DEPARTMENT OF ECONOMICS AND FINANCE

COLLEGE OF BUSINESS AND ECONOMICS

UNIVERSITY OF CANTERBURY

CHRISTCHURCH, NEW ZEALAND

How To Pick The Best Regression Equation: A Review And Comparison Of Model Selection Algorithms

Jennifer L. Castle, Xiaochuan Qin, and W. Robert Reed

WORKING PAPER

No. 13/2009

Department of Economics and Finance
College of Business and Economics
University of Canterbury
Private Bag 4800, Christchurch
New Zealand

WORKING PAPER No. 13/2009

How To Pick The Best Regression Equation: A Review And Comparison Of Model Selection Algorithms

Jennifer L. Castle[1], Xiaochuan Qin[2], W. Robert Reed[3†]

October 2009

Abstract: This paper reviews and compares twenty-one different model selection algorithms (MSAs) representing a diversity of approaches, including (i) information criteria such as *AIC* and *SIC*; (ii) selection of a "portfolio" or best subset of models; (iii) general-to-specific algorithms, (iv) forward-stepwise regression approaches; (v) Bayesian Model Averaging; and (vi) inclusion of all variables. We use coefficient unconditional mean-squared error (UMSE) as the basis for our measure of MSA performance. Our main goal is to identify the factors that determine MSA performance. Towards this end, we conduct Monte Carlo experiments across a variety of data environments. Our experiments show that MSAs differ substantially with respect to their performance on relevant and irrelevant variables. We relate this to their associated penalty functions, and a bias-variance tradeoff in coefficient estimates. It follows that no MSA will dominate under all conditions. However, when we restrict our analysis to conditions where automatic variable selection is likely to be of greatest value, we find that two general-to-specific MSAs, *Autometrics*, do as well or better than all others in over 90% of the experiments.

Keywords: Model selection algorithms, Information Criteria, General-to-Specific modeling, Bayesian Model Averaging, Portfolio Models, AIC, SIC, AICc, SICc, Monte Carlo Analysis, *Autometrics*

JEL Classifications: C52, C15

Acknowledgements: We acknowledge helpful comments from David F. Hendry, Jeffrey Wooldridge, participants at the 2008 Econometrics Society Australasian Meetings (Wellington, New Zealand) and the 2009 New Zealand Econometrics Study Group Meetings (Christchurch, New Zealand).

[1] Nuffield College, Oxford University, Oxford, UK
[2] Leeds School of Business, University of Colorado, Boulder, Colorado, USA
[3†] (Corresponding author) Department of Economics and Finance, University of Canterbury, Private Bag 4800, Christchurch 8140, NEW ZEALAND; email: bobreednz@yahoo.com.

WORKING PAPER No. 13/2009

How To Pick The Best Regression Equation: A Review And Comparison Of Model Selection Algorithms

I. INTRODUCTION

In many empirical applications, researchers face a choice of which variables to include in a regression model. Without some objective algorithm, non-systematic efforts may, at best, innocently miss superior specifications; or, at worst, strategically select results to support the researcher's preconceived biases. A substantial literature demonstrates that model selection matters. For example, many studies of economic growth find that results that are economically and statistically significant in one study are not robust to alternative specifications (cf. Levine and Renelt, 1992; Fernandez et al., 2001; Sala-i-Martin et al., 2004; Hoover and Perez; 2004; Hendry and Krolzig 2004). For these and related reasons, there is interest in automated model selection algorithms (MSAs) that can point researchers to the "best" model specification (Oxley, 1995; Phillips, 2005).

This study reviews and compares a large number of MSAs. In so doing, it addresses Owen's (2003, p. 622) call for evidence on the head-to-head performance of rival model selection methods. Our target audience is practitioners interested in using MSAs in their own research who seek guidance about which MSA(s) they should employ. The goal of this review is to identify factors which explain why MSAs succeed or fail in given data environments. While we make some tentative MSA recommendations, these are primarily meant to be suggestive, with the hope that they will stimulate further research on this subject.

As the list of all possible MSAs is uncountably large, we are forced to restrict ourselves to a subset of these. Even so, our comparison is extensive, consisting of twenty-one MSAs representing a number of different approaches to the model selection problem, including: (i) choosing a single best model based upon an information criterion (IC) such as the Akaike Information Criterion or the Schwarz Information Criterion (McQuarrie and Tsai, 1998); (ii) selection of a "portfolio" or best subset of models (Poskitt and Tremayne, 1985); (iii) general-to-specific algorithms such as the *Autometrics* package in PcGive (see Doornik, 2009a, for the former and Doornik and Hendry, 2007, for the latter); (iv) forward-stepwise regression approaches (see, e.g., Whittingham, Stephens, Bradbury, and Freckleton, 2006, and Doornik, 2008), (v) combination of models using Bayesian Model Averaging (Hoeting et al., 1999; Sala-i-Martin, 2004); and (vi) inclusion of all variables.

The literature on MSAs consists not only of a large number of alternative procedures, but also a variety of measures to determine "best" MSA performance. A non-exhaustive list of performance measures includes counts of the number of times the MSA "overfits" (selects too many variables), "underfits" (selects too few variables), or correctly picks the true DGP (cf. McQuarrie and Tsai, 1999); and predictive efficiency (Kuha, 2004; Burnham and Anderson, 2004).

The measure of estimator performance employed in this review is the unconditional mean-squared error (UMSE) of estimated coefficients. This measure allows us to conceptually decompose the performance of the respective MSAs into bias and variance components. Our review will show that this decomposition provides a useful framework for understanding MSA performance.

Our review will also show that an MSA's "effective penalty function" – that is, the "cost" the respective MSA attaches to the selection of an additional variable – is a key MSA attribute. Penalty functions are unique to IC MSAs. However, the "effective penalty function" can be measured by an MSA's null rejection frequency (denoted "gauge" by Castle, Doornik and Hendry, 2008). This attribute of MSAs plays an important role in determining whether a given MSA is likely to succeed or fail in particular data environments.

Not surprisingly, we find that no single MSA performs best in all data environments. However, our results are able to identify a set of MSAs that perform best in data environments that are likely to be of particular interest to practitioners. We characterize these data environments by two conditions. The first occurs when the researcher believes, on the basis of *a priori* judgment, that there are many more candidate than "relevant" variables, making it difficult to decide which ones to include. The second occurs when there is a substantial degree of DGP noise, so that many variables are on the edge of statistical significance. Under these conditions, we find that *Autometrics* performs as well or better than all other MSAs in over 90% of experiments.

II. A FRAMEWORK FOR COMPARING MODEL SELECTION ALGORITHMS

The Problem. Our analysis focuses on the following problem. We have a data set consisting of N observations on variables $Y, X_1, X_2, \ldots, X_L$. We assume that the data generating process (DGP) producing these observations is given by:

$$(1) \qquad Y_n = \gamma + \beta_1 X_{1n} + \beta_2 X_{2n} + \cdots \beta_L X_{Ln} + \varepsilon_n, \; n = 1,2,\ldots,N,$$

where K of the β's are nonzero and L-K are zero, $1 \le K \le L$; and the ε_n are i.i.d., with $\varepsilon_n \sim N(0,\sigma^2)$. We want to choose the "best" MSA, where "best" is defined as the MSA that results in the most accurate estimates of the β's. We define this more precisely below.

The Model Selection Algorithms (MSAs). We study twenty-one different MSAs. These are listed in TABLE 1, along with a brief description. The first four are based on information criteria (IC). While there are many information criteria, most of these are asymptotically related to either the Akaike Information Criterion (*AIC*) or the Schwarz Information Criterion (*SIC*) (Weakliem, 2004). Both the *AIC* and the *SIC* have the same general form: $-2\ell +$ *Penalty*, where ℓ is the maximized value of the log-likelihood function for the given specification, and *Penalty* is a function that monotonically increases in the number of coefficients to be estimated. In both cases, smaller is better, and the specification with the smallest *AIC/SIC* value is considered to be "best." The *SIC* generally penalizes the inclusion of parameters more harshly than the *AIC*, and thus favors more parsimonious models.

The *AIC* and *SIC* have asymptotic justification. The *SIC* is consistent. That is, if the true DGP is included among the set of candidate models, the *SIC* will select the true DGP with probability approaching one as the sample size increases. The *AIC* is asymptotically efficient but not consistent (see Hannan and Quinn, 1979) as it assumes that the true DGP is not included in the set of candidate models. It selects the model having the smallest expected prediction error with probability approaching one as the sample size increases (Kuha, 2004).

It is well-known that both the *AIC* and *SIC* tend to "overfit" (i.e., include more variables than the DGP) in small samples. As a result, small-sample corrections for these have been developed by Hurvich and Tsai (1989) and McQuarrie (1999), respectively. These are denoted in TABLE 1 as *AICC* and *SICC*, where the last "C" denotes that it is the "corrected" version of the respective information criterion.

In the context of our analysis, the procedure for identifying the "best" coefficient estimates for these MSAs is as follows: IC values are calculated for all 2^L possible models.[1] Coefficient estimates are taken from the model with the lowest IC value. If a variable does not appear in that model, then the associated estimate of that coefficient is set equal to zero.

The next eight MSAs are based on the idea of selecting – not a single "best" model – but a "portfolio" of models that are all "close" as measured by their information criterion (IC) values. Poskitt and Tremayne (1987) derive a measure based on the posterior odds ratio, $\Re_m = exp\left[-\frac{1}{2}\left(IC_{min} - IC_m\right)\right]$, where IC_{min} is the minimum *IC* value among all 2^L models, and IC_m is the value of the respective *IC* in model *m*, $m=1,2,\ldots,2^L$. They suggest forming a portfolio of models all having $\Re_m \leq \sqrt{10}$. Alternatively, Burnham and Anderson (2004) suggest a threshold $\Re_m$ value of *2*. Our study considers both values. The MSAs $AIC < 2$, $AICC < 2$, $SIC < 2$, and $SICC < 2$ each construct portfolios of models that have *AIC*, *AICC*, *SIC*, and *SICC* values that lie within 2 of the minimum value model. The next four MSAs ($AIC < \sqrt{10}$, $AICC < \sqrt{10}$, $SIC < \sqrt{10}$,

[1] The intercept, γ, is fixed to enter all models.

and $SICC < \sqrt{10}$) do the same for models lying within $\sqrt{10}$ of the respective minimum value model.

The procedure for identifying "best" coefficient estimates for these MSAs is: Coefficient estimates are set equal to zero for variables that never appear in the portfolio. For variables that appear at least once in the portfolio of models, the respective coefficient estimates are calculated as the arithmetic average of all nonzero coefficient estimates.

The next three MSAs use an automated general-to-specific (AUTO) regression algorithm. These are taken from the *Autometrics* program available in PcGive 12 (see Doornik, 2009). *Autometrics* undertakes a multi-path tree search, commencing from the general model with all potential variables, and eliminates insignificant variables while ensuring a set of pre-specified diagnostic tests are satisfied in the reduction procedure, checking the subsequent reductions with encompassing tests.[2] While the *Autometrics* program allows researchers the freedom to set their preferred significance level, our analyses focus on 1% and 5% (*AUTO_1%* and *AUTO_5%*), as these are most common in the applied economics literature; and on $\frac{1.6}{N^{0.9}} \cdot 100\%$ (*AUTO_Variable*) to adjust the significance level for large sample sizes (see Hendry, 1995, p.490).[3]

Next are three forward-stepwise (FW) algorithms. The particular versions that we employ also come from PcGive 12 and use the same three significance levels as the preceding AUTO algorithms (*FW_1%*, *FW_5%*, and *FW_Variable*). Variables are added

[2] For the results reported in this paper, *Autometrics* bias corrects the coefficient estimates of the retained variables for the bias induced by model selection using a 2-step correction procedure, see Hendry and Krolzig (2005).

[3] $\frac{1.6}{N^{0.9}} \cdot 100\% \cong 5\%$ when $N=47$, and $\frac{1.6}{N^{0.9}} \cdot 100\% \cong 1\%$ when $N=281$.

to the model in order of significance, one at a time, until no further significant regressors are found. If included variables become insignificant as others are added, they are removed from the model. Both the AUTO and FW algorithms produce a single-best model and assign a coefficient estimate of zero to those variables that are not retained in the final model.

The next two MSAs are examples of Bayesian Model Averaging (Hoeting, Madigan, Raftery, and Volinsky, 1999). In Bayesian Model Averaging, a composite model is constructed by taking a weighted average of a set of models, which might consist of all possible models, with weights consisting of the posterior model probabilities. In the composite model, each of the variable coefficients equals the weighted average of the individual estimated coefficients for that variable. The model weights employ the maximized value of the corresponding log-likelihood functions. The two versions we analyze are: (i) *LLWeighted_All*, which uses the full set of 2^L models to construct weighted average coefficient estimates; and (ii) *LLWeighted_Selected*, which restricts itself to the set of all 2^{L-1} models where the given variable is included in the model.

