Least Squares Forecast Averaging

Bruce E. Hansen*
University of Wisconsin†

www.ssc.wisc.edu/~bhansen

March 2006

Abstract

This paper proposes a new method of forecast combination based on the method of Mallows Model Averaging (MMA). The MMA weights are asymptotically mean-square optimal for estimation of regression coefficients, and therefore should be have good mean-square properties for point forecasting. We show how to compute MMA weights in forecasting settings, and investigate the performance of the method in simple static and dynamic simulation environments. We find that the MMA forecasts are strong competitors to BIC and equal-weight forecast combinations – two of the strongest known methods for forecast combination.

*Research supported by the National Science Foundation.
†Department of Economics, 1180 Observatory Drive, University of Wisconsin, Madison, WI 53706

1 Introduction

Forecast combination has a long history in econometrics. While a broad consensus is that forecast combination improves forecast accuracy, there is no consensus concerning how to form the forecast weights. The most recent literature has focused on two particularly appealing methods – simple averaging and Bayesian averaging. The simple averaging method simply picks a set of models and then gives them all equal weight for all forecasts. The Bayesian averaging method computes forecast weights as a by-product of Bayesian model averaging.

This paper introduces a simple method appropriate for linear models estimated by least-squares. The method is to construct forecast combinations using the weights computed by Mallows' Model Averaging (MMA), the weights which minimize a generalized Mallows' criterion introduced in Hansen (2006). MMA weights are asymptotically optimal with respect to mean-square loss, and are thus expected to produce good forecast combination weights, at least when evaluated using mean-square forecast loss.

As mentioned above, two powerful existing methods for forecast combination are simple averaging and Bayesian averaging. Both have been shown to be extremely versatile and successful in applications. Yet neither is inherently satisfying. Simple averaging only makes sense if the class of models under consideration is reasonable. If a terrible model is included in the class of forecasting models, simple averaging will pay the penalty. This induces an inherent arbitrariness, and thus the method is incomplete unless augmented by a description of how the initial class of models is determined, which destroys the inherent simplicity of the method.

On the other hand, the fact that Bayesian Model Averaging relies on priors (over the class of models and over the parameters in the models) means that this method suffers from the arbitrariness which is inherent in prior specification. Futhermore, the BMA paradigm is inherently misspecified. It is developed under the assumption that the truth is one finite-dimensional parametric model out of a class of models under consideration. The goal is to find the "true" model out of this class. This paradigm and goal is inherently misspecified and misguided, as it is more appropriate to think of models as approximations, and that the "true" model is more complex than any of the models in our explicit class. When we fit models, we balance specification error (bias) against overparameterization (variance). The correct goal is to define the object of interest (such as forecast mean-squared-error) and then evaluate methods based on this criterion, without assuming that we necessarily

have the correct model.

Mallows' Model Averaging takes exactly this approach. The goal is to obtain the set of weights which minimizes the forecast mean-squared-error (MSE) over the set of feasible forecast combinations. The generalized Mallows' criterion is an estimate of the forecast MSE, and the weights which minimize this criterion are asymptotically optimal in some settings.

The Mallows' criterion for model selection was introduced by Mallows (1973) and its asymptotic optimality studied by Shibata (1980, 1981, 1983), Li (1987), and Lee and Karagrigoriou (2001). The Mallows criterion is similar to the information criterion of Akaike (1973). Akaike (1979) proposed using the exponentiated AIC as model weights, and this suggestion was picked up by Buckland et. al. (1987) and Burnham and Anderson (2002) who propose model averaging based on exponentiated AIC weights. Hjort and Claeskens (2003) introduced a general class of frequentist model average estimators, including methods similar to Mallows' model averaging.

The Bayesian information criterion was introduced by Schwarz (1978) as a method for model selection. There is a large literature on Bayesian Model Averaging; see the review by Hoeting et. al. (1999). Some applications in econometrics include Doppelhofer, Miller and Sala-i-Martin (2000), Brock and Durlauf (2001), Avramov (2002), Fernandez, Ley and Steel (2001a,b), Garratt, Lee, Pesaran and Shin (2003), and Brock, Durlauf and West (2003).

The idea of forecast combination was introduced by Bates and Granger (1969) and spawned a large literature. Some excellent reviews include Clemen (1989), Diebold and Lopez (1996), Hendry and Clements (2002), Timmermann (2006) and Stock and Watson (2006). The idea of using Bayesian model averaging for forecast combination was pioneered by Min and Zellner (1993) and its usefulness recently demonstrated by Wright (2003ab). Stock and Watson (1999, 2004, 2005) have provided detailed empirical evidence demonstrating the gains in forecast accuracy through forecast combination, and in particular have demonstrated the success of simple averaging (equal weights) along with Bayesian model averaging.

The plan of the paper is simple. Section 2 introduces the framework and the linear forecasting models. Section 3 presents the generalized Mallows' criterion and the MMA forecast combination. Section 4 presents the results of a simulation experiment using a static regression example. Section 5 presents the results of a simulation experiment using a simple dynamic model. A conclusion follows.

2 Model Forecasts

Consider the problem of constructing a forecast f_{n+1} of a target variable y_{n+1} conditional on the vector $x_n \in R^k$. We restrict attention to forecasts which are linear in x_t, thus $f_{n+1} = x_n'\beta$ for some $\beta \in R^k$. A forecasting model consists of a set of restrictions on β, most typically restricting some elements to equal zero. We assume that these restrictions are linear, and so for the $m'th$ forecasting model we write the restriction as $S_m'\beta = c_m$ where S_m is $r_m \times k$. Let $k_m = k - r_m$ denote the number of free parameters in model m. We assume that there are M forecasting models under consideration. For example, in univariate forecasting the models might be the linear autoregressions AR(0), AR(1), ..., AR(k) in which case $M = k + 1$.

Define the sum of squared error function

$$S_n(\beta) = \sum_{t=1}^{n-1} \left(y_{t+1} - x_t'\beta\right)^2.$$

The restricted least squares estimator for the m'th model is

$$\hat{\beta}_m = \underset{S_m'\beta = c_m}{\operatorname{argmin}} S_n(\beta).$$

The $m'th$ model forecast is $\hat{f}_{n+1}(m) = x_n'\hat{\beta}_m$. In the case where model m simply excludes a subset of the variables x_t, this forecast is the standard least-squares forecast from the regression of y_{t+1} on the included variables.

A forecast combination takes a weighted average of the individual forecasts. Let $W = (w_1, ..., w_M)'$ be a vector of non-negative weights which sum to one. A forecast combination takes the form

$$\begin{aligned} \hat{f}_{n+1}(W) &= \sum_{m=1}^{M} w_m \hat{f}_{n+1}(m) \\ &= W'\hat{f}_{n+1} \end{aligned}$$

where

$$f_{n+1} = \begin{pmatrix} \hat{f}_{n+1}(1) \\ \hat{f}_{n+1}(2) \\ \vdots \\ \hat{f}_{n+1}(M) \end{pmatrix}$$

is the vector of model forecasts.

Given the linear structure, the forecast combination can be written as

$$\hat{f}_{n+1}(W) = x_n' \hat{\beta}(W)$$

where

$$\hat{\beta}(W) = \sum_{m=1}^{M} w_m \hat{\beta}_m$$

is a weighted average of the parameter estimates from the individual models.

3 Generalized Mallows' Criterion

Our question is how to select the weights w_m. Consider the criterion of mean-square loss. Let β_0 denote the population projection of y_{t+1} on x_t, and define the projection error $e_{t+1} = y_{t+1} - \beta_0' x_t$, the error variance $\sigma^2 = Ee_{t+1}^2$, and the expected design matrix $Q = Ex_t x_t'$. For fixed weight vector W, the mean-square loss from the combination forecast is

$$\begin{aligned} MSE(W) &= E\left(y_{n+1} - \hat{f}_{n+1}(W)\right)^2 \\ &= E\left(e_{n+1} - x_n'\left(\hat{\beta}(W) - \beta\right)\right)^2 \\ &= Ee_{n+1}^2 + \mathrm{tr}\left[E\left(\left(\hat{\beta}(W) - \beta\right)\left(\hat{\beta}(W) - \beta\right)' x_n x_n'\right)\right] \\ &\simeq \sigma^2 + \mathrm{tr}\left[E\left(\left(\hat{\beta}(W) - \beta\right)\left(\hat{\beta}(W) - \beta\right)'\right) Q\right] \end{aligned}$$

where the final line makes the approximation that $\hat{\beta}(W)$ is approximately independent of x_n. This final expression is the mean-squared-error (MSE) of the combination estimator $\hat{\beta}(W)$, weighted by the design matrix Q. It follows that the weight vector W which minimizes the weighted MSE of $\hat{\beta}(W)$ is the same as that which minimizes the mean-square forecast loss.

The problem of selecting weights to minimize an unbiased estimate of MSE of $\hat{\beta}(W)$ has been considered in Hansen (2006). The method is to select W to minimize the generalized Mallows' criterion

$$C_n(W) = S_n\left(\hat{\beta}(W)\right) + 2\hat{\sigma}^2 \sum_{m=1}^{M} w_m k_m \tag{1}$$

where $\hat{\sigma}^2$ is an estimate of σ^2. The standard choice is

$$\hat{\sigma}^2 = \frac{1}{n-k} S_n(\hat{\beta}) \tag{2}$$

where $\hat{\beta}$ is the unrestricted least-squares estimator.

The Mallows' Model Average (MMA) weight vector $\hat{W}$ is the set which minimize $C_n(W)$ subject to the feasible constraints on W :

$$\hat{W} = \underset{W \in \Omega}{\operatorname{argmin}}\, C_n(W) \tag{3}$$

where

$$\Omega = \left\{ W \in [0,1]^m : \sum_{m=1}^{M} w_m = 1 \right\}, \tag{4}$$

the unit simplex in R^M. Given $\hat{W}$, we define the Mallows' Model Average (MMA) forecast

$$\hat{f}_{n+1}\left(\hat{W}\right) = \hat{W}' \hat{f}_{n+1}. \tag{5}$$

In non-dynamic models $C_n(W)$ and $\hat{W}$ have desirable properties, as discussed in Hansen (2006). First, if $\hat{\sigma}^2$ is an unbiased estimate of σ^2, then $En^{-1}C_n(W) = MSE(W)$ and thus the criterion is an unbiased estimator of the weighted MSE. Second, if $\hat{\sigma}^2$ is consistent for σ^2, then $\hat{W}$ is asymptotically optimal in the sense that if Ω is restricted to any discrete grid (but with M unbounded), then

$$\frac{MSE(\hat{W})}{\inf_{W \in \Omega} MSE(W)} \to_p 1 \tag{6}$$

as $n \to \infty$. This means that with respect to the MSE loss function, the MMA weight vector $\hat{W}$ is asymptotically equivalent to the infeasible optimal weight vector. It follows that the MMA weights are asymptotically optimal forecast combination weights (in non-dynamic models).

