Prediction of Linear Models: Application of Jackknife Model Averaging

By Asterios Geroukis

Department of Statistics
Uppsala University

Supervisor: Per Johansson

2016
Abstract

When using linear models, a common practice is to find the single best model fit used in predictions. This on the other hand can cause potential problems such as misspecification and sometimes even wrong models due to spurious regression. Another method of predicting models introduced in this study as Jackknife Model Averaging developed by Hansen & Racine (2012). This assigns weights to all possible models one could use and allows the data to have heteroscedastic errors. This model averaging estimator is compared to the Mallows’s Model Averaging (Hansen, 2007) and model selection by Bayesian Information Criterion and Mallows’s Cp. The results show that the Jackknife Model Averaging technique gives less prediction errors compared to the other methods of model prediction. This study concludes that the Jackknife Model Averaging technique might be a useful choice when predicting data.

Keywords
## Contents

1. Introduction .................................................................................................................. 1
2. Jackknife model averaging and model selection ............................................................... 2
3. Method ............................................................................................................................ 5
   3.1 Data 1 – Wages ........................................................................................................ 6
   3.2 Data 2 – Economic Growth ...................................................................................... 7
   3.3 Evaluation ............................................................................................................... 8
4. Results .......................................................................................................................... 9
   4.1 Wage data set ........................................................................................................... 9
   4.2 Growth data set ...................................................................................................... 10
5. Conclusions ................................................................................................................. 11
6. References .................................................................................................................... 12
1. Introduction

One of several aspects in good decision-making in public policy and in business is accurate forecasting of the future. In order to bring about the best outcomes, a government or a firm must correctly anticipate the most likely future states of the world. In many situations the forecasts are being based on statistical models where a potential outcome, Y, is regressed on a set of predetermined regressors, X. The problem in this avenue is in determining the model. In the linear regression models this amounts to choose which covariates to add and how to add them in the model. A common practice is to find the single best model using different information criteria. This might not be the most advantageous way for model prediction.

There is a risk of overfitting which means adding spurious relations which may distort the prediction. The risk of overfitting increases with the sample size of the data and with the number of predetermined variables available.

An alternative approach to estimate the best model is to build prediction on an average of a number of possible (approximatively) models. In these models the problem consists in choosing the weights to different candidate models. The most common used model averaging is the Bayesian Model Averaging (BMA) and is well described in Hoeting et al (1999). Frequentist model averaging has recently gained interest though and is still under rigorous development with new studies coming out every year. Hansen (2007) provided a new estimator based on the Mallows’s criterion called Mallow model averaging (MMA). Hansen (2008) subsequently provided forecasts of it showing its potential. The problem of MMA is that it does not allow for heteroscedastic errors. As a means to overcome this problem Hansen & Racine (2012) suggested the Jackknife model averaging (JMA).

In this study I investigate JMA. I then take use of two data sets on wages and Growth and compare the (relative) performance of the JMA, MMA and two more simpler methods: one suggested by Schwarz (1978) and then an idea stemming from Mallows (1973).

The outline of the study is that I will first go through the theoretical background of JMA, followed by a brief explanation of MMA and the two methods suggested in Schwarz (1978) and Mallows (1973). I will then explain the whole methodology behind this study, explaining the data sets I will be using and also how I will evaluate the results. I will then later on run the JMA on a training set to find the weights and values of the estimates. These will then be used
in a test set to see the average squared prediction error (ASPE) and the relative efficiency compared to the other methods of model averaging. The results will be provided and the study will be concluded with other methods, focusing specifically on JMA and its prediction errors.

2. Jackknife model averaging and model selection

The following notations will be the same as used by Hansen & Racine (2012) and Zhang et al (2013), as this is most convenient for the reader. We have to first consider the data generating process in an OLS-setting so that it could be written as:

\[ y_i = \mu_i + e_i \]  

