w}), \tag{18}$$

where the set of possible weight vectors $\mathcal{H}_n$ is given by (11). Note that (17) is a quadratic function in $\boldsymbol{w}$ since $CV_n(\boldsymbol{w}) = \boldsymbol{w}^T\widetilde{\boldsymbol{u}}^T\widetilde{\boldsymbol{u}}\boldsymbol{w}/n$ with $\widetilde{\boldsymbol{u}} = (\widetilde{\boldsymbol{u}}_1, \ldots, \widetilde{\boldsymbol{u}}_m)$ and the Jackknife residuals $\widetilde{\boldsymbol{u}}_i = \boldsymbol{y} - \widetilde{\boldsymbol{P}}_i\boldsymbol{y}$. Due to the inequality constraints, this is a quadratic programming problem which is solved in the R language by the quadprog package.

Zhang, Wan, and Zou (2013, Section 3) derive conditions for dependent processes that guarantee the optimality of the JMA procedure where optimality is defined as

$$\frac{L_n(\widehat{\boldsymbol{w}}_{jma})}{\inf\limits_{\boldsymbol{w}\in\mathcal{H}_n} L_n(\boldsymbol{w})} \xrightarrow{p} 1, \tag{19}$$

where the quadratic loss is given by $L_n(\boldsymbol{w}) = (\boldsymbol{\mu} - \widehat{\boldsymbol{y}}_{ma}(\boldsymbol{w}))^T(\boldsymbol{\mu} - \widehat{\boldsymbol{y}}_{ma}(\boldsymbol{w}))$. These conditions include stationary homoskedastic finite-order AR processes.

Due to the quadratic nature of the loss function, no single weight will be estimated as exactly zero and therefore Jackknife model averaging is not designed for model selection. This is in contrast to the approach of the next subsection.

2.6 MCS-based Model Selection

The original aim of the MCS procedure is to estimate the set of superior models $\mathcal{M}^*$ given by (8) and thus to eliminate all inferior models from $\mathcal{M}^0$. In contrast to standard model selection procedures, it additionally allows to asymptotically control the family-wise error rate in the sequence of tests constituting the MCS procedure and thus to estimate the set $\mathcal{M}^*$ with a certain level of confidence, at least asymptotically. The final set should allow the conclusion that *a priori* it can be expected that, on average, all of the truly superior models are found in the estimated set $(1-\alpha)\%$ of the time, where α was fixed in advance.

To eliminate inferior objects from the set $\mathcal{M}^0$ a sequential testing procedure is used based on (7) with the following pair of hypotheses[1]

$$H_{0,\mathcal{M}} : \mu_{ij} \leq 0 \text{ for all } i,j \in \mathcal{M}, \mathcal{M} \subseteq \mathcal{M}^0, \tag{20}$$

$$H_{A,\mathcal{M}} : \mu_{ij} > 0 \text{ for some } i,j \in \mathcal{M}, \mathcal{M} \subseteq \mathcal{M}^0. \tag{21}$$

The hypotheses in (20) and (21) are indexed by $\mathcal{M}$ which emphasises the fact that both hypotheses refer to some generic set $\mathcal{M}$ that is a subset of $\mathcal{M}^0$ and may have been obtained by previous applications of the MCS testing procedure. The hypothesis $H_{0,\mathcal{M}}$ is true when $\mathcal{M} = \mathcal{M}^*$. The alternative is true when $\mathcal{M} = \mathcal{M}^* \cap (\mathcal{M}^0 \setminus \mathcal{M}^*)$. These

[1] Hansen, Lunde, and Nason (2011) specify the hypothesis slightly differently with $\mu_{ij} = 0$ versus $\mu_{ij} \neq 0$. This amounts to the same, however, since $\mu_{ij} \leq 0$ for all $i,j \in \mathcal{M}$ implies $\mu_{ij} = 0$.

two hypotheses can be used to find an estimate of $\mathcal{M}^*$ in a sequential manner: if $H_{0,\mathcal{M}}$ can be rejected, then remove a model from $\mathcal{M}$ and apply the hypothesis test again to the remaining models until the null hypothesis can no longer be rejected. The remaining set of models estimates $\mathcal{M}^*$ and is called Model Confidence Set (MCS) and denoted by $\widehat{\mathcal{M}}^*_{1-\alpha}$.

The algorithm to obtain the MCS is a sequential testing procedure in which two alternating tests are carried out. Let $\delta_{\mathcal{M}}$ be a binary variable that is associated with a suitable test for the null hypothesis (20) and which equals 1 if $H_{0,\mathcal{M}}$ is rejected and 0 if it is not rejected. Further, let $e_{\mathcal{M}}$ be the model that is removed if $\delta_{\mathcal{M}} = 1$. Hansen, Lunde, and Nason (2011) call $\delta_{\mathcal{M}}$ and $e_{\mathcal{M}}$ "equivalence test" and "elimination rule", respectively. Equipped with these tools the MCS procedure can be stated as follows.

Algorithm 2.1. (MCS procedure)

0. Start with $\mathcal{M} = \mathcal{M}^0$.
1. Test $H_{0,\mathcal{M}}$ using $\delta_{\mathcal{M}}$ at level α.
2. If $\delta_{\mathcal{M}} = 0$, set $\widehat{\mathcal{M}}^*_{1-\alpha} = \mathcal{M}$ and stop.
   If $\delta_{\mathcal{M}} = 1$, use $e_{\mathcal{M}}$ to remove a model and repeat from step 1.

For establishing that $\widehat{\mathcal{M}}^*_{1-\alpha}$ has an asymptotic coverage probability of $(1-\alpha)$, Hansen, Lunde, and Nason (2011, Assumption 1) state the following requirements. The equivalence test and elimination rule must be 'well behaved' in the sense that, asymptotically, (a) $H_{0,\mathcal{M}}$ is only rejected with probability less than or equal to α when it is true, (b) $H_{0,\mathcal{M}}$ is rejected with probability converging to one when it is false and (c) the probability of eliminating a superior model when $H_{0,\mathcal{M}}$ is false converges to zero. Assumptions (a) and (b) are relatively standard for conventional statistical hypothesis tests. Assumption (c) needs to be confirmed for any elimination rule that will be considered. Theorem 1 in Hansen, Lunde, and Nason (2011) shows that the coverage property of the MCS algorithm is asymptotically guaranteed as well as that the probability of the selected set to contain any inferior model approaches zero asymptotically. In other words, $\widehat{\mathcal{M}}^*_{1-\alpha}$ asymptotically includes all superior but no inferior models at the given confidence level.

Note that in this setting there is no allowance for the type I error to accumulate because asymptotically the sequential testing procedure ensures that the "first time a superior model is questioned by the elimination rule is when the equivalence test is applied to $\mathcal{M}^*$" (Hansen, Lunde, and Nason, 2011, p. 460). Thus, the FWER is asymptotically controlled at α. In the special case when $\mathcal{M}^*$ contains only a single model, Corollary 1 in Hansen, Lunde, and Nason (2011) states that the probability that $\widehat{\mathcal{M}}^*_{1-\alpha} = \mathcal{M}^*$ converges to one.

Since all stated assumptions concern the asymptotic behavior, the MCS procedure may well be oversized in finite samples. HLN devise a formal concept (HLN, Definition 3) which they call 'coherency' between the equivalence test and elimination rule. It requires that as long as there are inferior models in the set, the probability of removing a superior model must not be larger than in the case when there is no inferior model in the

set. In practice the assumption restricts the space of possible $\delta_{\mathcal{M}}$, $e_{\mathcal{M}}$ combinations to those where a rejection implies enough evidence that a specific model is inferior and can be eliminated. While the coherency requirement cannot assure that the family-wise error rate is controlled at α in finite samples, it contains the probability of removing superior models to an acceptable degree.

Next we state the specific equivalence and elimination rules used in Section 4. They are based on estimating the expected difference in losses $\mu_{ij} = E[d_{t,ij}]$ with $d_{t,ij} \equiv L_{t,i} - L_{t,j}$ underlying the set of superior models (7). Hansen, Lunde, and Nason (2011, Assumption 2) assume that the loss differences $d_{t,ij}$ are strictly stationary, α-mixing and fulfil some moment condition. This assumption will be met by our DGPs in our Monte Carlo study.

To estimate the expected losses μ_{ij}, the original sample $(y_t, \boldsymbol{x}_t)$, $t = 1, 2, \ldots, n$, is split into an estimation sample, $t = 1, 2, \ldots, n_e$, and a valuation sample, $t = n_e + 1, \ldots, n$. The former is used to obtain $\hat{\boldsymbol{\beta}}_i$ and the latter allows to estimate μ_{ij} by the relative sample loss statistic $\bar{d}_{ij} \equiv \bar{L}_i - \bar{L}_j$ with $\bar{L}_i = (n - n_e)^{-1} \sum_{t=n_e+1}^{n} L_{t,i}$ where we define $L_{t,i}$ by the quadratic loss of the one-step ahead prediction error $L_{t,i} = (y_t - \hat{y}_{t|t-1,i})^2$. In our study we apply the T-max and the T-min statistic as two alternatives for the equivalence test. Both are based on the relative sample loss statistic $\bar{d}_{i\cdot} \equiv \bar{L}_i - \bar{L}_\cdot$ with $\bar{L}_\cdot \equiv m^{-1} \sum_{i \in \mathcal{M}} \bar{L}_i$ and the following t-statistic

$$t_{i\cdot} \equiv \frac{\bar{d}_{i\cdot}}{\sqrt{\widehat{var}(\bar{d}_{i\cdot})}} \tag{22}$$

and are given by

$$T_{\max,\mathcal{M}} \equiv \max_{i \in \mathcal{M}} t_{i\cdot}, \tag{23}$$

$$T_{\min,\mathcal{M}} \equiv \min_{i \in \mathcal{M}} t_{i\cdot}. \tag{24}$$

The T-max statistic (23) fulfills the coherency rule and is recommended by Hansen, Lunde, and Nason (2011) for empirical work on the basis of their simulation results. However, by redoing their simulations Aka (2014) found that their simulation results were actually based on the T-min statistic (24). While the latter has good power, it violates the coherency condition stated above.[2] For the latter reason we include it in our simulation setup.