While we have not been able to show that the optimality result (6) extends to dynamic models, there is no reason to expect it to fail. Regardless, we propose using the MMA weights (3) for forecast combination.

As there is no closed-form solution to (3), the MMA weights (3) must be computed numerically. For this calculation, it is convenient to write (1) in the following form. Let $\hat{e}_m$ be the $n \times 1$ residual vector from the m'th model, let $\bar{e} = (\hat{e}_1, ..., \hat{e}_M)$ be the $n \times M$ matrix collection of these residuals, and let $K = (k_1, ..., k_M)'$ be the $M \times 1$ vector of the number of parameters in the M models. Then (1) can be written as

$$C_n(W) = W'\bar{e}'\bar{e}W + 2\hat{\sigma}^2 K'W \tag{7}$$

which is linear-quadratic in W. The solution (3) minimizes (7) subject to the nonnegativity and summation constraints (4). This is a classic quadratic programming problem, for which numerical algorithms are readily available. (For example, in the GAUSS programming language the procedure QPROG is appropriate.)

4 Finite Sample Investigation – Static Model

We now investigate the finite sample MSE of the our model average estimator in a simple simulation experiment. To keep the analysis simple and focused we start with the context of a random sample (independent and identically distributed observations).

The setting is the infinite-order regression

$$y_i = \sum_{j=1}^{\infty} \theta_j x_{ji} + e_i$$

where (y_i, x_i) are iid. We set $x_{1i} = 1$ to be the intercept, and the remaining x_{ji} are iid $N(0,1)$. The error e_i is $N(0,1)$ and independent of x_i. The parameter are determined by the rule $\theta_j = c\sqrt{2\alpha}j^{-\alpha-1/2}$. The population $R^2 = c^2/(1+c^2)$ is controlled by the parameter c.

The sample size is varied between $n = 50$, 100, and 200. The parameter α is varied between .5, 1.0, 1.5, and 2.0. The larger α implies that the coefficients θ_j decline more quickly with j. The number of models M is determined by the rule $M = 3n^{1/3}$ (so $M =$ 11, 14, and 18 for the three sample sizes). The coefficient c was selected to control the population R^2 to vary on a grid between 0.1 and 0.9.

We compare three estimators: AIC selection (AIC), weighted BIC (WBIC), and Mallows' Model Averaging (MMA).

The AIC forecast is $\hat{f}_{n+1}(\hat{m})$ where $\hat{m} = \text{argmin}_m AIC_m$ and

$$AIC_m = n \ln \left(n^{-1} S_n \left(\hat{\beta}_m \right) \right) + 2k_m$$

is the Akaike Information Criterion (AIC) for model m.

The WBIC forecast is $\sum_{m=1}^{M} w_m^{BIC} \hat{f}_{n+1}(m)$ where

$$w_m^{BIC} = \frac{\exp\left(-\frac{1}{2} BIC_m\right)}{\sum_{j=1}^{M} \exp\left(-\frac{1}{2} BIC_j\right)}$$

and

$$BIC_m = n \ln \left(n^{-1} S_n \left(\hat{\beta}_m \right) \right) + k_m \ln n$$

is the Bayesian Information Criterion (BIC) for model m. The WBIC forecast is approximately the Bayesian Model Average (BMA) forecast arising from equal model priors and diffuse coefficient priors.

The MMA forecast is (5) with $\hat{\sigma}^2$ in (1) computed by (2) with $k = M$.

We compare the three forecasting methods based on out-of-sample mean-square forecast error (MSE). We do this by computing averages across 100,000 simulation draws. For each parameterization, we normalize the MSE by dividing by the MSE of the infeasible optimal least-squares estimator (the MSE of the best-fitting model m).

The MSE calculations are displayed in Figures 1 to 3 for $n = 50$, 100 and 200, respectively. In each figure, the four panels correspond to the four values of α. In each panel, MSE is displayed on the y-axis, and the population R^2 on the x-axis. The three lines correspond to the three estimators. The dashed, dotted, and solid lines correspond to AIC, WBIC, and MMA, respectively.

First, we observe that the MMA forecast uniformly dominates the AIC forecast. In most cases, the difference is quite large. Second, neither MMA nor WBIC uniformly dominates the other. WBIC achieves lower MSE when R^2 is low and/or α is large. These are cases where a small number of included regressors are optimal, and WBIC tends to put more weight on the smaller models than MMA. On the other hand, MMA achieves lower MSE when α is small, R^2 is large, and/or n is large.

5 Finite Sample Investigation – Dynamic Model

In this section we extend the simulation experiment of the previous section to the context of a simple dynamic model. The setting is a univariate time-series y_t, which is generated by the m'th-order moving average process

$$y_t = (1 + \psi L)^m e_t$$

where $e_t \sim N(0, 1)$. The parameter ψ is varied on a grid between 0.1 and 0.9 and m is varied among $\{1, 2, 3, 4\}$. Again the sample size is varied between $n = 50$, 100, and 200. The forecasting models are AR(p) models, with p ranging from 0 to $M = 2n^{1/3}$. We compare four forecast combination methods, the three considered in the previous section (AIC, WBIC and MMA), and also equal weighting ($w_m = 1/M$).

The results are presented in Figures 4, 5, and 6, and $n = 50$, 100, and 200, respectively. In each figure, the four panels correspond to $m = 1$, 2, 3, and 4. MSE is displayed on the y-axis and ψ on the x-axis. The long dashes, dotted, short dashes, and solid lines correspond to AIC, WBIC, Equal weighting and MMA, respectively.

The figures show that the equal weighting estimator is the least robust. In some cases it is highly efficient relative to the others (e.g. $n = 50$, $m = 1$ and $\psi > .3$) but in many cases it is extremely inefficient. We also see, as in the previous section, that the MMA forecasts tend to achieve lower MSE than the AIC forecasts, although the ranking is not uniform. Comparing WBIC and MMA, we see that again there is not a strict ranking between the two methods. WBIC tends to achieve lower MSE when m and ψ are small. Once again this reflects the tendency of WBIC to put more weight on lower-dimensional models than MMA.

6 Conclusion

This paper has introduced a method of forecast combination based on Mallows' Model Averaging. The MSE performance of the MMA forecast combination is quite robust across simulation environments, and a strong competitor to BIC weighting. BIC weighting does well in regressions with small sample sizes and low regression signal, but otherwise the MMA weights produce better forecasts. Further investigations will be helpful to determine the relative merits of these competing procedures.

References

[1] Akaike, H. (1973): "Information theory and an extension of the maximum likelihood principle." In B. Petroc and F. Csake, eds., *Second International Symposium on Information Theory.*

[2] Akaike, H. (1979): "A Bayesian extension of the minimum AIC procedure of autoregressive model fitting," *Biometrika*, 66, 237-242.

[3] Avramov, D. (2002): "Stock return predictability and model uncertainty," *Journal of Finance,* 64, 423-458.

[4] Bates, J.M. and C.M.W. Granger (1969): "The combination of forecasts," *Operations Research Quarterly*, 20, 451-468.

[5] Brock, William and Stephen Durlauf (2001): "Growth empirics and reality," *World Bank Economic Review*, 15, 229-272.

[6] Brock, William, Stephen Durlauf, and Kenneth, D. West (2003): "Policy analysis in uncertain economic environments," Brookings Papers on Economic Activity, 1, 235-322.

[7] Buckland, S.T., K.P. Burnhamn and N.H. Augustin (1997): "Model Selection: An Integral Part of Inference," *Biometrics*, 53, 603-618.

[8] Burnham, Kenneth P. and David R. Anderson (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach.* Springer.

[9] Clemen, R.T. (1989): "Combining forecasts: A review and annotated bibliography," *International Journal of Forecasting*, 5, 559-581.

[10] Diebold, F. X. and J. A. Lopez (1996): "Forecast evaluation and combination," in Maddala and Rao, eds., *Handbook of Statistics,* Elsevier.

[11] Doppelhofer, G. R. Miller and X. Sala-i-Martin (2000): "Determinants of long-term growth: A Bayesian averaging of classical estimates (BACE) approach," working paper.

[12] Fernandez, C. E. Ley and M.F.J. Steel (2001a): "Benchmark priors for Bayesian model averaging," *Journal of Econometrics,* 100, 381-427.

[13] Fernandez, C. E. Ley and M.F.J. Steel (2001b): "Model uncertainty in cross-country growth regressions," *Journal of Applied Econometrics*, 16, 563-576.

[14] Garratt, A., K. Lee, M.H. Pesaran, and Y. Shin (2003): "Forecasting uncertainties in macroeconomic modelling: An application to the UK economy," *Journal of the American Statistical Association*, 98, 829-838.

[15] Hansen, Bruce E. (2006): "Least Squares Model Averaging," working paper, University of Wisconsin.

[16] Hendry, D.F. and M. P. Clements (2002): "Pooling of forecasts," *Econometrics Journal*, 5, 1-26.

[17] Hjort, Nils Lid and Gerda Claeskens (2003): "Frequentist Model Average Estimators," *Journal of the American Statistical Association*, 98, 879-899.

[18] Hoeting, Jennifer A., David Madigan, Adrian E. Raftery and Chris T. Volinsky (1999): "Bayesian Model Averaging: A Tutorial," *Statistical Science*, 14, 382-417.

[19] Lee, Sangyeol and Alex Karagrigoriou (2001): "An asymptotically optimal selection of the order of a linear process," *Sankhya*, 63, Series A, 93-106.

[20] Li, Ker-Chau (1987): "Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete Index Set," *Annals of Statistics*, 15, 958-975.

[21] Mallows, C.L. (1973): "Some comments on C_p," *Technometrics*, 15, 661-675.