Where \( y = (y_1, ..., y_n)' \), \( \mu = (\mu_1, ..., \mu_n)' \), and \( e = (e_1, ..., e_n)' \). Now it is assumed that \( E(e|X) = 0, E(y|X) = \mu \) and \( \text{Var}(e|X) = \Omega \), also noting that \( \Omega \) is a positive definite symmetric matrix. To do the model averaging of these sets, we set up \( M_n \) to be the number of models to average and the set of linear estimators for \( \mu \) is \( \{\hat{\mu}_1, ..., \hat{\mu}_M \} \). As this is for a linear model, it is important to follow this as it is desired that \( \hat{\mu}^m \) is the m'th estimator which can be rewritten as \( P_m y \), where \( P_m = X^m(X^m'X^m)^{-1}X^m' \). Since this is proposed by Hansen & Racine (2012) to be a linear model, \( P_m \) has good properties as it is only dependent on \( X \). This is important in the optimality theory used in Hansen & Racine (2012). Now consider the weight vector \( w = (w^1, ..., w^M) \) from a non-negative \( \mathbb{R}^M \) unit simplex:

\[ \mathcal{H}_n = \left\{ w \in \mathbb{R}^M: w^m \geq 0, \sum_{m=1}^{M_n} w^m = 1 \right\} \]

Since this then used in the averaging estimator of \( \mu \), that is obtained using the rest of the set linear estimators. It can be written like this:

\[ \hat{\mu}(w) = \sum_{m=1}^{M_n} w^m P_m y = P(w)y \]

From here it’s the important assumptions and proofs of Hansen & Racine (2012) which are to show how the JMA works. The vector weights should always be non-negative and should come from the subset \( \mathcal{H}_n \) which Hansen & Racine argued should contain discrete weights from the set of \( \left\{0, \frac{1}{N}, \frac{2}{N}, ..., 1\right\} \) for some positive integer of N. This however can be changed.

---

1 Readers are recommended to read this paper for further information about their proofs of how the JMA works under their asymptotic optimality and how the asymptotic optimality looks like.
by using the JMA-estimator proposed by Zhang et al. (2013) based on the works by Wan et al. (2010) which neither had a requirement of \( \Omega \) being a diagonal matrix as Hansen & Racine (2012) restricted it to be, allowing the errors to be autocorrelated.

The estimator \( \tilde{\mu}^m \) is now used for \( \mu \) as the jackknife estimation from a delete-one cross-validation is applied. This is the estimator one gets after removing \( i \)th observation from \((y_i, x_i)\) which will then yield \( \tilde{P}_m \) where the diagonal is now filled with zeros. Therefore we can rewrite \( \tilde{\mu}^m \) as \( \tilde{P}_m y \) now. As Hansen & Racine (2012) used a squared error loss criterion for choosing the specific weight vector to use in the averaging estimator, it should be the weight vector \( \hat{w} = \arg \min_{w \in \mathcal{H}_n} CV_n(w) \), where:

\[
CV_n(w) = (y - \tilde{\mu}(w))'(y - \tilde{\mu}(w))
\]

By substituting and replacing \( w \) in 2.1.3 with \( \hat{w} \), the JMA estimator of \( \mu \) is \( \tilde{\mu}(\hat{w}) \). By seeing equation 2.1.4 means that \( CV_n(w) \) could be written as \( w'\hat{e}'\hat{e}w \), which is a quadratic function of \( w \). It is relatively effortless to do the minimization of \( CV_n(w) \) with a software package such as R (Package “quadprog”).

From this, the weights are then applied in the last model. Yielding the following linear model;

\[
\tilde{\mu}(\hat{w}) = (w^1 + \cdots + w^{M_1})X_1\hat{\beta}_1 + (w^2 + \cdots + w^{M_2})X_2\hat{\beta}_2 + \cdots + w^{M_n}X_{M_n}\hat{\beta}_{M_n}
\]

Hence Hansen & Racine (2012) show that the nested averaging estimator will give less weights to the different \( X_j\hat{\beta}_j \). Therefore the order of \( X_j \) matters for the calculations, which in return is a weakness for JMA. Hansen (2007) suggested that the ordering of the regressors could be done in groups. However the interpretation of Hansen & Racine (2012) on how relevant the regressors are could be misunderstood if one is not sure of which groups to order the regressors by. Therefore JMA is very sensitive to the orders, which could yield different estimators depending on the order of the regressors. This could raise the problem of how these regressors’ should be ordered. The explanation of relevancy is not well defined and could yield misinterpretations as either partial adjusted \( R^2 \), partial information criteria weights or what previous knowledge tells. If one has prior knowledge of the relevancy and model outline, then it could argued that BMA would be a suitable way to use model averaging. This is due the fact of having priori distribution from which one can create posterior estimates in the Bayesian way. Since the BMA is most abundantly used and that FMA is still under development there could exist more incitement for people to use BMA rather than FMA.
Another drawback pointed out by Hansen & Racine (2012) is the lack of knowledge of the distribution of JMA. As there is no knowledge of the distribution, inference is therefore not a possibility to use for the JMA for inference. But since the suggestion of using a FMA has been used for a while, Buckland et al. (1997) suggested the use of either AIC or BIC where the weights are then calculated by the different information criteria. The suggestion of the weights could be done as following;