The final MSA (*ALLVARS*) selects the full set of potential variables for inclusion in the "final model." As should be apparent, the great disparity in approaches underlying these MSAs makes it difficult to analytically compare the performance of all twenty-one MSAs, and this is all the more true with respect to their performance in finite samples. Hence our analysis turns to Monte Carlo experiments.

Monte Carlo Experiments and the Performance Measure. Our experiments all use the DGP, $Y_n = \gamma + \beta_1 X_{1n} + \beta_2 X_{2n} + \cdots \beta_L X_{Ln} + \varepsilon_n$, $n = 1,2,...,N$, where $\gamma = 5$,

$\beta_1 = \beta_2 = \cdots = \beta_K = 1$, $\beta_{K+1} = \beta_{K+2} = \cdots = \beta_L = 0$, $1 \le K \le L$. The X_k's are i.i.d. and standard normally distributed. They are fixed both within and across experiments. We abstract from correlated data in this set of experiments but we do not correct the fixed regressors for sample correlations. The ε's are i.i.d. and normally distributed with mean *0* and standard deviation, σ^2. σ^2 is fixed within an experiment but variable across experiments – as will shortly be described.[4] Each experiment has *K* "relevant" variables and *L-K* "irrelevant" variables, with relevancy defined according to whether that variable has a nonzero coefficient in the DGP.

We use unconditional mean-squared error (UMSE) of the coefficient estimates to compare the performances of the preceding MSAs.[5] Each experiment consists of 1000 simulated data sets/replications *r*. For each of these, and for each MSA, we produce a set of estimates, $\left(\hat{\beta}_{1,r}^{MSA}, \hat{\beta}_{2,r}^{MSA}, ..., \hat{\beta}_{L,r}^{MSA}\right)$.[6] These are used to calculate experiment-, MSA-, and coefficient-specific UMSE values as follows:

$$(2) \qquad UMSE_k^{MSA} = \frac{\sum_{r=1}^{1000}\left(\hat{\beta}_{k,r}^{MSA} - \beta_k\right)^2}{1000}, \; k = 1,2, \ldots, L.$$

Because UMSE is not generally comparable across coefficients, we assign a coefficient-specific ranking from *1* to *21*, with the MSA producing the lowest UMSE for that coefficient receiving a rank of *1*, the MSA with the next smallest UMSE receiving a rank of *2*, and so on. These rankings are then averaged across all *L* coefficients to produce an

[4] The AUTO and FW MSAs were run using a different random number generator (available in Ox, see Doornik, 2007) so the draws of ε_n will differ between the 6 AUTO and FW MSAs and the 15 others. The fixed regressors, X_k, are identical across all MSAs and experiments. The effect of different random number generators is minimal across 1000 replications.

[5] Earlier analyses also compared MSA performance based on mean absolute deviations. We found little difference between these two performance measures and thus only report the UMSE results.

[6] The intercept is omitted in the calculations as it is imposed in the selected model for all MSAs.

overall MSA ranking for that experiment. For example, if $L = 5$ and a given MSA has individual coefficient rankings $\{10,10,12,13,10\}$, this MSA would receive an average rank of *11* for that experiment.[7]

There are several advantages to using UMSE as a measure of MSA performance. First, it coincides with a key goal of estimation: that of producing accurate coefficient estimates. Other performance measures, such as predictive efficiency, may accept biased estimates of individual coefficients as long as accurate predictions are produced.[8] However, in many applications, such as policy analysis, the sizes of the individual coefficients are the object of measurement.

With respect to coefficient mean-squared error, there is some question whether one should focus on (i) both relevant and irrelevant variables, or (ii) just the set of relevant variables. Alternatively, one could focus on just the set of included variables (the conditional mean-squared error) as this is the observed model in any empirical application. The choice among these alternative performance measures comes down to the researcher's loss function. In this study, we assume the researcher attaches equal loss to misestimating (i) relevant and (ii) irrelevant variables; and (i) included and (ii) excluded variables.

To put this back into a policy analysis framework, our study assumes that there is equal loss to falsely attributing a policy impact to an irrelevant variable and falsely

[7] Ties were handled as follows. Let the MSAs be ranked in ascending order, MSA_1, MSA_2, … ,MSA_j, MSA_{j+1}, …, MSA_{j+m}, …, MSA_{21}; and suppose MSA_{j+1} to MSA_{j+m} are tied. Each of these receive rank $\sum_{i=1}^{m}(j+i)\Big/m$.

[8] The difference between these two measures can be considerable when there is substantial multicollinearity. When this occurs, omitted variable bias may cause coefficients to differ substantially from their population values with little cost in predictive accuracy.

concluding there is no policy impact for a relevant variable.[9] Policy-makers may well attach different weights to retaining/omitting relevant and irrelevant variables. However, many of our qualitative results will still be valid as long as positive loss is attached to misestimating both relevant and irrelevant variables.

Another reason for using UMSE is that it can be decomposed into (i) bias and (ii) variance components. Some of the MSAs are weak on one but strong on the other, so that their relative performance depends on tradeoffs between these two components. As we demonstrate, this provides insights about the conditions under which particular MSAs are likely to be effective.

Factors Affecting the Relative Performance of MSAs. In order to gain a better understanding of the determinants of MSA performance, we study four factors: (i) K, the number of relevant variables in the DGP (holding L constant); (ii) N, the total number of observations; (iii) L, the total number of variables available for selection in a given experiment; and (iv) ψ, the "non-centrality parameter," which provides a measure of DGP noise.[10] Each of these is briefly explained below.

We expect that MSA performance will systematically vary with K. Specifically, we expect that MSAs that tend to underfit (overfit) will perform relatively well when there are few (many) relevant variables in the DGP. To investigate this for given L, we run L consecutive experiments where K starts at 1 and progresses through L. As discussed below, the variance of the error term, σ^2, is adjusted as K increases to hold constant the expected value of the t-statistics for the relevant variables.

[9] One could also divide relevant variables into "policy" and "control" variables, where the researcher is only concerned about accurate estimation of the coefficients for "policy" variables. Seen from this perspective, our experiments implicitly assume that all relevant variables are "policy" variables.

[10] See McQuarrie and Tsai (1998) for the importance of "signal-to-noise" ratio as a determinant of MSA performance for IC algorithms.

We also expect that MSA performance will systematically vary with N since some of the MSAs have established asymptotic properties. Specifically, *SIC, SICC*, and *Autometrics* all select the true DGP with probability *1* in the limit as sample size increases. This should translate into desirable UMSE performance for sufficiently large N. Accordingly, we set $N = 75, 150, 500$, and 1500.

As robustness checks, we also vary the total number of candidate variables, L, and the amount of noise disguising the true DGP. L is set equal to *5, 10,* and *15*. While larger values would be desirable, we are limited by computational constraints since many of the MSAs require estimation of all possible 2^L models. This equates to *32,768* possible models when $L=15$, and demonstrates the infeasibility of many MSAs when L is large.[11]

For our measure of DGP noise we considered using model R^2 values. This, however, has an undesirable consequence. Given our experimental design, an increase in K causes model R^2 to increase when σ^2 is held constant. If we compensate by increasing σ^2 to hold R^2 constant, we will lower the average sample *t*-statistic for relevant variables; or to state it differently, we will lower the retention rates associated with significance-driven variable selection. Instead, we use the "non-centrality parameter", $\psi \equiv E[t]$, as our measure of DGP noise.

A *t*-test of $H_0 : \beta_k = 0$ is given by $t_k = \frac{\hat{\beta}_k}{\sigma_{\hat{\beta}_k}}$. If the X_k's are i.i.d., then $\sigma^2_{\hat{\beta}_k} = \frac{\sigma^2}{N\sigma^2_{X_k}}$. It follows that $\psi_k = \frac{\beta_k}{\sqrt{\sigma^2 / N\sigma^2_{X_k}}}$. In our experimental design, $\beta_k = 1$ and

[11] This confers a computational advantage for those MSAs that don't require estimation of all possible models. For example, *Autometrics* can handle more variables than observations so L is not constrained, even by N (see Doornik, 2009b).

$\sigma_{X_k}^2 = 1$. Thus, the variance of the error term in the DGP can be adjusted to produce target values of ψ according to the relationship,

$$(3) \qquad \sigma^2 = \frac{N}{\psi^2}.$$

ψ has the attractive property that it is independent of *K* and *L* for a given sample size. As a result, a given value of ψ represents the expected value of the sample *t*-statistic for any of the relevant variables – no matter the model specification.[12]

In summary, our experimental framework is designed to compare MSA performance across a wide variety of simulated data environments. We produce a total of 360 experiments spanning a wide range of values for *K*, *L*, *N*, and ψ (cf. TABLE A2 in the Appendix).

III. RESULTS

Overall Performance of MSAs. TABLE 2 summarizes our overall findings. The first two columns report mean and median rankings for all 360 experiments. The next two columns report minimum and maximum rankings. A smaller rank indicates better performance, with *1* being best. The individual unit of observation is the experiment.

For example, the mean ranking for *SIC* over all 360 experiments is *10.6*. The best ranking achieved by this MSA in any one experiment is *4.7* (for the experiment *K=3*, *L=10*, *N=1500*, and $\psi=6$). This number is itself an average rank over the 10 coefficients

[12] The power to reject the null hypothesis $H_0 : \beta_k = 0$ can be calculated as a function of ψ and α by $P(t \geq c_\alpha \mid E[t] = \psi) \approx P(t - \psi \geq c_\alpha - \psi \mid H_0)$, where c_α is the critical value for a given significance level, α. The associated retention rates are largely independent of *N*, except to the extent that *N* affects the critical value, c_α. TABLE A1 records powers for a single *t*-test for different values of ψ and α when *N=75*. For example, there is a *16%* probability of retaining a variable with a non-centrality of *1* using a significance level of α*=5%*, which increases to *50%* for $\psi=2$ and *100%* for $\psi=6$. This provides a benchmark against which to compare each MSA's performance.

in that experiment. The worst ranking achieved by this MSA is *18.1* (for the experiment *K=10*, *L=10*, *N=1500*, and ψ*=2*). TABLE 2 ranks the 21 MSAs in descending order, with the best MSA (as measured by mean rank) listed first.

In terms of overall performance, the top three MSAs, as measured by both mean and median rankings, are the three *Autometrics* MSAs. The best of the three, *AUTO_5%*, has an average ranking a full rank better than its next best, non-*Autometrics* competitor.

Moving further down the table, we see that portfolio MSAs sometimes perform better than their non-portfolio analogs (cf. $AICC < \sqrt{10}$ and $AICC < 2$ versus *AICC*) and sometimes worse (cf. *SIC* versus $SIC < \sqrt{10}$ and $SIC < 2$). We also find that model averaging over all possible models (*LLWeighted_All*) is generally superior to model averaging over only those models in which the respective variable appears (*LLWeighted_Selected*), even though the former produces biased coefficient estimates. That being said, there are data environments where *LLWeighted_Selected* does better.

The worst-performing MSA is *ALLVARS*. Accordingly, we can conclude that it is not a good idea – as a general strategy – to include all potential variables in a regression specification.

The wide range of minimum and maximum values makes it very clear that no single MSA always performs best, or worst. For example, consider *ALLVARS*. While it generally performs poorly, it does better than any other MSA when all the candidate variables are relevant (*K=L*) because the estimated model is the DGP for this specification.[13]

[13] The median ranking for *ALLVARS* over the 36 experiments where *K=L* is *1.20*. The next closest MSA has a median rank of 3.15.

Care must be exercised in interpreting these rankings as they incorporate sampling error. We can calculate standard error bands for the MSAs to assess whether the UMSEs are significantly different for each MSA. Formal statistical tests such as Diebold and Mariano (1995) quickly become infeasible as there are *3900* UMSEs per MSA in the set of experiments conducted. Instead, we compare the average UMSE for each MSA ($\frac{1}{3900}\sum_{r=1}^{3900} UMSE_r^{MSA} = \overline{UMSE}^{MSA}$) against the average UMSE for all *21* MSAs ($\frac{1}{81900}\sum_{r=1}^{3900}\sum_{j=1}^{21} UMSE_r^{MSA_j} = \overline{UMSE}$).

Generally UMSEs cannot be directly compared but we use a result by Rao (1952, p.214) that states that the variance of $ln(UMSE)$ is independent of UMSE, enabling comparisons across log UMSEs. We record $ln\left(\overline{UMSE}^{MSA}\right)$ against $ln\left(\overline{UMSE}\right)$ in FIGURE 1.[14] Standard error bands can be computed around the mean using $\sigma^2_{ln\left(\overline{UMSE}\right)} = \frac{2}{M}$ asymptotically, where M is the number of replications, so the standard error bands are given by $\pm 2\sqrt{2/1000} = \pm 0.089$.