The corresponding elimination rules are given by

$$e_{\mathcal{M},T_{max}} = \arg\max_{i \in \mathcal{M}} t_{i\cdot}, \tag{25}$$

[2]We thank Peter Hansen and Asger Lunde for providing us with the source code of their Ox package. When redoing the simulations with the maximum statistic, the power turned out to be worse than for the maximum range statistic $T_{R,\mathcal{M}}$ which was also suggested by Hansen, Lunde, and Nason (2011) but requires $m(m-1)/2$ instead of $m - 1$ comparisons in case of the T-max and T-min statistics. Details can be found in Aka (2014, Section 4.1).

$$e_{\mathcal{M},T_{min}} = \arg\max_{i \in \mathcal{M}} t_i. \tag{26}$$

Both test statistics exhibit nonstandard distributions under the null hypothesis $H_{0,\mathcal{M}}$ which HLN approximate using a circular block bootstrap which also allows to compute $\widehat{Var}(\bar{d}_{i.})$. Details are given in the supplement of Hansen, Lunde, and Nason (2011). This completes the MCS procedure. In sum, the MCS procedure requires to choose a significance level α, an equivalence and elimination rule, the ratio $r = n_e/n$ of dividing the sample into the estimation and evaluation sample, the number of bootstrap replications and the block size for the bootstrap.

Once the MCS $\widehat{M}^*_{1-\alpha}$ is estimated and contains more than one model, which is typically the case, one has to decide how to proceed. Since according to the null hypothesis $H_{0,\mathcal{M}}$ all models exhibit the identical lowest risk, one can argue to use all models contained in the MCS $\widehat{\mathcal{M}}^*_{1-\alpha}$ in an identical way. This suggests to do model averaging and compute $\hat{b}_{ma}(\widehat{\boldsymbol{w}}_{T_{mcs}})$, $T_{mcs} \in \{T_{max}, T_{min}\}$, using a weight vector $\widehat{\boldsymbol{w}}_{T_{mcs}}$ that assigns equal weights to models in $\widehat{\mathcal{M}}^*_{1-\alpha}$ and zero weights to all other models:

$$\widehat{w}_{T_{mcs},i} \equiv \begin{cases} \#(\widehat{\mathcal{M}}^*_{1-\alpha})^{-1} & \text{if } i \in \widehat{\mathcal{M}}^*_{1-\alpha}, \\ 0 & \text{otherwise,} \end{cases} \tag{27}$$

$i = 1, 2, \ldots, m_0$, where $\#(\cdot)$ denotes the number of elements.

2.7 Shrinkage Methods

Shrinkage methods can also be called penalized estimation or estimation with a regularization term. The regularization limits the flexibility of the parameters and therefore allows to estimate models with a large number of parameters such as, for example, the largest model in $\mathcal{M}^0$ indexed by m_0. Specific shrinkage methods differ w.r.t. the basic estimator and the regularization term. The degree of regularization is controlled by the regularization parameter λ which has to be estimated. In the following we consider regularization of the OLS estimator.

The Ridge estimator $\hat{\boldsymbol{\beta}}_{m_0,ridge}$ is obtained by summing over squared parameter values (except the constant) and is available in matrix form

$$\hat{\boldsymbol{\beta}}_{m_0,ridge}(\lambda) = \underset{\boldsymbol{\beta}_{m_0}}{\operatorname{argmin}} ||\boldsymbol{y} - \boldsymbol{X}_{m_0}\boldsymbol{\beta}_{m_0}||^2 + \lambda \sum_{j=2}^{k_{m_0}} \beta^2_{m_0,j} \tag{28}$$

$$= \left(\boldsymbol{X}'_{m_0}\boldsymbol{X}_{m_0} + \lambda \boldsymbol{I}\right)^2 \boldsymbol{X}'_{m_0}\boldsymbol{y} \tag{29}$$

Increasing the regularization parameters λ implies a more restrictive estimator which may lead to larger bias and smaller variance. Since, independently from the value of the regularization parameter λ, the estimated $\hat{\beta}_{m_0,j}$ are different from zero with probability one, no model selection is conducted and all parameters of the largest model m_0 are

estimated. Therefore the Ridge estimator may be viewed as a restricted estimator of the largest model m_0 which is the encompassing model (5) if $\boldsymbol{\beta}_{m_0} = \boldsymbol{\beta}_{m_{all}}$.

The Lasso estimator, in contrast, is defined by summing over absolute values of the parameters values (except the constant)

$$\hat{\boldsymbol{\beta}}_{m_0,lasso}(\lambda) = \underset{\boldsymbol{\beta}_{m_0}}{\operatorname{argmin}} \frac{1}{2}||\boldsymbol{y} - \boldsymbol{X}_{m_0}\boldsymbol{\beta}_{m_0}||^2 + \lambda \sum_{j=2}^{k_{m_0}} |\beta_{m_0,j}| \tag{30}$$

Using absolute values allows for the possibility that some elements of $\hat{\boldsymbol{\beta}}_{m_0,lasso}$ (except the constant) can be estimated to be exactly zero which implies model selection. How many and which parameters are set to zero depends on the regularization parameter λ, among other things. The larger λ, the more zeros may occur. Note that the selected model may not be included in $\mathcal{M}^0$ except if it coincides with the complete set of possible models $\mathcal{M}^{all} = \{1, 2, \ldots, m_{all}\}$, see Section 2.1. Let $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_l, \ldots, \lambda_g)$ denote a $(g \times 1)$ vector of increasing λ-values. Then we denote the model selected by λ_l by $\hat{j}_{\lambda_l} \in \mathcal{M}^{all}$ and the corresponding Lasso estimator by $\hat{\boldsymbol{\beta}}_{\hat{j},lasso}(\lambda_l)$ or $\hat{\boldsymbol{b}}_{\hat{j}_{\lambda_l},lasso}$. It may well happen that different λ_l select the same model. Thus the number of different models selected $j_{\lambda_l}, l = 1, 2, \ldots, g$ may be smaller than g. We denote the set of different models implied by the vector of regularization vectors $\boldsymbol{\lambda}$ by

$$\widehat{\mathcal{M}}^0_{lasso}(\boldsymbol{\lambda}) \equiv \{j \in \mathcal{M}^{all} : j = \hat{j}_{\lambda_l}, l = 1, 2 \ldots, g\} \tag{31}$$

In order to select λ from $\boldsymbol{\lambda}$ we follow Hansen (2016, Section 4.1) and use the default settings in the R package **glmnet** and 5-fold cross-validation in **cv.glmnet**, see Hastie, Tibshirani, and Friedman (2009, Section 7.10) for an introduction to K-fold cross-validation. We therefore obtain the Lasso estimate exhibiting lowest estimated risk given $\boldsymbol{\lambda}$ by $\hat{\boldsymbol{\beta}}_{\hat{j}_{\hat{\lambda}},Lasso}(\hat{\lambda})$ or $\hat{\boldsymbol{b}}_{\hat{j}_{\hat{\lambda}},lasso}$. The same procedure for estimating λ is used for the ridge estimator (28) which delivers $\hat{\boldsymbol{\beta}}_{m_0,ridge}(\hat{\lambda})$.

As Hansen (2016) we include the post Lasso estimator of Belloni and Chernozhukov (2013) that re-estimates $\hat{\boldsymbol{\beta}}_{\hat{j}_{\hat{\lambda}},lasso}(\hat{\lambda})$ for the same model using OLS. We denote this parameter estimate by $\hat{\boldsymbol{\beta}}_{\hat{j}_{\hat{\lambda}},postlasso}$. Note that all estimators of this section estimate a single model but differ in the effect of the regularization.

2.8 Two New Suggestions

In Section we propose to apply Jackknife model averaging described in Section 2.5 to specific subsets of either the initial model set chosen $\mathcal{M}^0$ or the complete model set $\mathcal{M}^{all}$ in order to combine the advantages of the Jackknife model averaging with the methods delivering the subsets.