\[ w_m = \frac{\exp(-I_m/2)}{\sum_{i=1}^{M} \exp(-I_i/2)}, \quad i = 1, 2, ..., M \tag{2.1.6} \]

Where \( I_i \) is the different information criterion generated from the formula of BIC. So either the \( I_i \) is then described by;

\[ BIC: I_i = -2 \log(L) + p \log(n) \tag{2.1.7} \]

Where \( L \) is the likelihood function, \( p \) the number of regressors in the model and \( n \) the number of observations. By the different values of the information criteria, the regressors therefore gain certain weights calculated by these weights as shown in 2.1.6. Hansen & Racine (2012) developed JMA due to the fact that these methods does not allow the presence of heteroscedasticity. Hansen (2007) already proposed another technique following equation 2.1.4. but used a penalty of \( 2\sigma^2 K(W) \) as the criterion to follow instead of the usual weight criteria of AIC or BIC, however the same problem follows here. The JMA is argued to be a better fit for model averaging as MMA does not allow heteroscedasticity (Hansen & Racine, 2012). By further following the study of Hansen & Racine (2012) and Zhang et al. (2013) the MMA will be approximately the same as JMA when having larger sample sizes.

The MMA is approximately written the same as the JMA due to its similar criteria of minimizing the Mallows’s \( C_p \) and the same assumptions under the JMA methods, except for its weakness to heteroscedastic errors (Hansen & Racine, 2012). The MMA is also written under the assumption of \( M \) (the largest amounts of models used) will be fixed when \( n \to \infty \) or that \( M = M_n \) diverges into infinity (Hansen, 2007). This could very well be problematic under finite samples when it is not useful to approximate to infinity (such as small sample sizes).

The biggest difference between MMA and JMA is that JMA is based on the Jackknife technique with leaving out one observation for cross-validation and JMA also allows heteroscedastic errors. As in equation 2.1.2, MMA is also based on vector weights summing to one in the same unit simplex. Hansen (2007) argues that the problems otherwise could yield ill-behaved vectors with bad forecasting ability. The Mallows’s \( C_p \) is based by
Mallows’s work (1973) on model selection. This model selection is based on the residual sum of squares of the model in question and is defined as:

\[ C_p = \frac{1}{\hat{\sigma}^2} \text{RRS}_p - n + 2p \]  

2.1.8

Where the \( \text{RRS}_p \) is defined as the residual sum of squares of the estimator and the real values. \( \hat{\sigma}^2 \) is the estimated version of the real \( \sigma^2 \). Under these circumstances, the suggestion is to base the model selection on which model that actually minimizes the standard errors of its estimation. This is quite similar to the AIC and under larger samples, it is approximately the same as the AIC (Boisbounon et al. 2014).

3. Method

In this section I will go through the process of how the results and analysis will be done. I have decided to apply this on real data used in other applications. The reason for this is due to the lack of applied studies on the JMA and especially under potential misspecified orders. The orders of how JMA should be applied are mentioned to be grouped and also how relevant the different regressors are (Hansen, 2007; Hansen, 2008; Hansen & Racine, 2012; Hansen, 2014). How relevant the regressors are not well specified, therefore the orders of the regressors will be calculated by the adjusted \( R^2 \) by the different subsets. This is used by R package’” relaimpo” (Grömping, 2006) and the method of ”lmg” in the same package. The ”lmg” method is based on the \( R^2 \) on the different regressors’. The reason why I use this instead of the ”pmvd” which is based on the marginal variance is simply because of the “pmvd” method does not work under the global R package and has to specifically be used in Europe. Both methods take time in their computations but as ”lmg” is based \( R^2 \), it also shows how relevant the different regressors are in the model, which is a requirement for JMA. However \( R^2 \) is somewhat of a criticized explanation of relevancy in many cases but if this is based on some information criterion (AIC, BIC, Mallows’s \( C_p \) etc.), it could bias the results in favour of a certain model (as they could be based on some of these information criteria). So I decide to use \( R^2 \) as the best criterion for regressor relevancy.