FIGURE 1 reports the results of testing for differences in the UMSEs of the respective MSAs. It is apparent that most UMSEs are not statistically different from each other. However, four MSAs lie outside the ±*2σ* bands: *LLWeighted_All*, *AUTO_5%*,

[14] More formally, the result by Rao states that if $u_i \sim \text{IN}\left[\mu, \sigma_u^2\right]$ with an unbiased estimate of the sample variance given by $\hat{\sigma}^2 = \frac{1}{M-1}\sum_{i=1}^{M}\left(x_i - \mu\right)^2$, then $V\left[\ln \hat{\sigma}_u^2\right]$ is independent of σ_u^2. Furthermore, the UMSE of the log variance is given by $UMSE\left[ln\hat{\sigma}_u^2\right] = \frac{1}{M}\sum_{i=1}^{M}\left[\left(ln\hat{\sigma}_u^2 - ln\sigma_u^2\right)^2\right]$ which is $\frac{2}{M}$ asymptotically.

ALLVARS, and *FW_1%*. In contrast, we would expect only one MSA to exceed the bands if the null that all methods are equally good were true.

It is noteworthy that there is relatively little correspondence between the $ln\left(\overline{UMSE}^{MSA}\right)$ values in FIGURE 1 and the rankings in TABLE 2. The explanation for this discrepancy lies in the distribution of UMSE values. For example, *FW_1%* has a significantly higher $ln\left(\overline{UMSE}\right)$ value than the other MSAs, and yet is ranked 4th in TABLE 2. Further investigation reveals that the distribution of UMSEs for *FW_1%* exhibits a bimodal distribution with a small mass at very large UMSEs when ψ is low and K is relatively large.[15] This illustrates a shortcoming of using rankings to summarize UMSE performance, though given the noncomparability of UMSEs across experiments, we are left with little alternative. Nevertheless, while the subsequent discussion focuses on rankings, we shall show that the key results also hold true when evaluated using $ln\left(\overline{UMSE}^{MSA}\right)$ values.

A Bias-Variance Framework for Understanding Relative Performance of MSAs. As noted above, measures of overall performance mask substantial differences between MSAs across different data environments. TABLE 3 illustrates the important role that K plays in determining MSA performance. It compares rankings for two IC algorithms (*AIC* and *SIC*) as K changes, holding L, N, and ψ constant (here set equal to $L=10$, $N=75$, and $\psi=2$). Columns (1) and (4) report the average rank (over the 10 coefficients) for each of the respective experiments (where each experiment consists of 1000 replications).

[15] Leeb and Pötscher (2005) show that the post-selection distribution is highly bimodal for low non-centralities, with many significant 'wrong-signed' estimates, which would adversely affect UMSE.

Columns (2/3) and (5/6) decompose these into average ranks over relevant and irrelevant variables.

When the number of relevant variables is relatively small, *SIC* outperforms *AIC*. As *K* increases, *SIC* monotonically loses ground to *AIC*. When *K=5*, the relative rankings of the two MSAs switch positions, with *AIC* outperforming *SIC*. Note that average performance within the sets of irrelevant and relevant variables is little affected by increases in *K*.

SIC outperforms *AIC* on irrelevant variables (cf. Columns 2 and 5). *AIC* outperforms *SIC* on relevant variables (cf. Columns 3 and 6). The switch in relative performance occurs because of changes in the weights of these two components. When there are many irrelevant variables and few relevant variables, *SIC*'s advantage on the former causes its overall performance to dominate *AIC*. As *K* increases, *AIC*'s advantage on relevant variables allows it to overtake *SIC*.

The explanation for *SIC*'s advantage (disadvantage) on irrelevant (relevant) variables must be due to the penalty function, since this is the only characteristic that distinguishes the two MSAs. *SIC* has a larger penalty function than *AIC* and therefore selects, on average, fewer irrelevant variables. Both *SIC* and *AIC* produce unbiased coefficient estimates for irrelevant variables. However, *SIC*-selected models will have lower variance since omitted variables are assigned coefficient values of *0*. Of course, *SIC* also admits fewer relevant variables. This biases coefficient estimates of the relevant variables since their population values are nonzero. Therefore, *SIC*'s larger penalty function harms its performance with respect to relevant variables.

In summary, a larger penalty function decreases the variance associated with irrelevant variables while also biasing coefficient estimates of relevant variables. We conjecture that this tradeoff between variance and bias is a key component of the relationship between MSA performance and *K*.

While we cannot demonstrate this conjecture analytically, our experiments allow us to investigate it empirically. The four IC MSAs can be strictly ordered in terms of increasing penalty functions: *AIC* < *AICC* < *SIC* < *SICC*. The preceding analysis leads us to two predictions regarding MSA performance with respect to *K*:

> ***Prediction #1 (K fixed):*** **If penalty functions are a key determinant of MSA performance, there should be a clear rank-order relationship between *AIC/AICC/SIC/SICC* for given *K*.**

> ***Prediction #2 (K variable):*** **If MSA with larger penalty functions are most advantaged (disadvantaged) with many irrelevant (relevant) variables, then $\frac{\Delta Rank}{\Delta K}$ (holding *L* constant) should display a clear rank-order relationship between *AIC/AICC/SIC/SICC.***

FIGURE 3 reports the performance results for all 180 experiments where *N=75* (cf. TABLE A1). The vertical axes report MSA rankings (from *1* to *21*). The horizontal axes are ordered by *K* (from *1* to *L*). There are three columns of figures, corresponding to *L = 5, 10,* and *15*; and six rows for ψ from *1* to *6* (with DGP noise greatest for smallest ψ). The four boldfaced lines indicate the rankings for *AIC/AICC/SIC/SICC*, with the dotted lines becoming increasingly solid for IC with larger penalty functions. The performances of the other seventeen MSAs are indicated by dotted, non-boldfaced lines.

Visual inspection confirms that the experimental results provide strong support in favor of both predictions. *141* of the *180* experiments (approximately *78%*) represented

in FIGURE 3, are characterized by a clear rank order for *AIC/AICC/SIC/SICC,* with either $AIC \leq AICC \leq SIC \leq SICC$ or $AIC \geq AICC \geq SIC \geq SICC$. The results are somewhat weaker, but still strong, for the additional 180 experiments that study *N = 150, 500,* and *1500* (cf. FIGURE A1 in the Appendix). Over all 360 experiments, *257* (approximately *71%*) satisfy *Prediction #1*. Note that these results make no allowance for sampling error.

A strong test of *Prediction #2* is $AIC \geq AICC \geq SIC \geq SICC$ (for *K=1*) and $AIC \leq AICC \leq SIC \leq SICC$ (for *K=L*). *13* of the *18* graphs in FIGURE 3 (approximately 7*2%*) satisfy this test. When one includes the additional eighteen graphs from FIGURE A1, the overall success rate is *25* of *36* (approximately 69%).

These results are consistent with the penalty function/bias-variance explanation of MSA performance. Inspection of FIGURE 3 indicates that other factors, such as DGP noise, also play a role. But they suggest that the bias-variance framework is useful for understanding the performance of other MSAs whose penalty function properties are not easily established analytically, such as the portfolio model MSAs.

Further insight into the performance of the MSAs can be gained by noting the relationship between the penalty function and the empirical non-null and null rejection frequencies, denoted potency and gauge (see Castle, Doornik and Hendry, 2008):

$$(4) \qquad potency^{MSA} = \frac{1}{K}\sum_{k=1}^{K} p_k^{MSA}$$

$$(5) \qquad gauge^{MSA} = \frac{1}{L-K}\sum_{k=K+1}^{L} p_k^{MSA}$$

where $p_k^{MSA} = \frac{\sum_{r=1}^{1000} 1\left(\hat{\beta}_{k,r}^{MSA} \neq 0\right)}{1000}$, $k = 1,2,\ldots,L$, is the retention rate. Potency reports the probability of retaining a relevant variable for a given MSA. Gauge reports the probability of retaining an irrelevant variable.

The gauge is informative as a measure of the penalty function. TABLE 4 reports the mean, median and standard deviation of the gauge for each of the MSAs over all experiments. The table orders MSAs from lowest to highest mean gauge; i.e., in order of decreasing penalty functions.[16] The increasing penalty functions for *AIC* < *AICC* < *SIC* < *SICC* are equivalent to decreasing gauges, as can be seen by both the mean and median gauge in TABLE 4. The mean gauge for *AIC* is *17%*, compared to *3%* for *SIC*. Hence, when there are many irrelevant variables and fewer relevant variables the tighter *SIC* criterion will outperform *AIC* and vice versa.

MSAs with larger penalty functions will also have lower potencies, or at least no higher potencies. FIGURE 2 records average potencies for the four IC MSAs averaged across the *360* experiments for each ψ. At low non-centralities the tighter criterion for *SIC* is most evident, reducing the potency relative to *AIC*, but at higher non-centralities the potencies converge towards unity. The IC potencies are close to the analytic retention probabilities (TABLE A2) when the gauge is matched to the nominal significance level, α. The higher retention rates of *AIC* versus *SIC* works to its advantage when there are many relevant variables and fewer irrelevant variables.

[16] Four MSAs have a gauge of unity for all experiments: *LLWeighted_All*, *LLWeighted_Selected*, *ALLVARS* and $AIC < \sqrt{10}$. In these cases, the notion of a "penalty function" is not well defined, because these MSA retain all variables by construction, albeit assigning different weights to estimated coefficients across models.

Progress towards identifying best MSAs. Having identified factors that affect MSA performance, it would be useful if our empirical results could provide guidance as to which MSA is most likely to produce the "best" model specification. Unfortunately, we know from TABLE 2 that no single MSA will be best in all circumstances. Further, we have shown this follows from the fact that the same penalty function behavior that confers an advantage to an MSA in one environment, will work to its disadvantage in another. However, since not all data environments are likely to be of equal interest to users of MSAs, we narrow our analysis to a subset of our experiments.

FIGURE 4 reports the same experiments as FIGURE 3, highlighting once again the effects of K, L, and ψ on MSA performance. For reasons discussed below, we now focus on the performance of *AUTO_1%*, which is represented by the solid, boldface line. All other MSAs are represented by dotted, non-boldfaced lines.

AUTO_1% generally performs very well when ψ is low (cf. the first three rows of FIGURE 4) and the ratio of relevant to irrelevant variables, $\frac{K}{L}$, is relatively small (cf. the lefthand side of each of the graphs in FIGURE 4). FIGURE 5 investigates the robustness of these results as sample size increases from $N = 75$ to $N = 150$, 500, and 1500. As before, the solid, boldfaced line represents *AUTO_1%*.

FIGURE 5 also highlights *AUTO_Variable*, which is represented by a dotted, boldfaced line. *AUTO_Variable* sets the significance level to $\frac{1.6}{N^{0.9}} \cdot 100\%$, as recommended by Hendry (1995) for large N. Note that $\frac{1.6}{N^{0.9}} \cdot 100\% > (<) 1\%$ when

$N < (>) 281$. A comparison of *AUTO_1%* and *AUTO_Variable* is enlightening given the earlier discussion on the relationship between MSA performance and *K*.

If we take the significance levels as a measure of the effective penalty functions associated with these MSAs (TABLE 4 confirms that the gauge is very close to the nominal significance level for AUTO), then FIGURE 5 is consistent with the bias-variance explanation that we previously used to explain the performance of *SIC* versus *AIC*: When $N < 281$, *AUTO_1%* has a larger penalty function than *AUTO_Variable* and thus performs better (worse) when *K* is relatively small (large). When $N > 281$, the positions are reversed as *AUTO_Variable* has the larger penalty function. Even so, there is relatively little difference in MSA performance between these two, even at large *N*.

We now identify a subset of our experiments that may be of particular interest to practitioners. In many situations, there will not be a need for automated routines to sort through alternative model selections. However, there are situations where automated selection can be of great value. Arguably, these will occur when the following two conditions hold:

1. The researcher believes, on the basis of *a priori* judgment, that there are many more candidate than relevant variables, making it difficult to decide which ones to select

2. There a substantial degree of DGP noise, so that many variables are on the edge of statistical significance

In the context of our simulations, and informed by FIGURES 4 and 5, we map these two conditions to (i) $\frac{K}{L} \leq 0.5$ and (ii) $\psi \leq 2$. TABLE 5 analyzes MSA performance for the 58 experiments where (i) half or less of the candidate variables are relevant and (ii) the sample *t*-statistics for the relevant variables have an expected value of either *1* or *2*.