First, we suggest to apply JMA to model confidence sets described in Section 2.6. There we proposed to use equal weights to all models in the MCS $\widehat{\mathcal{M}}^*_{1-\alpha}$ in order to

reflect the idea of the underlying null hypothesis. In light of the property of the estimator of $\mathcal{M}^*$ to possibly include inferior models when there is insufficient information in the data, using equal weights may give inferior models a too large weight. As JMA is known to perform quite well also if inferior models are in the model set $\mathcal{M}^0$, we propose to apply it to all models contained in the MCS $\widehat{\mathcal{M}}^*_{1-\alpha}$ which delivers the MCS-based model averaging estimator

$$\hat{\boldsymbol{b}}_{ma}(\widehat{\boldsymbol{w}}_{T_{mcs},jma}) \equiv \sum_{i \in \widehat{\mathcal{M}}^*_{1-\alpha}} \hat{\boldsymbol{b}}_i \widehat{w}_{T_{mcs},jma,i}, \quad T_{mcs} \in \{T_{max}, T_{min}\} \tag{32}$$

and MCS-based forecast combinations

$$\hat{y}_{t+h|t,fc}(\widehat{\boldsymbol{w}}_{T_{mcs},jma}) \equiv \sum_{i \in \widehat{\mathcal{M}}^*_{1-\alpha}} \hat{y}_{t+h|t,i} \widehat{w}_{T_{mcs},jma,i}, \quad T_{mcs} \in \{T_{max}, T_{min}\} \tag{33}$$

The same idea can be applied to the set of models $\widehat{\mathcal{M}}^0_{lasso}$ implied by a vector $\boldsymbol{\lambda}$ of regularization parameters when using Lasso estimation. This estimated set defined by (31) typically contains models with differing bias-variance trade-offs. So instead of picking a single model using K-fold cross-validation as in Section 2.7, one can apply JMA to the estimated initial set $\widehat{\mathcal{M}}^0_{lasso}$ from which one obtains the Lasso-based model averaging estimator

$$\hat{\boldsymbol{b}}_{ma}(\widehat{\boldsymbol{w}}_{lasso,jma}) \equiv \sum_{j \in \widehat{\mathcal{M}}^0_{lasso}} \hat{\boldsymbol{b}}_j \hat{w}_{lasso,jma,j} \tag{34}$$

and Lasso-based forecast combinations

$$\hat{y}_{t+h|t,fc}(\widehat{\boldsymbol{w}}_{T_{mcs},jma}) \equiv \sum_{j \in \widehat{\mathcal{M}}^0_{lasso}} \hat{y}_{t+h|t,j} \hat{w}_{lasso,jma,j}, \tag{35}$$

where the indexation of all models is given by $j \in \mathcal{M}^{all}$ and therefore $\widehat{\mathcal{M}}^0_{lasso} \in \mathcal{M}^{all}$. In this Section 2 we described and suggested 18 different possibilities to compute h-step ahead predictions. Their performance will be compared in Section 4 based on the setup described in the following section.

3 Design of Monte Carlo Simulation

3.1 Data Generation

This section describes the way we generate the artificial data which is used in the Monte Carlo simulation. We focus on univariate linear autoregressive processes as the data

generating process (DGP):

$$\alpha_0(L)y_t = \nu_0 + \varepsilon_t, \quad \varepsilon_t|y_{t-1}, y_{t-2}, \ldots \sim (0, \sigma_0^2), \quad t = 1, 2, \ldots, n, \tag{36}$$

$$\alpha_0(L) = \alpha_{1,0}L + \alpha_{2,0}L^2 + \cdots \alpha_{p_0,0}L^{p_0} \tag{37}$$

where $\alpha_0(L)$ denotes the p_0-order lag polynomial with the specific set of lags of the DGP. All models which are considered for fitting the data are finite-order autoregressive models with different lag polynomials $\alpha_i(L) = \alpha_{1,i}L + \alpha_{2,i}L^2 + \cdots \alpha_{p_i,i}L^{p_i}$ and possibly a constant ν_i.

One objective of our simulation exercise is to gauge the ability of the various methods to identify not just a maximum lag order p, but to recover a DGP with a strict subset of lags up to order p. We therefore select DGPs from the set of processes that have non-zero coefficients for lags one, six, and seven and zero coefficients for all others. To obtain DGPs that may be comparable to those found in applied work, we take into account three properties of any given process: the signal-to-noise ratio, the roots of the autoregressive polynomial $\alpha(L)$, and the frequency properties of the process. Denote the variance of y_t as σ_y^2 and the variance of the error term as σ_ε^2. The precision (i.e. the inverse of the variance) of the OLS estimator is increasing in $\sigma_y^2/\sigma_\epsilon^2$, which we call here the signal-to-noise ratio (SNR). We follow Zhang, Wan, and Zou (2013, p. 87) and convert this ratio to the probability limit of the more usual goodness-of-fit measure R^2 of the correct model which is $1 - 1/\text{SNR}$ and fix the latter in our simulations to control the finite sample estimation precision. We constrain the roots of the autoregressive polynomial to be greater than some value $\eta > 1$ to ensure stationarity. We further analyse the spectral density of any candidate process and disregard those that derive the majority of their variance from extremely high or low frequencies. For finding specific processes that fulfil all properties, we use a constrained optimisation algorithm. The objective function is given by the deviation of any process' R^2 from the targeted R^2, while the roots provide an inequality constraint. At this stage all processes are filtered such that they have the desired frequency properties. From the remaining processes we pick one at random.

In the simulation, we let R^2 vary between $0.2, 0.5$ and 0.8 and set $\eta = 1.1$, $\sigma_\epsilon^2 = 1$ and $\nu_0 = 0$ throughout. We thereby obtain processes as shown in Table 1. Each row presents one autoregressive process whereas the columns indicate the corresponding R^2, the length of the smallest root of the autoregressive lag polynomial and the values of the autoregressive coefficients at lags one, six and seven. The last three columns show the proportion of the variance attributable to frequencies below or equal to $\frac{1}{4}\pi$, $\frac{2}{4}\pi$ and $\frac{3}{4}\pi$. As the table shows, the inequality constraint is never binding as all three processes have their smallest root fairly close to, yet above, 1.1. The spectral densities indicate that as the R^2 increases there is also a shift in weight to lower frequencies.

Table 1: Characteristics of the three chosen data-generating processes

R^2	Smallest Root	Coefficients			Spectral Density		
		α_1	α_6	α_7	$\frac{1}{4}\pi$	$\frac{2}{4}\pi$	$\frac{3}{4}\pi$
0.2	1.105	0.214	-0.265	0.380	0.33	0.55	0.78
0.5	1.138	0.671	-0.422	0.342	0.73	0.85	0.93
0.8	1.131	0.855	0.338	-0.310	0.89	0.96	0.99

Notes: R^2 denotes the probability limit of R^2 for the corresponding correct model. The last three columns show the proportion of the variance attributable to frequencies below or equal to $\frac{1}{4}\pi$, $\frac{2}{4}\pi$ and $\frac{3}{4}\pi$. Values are rounded to third or second decimal place.

3.2 Choice of initial model set and auxiliary parameters

The encompassing model of the initial collection of models $\mathcal{M}^0$ is an AR(8) model including constant. The initial model set $\mathcal{M}^0$ contains all subset AR models obtained by considering all possible combinations of zero restrictions on the lagged variables. This delivers $m_0 = 2^8 = 256$ models in $\mathcal{M}^0$ since a constant is always included. We consider six small to medium sized samples with $n = 40, 60, 80, 100, 250, 500$ each with 50 burn-in observations. The h-step ahead forecasts are computed for $h = 1, 2, \ldots, 15$ and the impulse respones coefficients for $h = 1, 2, \ldots, 20$. The number of replications is 5000.

The JMA procedure is fully described in Section 2.5. The JMA weights $\hat{\boldsymbol{w}}_{jma}$ (18) are used for both model averaging and forecast combinations. The MCS algorithm presented in Section 2.6 requires to choose several auxiliary parameters. The significance level is set to 0.2. This may seem rather high but ensures decent power to eliminate inferior models in smaller samples. The base block length in the circular block bootstrap is set to 20 but is automatically modified by the function b.star in the R package **np**. The number of bootstrap replications is set to 1000 and the ratio of the estimation to the full sample determined by $n_e/n = 1e^{-1/5}$ which is rounded to the next larger integer. This formula reflects preliminary investigations of the authors that with increasing sample size a larger fraction of the data should be used for computing the losses by evaluating the one-step out-of-sample forecasts. Finally, both the T-max (23) and the T-min statistic (24) are used for testing equivalence. The equal weights as well as the JMA based weights are both used for model averaging and forecast combinations. The choice of the grid of regularization parameters and its optimal choice is described in Section 2.7.

4 Results

In this section we summarize the main findings from our Monte Carlo simulation using the setup described in Section 3.