The data I will be using in this study is described in the following subsections 3.1 and 3.2. JMA has already been under simulated studies but not applied on real data, the only study I have found which has used JMA on real data is when it was applied on current accounts by Urosevic et al (2012). That study used the different method of JMA which allows
autocorrelation. Since I will use the original version proposed by Hansen & Racine (2012). I cannot use autocorrelated data but nevertheless, the results should be the same as the base of the updated JMA proposed by Wan et al. (2010) and Zhang et al. (2013) are based on the origins of JMA proposed by Hansen & Racine (2012).

Besides the description of the different data sets, the subsection 3.3 will describe thoroughly how I will compare the methods and what kind of samples and sizes I use for the validation set where the predictions will be conducted on. The reason for using a cross-validation set before a training set is because the model averaging techniques are based on forecasting and predictions first and foremost. As there are not yet any R packages for the JMA, one could either construct the function by the formulae provided by Hansen & Racine (2012) or use the R code provided by Hansen at his website\(^2\). There is however an R package which has the MMA in it called “MAMI” developed by Schomaker & Heumann (2014) found at R-forge. The codes show signs of being computer efficient and provide the same results as Hansen’s code. However since this is not checked with the official R development team, there could exist bugs and misspecified codes which could result in skewed results or completely false results due to coding errors. In the following sections the data sets will be explained and further explain my reasoning behind using them.

3.1 Data 1 – Wages

This data was used in Hansen & Racine (2012) to show how well JMA performed in their predictions. The data comes from Woolridge (2003, p. 226). Hansen & Racine (2012) used different sample sizes for their training set and then forecasted it on a test set. They got the ASPE for that forecast and repeated the forecast 1000 times to then use the median of all the ASPE’s they got from the repeats. However this is not how I will do. I will construct training sets and then forecast it once. This is due to the fact that one person using this model would most likely forecast once and not repeat the forecast several times. This data has 526 observations and 23 regressors, (Hansen & Racine (2012) used 30 regressors with the application of interaction between the different regressors). The reason I use the raw data and no application of interaction regressors is because I want to differentiate my usage of the same data set as Hansen & Racine (2012) with potential lost information in the whole model. This could affect the estimators and also see how predictions will be affected by this. The different variables in this data set are;

\(^2\) Read about the code and other replication codes on model averaging at http://www.ssc.wisc.edu/~bhansen/progs/progs_ma.html
Log of hourly earnings; Years of education; Years of potential experience; Years with current employer; Non-white dummy; Female dummy; Married dummy; Number of dependents; SMSA living dummy; North central US living dummy; Southern US living dummy; Western living dummy; Construction worker dummy; Manufacturing worker dummy; Transport community worker dummy; Retail worker dummy; Services worker dummy; Professional service industry worker dummy; Professor occupation dummy; Clerical occupation dummy; Service occupation dummy; Experience squared; Year with current employer squared.

3.2 Data 2 – Economic Growth

The data set here consists of 41 different regressors for 72 different countries. This data set was used by Fernandez et al. (2001) in their application of BMA of the economic growth in different countries. This data set can be found in the R package "BMS" (Zeuger & Feldkircher, 2015). However the regressors used by Fernandez et al. (2001) are the following (in order of importance):

GDP level in 1960; Fraction of Confucian; Life expectancy; Equipment Investment; Sub-Saharan dummy; Fraction Muslim; Rule of Law; Number of Years open economy; Degree of capitalism; Fraction Protestant; Fraction GDP in mining; Non-Equipment Investment; Latin American dummy; Primary School Enrollment (1960); Fraction Buddhist; Black Market Premium; Fraction Catholic; Civil Liberties; Fraction Hindu; Primary Exports (1970); Political Rights; Exchange rate distortions; Age; War dummy; Size labor force; Fraction speaking foreign language; Fraction of population speaking English; Ethnolinguistic fractionalization; Spanish colony dummy; S.D. of black market premium; French colony dummy; Absolute latitude; Ratio workers to population; Higher education enrollment; Population Growth; British colony dummy; Outward Orientation; Fraction Jewish; Revolutions and coups; Public education share; Area (Scale).