[22] Min, C.-K. and A. Zellner (1993): "Bayesian and non-Bayesian methods for combining models and forecasts with applications to forecasting international growth rates," *Journal of Econometrics*, 56, 89-118.

[23] Schwarz, G. (1978): "Estimating the dimension of a model," *Annals of Statistics*, 6, 461-464.

[24] Shibata, Ritaei (1980): "Asymptotically efficient selection of the order of the model for estimating parameters of a linear process," *Annals of Statistics*, 8, 147-164.

[25] Shibata, Ritaei (1981): "An optimal selection of regression variables," *Biometrika*, 68, 45-54.

[26] Shibata, Ritaei (1983): "Asymptotic mean efficiency of a selection of regression variables," *Annals of the Institute of Statistical Mathematics*, 35, 415-423.

[27] Stock, J.H. and M. W. Watson (1999): "A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series," in Engle and White, eds., *Cointegration, Causality and Forecasting: A Festschrift for Clive W.J. Granger*, Oxford University Press.

[28] Stock, J.H. and M. W. Watson (2004): "Combination forecasts of output growth in a seven-country data set," *Journal of Forecasting*, forthcoming.

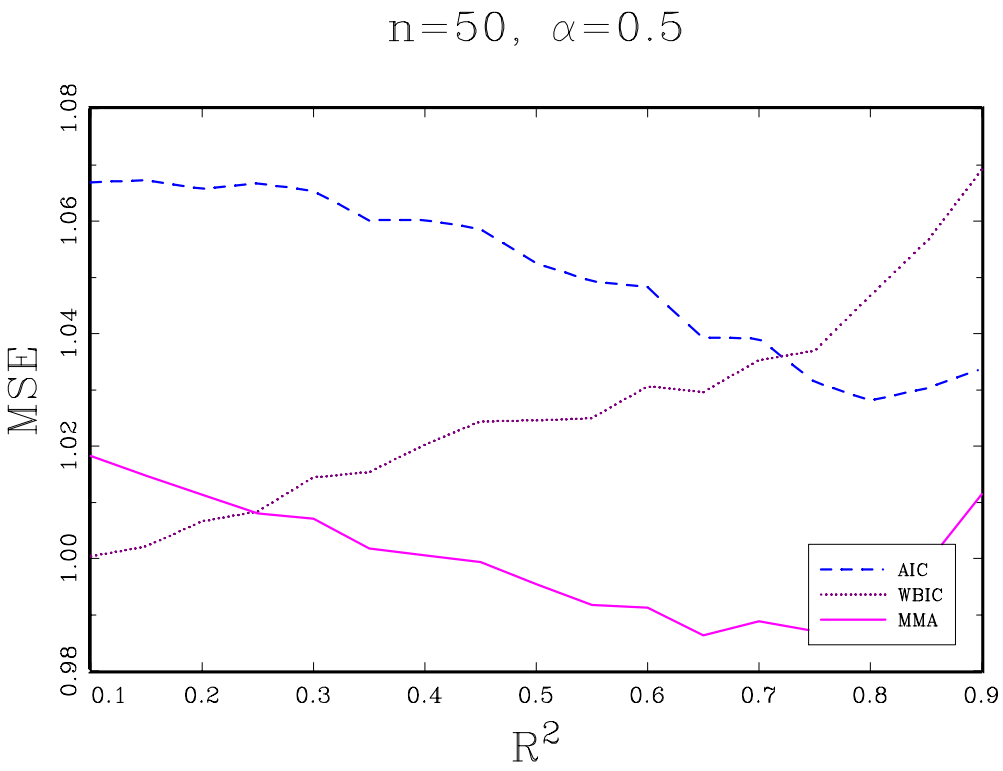
[29] Stock, J.H. and M. W. Watson (2005b):"An empirical comparison of methods for forecasting using many predictors," working paper, NBER.

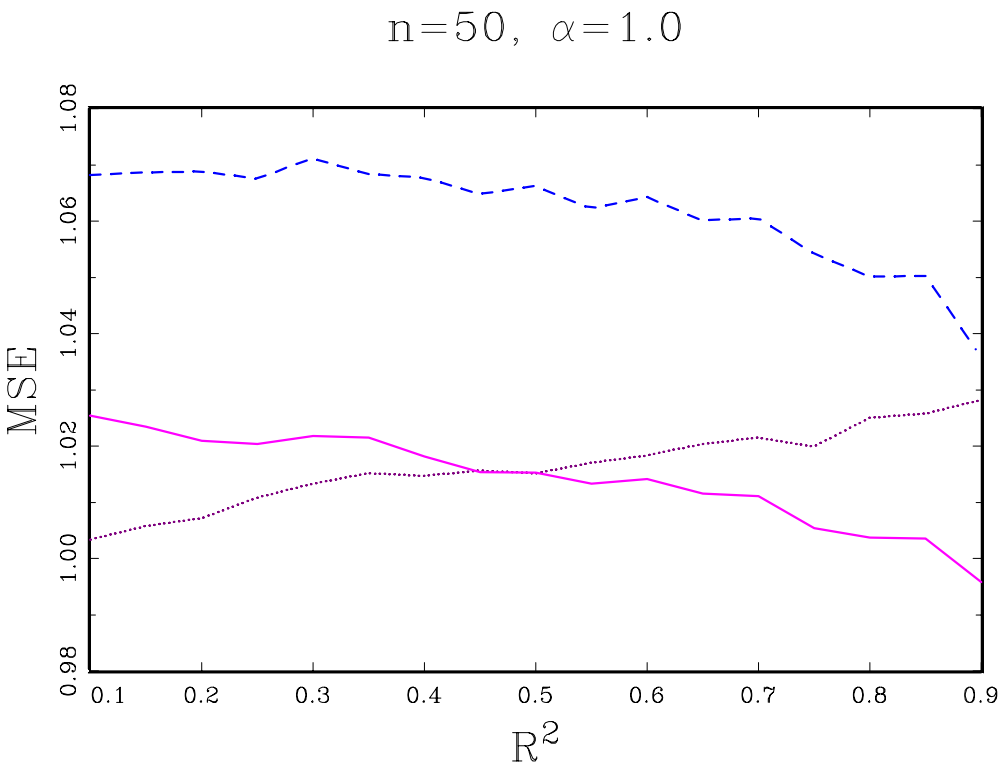
[30] Stock, J.H. and M. W. Watson (2006): "Forecasting with many predictors," in Elliott, Granger and Timmermann, eds., *Handbook of Economic Forecasting*, forthcoming, Elsevier.

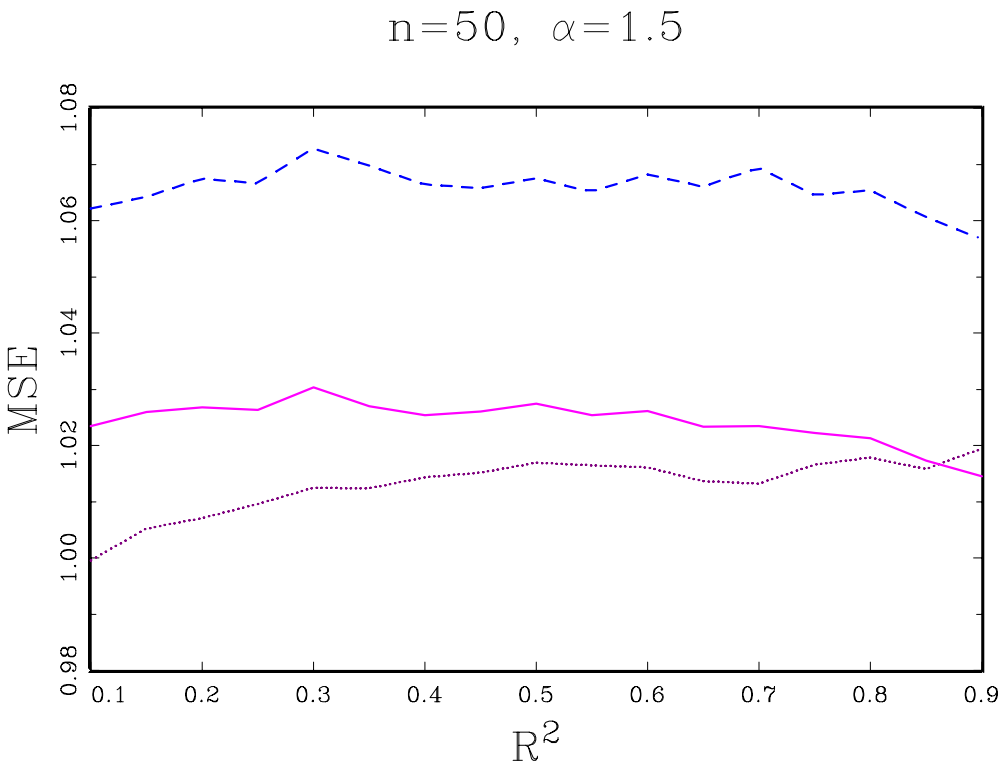
[31] Timmermann, Allan (2006): "Forecast Combinations," in Elliott, Granger and Timmermann, eds., *Handbook of Economic Forecasting*, forthcoming, Elsevier.