Panel A repeats the analysis of TABLE 2 for the restricted set of 58 experiments. As before, MSAs are ranked in decreasing order of performance. The three *Autometrics* MSAs are (again) the top performers, but this time *AUTO_1%* and *AUTO_Variable* are virtually tied for best. Substantially further back (over two full ranks higher), are the two forward-stepwise algorithms, *FW_1%* and *FW_Variable*. Still further back are the information criteria MSAs.

Another look at the superior performance of the *Autometrics* MSAs is provided by Panel B of TABLE 5. These results report the frequency at which the respective *Autometrics* MSAs perform as well or better than all other MSAs – where "as well or better" means that the respective MSA has a rank equal to or lower than all other, non-*Autometrics* MSAs. *AUTO_1%* did at least as well as all other non-*Autometrics* MSAs in *54* out of *58* experiments (*93.1%*). *AUTO_Variable* did at least as well in *53* of the *58* experiments (*91.4%*). We emphasize again that these results make no allowance for sampling error in the experiments.

We once again test for significant differences across the MSAs. FIGURE 6 records $ln\left(\overline{UMSE}^{MSA}\right)$ values for the 58 experiments with $\frac{K}{L} \leq 0.5$ and $\psi \leq 2$, along with the ±2σ bands. The AUTO MSAs are significantly better than the mean for these experiments, supporting the results in TABLE 5. Further, there is a close correspondence between the $ln\left(\overline{UMSE}^{MSA}\right)$ values in FIGURE 6 and the UMSE ranks in TABLE 5. Not only that, there is also close correspondence between the UMSE ranks in TABLE 5 and the mean gauge values in TABLE 4. Since gauge provides a measure of an MSA's

effective penalty function, this provides further support for our penalty function explanation of MSA performance.

While penalty function behavior appears to be a main driver of MSA performance when $\frac{K}{L} \leq 0.5$ and $\psi \leq 2$, it is noteworthy that the correspondence is not perfect. Each of the AUTO MSAs has a slightly higher gauge (i.e., a looser penalty function) than its FW analog due to searching many reduction paths, yet the AUTO MSAs outperform in each case (compare *AUTO_1%* with *FW_1%*, *AUTO_Variable* with *FW_Variable*, and *AUTO_5%* with *FW_5%* in TABLES 4 and 5). We conjecture that while penalty function behavior is the main determinant of MSA performance in these data environments, there are other factors. Specifically, as *Autometrics* applies bias correction after selection it will drive retained coefficient estimates near the critical value towards the origin. This supplements its penalty function behavior and may be the explanation for why *Autometrics* is able to dominate FW MSAs with identical nominal significance levels, despite having higher gauge.

IV. CONCLUSION

Whether automated model selection algorithms (MSAs) are desirable is a subject that elicits strong responses (see e.g., Hansen, 2005). This review does not take a position on this issue. Instead, its main goal has been to identify factors that explain why MSAs succeed or fail in given data environments.

We compare twenty-one different MSAs, representing a variety of approaches including (i) information criteria such as *AIC* and *SIC*; (ii) selection of a "portfolio" or best subset of models; (iii) general-to-specific algorithms, (iv) forward-stepwise

regression approaches; (v) Bayesian Model Averaging; and (vi) inclusion of all variables. We use unconditional mean-squared error (UMSE) as our performance measure.

Among other results, we find that many MSAs differ substantially in their performance with respect to relevant and irrelevant variables. As the ratio of relevant to irrelevant variables changes, so do the relative performances of the MSAs. We relate these performance differences to the effective penalty functions associated with adding variables. MSAs with large effective penalty functions tend to do well when there are few relevant variables and many irrelevant variables. This occurs because the benefit of omitting irrelevant variables (lowered variance from assigning non-selected variables a coefficient of zero) dominates the cost of omitting relevant variables (greater bias). As the ratio of irrelevant to relevant variables changes, so do the respective benefits and costs. This implies that no MSA will dominate in all circumstances. Even the worst MSA in terms of overall performance – the strategy of including all candidate variables – sometimes performs best (viz., when all candidate variables are relevant).

Our comparison of different MSAs highlights the fact that MSAs differ in the weights they place on type I and type II errors. MSAs with loose criterion place more weight on type II errors and are less concerned with type I errors, retaining irrelevant variables with a very high probability. MSAs with tight criterion place a lot of weight on type I errors, controlling the null-rejection frequency at a cost of failing to retain relevant variables when they have low non-centralities. It is this trade-off that is at the heart of MSA performance.

While not our main goal, this review also supplies a very tentative recommendation to practitioners seeking guidance as to which MSA is "best." In many

situations there will be little need for automated routines to sort through alternative model specifications. However, there are situations where automated selection can be of great value to practitioners. Arguably, these will occur when (i) the researcher believes, on the basis of *a priori* judgment, that there are many more candidate than relevant variables, making it difficult to decide which ones to select; and (ii) there is a substantial degree of DGP noise, so that many variables are on the edge of statistical significance.

When we restrict our analysis to experiments where (i) half or less of the candidate variables are relevant, and (ii) the sample *t*-statistics for the relevant variables have an average value less than or equal to *2*, we find that two *Autometrics* MSAs perform consistently better than all others: one uses a significance value of *1%*, the other adjusts the significance value according to sample size. These two MSAs did as well or better than all other MSAs in over 90% of the respective experiments.

While these results are promising, it needs to be emphasized that they arise in a rarefied testing environment. Among other restrictions, our simulations assume orthogonal explanatory variables and spherical error terms. It remains to be seen whether the superior performance of *Autometrics* carries over when these restrictions are relaxed. It is hoped that this review will stimulate further research along these lines.

REFERENCES

Burnham, Kenneth P. and David R. Anderson. "Multimodel Inference: Understanding the AIC and BIC in Model Selection." Sociological Methods & Research Vol. 33, No. 2 (November 2004): 261-304.

Castle, Jennifer. L., Jurgen. A. Doornik and David. F. Hendry. "Model Selection when there are Multiple Breaks", Department of Economics, Oxford University. Working Paper No. 407, 2008.

Diebold, Francis. X. and Robert. S. Mariano. "Comparing Predictive Accuracy". Journal of Business and Economic Statistics. Vol. 13 (1995): 253-263.

Doornik, Jurgen. A. An Object-Oriented Matrix Language Ox 5, London: Timberlake Consultants Press, 2007.

Doornik, Jurgen. A. "Encompassing and Automatic Model Selection". Oxford Bulletin of Economics and Statistics, Vol. 70, (2008): 915-925.

Doornik, Jurgen. A. 'Autometrics', in Castle, Jennifer. L. and Neil Shephard, (eds), The Methodology and Practice of Econometrics. Oxford: Oxford University Press, 2009a. Ch. 4, pp. 88-121.

Doornik, Jurgen. A. "Econometric Model Selection With More Variables Than Observations", Nuffield College Working Paper (2009b). Available at: http://www.ecore.be/Papers/1245057273.pdf

Doornik, Jurgen. A. and David F. Hendry. Empirical Econometric Modelling Using PcGive 12: Volume I. London: Timberlake Consultants Press, 2007.

Fernández, Carmen, Eduardo Ley, and Mark F. J. Steel. "Model Uncertainty in Cross-Country Growth Regressions." Journal of Applied Econometrics Vol. 16 (2001): 563-576.

Hannan, Edward. J. and Barry G. Quinn. "The determination of the order of an autoregression". Journal of the Royal Statistical Society, B, Vol. 41 (1979): 190-195.

Hansen, Bruce. E. "Challenges for Econometric Model Selection". Econometric Theory, Vol. 21, (2005): 60-68.

Hendry, David F. Dynamic Econometrics. Oxford: Oxford University Press, 1995.

Hendry, David F. and Hans-Martin Krolzig, "New Developments in Automatic General-to-specific Modelling". In Econometrics and the Philosophy of Economics, edited by B.P. Stigum, Princeton University Press, (2003): 379-419.

Hendry, David F. and Hans-Martin Krolzig, "We Ran One Regression." Oxford Bulletin of Economics and Statistics, Vol. 66, No. 5 (2004): 799-810.

Hendry, David F. and Hans-Martin Krolzig, "The Properties of Automatic GETS Modelling." The Economic Journal, Vol. 115 (March 2005): C32-C61.

Hoeting, Jennifer A., David Madigan, Adrian E. Raftery, and Chris T. Volinsky. "Bayesian Model Averaging: A Tutorial." Statistical Science Vol. 14, No. 4 (1999): 382-417.

Hoover, Kevin D. and Stephen J. Perez, "Truth and Robustness in Cross-country Growth Regressions." Oxford Bulletin of Economics and Statistics Vol. 66, No. 5 (2004): 765-798.

Hurvich, C. M. and C. L. Tsai. "Regression and Time Series Model Selection in Small Samples." Biometrika, Vol. 76 (1989): 297-307.

Kuha, Jouni. "AIC and BIC: Comparisons of Assumptions and Performance." Sociological Methods & Research Vol. 33, No. 2 (November 2004): 188-229.

Leeb, Hannes. and Benedikt. M. Pötscher. "Model Selection and Inference: Facts and Fiction." Econometric Theory, Vol. 21 (2005): 21-59.

McQuarrie, Allan D. "A Small-sample Correction for the Schwarz SIC Model Selection Criterion." Statistics & Probability Letters Vol. 44 (1999): 79-86.

McQuarrie, Allan D. R. and Chih-Ling Tsai. Regression and Time Series Model Selection. Singapore: World Scientific Publishing Co. Pte. Ltd., 1998.

Owen, P. Dorian. "General-to-Specific Modelling Using PcGets." Journal of Economic Surveys Vol. 17, No. 4 (2003): 609-628.

Oxley, L. T. "An Expert Systems Approach to Econometric Modelling." Mathematics and Computers in Simulation Vol. 39 (1995): 379-383.

Phillips, Peter C. B. "Automated Discovery in Econometrics." Econometric Theory Vol. 21, No. 1 (2005): 3-20.

Poskitt, D. S., and A. R. Tremayne. "Determining a Portfolio of Time Series Models." Biometrika Vol. 74, No. 1 (1987): 125-137.

Rao, Calyampudi R. Advanced Statistical Methods in Biometric Research. New York: John Wiley, 1952.

Sala-i-Martin, Xavier, Gernot Doppelhofer, and Ronald I. Miller, "Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach." American Economic Review Vol. 94, No. 4 (2004): 813-835.

Weakliem, David L. "Introduction to the Special Issue on Model Selection." Sociological Methods & Research Vol. 33, No. 2 (November 2004): 188-229.

Whittingham, Mark. J., Philip A. Stephens, Richard B. Bradbury, and Robert P. Freckleton, "Why do we still use stepwise modelling in ecology and behaviour?" Journal of Animal Ecology, Vol. 75, (2006): 1182–1189.

TABLE 1
Description of Model Selection Algorithms (MSAs)

Information Criterion (IC) Algorithms:		
1) AIC	$AIC = ln\left(\hat{\sigma}^2\right) + \frac{2\left(\widetilde{K}+2\right)}{N}$	$\hat{\beta}_k$ is the estimate of β_k in the model with the minimum *IC* value. If X_k does not appear in that model, $\hat{\beta}_k = 0$. NOTE: $\hat{\sigma}^2$ is the maximum likelihood estimate of the variance of the error term; $\widetilde{K}$ is the number of coefficients in the model excluding the intercept; and *N* is the number of observations.
2) AICC	$AICC = ln\left(\hat{\sigma}^2\right) + \frac{\left(N+\widetilde{K}+1\right)}{\left(N-\widetilde{K}-3\right)}$	
3) SIC	$SIC = ln\left(\hat{\sigma}^2\right) + \frac{\left(\widetilde{K}+1\right)\cdot ln(N)}{N}$	
4) SICC	$SICC = ln\left(\hat{\sigma}^2\right) + \frac{\left(\widetilde{K}+1\right)\cdot ln(N)}{\left(N-\widetilde{K}-3\right)}$	
Portfolio Algorithms:		
5) AIC < 2	$\hat{\beta}_k$ is the average value of β_k estimates from the portfolio of models that lie within a distance $\Re = 2$ of the respective minimum *IC* model, where $\Re_m = exp\left[-\frac{1}{2}\left(IC_{min} - IC_m\right)\right]$, IC_{min} is the minimum *IC* value among all 2^L models, and IC_m is the value of the respective *IC* in model *m*, $m=1,2,\ldots,2^L$. If X_k does not appear in any of the portfolio models, $\hat{\beta}_k = 0$.	
6) AICC < 2		
7) SIC < 2		
8) SICC < 2		
9) AIC < $\sqrt{10}$	Same as above, except $\Re = \sqrt{10}$.	
10) AICC < $\sqrt{10}$		
11) SIC < $\sqrt{10}$		
12) SICC < $\sqrt{10}$		

General-to-Specific Regression Algorithms (Autometrics):	
13) AUTO_1% ***14) AUTO_5%*** ***15) AUTO_Variable***	$\hat{\beta}_k$ is the estimate of β_k in the best model as selected by the Autometrics program in PcGive 12, with the significance level, α, set equal to *1%*, *5%*, and $\frac{1.6}{N^{0.9}} \cdot 100\%$, respectively. If X_k does not appear in that model, $\hat{\beta}_k = 0$. $\hat{\beta}_k$ is bias corrected using a two-step procedure.
Forward-Stepwise Regression Algorithms	
16) FW_1% ***17) FW_5%*** ***18) FW_Variable***	$\hat{\beta}_k$ is the estimate of β_k in the best model as selected by the Forward Stepwise program in PcGive 12, with the significance level, α, set equal to *1%*, *5%*, and $\frac{1.6}{N^{0.9}} \cdot 100\%$, respectively. If X_k does not appear in that model, $\hat{\beta}_k = 0$.
Bayesian Model Averaging Algorithms:	
19) LLWeighted_All	$\hat{\beta}_k$ is the weighted average value of β_k estimates over all 2^L models, where model weights are determined according to $\omega_m = \frac{\ell_m}{\sum_{m=1}^{2^L} \ell_m}$, $m=1,2,\ldots,2^L$, and ℓ is the maximized value of the log likelihood function for model *m*. For the 2^{L-1} models where X_k does not appear in any of the portfolio models, $\hat{\beta}_k = 0$.