This was used to estimate the growth of the countries but as previously mentioned, this data set will only be used for training and test sets for these different FMA estimators and model selection techniques. The interesting aspect of this data set is due to its low numbers of observations, this will yield small sample sizes for the training sets and even smaller sizes for the testing sets. As this data set is also heteroscedastic, it will give us an indication of how well JMA actually works on smaller data, as opposite to the data used by Hansen & Racine (2012) which had 526 observations which is relatively large compared to 72 observations in this data set.
3.3 Evaluation

The evaluation of the different test sets will be considering the minimized errors of the predictions. Therefore I will judge the different predictions by their variance and their squared error in terms of relative efficiency, which is the following

\[ \frac{V(\hat{\mu}_k)}{V(\hat{\mu}_{JMA})} = \text{relative efficiency} \]

3.3.1 If the efficiency is > 1 then it indicates that \( V(\hat{\mu}_{JMA}) \) has the best minimizing properties of the errors, if < 1 it indicates that \( V(\hat{\mu}_k) \) is the better estimator is the senses of minimizing the errors of the predictions. Because JMA and MMA are meant to minimize the errors of the predictions, I find that this is a good measurement of how efficient these estimators are. I will also go through all potential models in these data sets, meaning that they range from \( p = 1, \ldots, k \) where \( k \) is the number of potential regressors in the model. One could change the number of subsets into their own selection so that the number of potential models could be:

- **Model 1**: \( y_i = \beta_1 X_1 + e_i \)  
- **Model 2**: \( y_i = \beta_1 X_1 + \beta_3 X_3 + e_i \)  
- **Model 3**: \( y_i = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e_i \)

When ordering the nested subsets, the relative importance of the different models are already in their nested space. If one want to combine all the different models to see whether there is some effect, the selection subset-input in Hansens code is therefore not needed in this case. The models will be ordered in their relative importance and will use all possible regressors. One could therefore argue about the problems of overfitting and the potential consequences of this. However I want to keep the study similar to Hansen & Racine (2012), and include all potential regressors in a global model ordered in their relevancy. If the regressors are not relevant, the JMA weights will be 0 and therefore will construct coefficients equal to 0. The same will happen to MMA were the weights of the different regressors will be equivalent to 0 as well. Therefore the global model is of my specific interest to analyze the relative efficiency of the different techniques. However as I will include information criteria selection, I will use the best models from each sample to then use in the predictions. This is done by finding all potential models and then get the model which minimizes the information criterion in question. This will be used to forecast the rest of the test data set and compare the ASPE’s to JMA’s ASPE.
4. Results

4.1 Wage data set

For the data given in Wooldridge (2003, p. 226), I first sample 60 % to 85 % in steps of 5 % each time of the data set to be my training set. I then use the rest of the data to be the test set to analyze predictions. The sample sizes (n_{train}) therefore vary from 316, 342, 368, 395, 421 to 447 observations. The test set is then set to n − n_{train} = n_{test} which is where the predictions are conducted. The results of the sets are described in the following tables (4.1.1-4.1.6) with the relative efficiency of the prediction variance and the relative efficiency of the prediction error variance.

<table>
<thead>
<tr>
<th>Sample 60 %</th>
<th>MMA</th>
<th>Cp</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative Prediction</td>
<td>1.02</td>
<td>1.06</td>
<td>1.03</td>
</tr>
<tr>
<td>ASPE</td>
<td>1.00</td>
<td>1.01</td>
<td>1.02</td>
</tr>
</tbody>
</table>

Table 4.1.1. – Relative efficiency with a test set of 40 % of the total numbers of observed variables.

<table>
<thead>
<tr>
<th>Sample 70 %</th>
<th>MMA</th>
<th>Cp</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative Prediction</td>
<td>1.01</td>
<td>1.05</td>
<td>0.99</td>
</tr>
<tr>
<td>ASPE</td>
<td>1.00</td>
<td>1.01</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 4.1.3. – Relative efficiency with a test set of 30 % of the total numbers of observed variables.