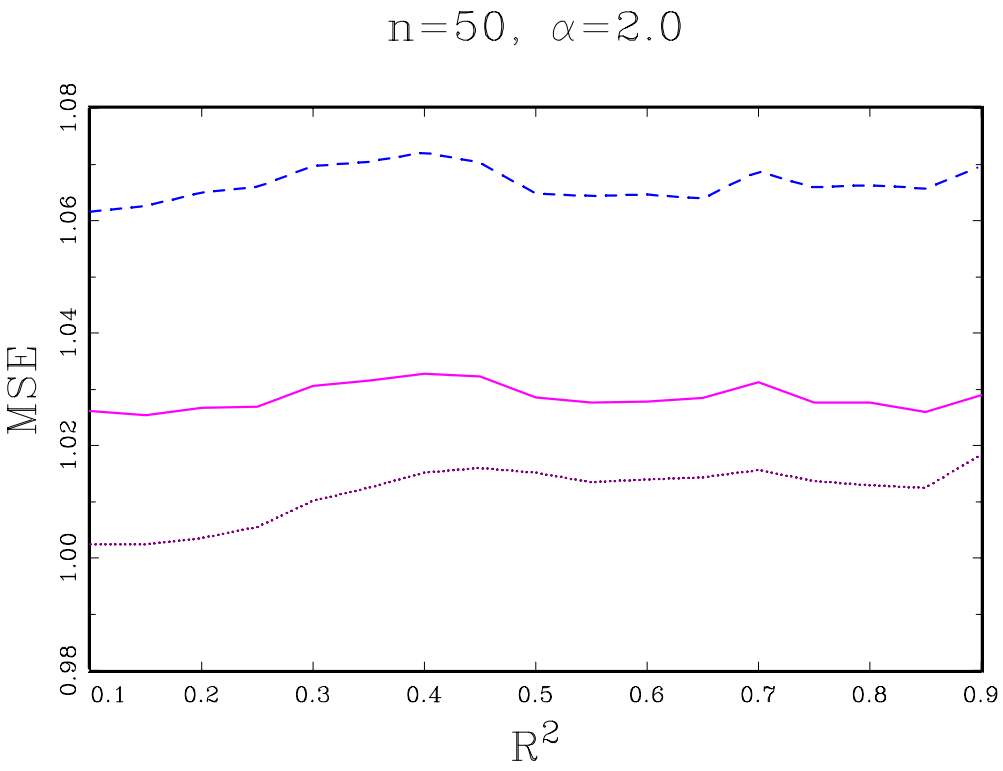
[32] Wright, Jonathan H. (2003a): "Bayesian model averaging and exchange rate forecasting," *Federal Reserve Board International Finance Discussion Papers*, 779.

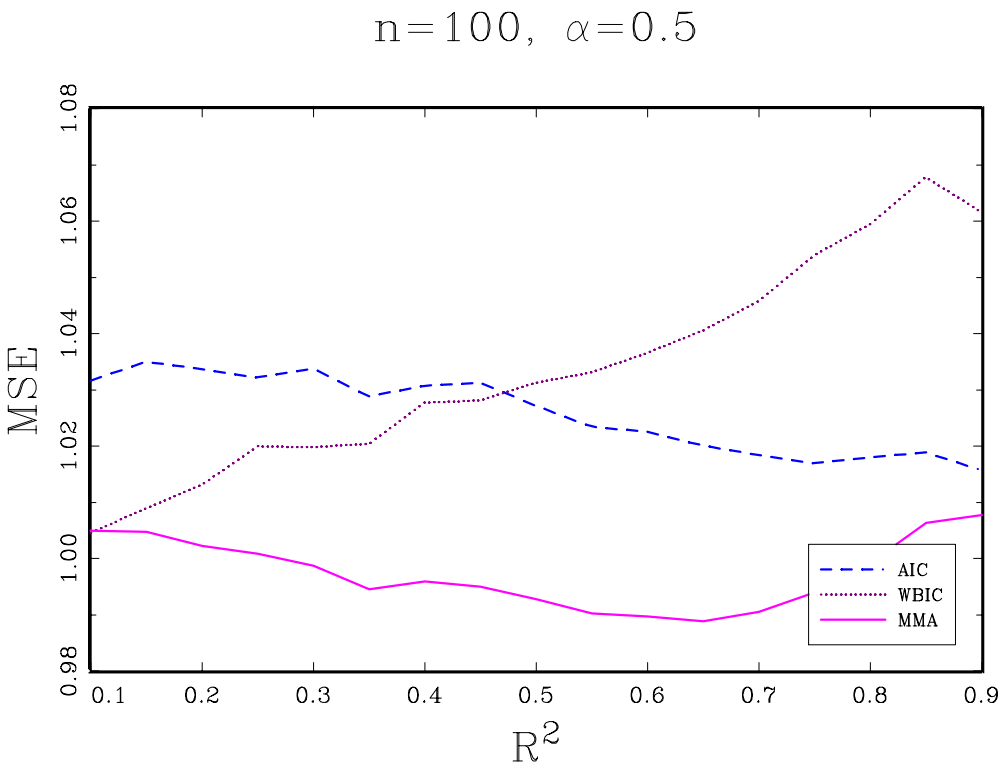
[33] Wright, Jonathan H. (2003b): "Forecasting US Inflation by Bayesian Model Averaging," *Federal Reserve Board International Finance Discussion Papers*, 780.