20) LLWeighted_Selected	$\hat{\beta}_k$ is the weighted average value of β_k estimates over the 2^{L-1} models where X_k is included in the regression equation. Model weights are determined according to $\omega_m = \frac{\ell_m}{\sum_{m \in \{\text{models that contain the variable } X_k\}} \ell_m}$.
<u>All Variables</u>:	
21) ALLVARS	$\hat{\beta}_k$ is the estimate of β_k in the specification in which all variables are included.

TABLE 2
Comparison of MSA Performance: All Experiments
(Sorted By Mean UMSE Rank in Ascending Order)

MSA	*Mean*	*Median*	*Minimum*	*Maximum*
AUTO_5%	9.4	9.4	3.7	17.6
AUTO_Variable	9.7	9.2	1.7	21.0
AUTO_1%	9.9	9.2	1.1	21.0
FW_1%	10.6	9.9	2.3	21.0
SIC	10.6	10.6	4.7	18.1
FW_Variable	10.8	9.8	3.4	21.0
SICC	10.9	10.3	4.0	19.2
SIC < 2	10.9	10.8	5.8	18.4
FW_5%	11.0	10.7	7.3	20.2
SICC < 2	11.1	11.0	5.4	18.8
AICC < $\sqrt{10}$	11.1	11.2	3.4	18.5
AICC < 2	11.1	11.5	3.3	15.6
SIC < $\sqrt{10}$	11.2	11.0	6.7	20.0
AIC < 2	11.2	11.8	2.6	16.9
LLWeighted_All	11.2	11.9	1.0	16.6
SICC < $\sqrt{10}$	11.3	11.1	5.6	20.1
AICC	11.3	11.2	3.7	19.2
AIC < $\sqrt{10}$	11.4	11.8	3.0	18.5
AIC	11.6	12.1	3.1	19.0
LLWeighted_Selected	11.8	12.5	1.9	19.7
ALLVARS	12.7	14.0	1.0	20.9

TABLE 3
Experimental Results for the Case: *L = 10, N = 75, ψ = 2*

Number of Relevant Variables (K)	*Mean Ranking of SIC Algorithm Over…*			*Mean Ranking of AIC Algorithm Over…*		
	All Variables (1)	*Irrelevant Variables (2)*	*Relevant Variables (3)*	*All Variables (4)*	*Irrelevant Variables (5)*	*Relevant Variables (6)*
1	8.0	7.1	16.0	13.6	14.1	9.0
2	8.9	7.0	16.5	13.2	14.3	9.0
3	9.9	7.1	16.3	12.7	14.3	9.0
4	10.7	7.0	16.3	12.1	14.2	9.0
5	11.7	7.4	16.0	11.4	14.2	8.6
6	12.7	7.5	16.2	10.6	13.8	8.5
7	13.5	7.3	16.1	10.0	14.3	8.1
8	14.6	8.0	16.3	9.4	15.0	8.0
9	15.4	8.0	16.2	8.6	14.0	8.0
10	16.2	---	16.2	8.0	---	8.0

TABLE 4
Comparison of MSA Gauge: All Experiments
(Sorted By Mean Gauge in Ascending Order)

MSA	*Mean*	*Median*	*Standard Deviation*
FW_1%	1.0%	1.0%	0.3%
AUTO_1%	1.4%	1.2%	0.7%
FW_Variable	2.2%	2.3%	1.5%
SICC	2.3%	2.4%	1.2%
AUTO_Variable	2.6%	3.0%	1.8%
SIC	3.4%	3.7%	2.0%
SIC < 2	4.8%	5.2%	2.4%
FW_5%	5.1%	5.0%	0.6%
AUTO_5%	5.4%	5.3%	0.9%
SICC < 2	7.2%	8.0%	4.2%
SICC < $\sqrt{10}$	8.1%	8.9%	4.0%
SIC < $\sqrt{10}$	12.3%	13.5%	7.1%
AICC	14.3%	14.5%	1.1%
AIC	17.4%	16.5%	2.4%
AICC < 2	36.1%	36.6%	4.9%
AIC < 2	45.3%	44.4%	2.5%
AICC < $\sqrt{10}$	85.4%	96.9%	18.5%
LLWeighted_All	100%	100%	0%
AIC < $\sqrt{10}$	100%	100%	0%
LLWeighted_Selected	100%	100%	0%
ALLVARS	100%	100%	0%

TABLE 5

Comparison of MSA Performance: Experiments Where $\psi \leq 2$ and $\frac{K}{L} \leq 0.5$

A. Comparison of UMSE Ranks (Sorted in Ascending Order of Mean UMSE Rank)

MSA	*Mean*	*Median*	*Minimum*	*Maximum*
AUTO_1%	4.6	4.0	1.1	10.6
AUTO_Variable	4.8	3.8	1.7	10.6
AUTO_5%	6.4	6.3	3.7	9.3
FW_1%	7.0	7.0	2.7	12.4
FW_Variable	7.9	7.9	3.4	12.4
SICC	8.5	8.0	5.0	12.3
SIC	9.2	9.2	5.2	12.1
SICC < 2	10.6	10.4	8.2	14.0
FW_5%	10.9	10.7	8.1	16.3
SIC < 2	11.1	10.8	8.6	14.5
LLWeighted_All	11.6	11.8	7.8	14.2
SICC < $\sqrt{10}$	11.9	11.5	10.5	15.2
SIC < $\sqrt{10}$	12.3	12.1	10.9	15.3
AICC	12.7	12.8	10.8	15.4
AIC	13.5	13.6	10.9	16.5
AICC < 2	13.7	14.0	11.0	15.6
AIC < 2	14.0	14.5	11.0	16.5
AICC < $\sqrt{10}$	14.2	14.2	10.9	18.3
AIC < $\sqrt{10}$	14.8	14.8	10.9	18.1
LLWeighted_Selected	14.9	15.0	10.5	19.2
ALLVARS	16.3	16.8	10.3	20.4

B. Percent of Experiments Where *Autometrics* MSAs Perform as Well or Better Than All Other MSAs

MSA	*Percent*
AUTO_1%	93.1
AUTO_Variable	91.4
AUTO_5%	46.6

NOTE: There are a total of 58 experiments where $\psi \leq 2$ and $\frac{K}{L} \leq 0.5$.

FIGURE 1
Log of the mean UMSE for each MSA with ±2σ bands across all 360 experiments (3900 MSEs per MSA)

FIGURE 2
Average potencies for the four IC MSAs for non-centralities from ψ=1 to ψ=6.

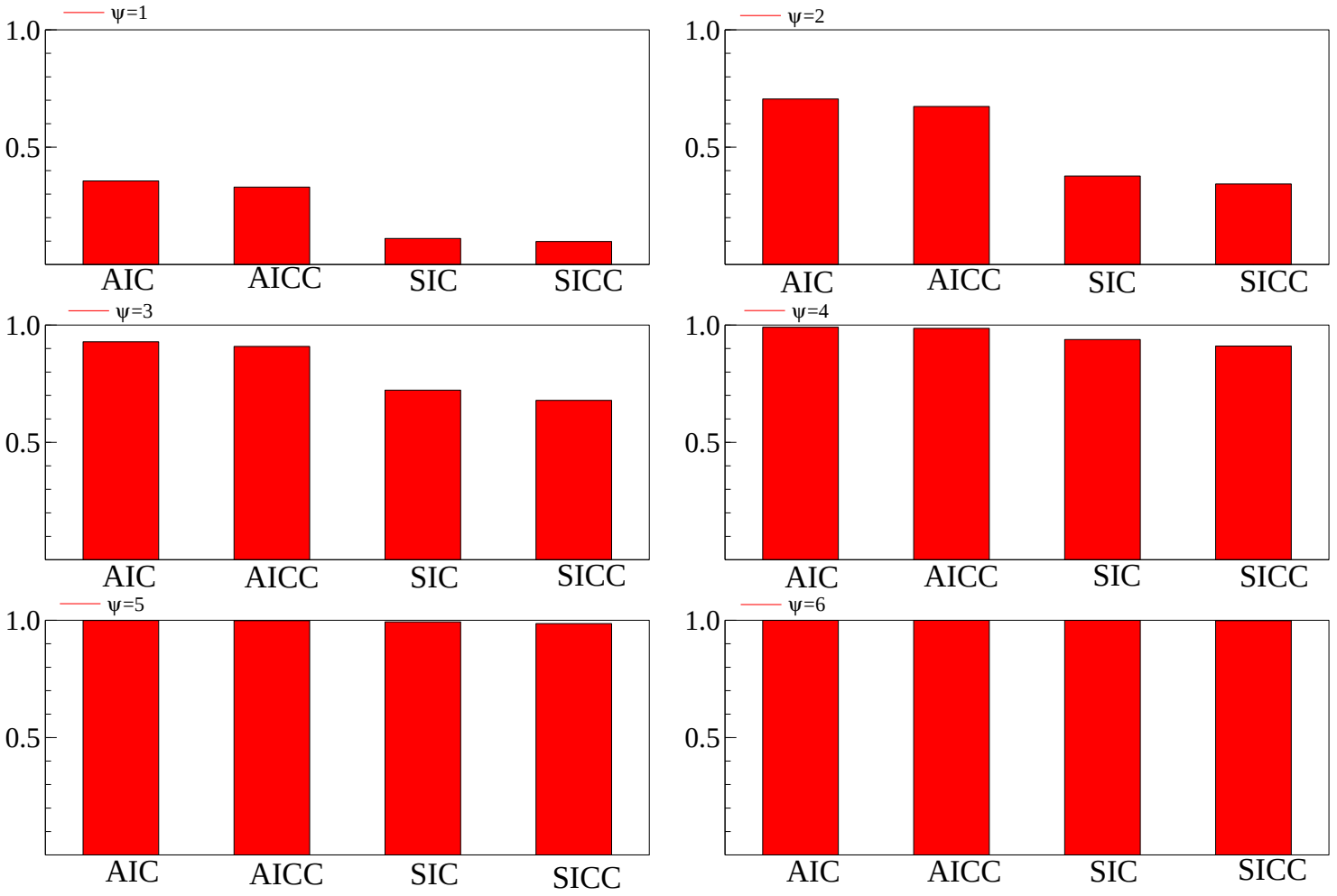
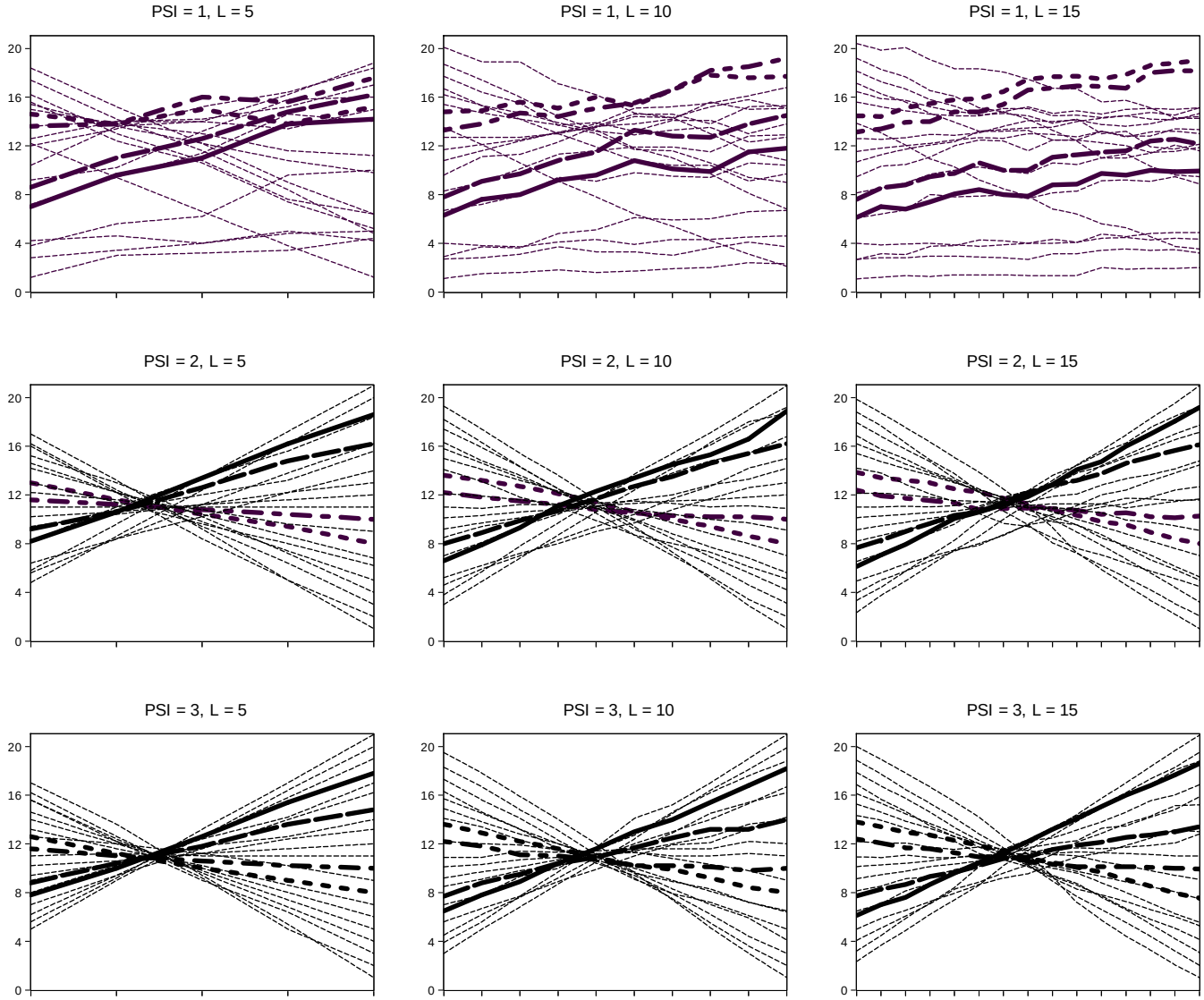