<table>
<thead>
<tr>
<th>Sample 80 %</th>
<th>MMA</th>
<th>Cp</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative Prediction</td>
<td>1.00</td>
<td>1.01</td>
<td>1.02</td>
</tr>
<tr>
<td>ASPE</td>
<td>1.00</td>
<td>1.05</td>
<td>1.01</td>
</tr>
</tbody>
</table>

Table 4.1.5. – Relative efficiency with a test set of 20 % of the total numbers of observed variables.

<table>
<thead>
<tr>
<th>Sample 85 %</th>
<th>MMA</th>
<th>Cp</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative Prediction</td>
<td>1.01</td>
<td>1.06</td>
<td>1.02</td>
</tr>
<tr>
<td>ASPE</td>
<td>1.00</td>
<td>1.05</td>
<td>1.02</td>
</tr>
</tbody>
</table>

Table 4.1.6. – Relative efficiency with a test set of 15 % of the total numbers of observed variables.

As seen in the tables (4.1.1-4.1.6), the MMA and JMA are quite similar in these cases of their relative efficiency. This is in line what Hansen & Racine (2012) found out, as larger sample sizes will give MMA better performances. The relative prediction variance for all sample sizes have a small advantage for the JMA with sample sizes of 85 %, 75 %, 70 %, 65 % and 60 % where all efficient values are larger than 1, indicating a better performance for the JMA. However the APSE errors are all equal to one, indicating that the JMA and MMA are equal in their efficiency of minimizing the APSE. Comparing the JMA’s efficiency to the information criteria model selection, the JMA has a large advantage to Mallows’s $C_p$. However it is not a
lot different to the BIC, with even efficiency at sample sizes of 70 % (table 4.1.3) and just a small advantage to the other sample sizes. This corresponds to the findings of Hansen & Racine (2012) found in their use of the wage data set. Even though they repeated the predictions 1000 times so it evened out with the BIC model selection having approximately the same efficiency as the JMA. I did not repeat the predictions which might be a reason why BIC got less efficiency. Another reason could also be the potential omitted variables in this case; the interaction variables which were used in Hansen & Racine (2012) might have useful information to the BIC model selection. On the other hand JMA seems to outperform the other choices of models in terms of lowering the prediction errors. This could be argued as JMA allowing the potential heteroscedasticity in this data set.

4.2 Growth data set

This data set is similarly sampled as in section 4.1., in steps of 5 % ranging from 60 % to 85 %. However this data set consists of fewer observations compared to the wage data set. The wage data set had 526 observations while this data set consists of only 72 observations. Hansen & Racine (2012) showed the problems of MMA under small sample sizes which is why I am going to investigate this further. The sample sizes here are the following 43, 47, 50, 54, 58 and 61 observations, which will give small amounts of observations to predict. The results are in tables’ 4.2.1-4.2.6.

<table>
<thead>
<tr>
<th>Sample 60 %</th>
<th>MMA</th>
<th>Cp</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative Prediction</td>
<td>6.20</td>
<td>43.86</td>
<td>75.55</td>
</tr>
<tr>
<td>ASPE</td>
<td>7.79</td>
<td>55.42</td>
<td>91.78</td>
</tr>
</tbody>
</table>

Table 4.2.1. – Relative efficiency with a test set of 40 % of the total numbers of observered variables.

<table>
<thead>
<tr>
<th>Sample 65 %</th>
<th>MMA</th>
<th>Cp</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative Prediction</td>
<td>4.54</td>
<td>6.92</td>
<td>3.59</td>
</tr>
<tr>
<td>ASPE</td>
<td>4.70</td>
<td>7.20</td>
<td>3.70</td>
</tr>
</tbody>
</table>

Table 4.2.2. – Relative efficiency with a test set of 55 % of the total numbers of observered variables.

<table>
<thead>
<tr>
<th>Sample 70 %</th>
<th>MMA</th>
<th>Cp</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative Prediction</td>
<td>1.32</td>
<td>2.53</td>
<td>3.93</td>
</tr>
<tr>
<td>ASPE</td>
<td>2.15</td>
<td>6.42</td>
<td>9.90</td>
</tr>
</tbody>
</table>

Table 4.2.3. – Relative efficiency with a test set of 30 % of the total numbers of observered variables.