For evaluating and comparing all 18 methods we use the following summary measures. They are all based on estimating the MSEP for h-step ahead predictions of y_{n+h} and the RMSE for impulse response functions ϕ_h by averaging across all $r = 1, 2, \ldots, R$ simulation runs

$$\widehat{MSEP}(y_{n+h}, h, s) = \sum_{r=1}^{R} \left(\hat{y}_{n+h|n,r,s} - y_{n+h,r}\right)^2, \tag{38}$$

$$\widehat{RMSE}(\phi_h, h, s) = \sqrt{\sum_{r=1}^{R} \left(\hat{\phi}_{h,r,s} - \phi_h\right)^2}, \tag{39}$$

where s denotes one of the methods listed in the first column of Table 2. To succinctly summarize the results, we average across all H horizons. To avoid scaling effects, we average over relative MSEP's by relating the MSEP for method s and horizon h to the corresponding MSEP when the DGP is known

$$\widehat{RelMSEP}(y_{t+h}, h, s) \equiv \frac{\widehat{MSEP}(y_{t+h}, h, s) - \widehat{MSEP}(y_{t+h}, h, DGP)}{\widehat{MSEP}(y_{t+h}, h, DGP)}, \tag{40}$$

$$\widehat{AvRelMSEP}(y_{t+h}, H, s) \equiv H^{-1} \sum_{h=1}^{H} \widehat{RelMSEP}(y_{t+h}, h, s). \tag{41}$$

Since there is no uncertainty about the impulse responses of the DGP, we use the RMSE of estimating the impulse response values based on the correct model i_{DGP}, which only includes the lags of the DGP, to obtain relative quantities

$$\widehat{RelRMSE}(\phi_h, h, s) \equiv \frac{\widehat{RMSE}(\phi_h, h, s) - \widehat{RMSE}(\phi_h, h, i_{DGP})}{\widehat{RMSE}(\phi_h, h, i_{DGP})}, \tag{42}$$

$$\widehat{AverRelRMSE}(\phi_h, H, s) \equiv H^{-1} \sum_{h=1}^{H} \widehat{RelRMSE}(\phi_h, h, s). \tag{43}$$

For comparing all 18 methods we adopt a two-stage procedure. First, we categorize the methods into three groups and look at each group individually. At the second stage we pool together the best performing methods from each group and judge their overall merits. The groups are formed as follows. The first group consists of those methods that perform model selection by placing all weight on a single model. For this group model averaging and forecast combinations are identical. The members of this group are listed in the top half of Table 2. The other two groups consist of those methods that place weight on more than one model and apply either model averaging (second group) or forecast combinations (third group). The bottom half of Table 2 lists both of these groups. At the

Table 2: Description of all methods used in simulation study.

Method	Model Averaging	Forecast Combination	Descriptions
aic	$\hat{\boldsymbol{\beta}}_{\hat{i}_{AIC}}$		Section 2.2
hq	$\hat{\boldsymbol{\beta}}_{\hat{i}_{HQ}}$		Section 2.2
lasso	$\hat{\boldsymbol{\beta}}_{\hat{j}_{\hat{\lambda}},lasso}(\hat{\lambda})$		Section 2.7
postlasso	$\hat{\boldsymbol{\beta}}_{\hat{j}_{\hat{\lambda}},postlasso}$		Section 2.7
ridge	$\hat{\boldsymbol{\beta}}_{m_0,ridge}(\hat{\lambda})$		Section 2.7
sc	$\hat{\boldsymbol{\beta}}_{\hat{i}_{SC}}$		Section 2.2
jma	$\hat{\boldsymbol{b}}_{ma}(\hat{\boldsymbol{w}}_{jma})$	$\hat{y}_{t+h\|t,fc}(\hat{\boldsymbol{w}}_{jma})$	Eq. (10), (18) / Eq. (13), (18)
lassojma	$\hat{\boldsymbol{b}}_{ma}(\hat{\boldsymbol{w}}_{JMA})$	$\hat{y}_{t+h\|t,fc}(\hat{\boldsymbol{w}}_{JMA})$	Eq. (34) / Eq. (35)
in mcs_t.max_h1h1	$\hat{\boldsymbol{b}}_{ma}(\hat{\boldsymbol{w}}_{T_{max}})$	$\hat{y}_{t+h\|t,fc}(\hat{\boldsymbol{w}}_{T_{max}})$	Eq. (10), (27) / Eq. (13), (27)
mcs_t.min_h1h1	$\hat{\boldsymbol{b}}_{ma}(\hat{\boldsymbol{w}}_{T_{min}})$	$\hat{y}_{t+h\|t,fc}(\hat{\boldsymbol{w}}_{T_{min}})$	Eq. (10), (27) / Eq. (13), (27)
mcsjma_t.max_h1h1	$\hat{\boldsymbol{b}}_{ma}(\widehat{\boldsymbol{w}}_{T_{max},jma})$	$\hat{y}_{t+h\|t,fc}(\widehat{\boldsymbol{w}}_{T_{max},jma})$	Eq. (32) / Eq. (33)
mcsjma_t.min_h1h1	$\hat{\boldsymbol{b}}_{ma}(\widehat{\boldsymbol{w}}_{T_{min},jma})$	$\hat{y}_{t+h\|t,fc}(\widehat{\boldsymbol{w}}_{T_{min},jma})$	Eq. (32) / Eq. (33)

second stage we choose two methods from each group. The criteria for this choice are:

i) one method has to be the best one for either large or small sample sizes for at least two DGPs and never be the worst one and

ii) in case there are more than two in step i) choose those with the better overall performance.

Having suitable measures and a structured procedure for comparison at hand, we look at the results for h-step ahead predictions and for impulse response analysis in turn.

4.1 Predictions

We illustrate the results by plotting the averaged relative MSEPs (41) for all sample sizes and DGPs. Figures 1 to 3 contain the first-stage results for each of the three groups. The three panels in each graph correspond to one of the three DGPs. Our performance measure, the averaged relative MSEPs, on the y-axis is plotted against the sample size on the x-axis. In all three graphs we see a clear trend towards zero as the sample size increases. This is as expected, because all methods produce consistent forecasts and thus the relative differences between the estimated forecasts and the DGP forecasts vanish as $n \to \infty$.

Looking at Figure 1 for the single-model methods a clear pattern emerges. Lasso and Ridge have an advantage at small sample sizes and perform between 5 and 15 percentage points better than post Lasso or ordinary information criteria. For larger sample sizes

Ridge looses its advantage, although the differences diminish, and the Schwarz criterion has a slight edge over the other methods. Thus, for this group we pick Lasso and Schwarz as the winners.

For the second, model-averaging group, Figure 2 reveals a similarly clear pattern. For smaller sample sizes and across DGPs, MCS using the T-max statistic has an advantage. Yet, with increasing population R^2 and sample size, JMA starts to dominate MCS and the other methods. Because of too little power, MCS based on the T-max statistic is not able to remove inferior models as accurately as the other methods are capable of when the sample size is large and the signal is low, which in turn leads to inferior forecasting performance probably because of the equal weighting of included inferior models. Vice versa, for small sample sizes and low population R^2, the estimated model confidence set using the T-max statistic is likely to be equivalent to averaging across the initial model set $\mathcal{M}^0$. So what this result tells us is that in those situations it is better to average across almost all models than to apply one of the other averaging techniques. With the exception of MCS using equal weights, all other methods result from applying, first, a model selection tool, like Lasso on a grid of penalization terms or MCS, and subsequently applying JMA to the remaining (smaller) group of models. So another insight is that for forecasting purposes forming these combinations does not yield an advantage over simply applying JMA to $\mathcal{M}^0$. The combinations are either inferior or equivalent to JMA itself but never superior. Thus, for the second group we pick MCS based on the T-max-statistic and JMA as winners.

For the DGPs that we have chosen, model averaging and forecast combinations do not make a large difference with respect to forecasting, as can be seen when comparing figures 2 and 3. The ordering of the third group is therefore the same as in the second group and we pick MCS based on the T-max-statistic and JMA as winners.

Figure 4 depicts the key results. All six winners of the first-stage compete against each other, which means that some of the methods appear as their forecast combination (fc) version as well as model averaging (ma) version. However, as noted before, the respective versions do not make a difference for the chosen DGPs. Thus, the overall results indicate that using MCS or, equivalently, weighting each model in $\mathcal{M}^0$ equally is useful in small samples, whereas for a higher observation-to-parameter ratio JMA is the best performing method.

4.2 Impulse Response Analysis

The results for impulse response analysis are not as clear-cut as for predictions. The ranking of methods is more heterogeneous, depending on sample size and DGP. The bottom line is, however, that for small samples MCSJMA with the T-max-statistic performs fairly well and in larger samples the Schwarz criterion is the best choice across DGPs. Another dominant feature is that several averaged relative ratios increase with sample size. This is related to the fact that the denominator in (42) approaches zero with $n \to \infty$ while the

denominator in (40) converges to a constant.

The group of single-model methods in Figure 5 indicates a similar performance pattern as in the prediction exercise. In small samples, Lasso and Rigde do well, in larger samples Schwarz outperforms all others. This is consistent across DGPs. In large samples, AIC, HQ and post Lasso perform between 20 and 200 percentage points worse than Schwarz does and are therefore markedly inferior by our measure. The RMSE of Ridge seems to converge at a completely different rate than that of the true model. We therefore rank it lower than Lasso and choose Lasso and Schwarz as winners.

When it comes to model averaging (Figure 6), there is no clear indication which of the methods are performing consistently well across DGPs at either large or small sample sizes. In small samples MCS based on the T-max-statistic performs well for population $R^2 = 0.2$ and 0.5 but shows very poor performance for $R^2 = 0.8$. Similarly, Lasso-JMA and JMA do well for some signal-to-noise ratios, but not for all, so there is no conclusive evidence. Methods without large deviations away from zero, on the other hand, are given by the MCS using the T-min-statistic and by combining MCS and JMA, even though they do not sizeably outperform other methods. These methods are therefore fairly robust tools, which we will pick for later comparison.

Comparing model averaging and forecast combinations yields some difference, especially for smaller sample sizes. In particular for a population $R^2 = 0.8$ one has a substantially worse performance of the methods including (almost) all models such as JMA or MCSJMA with the T-max-statistic. This can be attributed to fact that we cannot use weights optimal for combining impulse responses. This is in contrast to out-of-sample predictions, where in our case the two approaches hardly made a difference. Nevertheless, the ranking between the methods remains similar. We have, therefore, heterogeneous results across DGPs and sample sizes for all methods except the combination of MCS and JMA, which is again fairly robust across DGPs and we therefore pick both versions for the next round.