FIGURE 3
Rankings of MSAs as a Function of *K*, *ψ*, and *L* (*N* = 75)



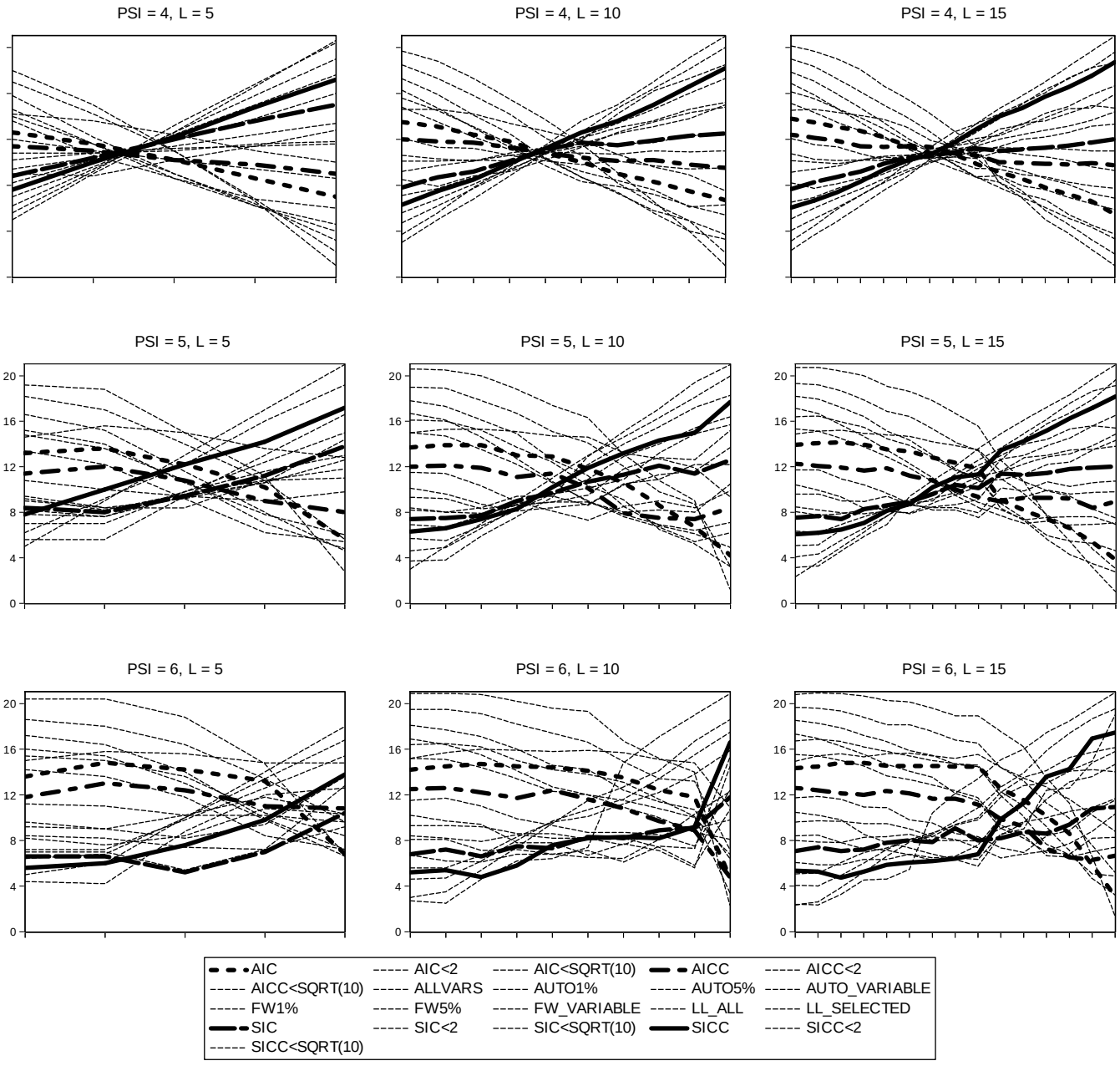
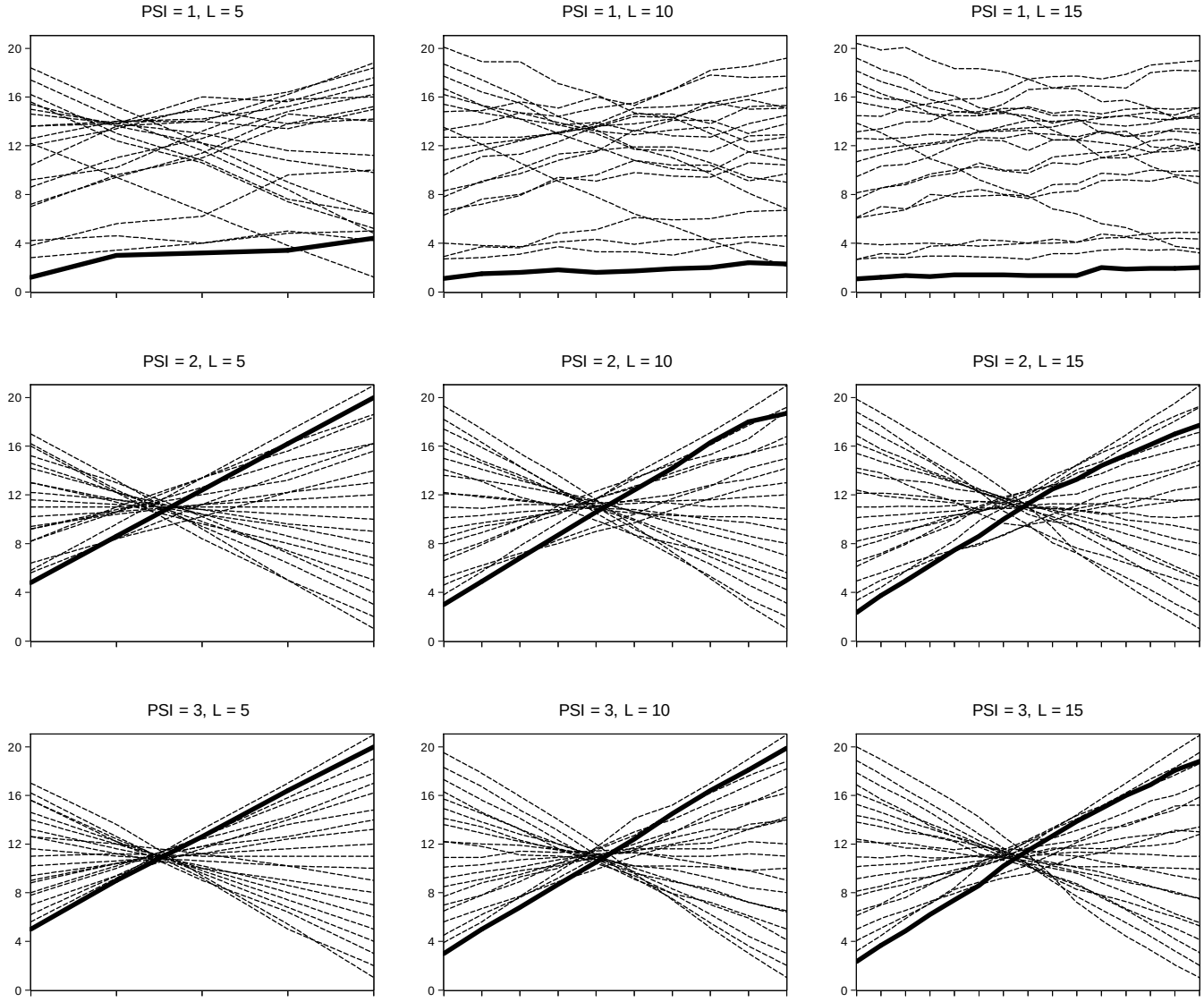

PSI = 4, L = 5
PSI = 4, L = 10
PSI = 4, L = 15
PSI = 5, L = 5
PSI = 5, L = 10
PSI = 5, L = 15
PSI = 6, L = 5
PSI = 6, L = 10
PSI = 6, L = 15
0
4
8
12
16
20
AIC
AIC<2
AIC<SQRT(10)
AICC
AICC<2
AICC<SQRT(10)
ALLVARS
AUTO1%
AUTO5%
AUTO_VARIABLE
FW1%
FW5%
FW_VARIABLE
LL_ALL
LL_SELECTED
SIC
SIC<2
SIC<SQRT(10)
SICC
SICC<2
SICC<SQRT(10)

FIGURE 4
Rankings of MSAs as a Function of *K*, ψ, and *L* (*N* = 75)