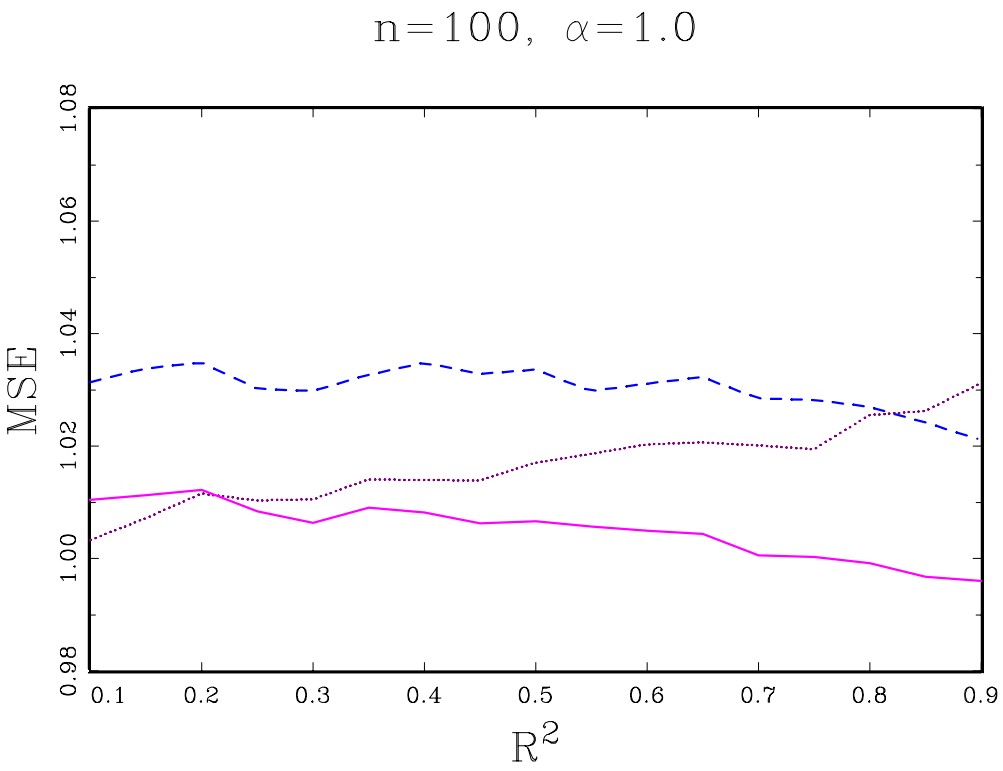

n=50, α=0.5
MSE
R2
AIC
WBIC
MMA

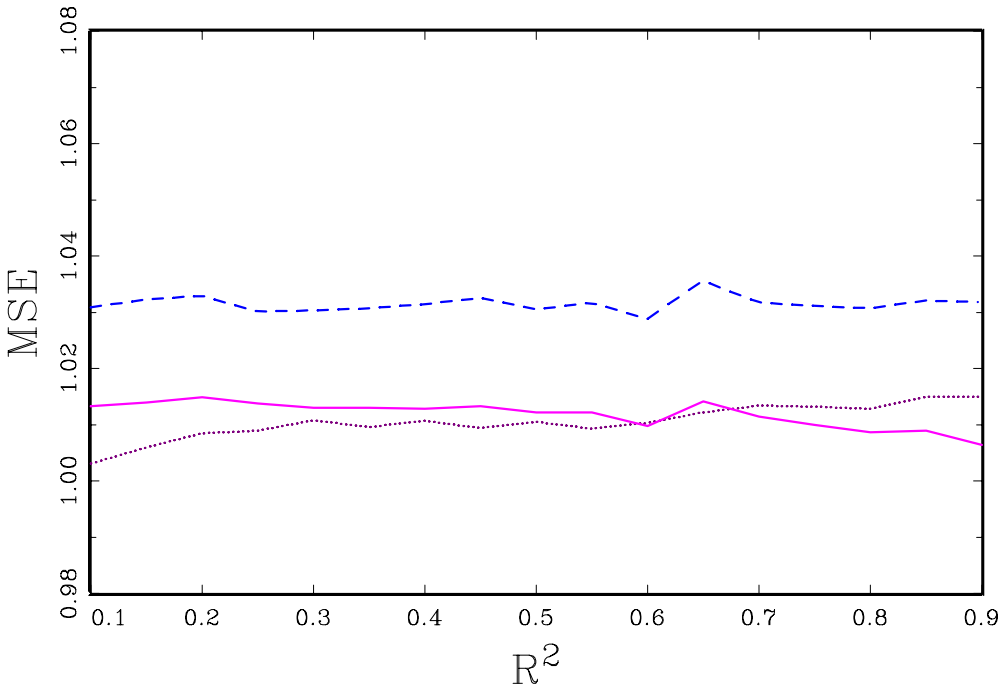

n=50, α=1.0
MSE
R2

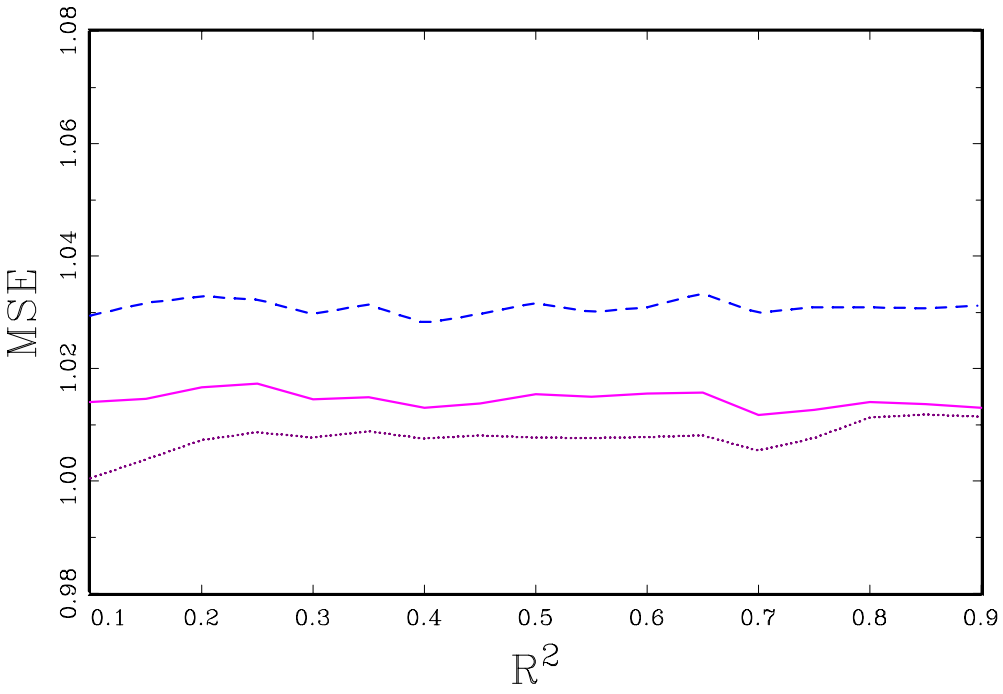

n=50, α=1.5
MSE
R2

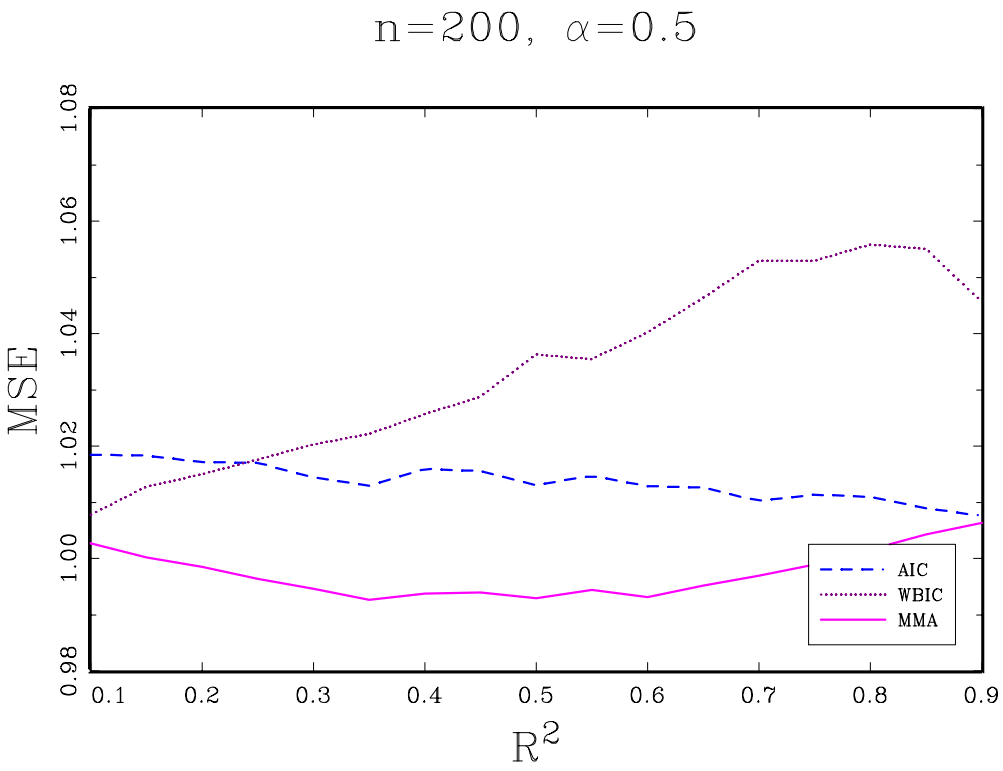

n=50, α=2.0
MSE
R2

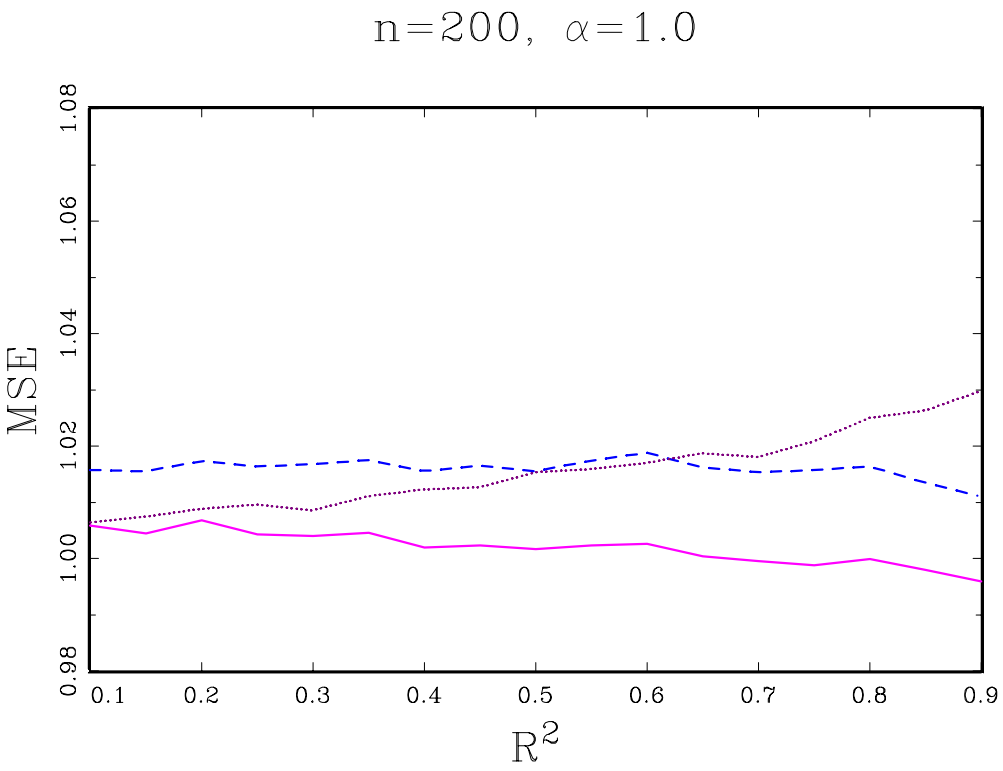

n=100, α=0.5
MSE
R2
AIC
WBIC
MMA