<table>
<thead>
<tr>
<th>Sample 75 %</th>
<th>MMA</th>
<th>Cp</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative Prediction</td>
<td>2.56</td>
<td>102.70</td>
<td>32.78</td>
</tr>
<tr>
<td>ASPE</td>
<td>9.61</td>
<td>512.97</td>
<td>160.10</td>
</tr>
</tbody>
</table>

Table 4.2.4. – Relative efficiency with a test set of 25 % of the total numbers of observered variables.

<table>
<thead>
<tr>
<th>Sample 80 %</th>
<th>MMA</th>
<th>Cp</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative Prediction</td>
<td>2.15</td>
<td>13.60</td>
<td>3.70</td>
</tr>
<tr>
<td>ASPE</td>
<td>4.37</td>
<td>39.45</td>
<td>7.73</td>
</tr>
</tbody>
</table>

Table 4.2.5. – Relative efficiency with a test set of 20 % of the total numbers of observered variables.

<table>
<thead>
<tr>
<th>Sample 85 %</th>
<th>MMA</th>
<th>Cp</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative Prediction</td>
<td>1.21</td>
<td>1.14</td>
<td>1.14</td>
</tr>
<tr>
<td>ASPE</td>
<td>1.75</td>
<td>2.20</td>
<td>2.20</td>
</tr>
</tbody>
</table>

Table 4.2.6. – Relative efficiency with a test set of 15 % of the total numbers of observered variables.
As Hansen & Racine (2012) argued, the MMA has huge disadvantages under smaller sample sizes. The efficiency of JMA is largely above 1 in many of the cases, this is most likely due to the troubles of MMA under heteroscedastic errors which might be present in this data set. Another indication of why the MMA is unsuitable in these small samples is because of the assumptions of the $M$ will be fixed when $n \to \infty$ (Hansen, 2007). These sample sizes are too small to approximate that $n \to \infty$ and could be a reason as to why MMA generates large errors in its predictions. If compared to the model selection based on the information criteria, JMA shows a lot more promising results than compared to the wage data set. Here the efficiency is at a large advantage compared to Mallows’s $C_p$ and BIC. The large errors made by the information criteria under these small samples and smaller numbers of observations to predict is shown in the tables 4.2.1:4.2.6, especially in table 4.2.4 were the Mallows’s $C_p$ has a prediction error of 512.97 times larger errors than the JMA. The BIC also has huge errors compared to the JMA with errors, specifically 160.10 times more than the JMA (table 4.2.4). This is most likely due to potential problems with the information criteria under smaller samples and heteroscedastic errors which will in turn increase the errors of the estimators. The properties of the JMA seem to be suitable for these small sample sizes compared to other methods. Especially when considering the presence of heteroscedastic errors in the model. The data in 4.1 was not severely affected by heteroscedasticity and it is seen by the relative efficiency of the different methods compared to the JMA. When having smaller sizes to predict, the JMA seems to have the best predictability due to its robustness to heteroscedasticity.

5. Conclusions

In this study I introduced the JMA and discussed its properties. I compare it to MMA, BIC and Mallows’s $C_p$ as an alternative to predict data based on different training sets. I used the wage data set used by Hansen & Racine (2012) and also the data set used by Fernandez et al (2001). The JMA outperformed the other selection tools in its predictions with lower forecasting errors as previously shown in the study of Hansen & Racine (2012). It also seems to have better forecasting properties for smaller sample sizes, which was seen in the growth data set by Fernandez et al (2001). JMA therefore could be used an interesting tool for forecasting and model selection as it allows for heteroscedastic errors. Especially with the
updated version by Zhang et al (2013) as it allows JMA to use data with autocorrelation. As pointed out by Hansen & Racine (2012), this still needs further investigation, especially for inference as JMA’s distribution is still unknown. It could also be of interest to investigate JMA to regular Bayesian methods and especially for smaller samples and observations, especially with lower amounts of explained variation as Zhang et al (2013) found the potential problems of JMA and MMA when this happens. JMA shows very promising results even under smaller sample sizes and could be of interest as a method of forecasting or prediction.

6. References