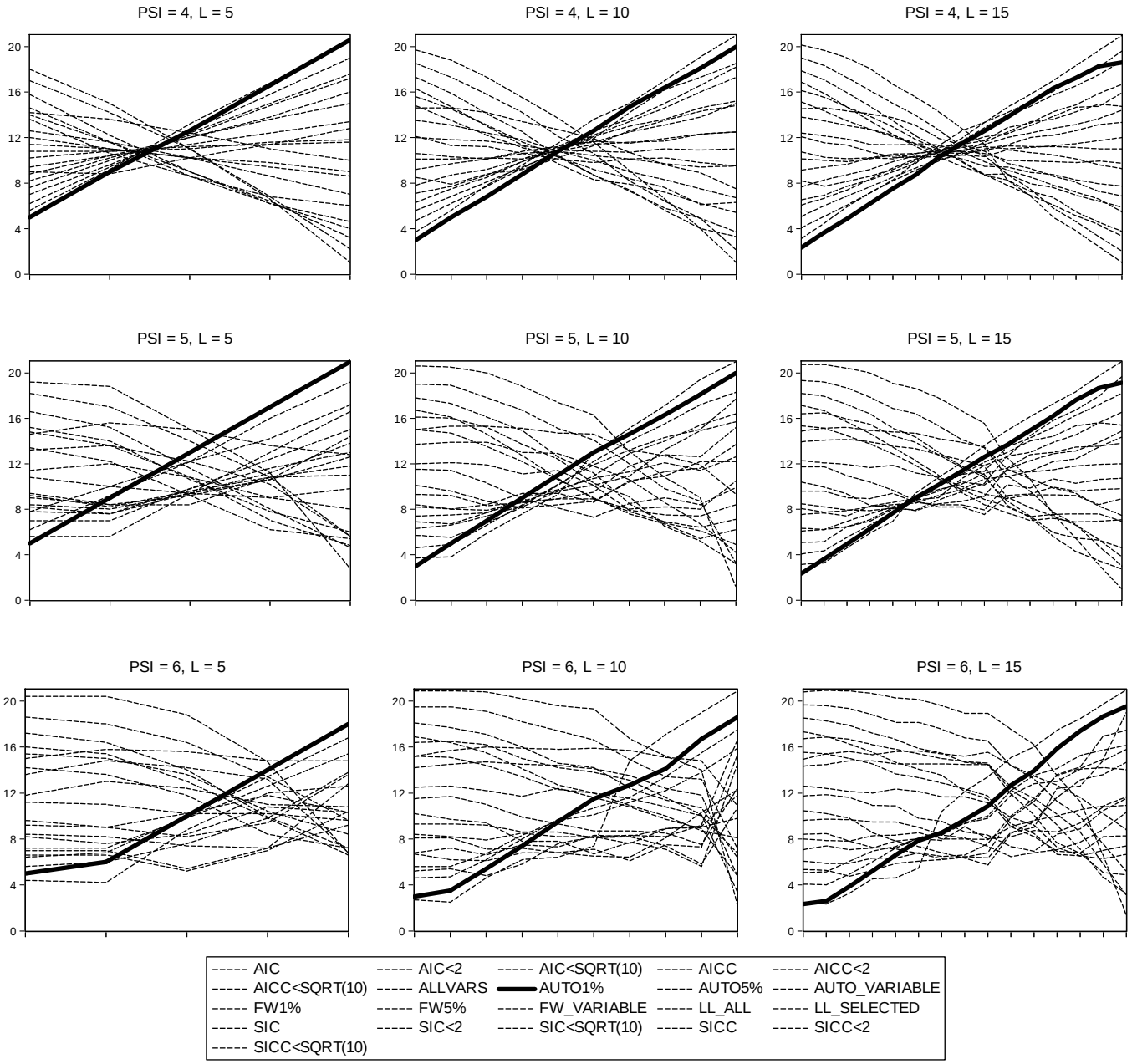
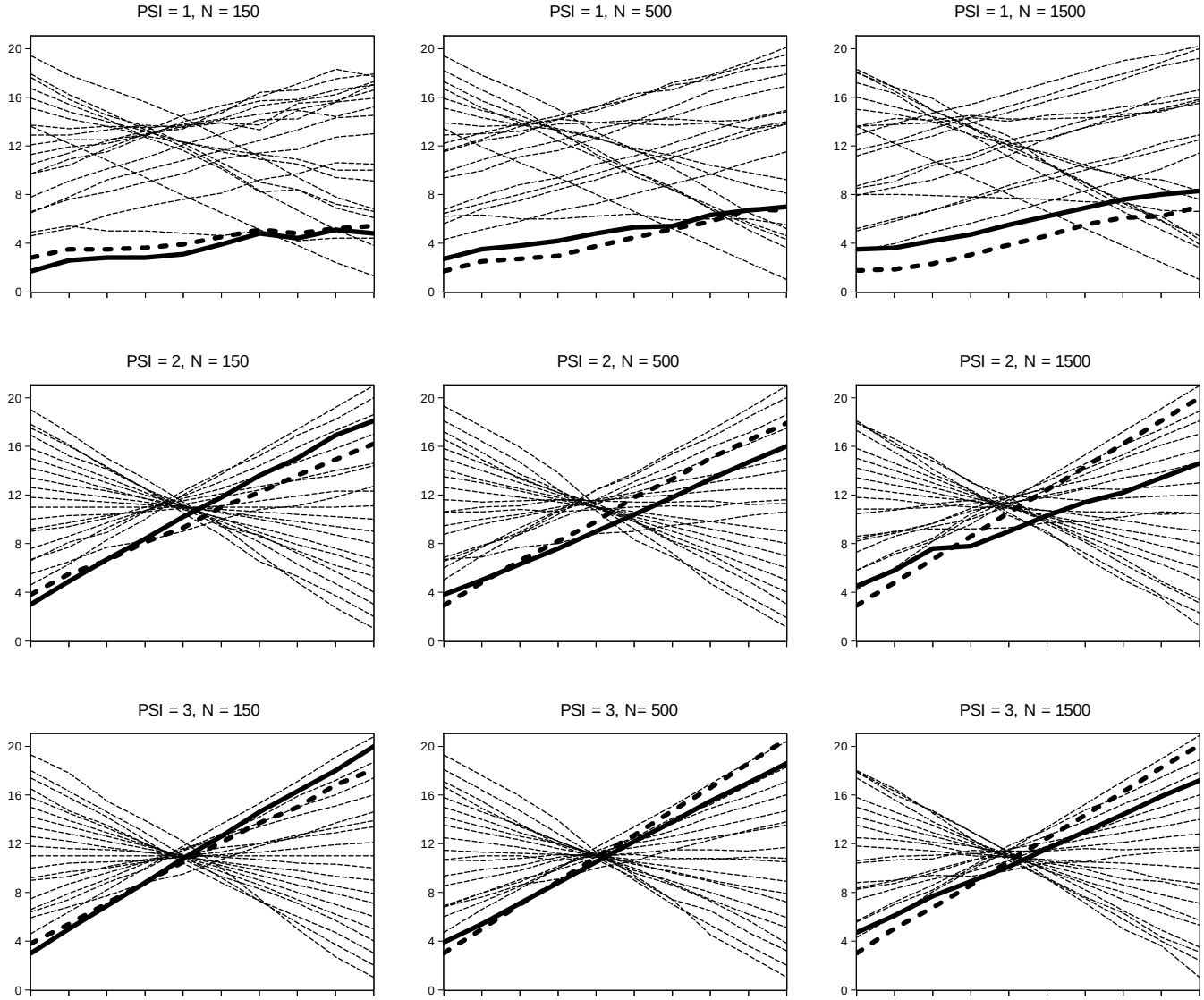

PSI = 4, L = 5
PSI = 4, L = 10
PSI = 4, L = 15
PSI = 5, L = 5
PSI = 5, L = 10
PSI = 5, L = 15
PSI = 6, L = 5
PSI = 6, L = 10
PSI = 6, L = 15
0
4
8
12
16
20
AIC
AIC<2
AIC<SQRT(10)
AICC
AICC<2
AICC<SQRT(10)
ALLVARS
AUTO1%
AUTO5%
AUTO_VARIABLE
FW1%
FW5%
FW_VARIABLE
LL_ALL
LL_SELECTED
SIC
SIC<2
SIC<SQRT(10)
SICC
SICC<2
SICC<SQRT(10)

FIGURE 5
Rankings of MSAs as a Function of *K*, *ψ*, and *N* (*L* = 10)



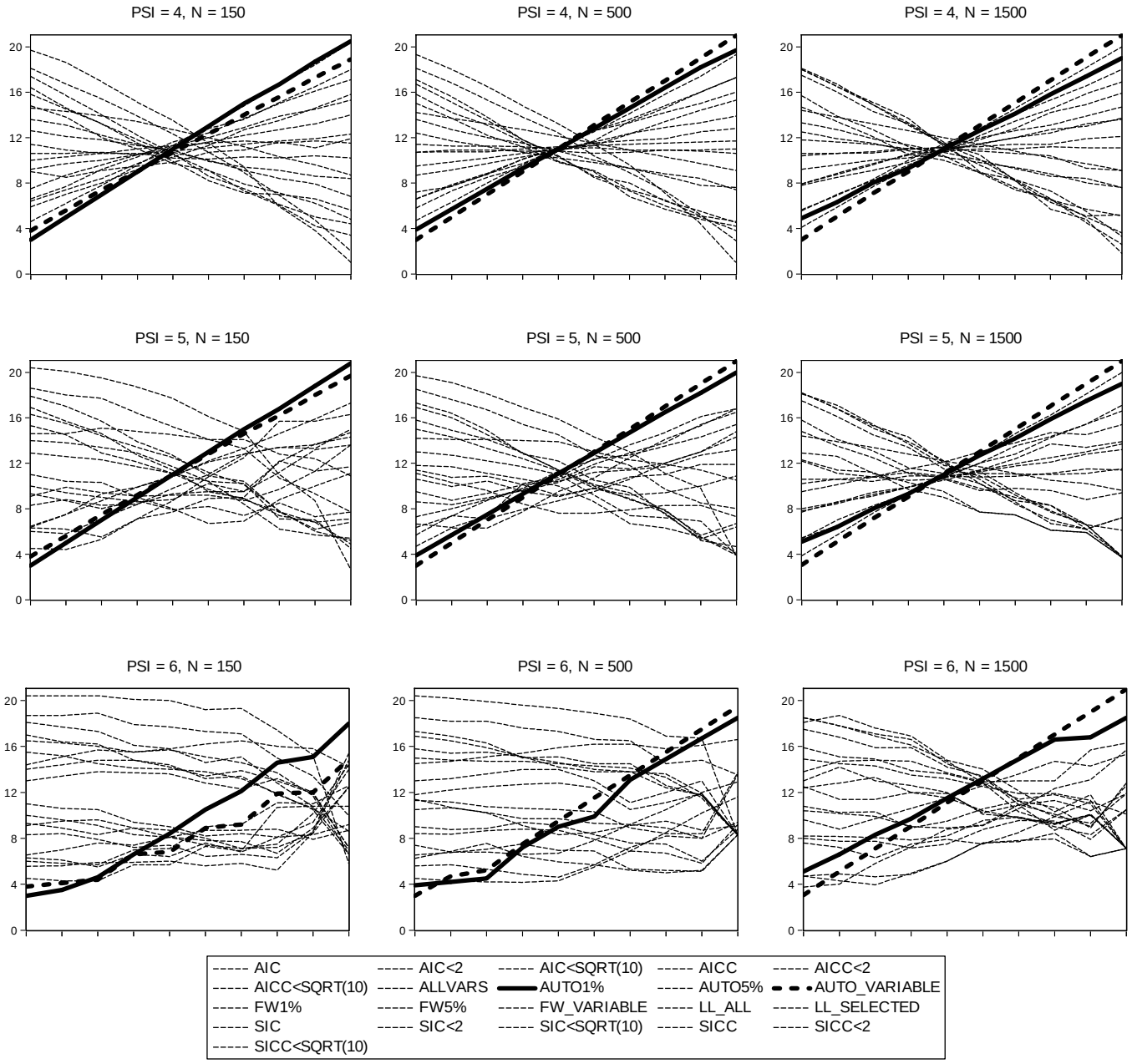

PSI = 4, N = 150
PSI = 4, N = 500
PSI = 4, N = 1500
PSI = 5, N = 150
PSI = 5, N = 500
PSI = 5, N = 1500
PSI = 6, N = 150
PSI = 6, N = 500
PSI = 6, N = 1500
0
4
8
12
16
20
AIC
AIC<2
AIC<SQRT(10)
AICC
AICC<2
AICC<SQRT(10)
ALLVARS
AUTO1%
AUTO5%
AUTO_VARIABLE
FW1%
FW5%
FW_VARIABLE
LL_ALL
LL_SELECTED
SIC
SIC<2
SIC<SQRT(10)
SICC
SICC<2
SICC<SQRT(10)

FIGURE 6
Log of the mean UMSE for each MSA across 58 experiments where K/L≤0.5 and ψ≤2 with ±2σ bands