Overall, we see in Figure 8 that when the sample size is relatively large, using the Schwarz criterion for selecting models in the context of impulse response analysis dominates all other methods. Especially when faced with a low-signal stochastic process and when the observation-to-parameter ratio is above 30, does Schwarz outperform other approaches by 20 to 200 percentage points. In contrast, when the observation-to-parameter ratio is small, so between 4 and 12, then Schwarz is dominated by essentially all other methods. Our results indicate that in small sample situations relying on model averaging via MCSJMA using the T-max-statistic is the most robust method across DGPs.

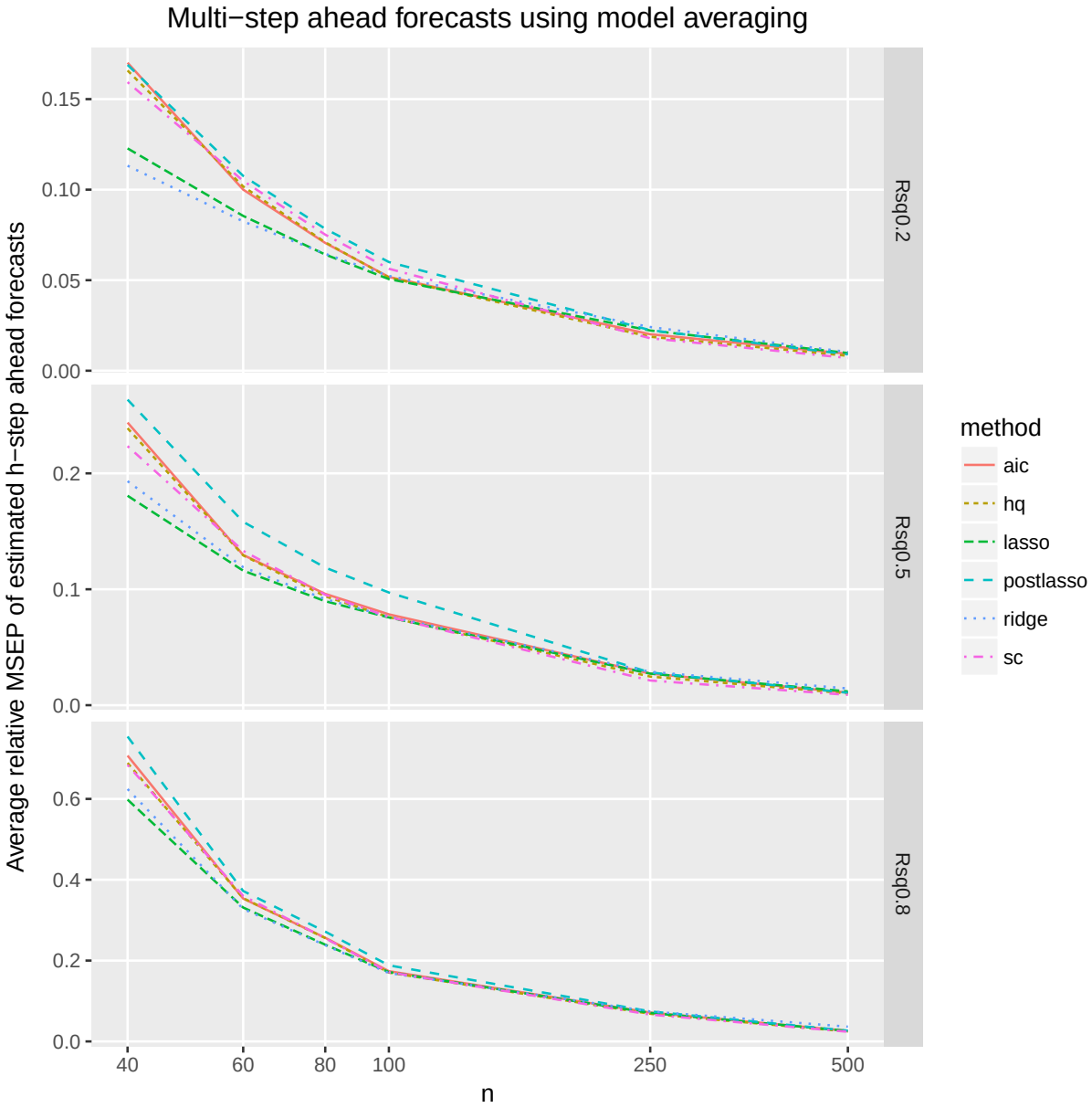

Figure 1: Averaged relative MSEPs for h-step ahead predictions for methods selecting and estimating a single model

Notes: Each line shows the average relative mean squared error of prediction (41) and is based on $R = 5000$ replications. The DGPs are AR(8) processes with zero restrictions and differ w.r.t. their signal-to-noise ratio. Their specification is given in Table 1. References for the methods used are given in Table 2. The various auxiliary parameters chosen for some methods are described in Section 3.2. All simulations are carried out in R.

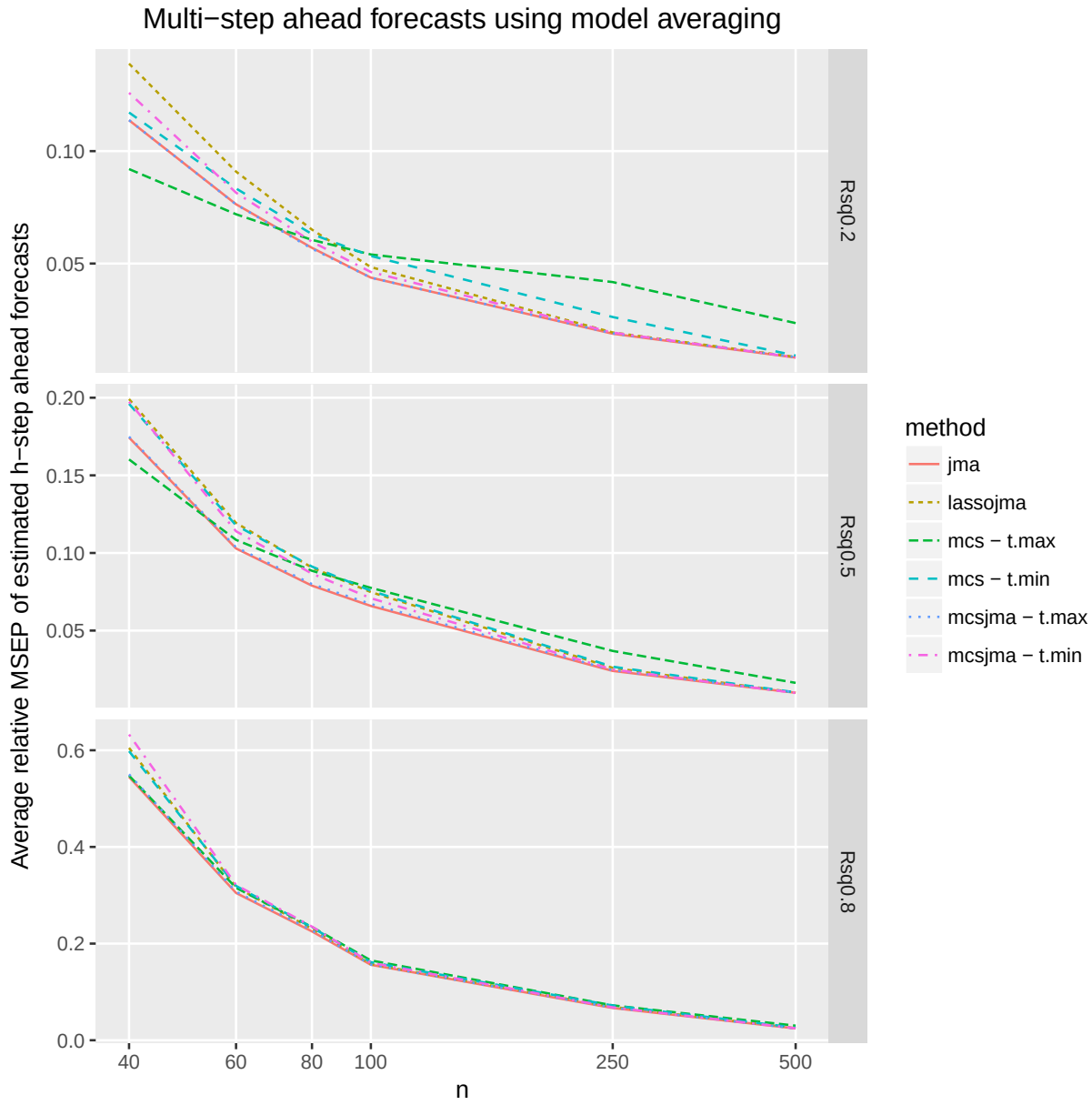

Figure 2: Averaged relative MSEPs for h-step ahead predictions using methods using model avering

Notes: see Figure 1.