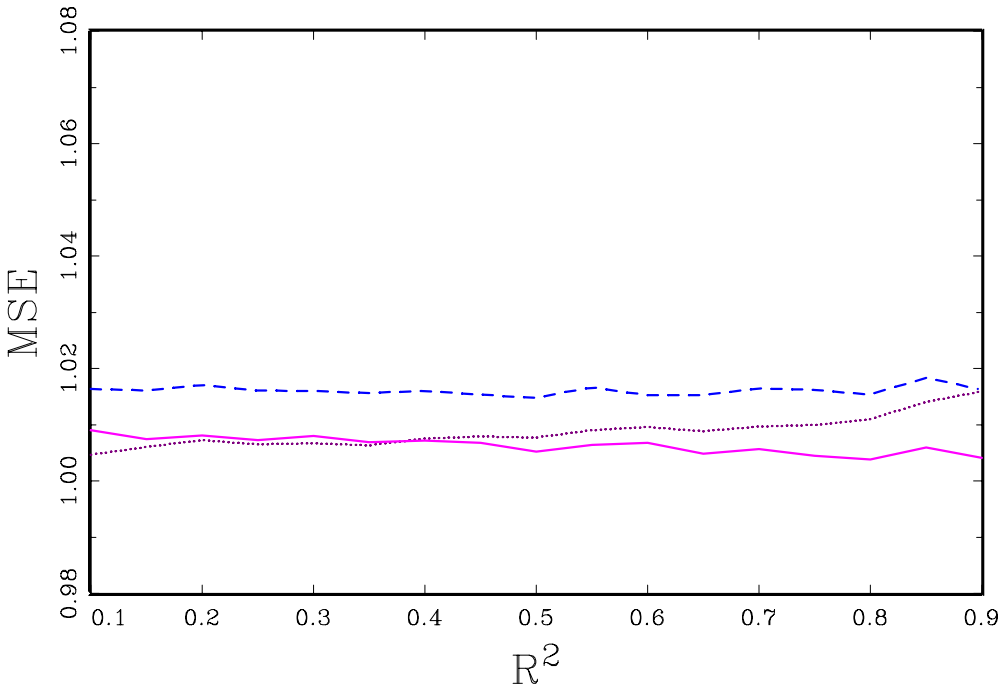

n=100, α=1.0
MSE
R2

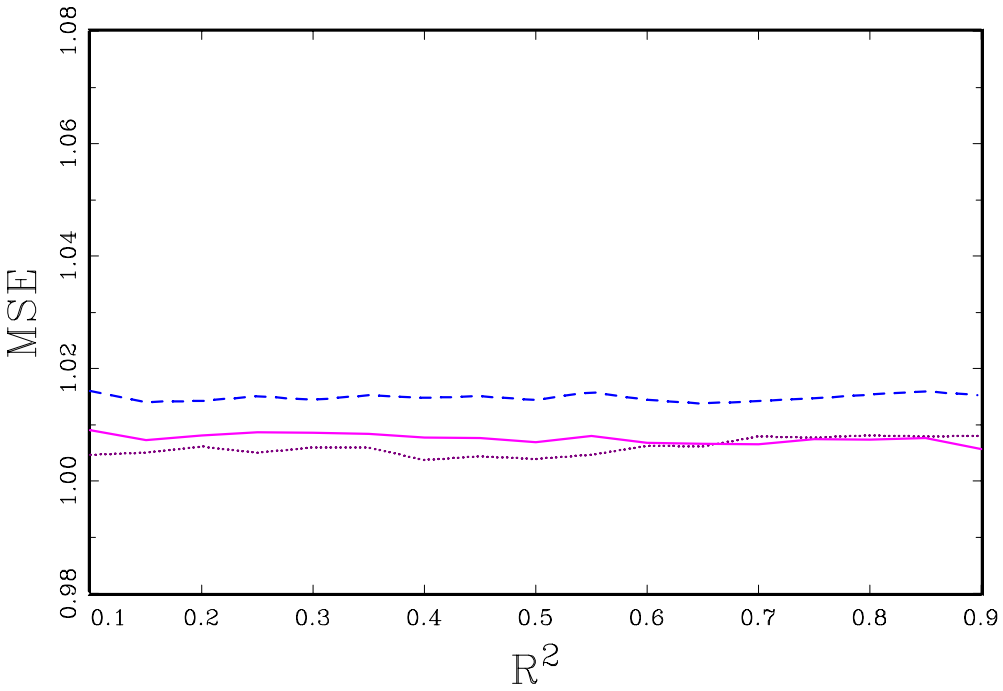

n=100, α=1.5
MSE
R2

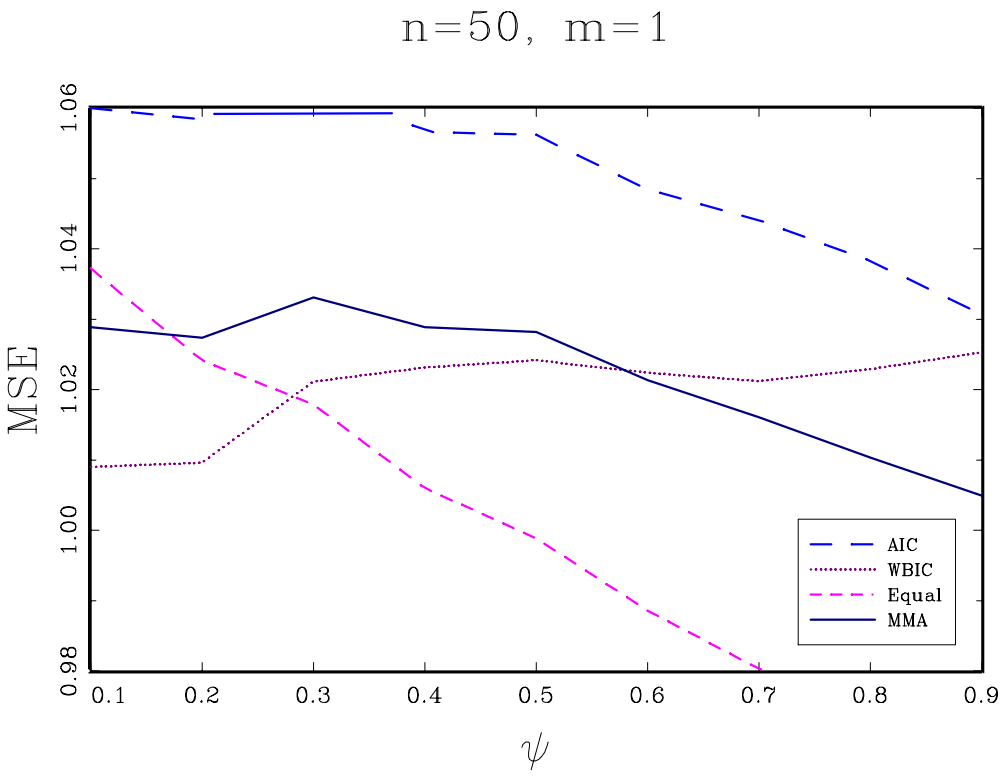

n=100, α=2.0
MSE
R2

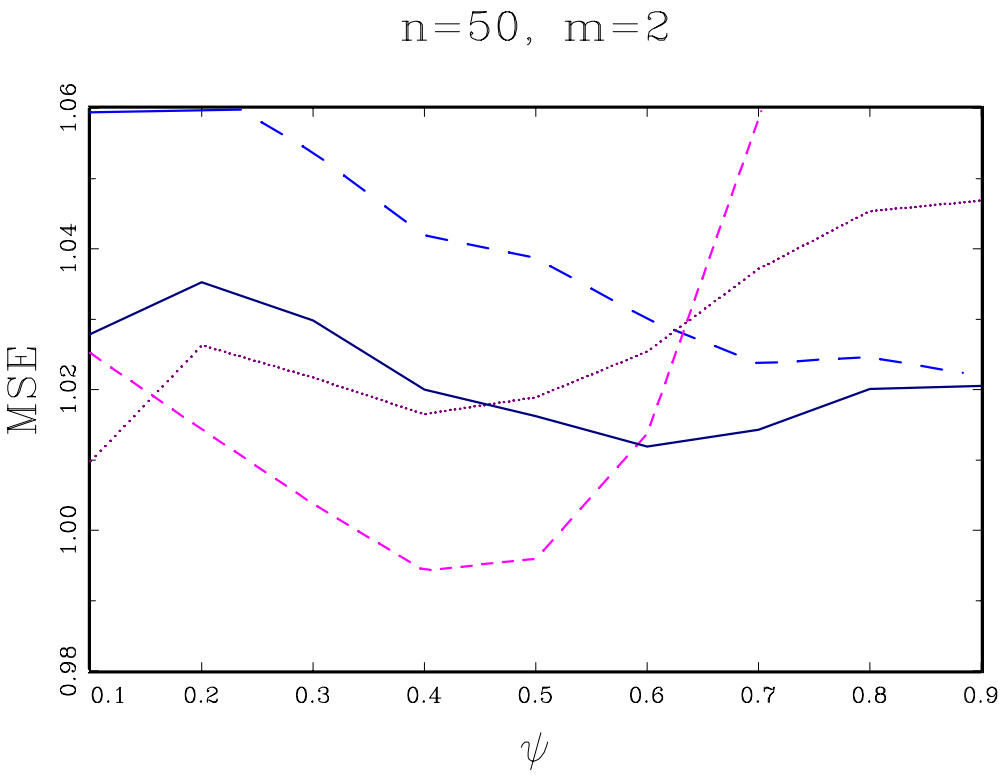

n=200, α=0.5
MSE
R2
AIC
WBIC
MMA

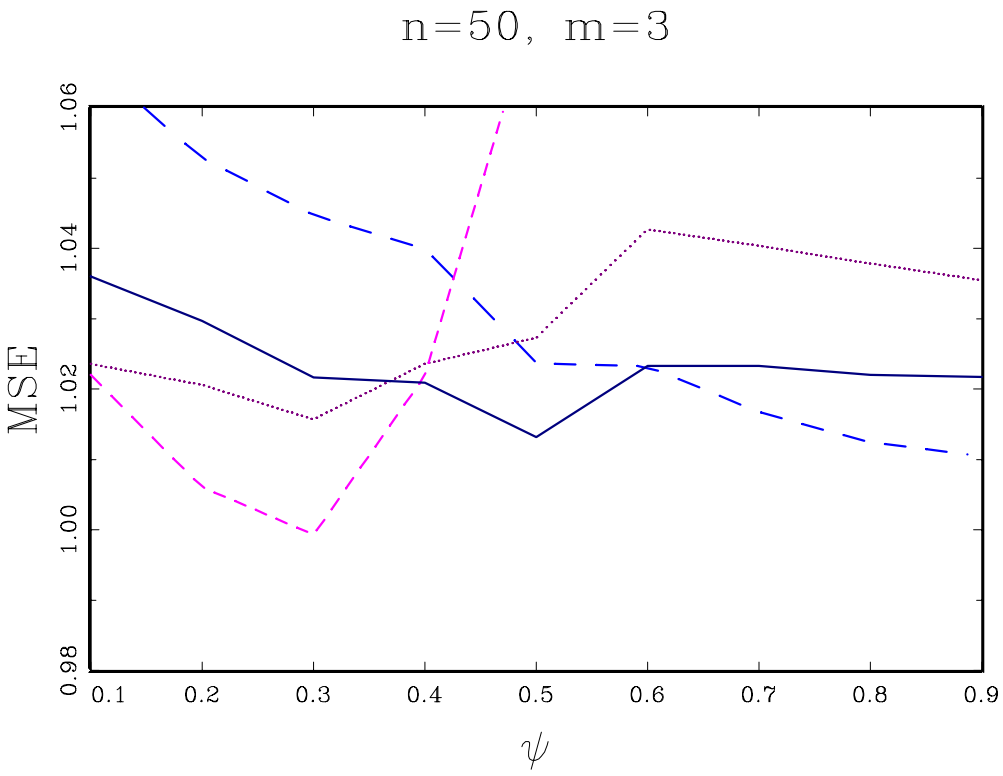

n=200, α=1.0
MSE
R2