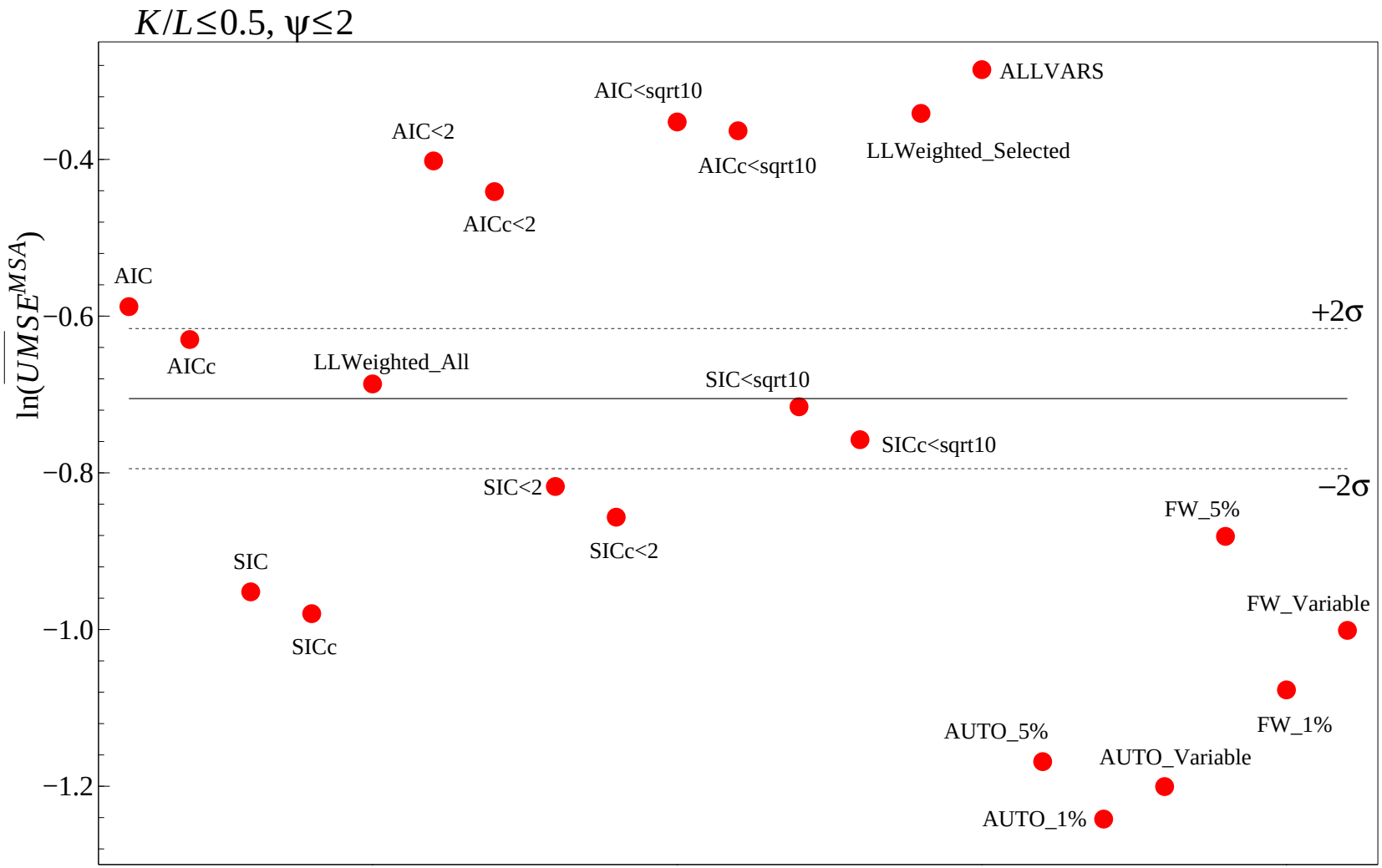


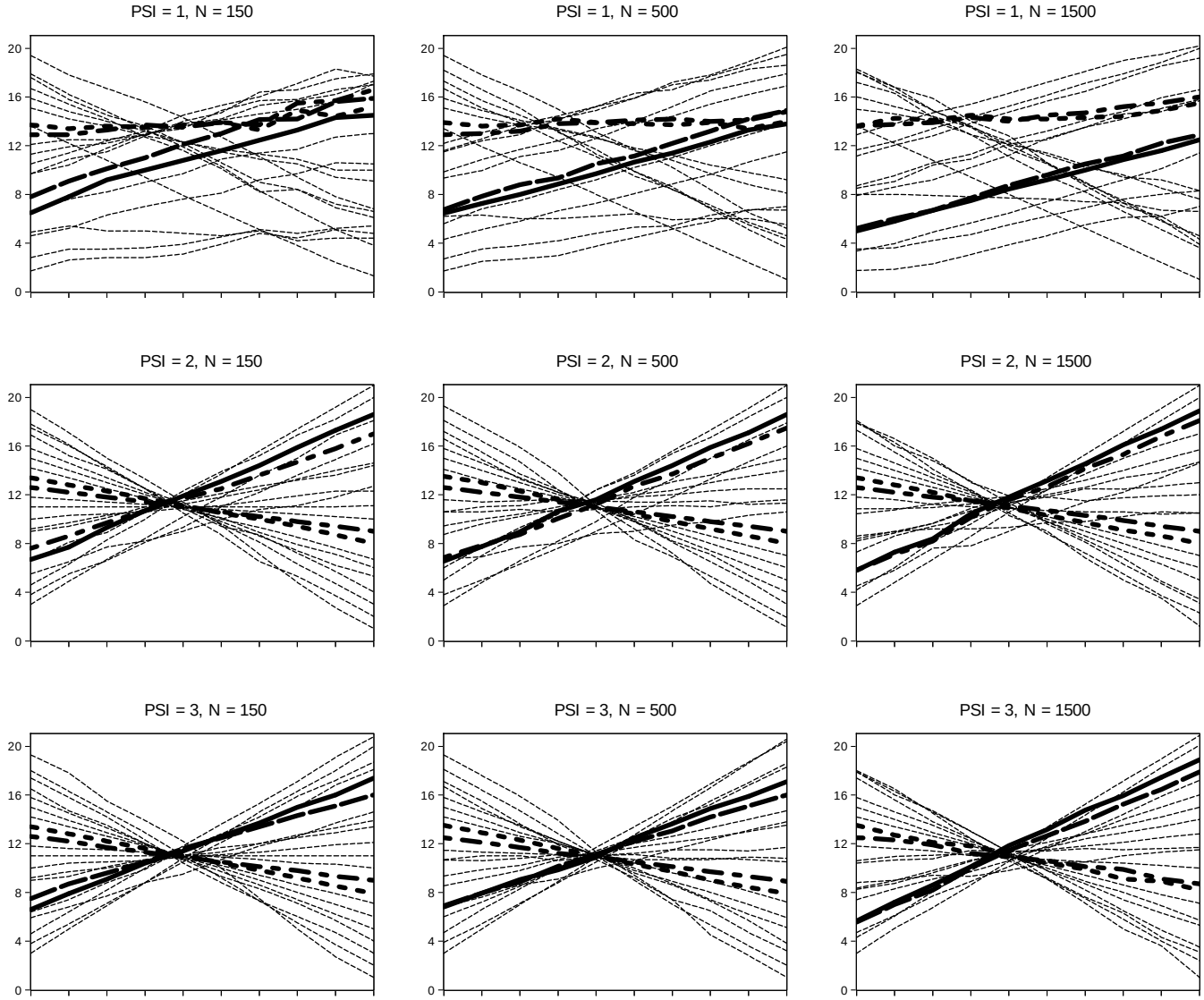
TABLE A1
Retention Probabilities as a Function of ψ and α (for *N=75)*

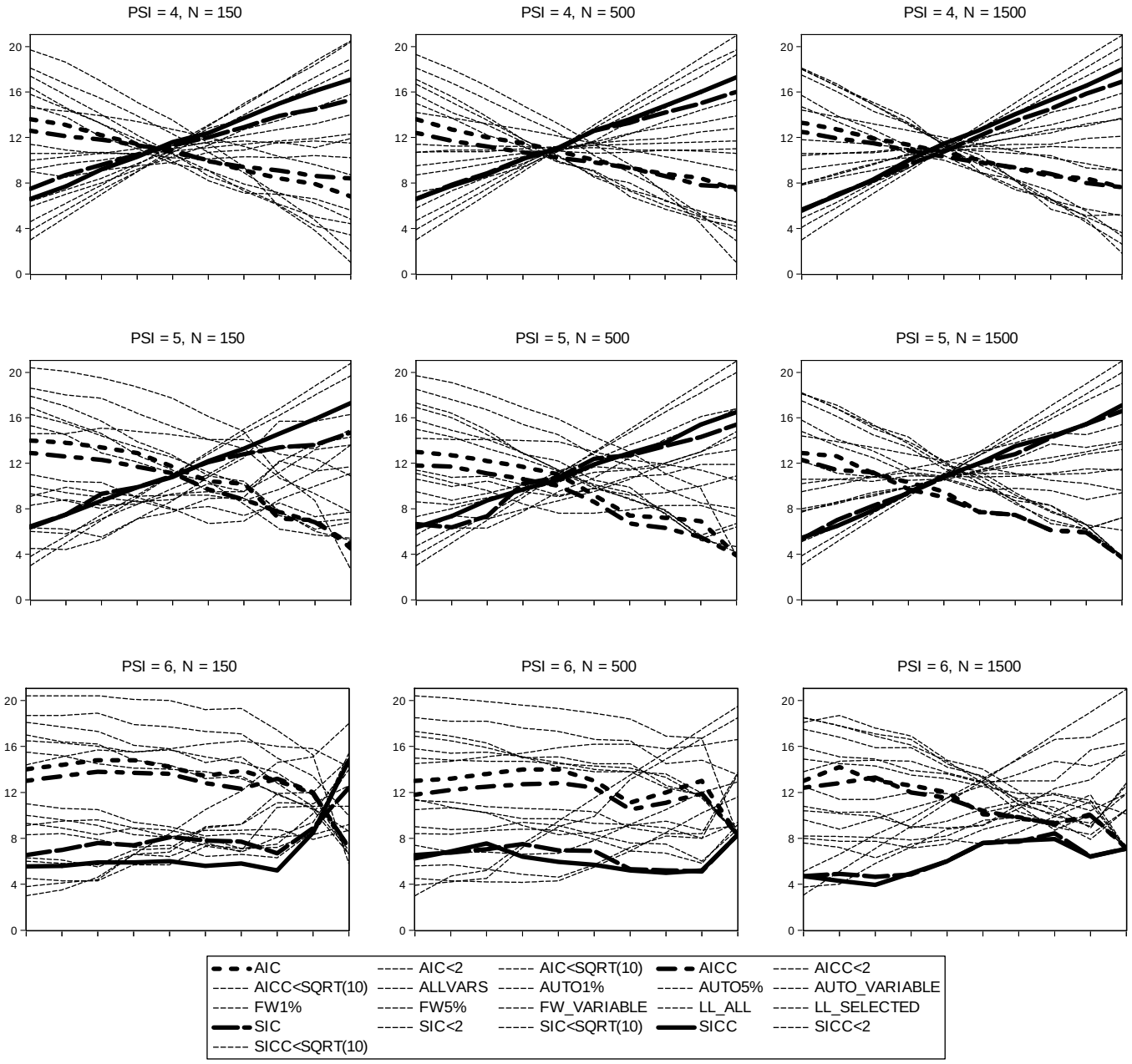
ψ_k	$P(t_k \geq c_\alpha \mid E[t_k] = \psi)$			
	$\alpha = 50\%$	**$\alpha = 20\%$**	***$\alpha = 5\%$***	***$\alpha = 1\%$***
1	62.6%	38.5%	16.1%	5.0%
2	90.7%	76.0%	50.3%	26.0%
3	99.0%	95.6%	84.3%	63.9%
4	100%	99.7%	97.8%	91.3%
5	100%	100%	99.9%	99.1%
6	100%	100%	100%	100%

TABLE A2
Total Number of Experiments by ψ and N

	N=75	***N=150***	***N=500***	***N=1500***	***TOTAL***
ψ=1	L=5,10,15 (30 experiments)	L=10 (10 experiments)	L=10 (10 experiments)	L=10 (10 experiments)	60
ψ=2	L=5,10,15 (30 experiments)	L=10 (10 experiments)	L=10 (10 experiments)	L=10 (10 experiments)	60
ψ=3	L=5,10,15 (30 experiments)	L=10 (10 experiments)	L=10 (10 experiments)	L=10 (10 experiments)	60
ψ=4	L=5,10,15 (30 experiments)	L=10 (10 experiments)	L=10 (10 experiments)	L=10 (10 experiments)	60
ψ=5	L=5,10,15 (30 experiments)	L=10 (10 experiments)	L=10 (10 experiments)	L=10 (10 experiments)	60
ψ=6	L=5,10,15 (30 experiments)	L=10 (10 experiments)	L=10 (10 experiments)	L=10 (10 experiments)	60
TOTAL	180	60	60	60	360

FIGURE A1
Rankings of MSAs as a Function of *K*, *ψ*, and *N* (*L* = 10)




PSI = 4, N = 150
PSI = 4, N = 500
PSI = 4, N = 1500
PSI = 5, N = 150
PSI = 5, N = 500
PSI = 5, N = 1500
PSI = 6, N = 150
PSI = 6, N = 500
PSI = 6, N = 1500
0
4
8
12
16
20
AIC
AIC<2
AIC<SQRT(10)
AICC
AICC<2
AICC<SQRT(10)
ALLVARS
AUTO1%
AUTO5%
AUTO_VARIABLE
FW1%
FW5%
FW_VARIABLE
LL_ALL
LL_SELECTED
SIC
SIC<2
SIC<SQRT(10)
SICC
SICC<2
SICC<SQRT(10)