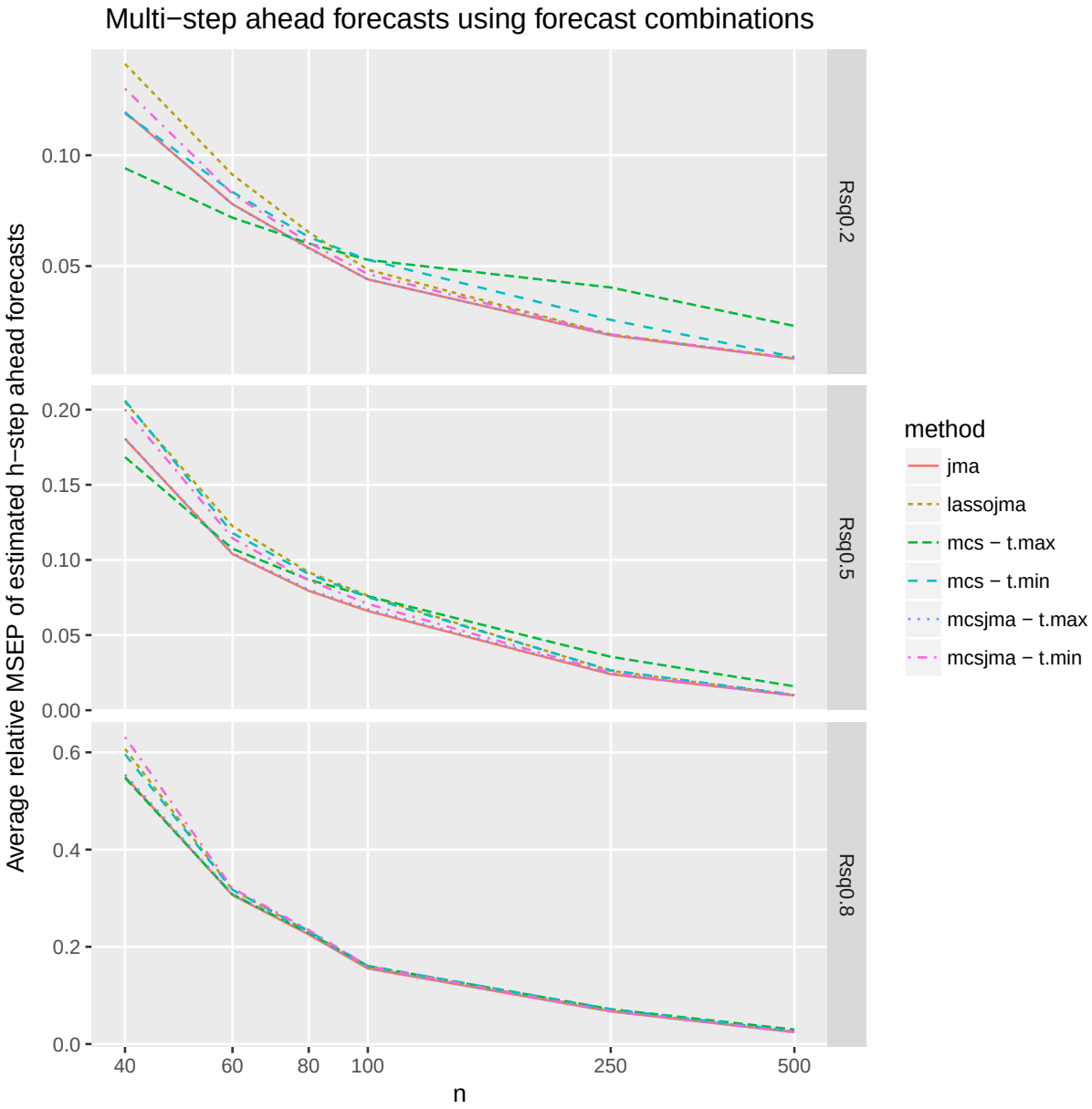

Figure 3: Averaged relative MSEPs for h-step ahead predictions for methods using forecast combinations

Notes: Notes: see Figure 1.

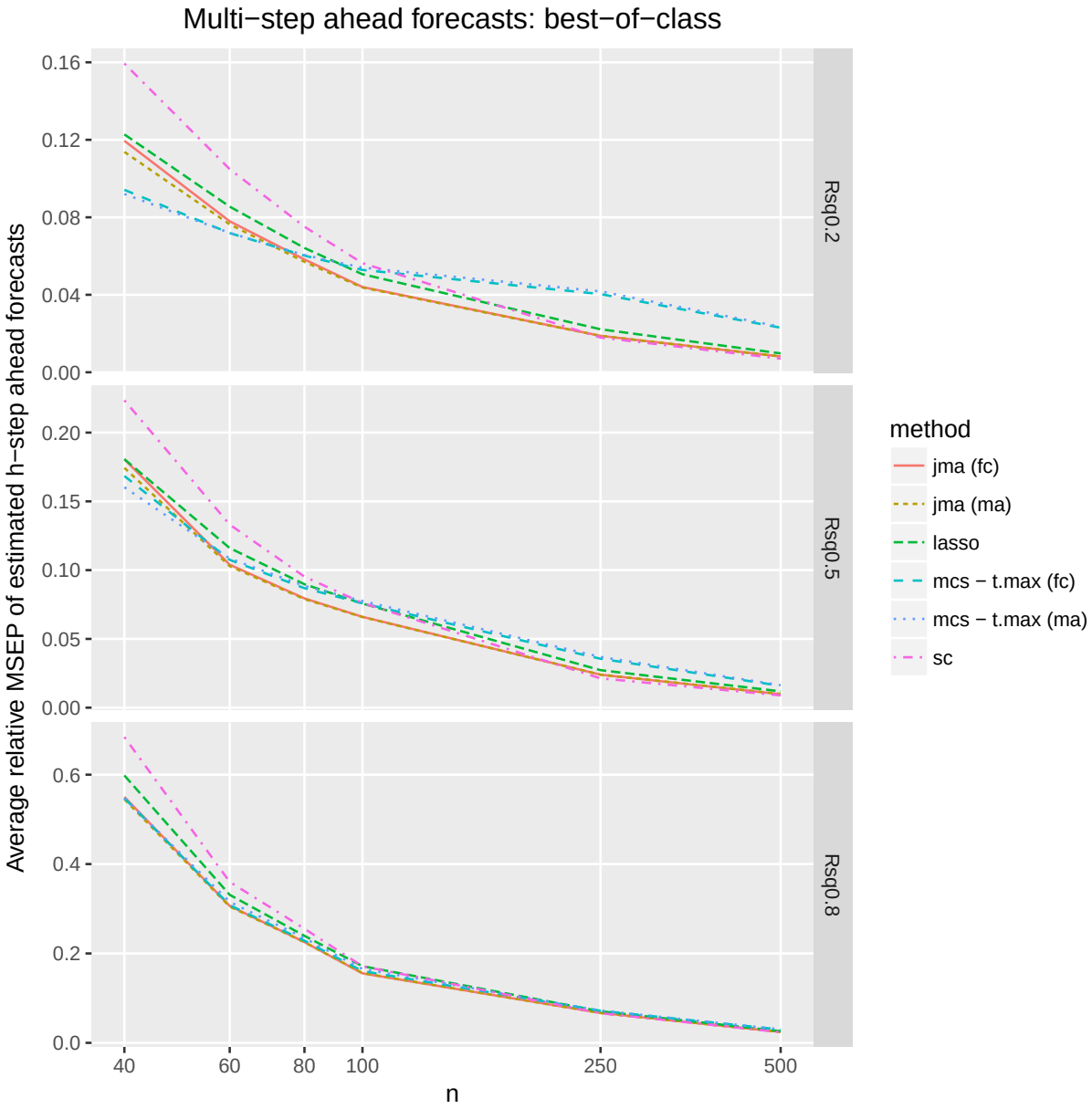

Figure 4: Averaged relative MSEPs for h-step ahead predictions for best performing methods.

Notes: Notes: see Figure 1.