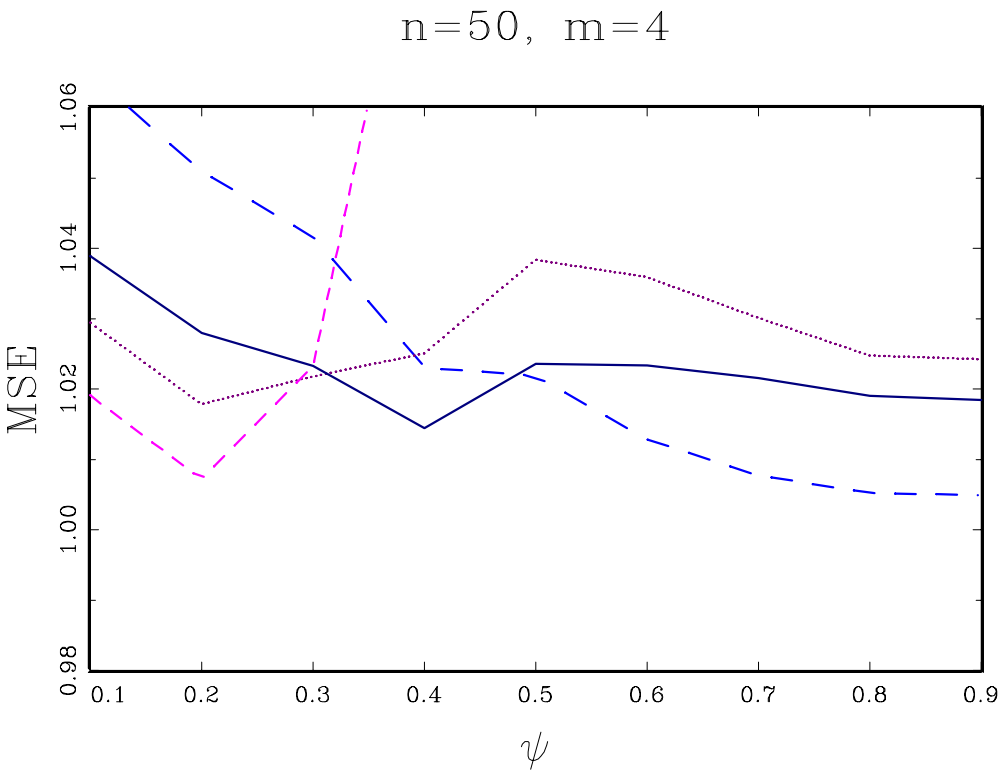

n=200, α=1.5
MSE
R2

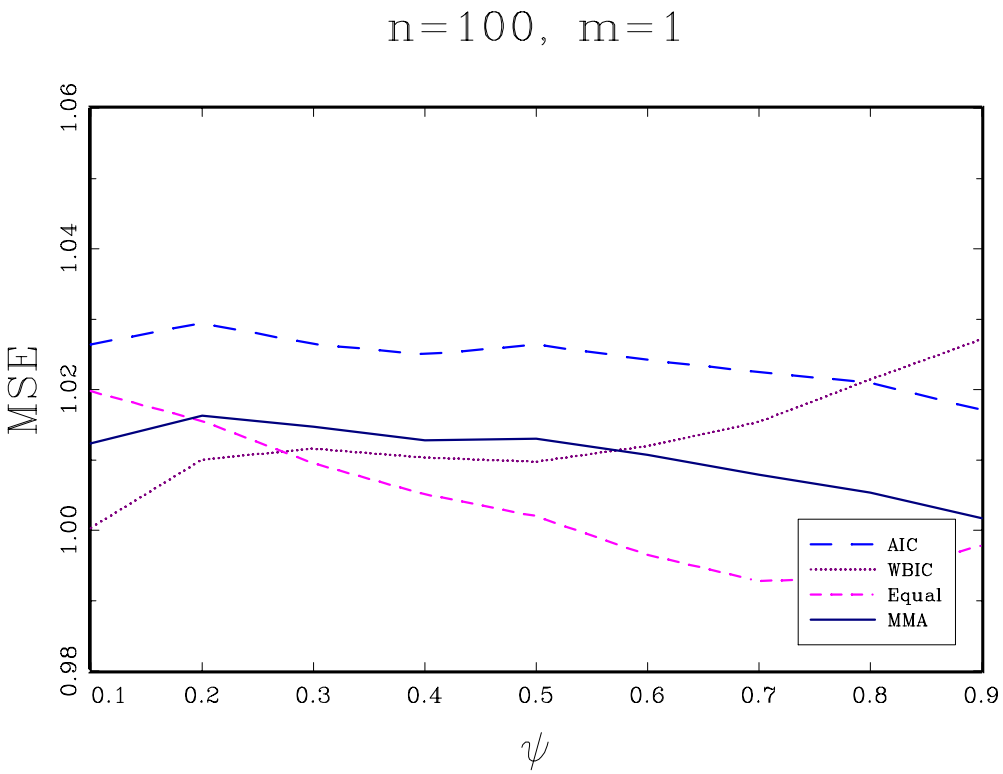

n=200, α=2.0
MSE
R2

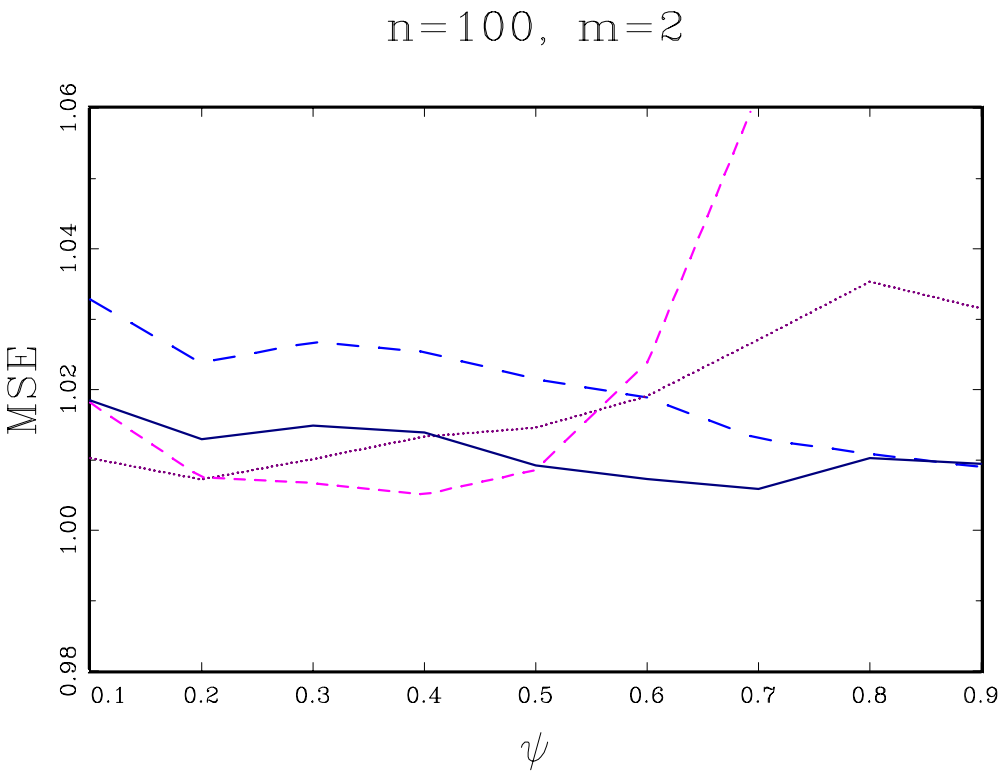

n=50, m=1
MSE
ψ
AIC
WBIC
Equal
MMA

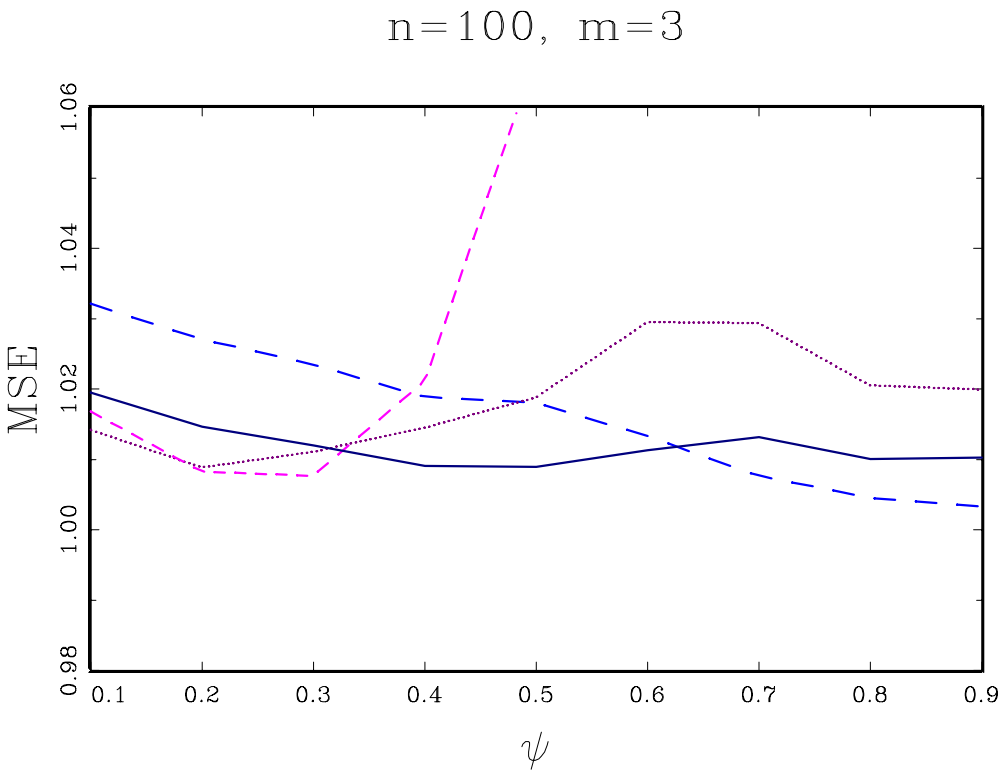

n=50, m=2
MSE
ψ

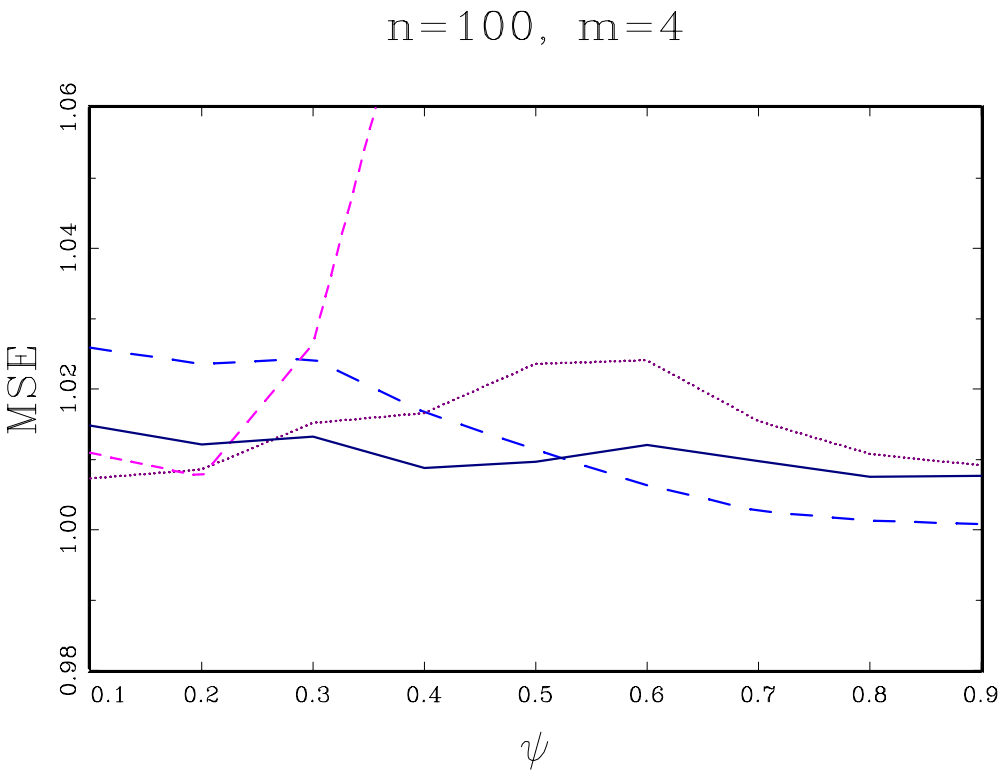

n=50, m=3
MSE
ψ

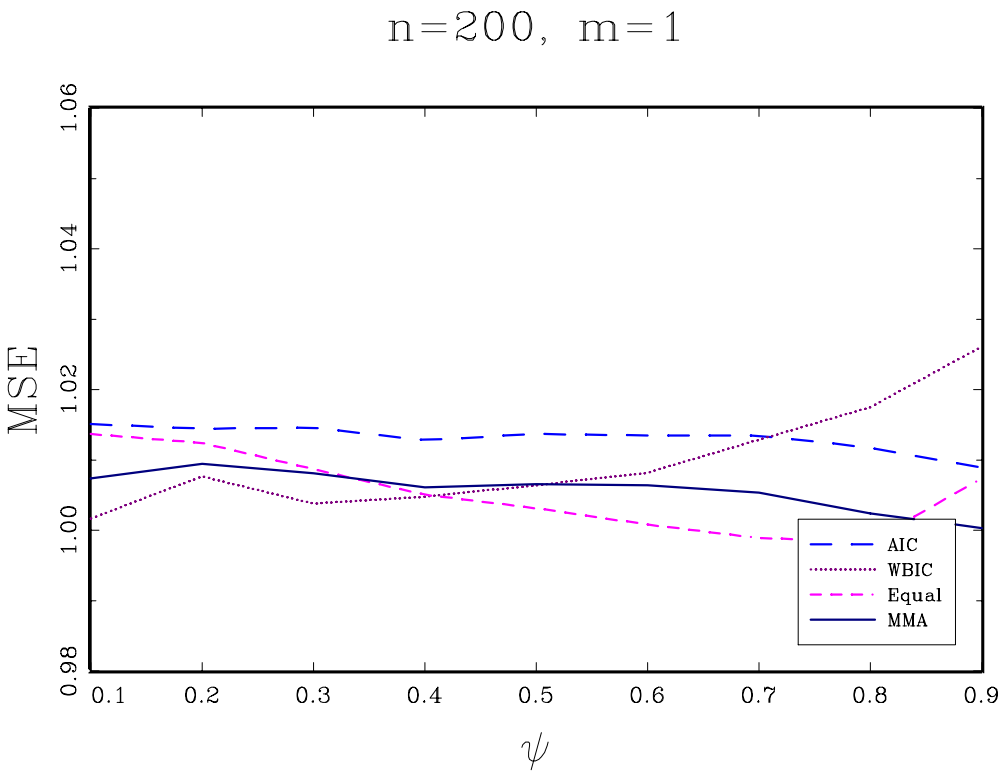