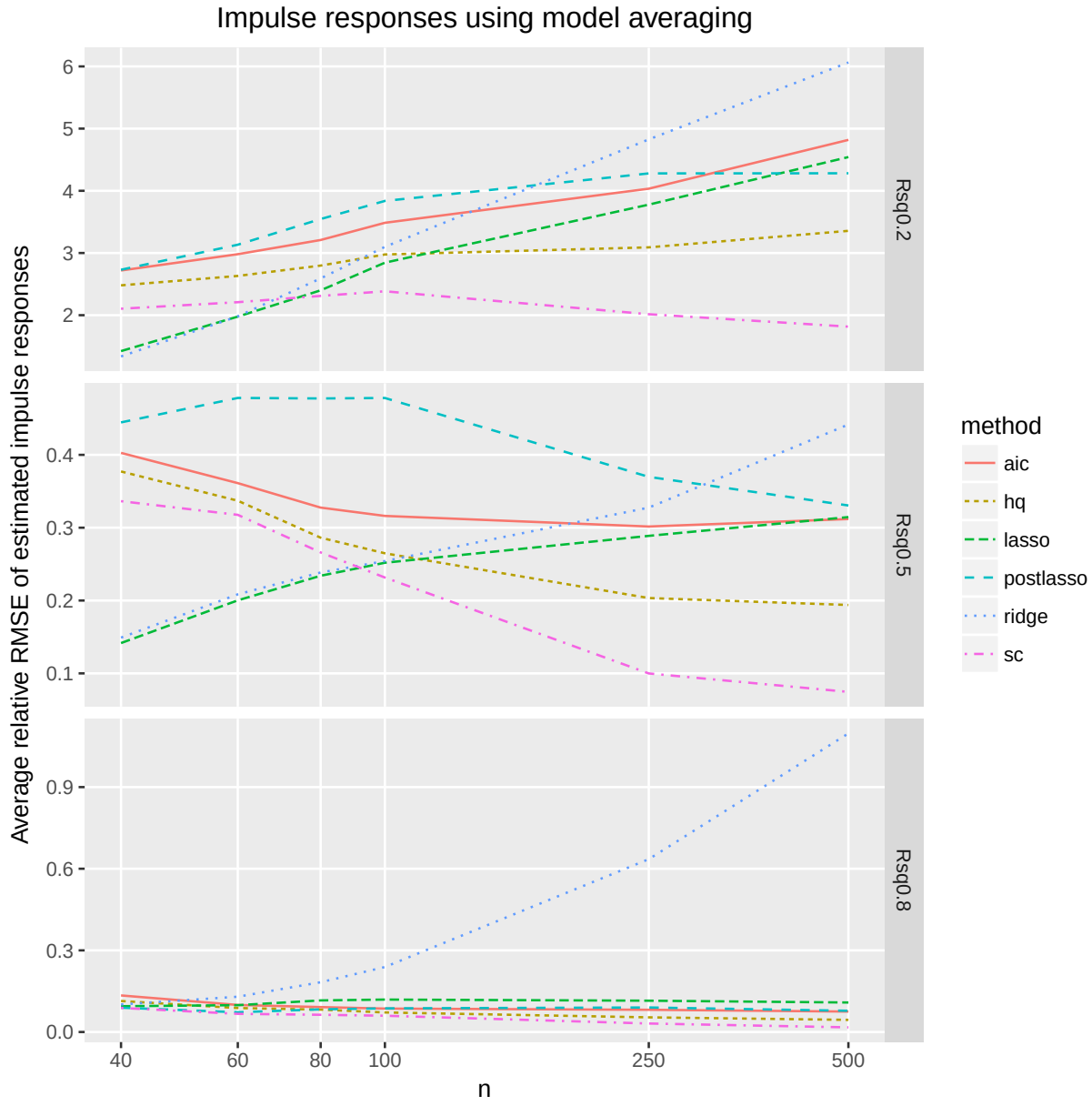

Figure 5: Averaged relative MSEs for impulse responses for methods selecting and estimating a single model

Notes: Each line shows the average relative mean squared error of (43) and is based on $R = 5000$ replications. The DGPs are AR(8) processes with zero restrictions and differ w.r.t. their signal-to-noise ratio. Their specification is given in Table 1. References for the methods used are given in Table 2. The various auxiliary parameters chosen for some methods are contained in Section 3.2. All simulations are carried out in R.

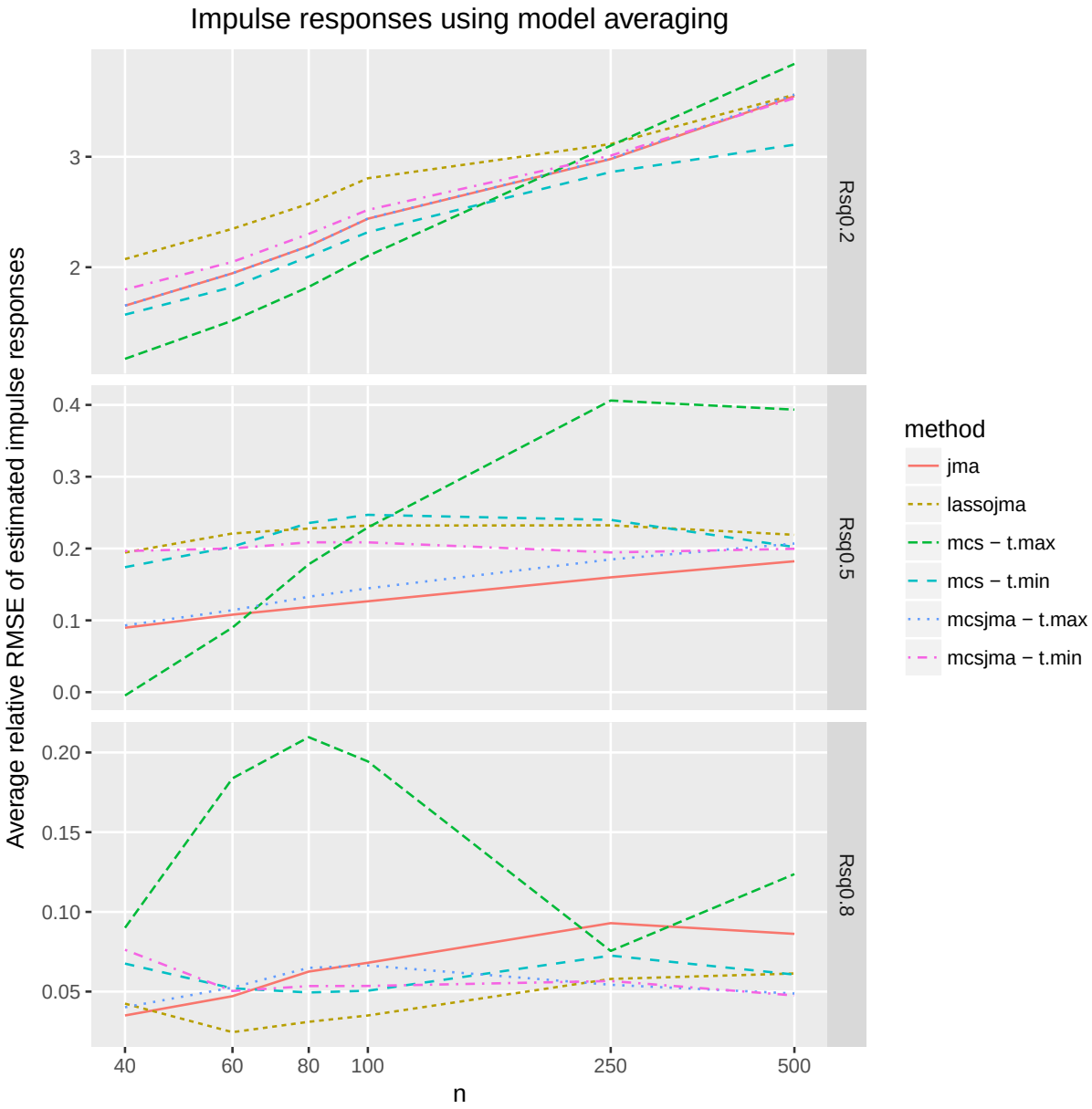

Figure 6: Averaged relative MSEPs for h-step ahead predictions using methods using model avering

Notes: see Figure 5.

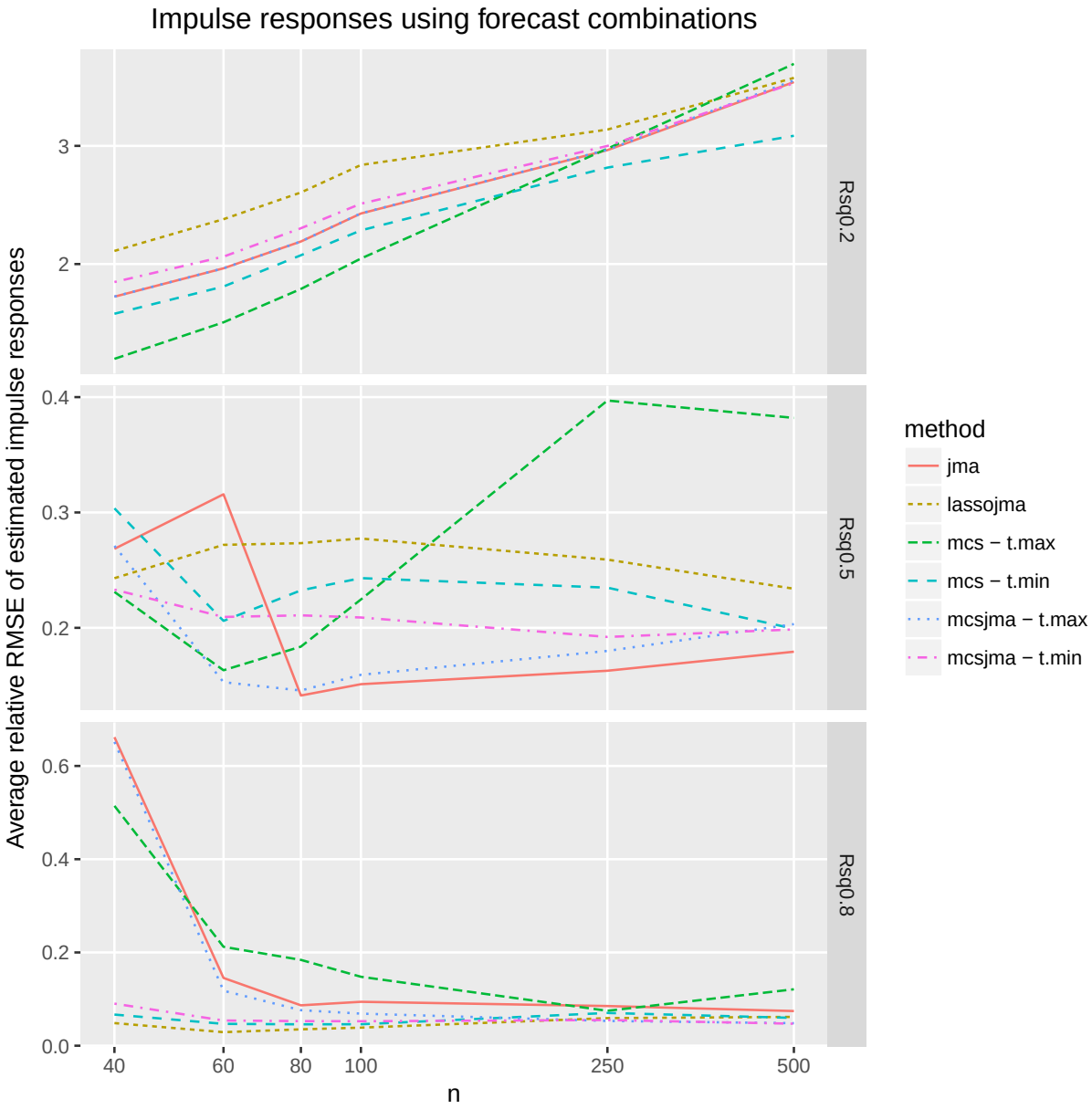

Figure 7: Averaged relative MSEPs for h-step ahead predictions for methods using forecast combinations

Notes: see Figure 5.