n=50, m=4
MSE
ψ

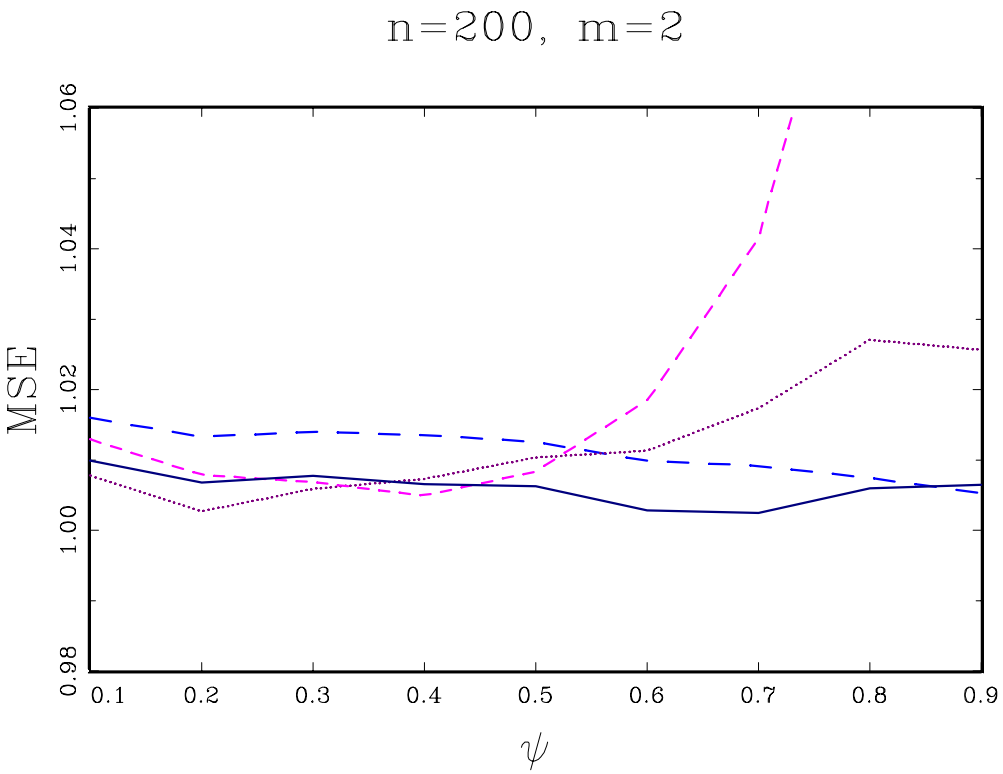

n=100, m=1
MSE
ψ
AIC
WBIC
Equal
MMA

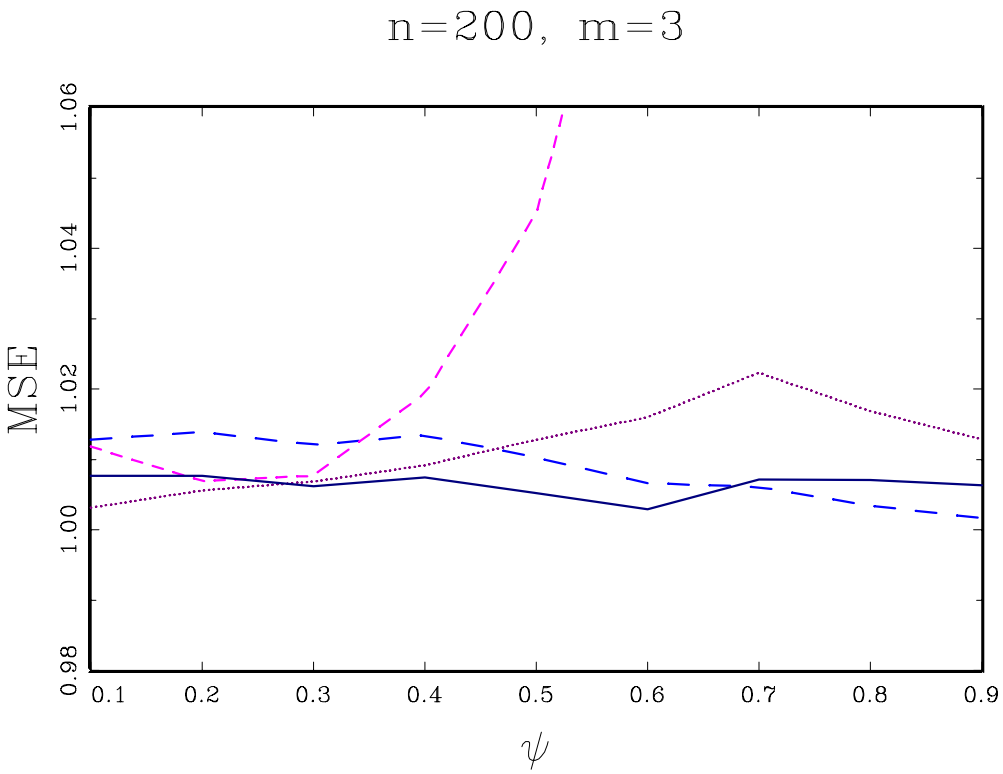

n=100, m=2
MSE
ψ

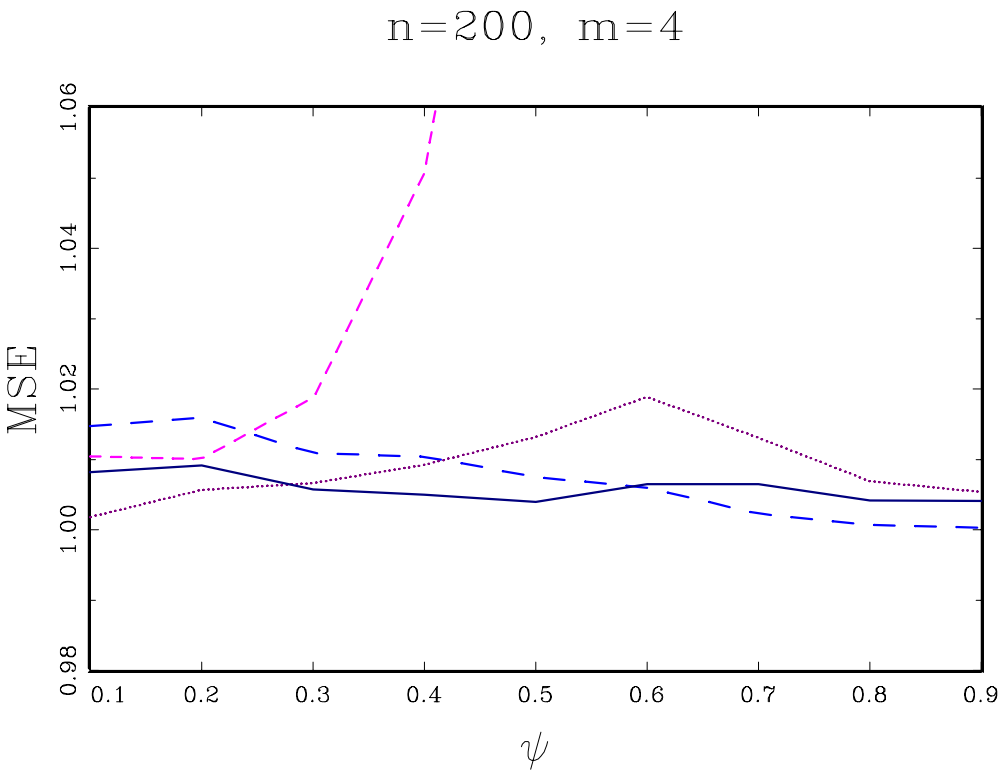

n=100, m=3
MSE
ψ


n=100, m=4
MSE
ψ


n=200, m=1
MSE
ψ
AIC
WBIC
Equal
MMA


n=200, m=2
MSE
ψ


n=200, m=3
MSE
ψ


n=200, m=4
MSE
ψ