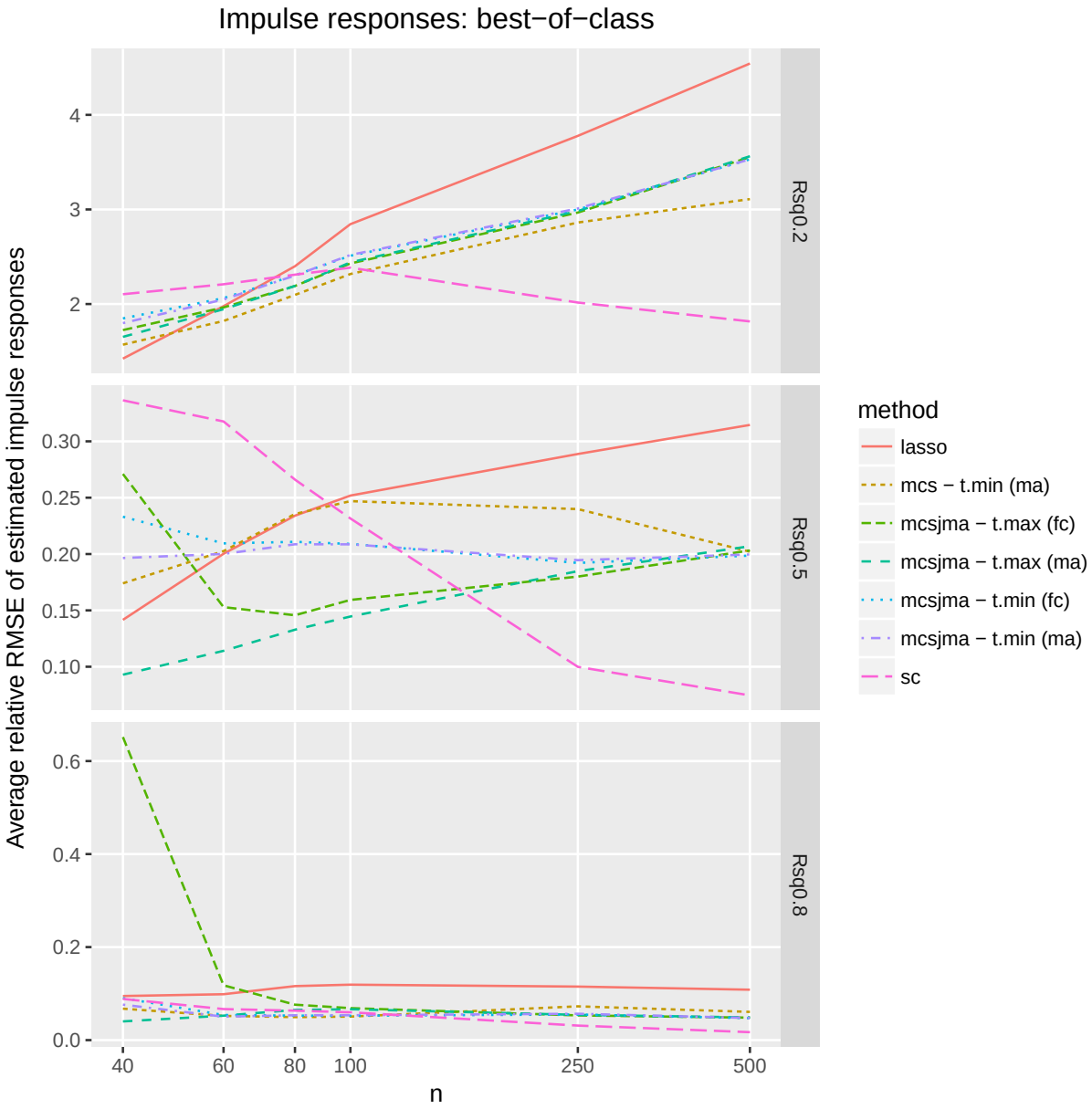

Figure 8: Averaged relative MSEPs for h-step ahead predictions for methods using forecast combinations

Notes: see Figure 5.

5 Conclusion

An important advantage of using model confidence sets is the asymptotic control of the family-wise error rate. In this paper we propose ways how to use the potentialy large number of models in the model confidence set for further analysis. One possibility is to apply model averaging to all models in the model confidence set. This amounts to averaging across the parameters of each estimated model in the set. Another possibility is to compute forecast combinations, or more generally, to combine the quantities of interest computed for each model in the model confidence set. As weighting schemes we consider equal weights and Jackknife model averaging.

In order to evaluate these suggestions, we conduct an extensive Monte Carlo simulation using three autoregressive processes, six sample sizes and in total 18 competing methods, among them Jackknife model averaging as well as standard model selection methods and the Lasso.

Our Monte Carlo results suggest that using the Schwarz criterion works well in larger samples but may perform poorly in small samples, in particular for impulse response estimation. In the latter case, estimating the MCS with the T-max statistic and applying JMA to the remaining models turns out to be a robust strategy that is found among the best strategies in small samples and performs comparable to the best competitors excluding Schwarz in larger samples. It is also found that for computing impulse responses, model averaging is superior to combining impulse responses of each model in the MCS.

Subsequent research could investigate the potential of the proposed methods for lag selection in multivariate time series models.

References

Aiolfi, M., C. Capistrán, and A. Timmermann (2011), "Forecast Combinations," in: M. P. Clements and D. F. Hendry (eds.), *The Oxford Handbook of Economic Forecasting*, chap. 12, Oxford University Press, pp. 355–390.

Aka, N. (2014), "Model Selection Sets: Theory, Simulations and an Application," Master thesis, University of Regensburg.

Belloni, A., and V. Chernozhukov (2013), "Least squares after model selection in high-dimensional sparse models," *Bernoulli*, 19(2), 521–547, doi:10.3150/11-BEJ410, URL `http://dx.doi.org/10.3150/11-BEJ410`.

Breiman, L. (1996), "Stacked regressions," *Machine Learning*, 24(1), 49–64, doi:10.1007/BF00117832, URL `http://dx.doi.org/10.1007/BF00117832`.

Claeskens, G., and N. L. Hjort (2008), *Model Selection and Model Averaging*, Cambridge University Press, Cambridge, URL `http://ideas.repec.org/b/cup/cbooks/9780521852258.html`.

Demetrescu, M., U. Hassler, and V. Kuzin (2011), "Pitfalls of post-model-selection testing: experimental quantification," *Empirical Economics*, 40(2), 359–372, doi:10.1007/s00181-009-0334-2, URL http://dx.doi.org/10.1007/s00181-009-0334-2.

Elliot, G., and A. Timmermann (2016), *Economic Forecasting*, Princeton University Press.

Hansen, B. (2016), "The Risk of James-Stein and Lasso Shrinkage," *Econometric Reviews*.

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

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

Hastie, T., R. Tibshirani, and J. Friedman (2009), *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*, Springer.

Leamer, E. E. (1978), *Specification Searches: Ad Hoc Inference with Nonexperimental Data*, Wiley.

Leeb, H., and B. M. Pötscher (2005), "Model Selection and Inference: Facts and Fiction," *Econometric Theory*, 21, 21–59, doi:10.1017/S0266466605050036.

Moral-Benito, E. (2015), "Model Averaging In Economics: An Overview," *Journal of Economic Surveys*, 29(1), 46–75, doi:10.1111/joes.12044, URL http://dx.doi.org/10.1111/joes.12044.

Romano, J. P., and M. Wolf (2005), "Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing," *Journal of the American Statistical Association*, 100(469), 94–108, doi:10.1198/016214504000000539.

Theil, H. (1957), "Specification Errors and the Estimation of Economic Relationships," *Review of the International Statistical Institute*, 25(1/3), 41–51.

Timmermann, A. (2006), "Forecast Combinations," in: G. Elliot, C. W. Granger, and A. Timmermann (eds.), *Handbook of Economic Forecasting*, vol. 1, chap. 4, Elsevier, pp. 135–196.

Wolpert, D. H. (1992), "Stacked generalization," *Neural Networks*, 5(2), 241 – 259, doi:http://dx.doi.org/10.1016/S0893-6080(05)80023-1, URL http://www.sciencedirect.com/science/article/pii/S0893608005800231.

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