Frequentist Model Averaging for the Ordered Probit and Nested Logit Models

Longmei Chen[1], Alan T.K. Wan[1,2], Geoffrey K.F. Tso[1,2], Xinyu Zhang[3]

[1] *College of Business, City University of Hong Kong, Kowloon, Hong Kong*

[2] *Centre for Social Media, Marketing and Business Intelligence, City University of Hong Kong, Kowloon, Hong Kong*

[3] *Centre for Forecasting Science, Chinese Academy of Sciences, Beijing, China*

Abstract

Model selection, which affects all stages of inference, is of special relevance to empirical sociological research that makes use of statistics. Typically, a sociologist would select a single model from the pool of all possible models, proceed as if the selected model were known at the outset, and neglect the uncertainty in the model selection step. Recent statistical literature has advocated model averaging as an alternative to model selection. Model averaging assigns weights to different models. These weights are then used to produce average estimates of the unknown parameters. There is ample evidence to indicate that model averaging frequently produces more efficient estimators than model selection. Model averaging also guards against the selection of very poor models. The purpose of this paper is to introduce model averaging to the sociological research community through the ordered probit and nested logit models that are widely used as analytical platforms in empirical sociological research. We examine a range of averaging and selection schemes for these models. Our results show that although neither averaging nor selection is a uniformly better strategy than the other, selection results in the poorest estimates far more frequently than averaging does, and more often than not, averaging yields superior estimates to selection. Among the averaging methods considered, the one based on a smoothed version of the Bayesian Information criterion frequently produces the most accurate estimates. The proposed methodology is applied to two real data sets related to sociology.

Keywords: hit rate, model averaging, model selection, Monte Carlo, nested logit, ordered probit, screening

1. Introduction

Many sociological studies use logit/probit models to analyse data in which the outcome variable has two or more discrete categories. There are many types of logit/probit models available for the analysis of discrete responses. Multinomial (Ordered) logit/probit are the most common models for nominal (ordinal) discrete outcomes. When there are only two responses, the multinomial and ordered logit/probit models simplify to the traditional binary logit/probit models. There is also the closely related technique of nested logit, which is widely used when individuals make their decisions in a sequential fashion. Excellent discussions of the state of the art of these models tailored for sociologists are given by Fullerton (2009), Gayle and Lambert (2009) and Long (2012). Methodological matters related to logit/probit have been and continue to be an important topic of research among quantitative sociologists, as evidenced by the number of papers in leading journals in sociological methodology on approaches to estimating and analysing logit/probit models; see, for example, Allison (1999a), Cheng and Long (2007), Williams (2009), Karlson, Holm and Breen (2012), and Breen, Karlson and Holm (2013).

A normal practice when applying logit/probit models, in much the same way when applying any other statistical models, is that one would try many models, each containing a different combination of regressors, and eventually select the "winner" of all models considered and report results based on this single winning model. Model selection methods typically entail the cumbersome step of pre-testing to decide which regressors to retain and which to drop. It is also common to select a model by information or out-of-sample prediction criteria. These criteria are strongly advocated by Kuha (2004), Stine (2004) and Weakliem (2004) in a special 2004 issue of *Sociological Methods and Research* devoted to model selection. The search for the winning model recognises the existence of multiple plausible models, implying a level of uncertainty associated with the choice of model. However, practitioners typically neglect this uncertainty when making inference contingent on the chosen best model, resulting in underestimated variances and over-optimistic inference (Giles and Giles, 1993; Danilov and Magnus, 2004; Leeb and Pötscher, 2005). It is also well-known that many model selection techniques can be highly influenced by slight variations in data (Yuan and Yang, 2005).

One way to circumvent model uncertainty is to replace the practice of discontinuously switching between models by smoothly interpolating the different models. The latter strategy is known as model averaging or combining. A model averaged estimate of a parameter is a weighted mean of a set of single model estimates for the

parameter. The weights reflect the degrees to which different models are trusted. The model averaging estimator has a distribution that is unconditional on the model selected, and provided that one works with this distribution, inference after averaging will not suffer from the same distortions associated with selection. Bayesian model averaging (BMA) has been widely applied in many disciplines. Examples of its applications in sociology and political science include Raftery (1995), Montgomery and Nyhan (2010), and Deller, Amiel and Deller (2011). While BMA provides a formal approach for incorporating prior knowledge, any poor handling of prior distributions can lead to undesirable behaviour of the model average estimator. In recent years, frequentist model averaging (FMA) has also garnered interests. FMA precludes the need to specify any prior distribution, although how to determine an optimal weight choice by a data-driven method is a challenge for the frequentist formulation.

Compared to the vast BMA literature, the literature on FMA is more recent, nonetheless a great deal of work has been done on developing model weighting schemes for FMA estimators and the investigation of their properties. The most common approach is to construct weights based on information criterion values obtained from the different models. This is the approach taken by Buckland, Burnham and Augustin (1997) and Hjort and Claeskens (2006). Hansen (2007) proposed a least squares FMA estimator with weights selected by minimising the Mallows' criterion. Liang, Zou, Wan and Zhang (2011) developed a model weighting mechanism based on minimisation of the trace of an unbiased estimator of the FMA estimator's mean square error. In heteroskedastic error settings, Hansen and Racine (2012)'s jacknnife model averaging estimator has been shown to achieve the smallest possible expected squared errors over a broad class of estimators. This estimator selects weights by minimising a delete-one cross validation criterion, and is the first FMA estimator that explicitly allows for heteroskedasticity. Recently, Zhang, Wan and Zou (2013) extended the jackknife averaging technique to models with dependent data. Some studies have shown that simple equally weighted average estimator can sometimes perform well (Wan, Zhang and Wang, 2014). Others have shown that screening out the very poor models prior to combining can often yield superior estimates (Yuan and Yang, 2005). The state of art in this rapidly expanding field is summarised in Claeskens and Hjort (2008), Wang, Zhang and Zou (2009), Moral-Benito (2015) and Liu (2015).

FMA has been successfully applied in many other disciplines including biomedical sciences (Claeskens, Croux and Van Kerckhoven, 2006), climatology (Duan and Mei, 2014), ecology (Johnson and Omland, 2004), health economics (Jackson, Thompson

and Sharples, 2009), growth economics (Amini and Parmeter, 2012), and tourism study (Wan and Zhang, 2009). It has not come into usage within the discipline of sociology although arguments in favour of combining models over selecting a single model in sociological research have been put forward by Burnham and Anderson (2004). As discussed before, logit/probit models are frequently used for data analysis in sociological studies. Studies that have considered FMA for logit models include Claeskens, Croux and Van Kerckhoven (2006) and Wan, Zhang and Wang (2014). Focusing on the binary logit model, Claeskens, Croux and Van Kerckhoven (2006) constructed model weights based on a focused information criterion (FIC). The FIC, proposed by Hjort and Claeskens (2003), is a model selection criterion aimed at minimising the mean squared error of the estimator of the focus parameter. Wan, Zhang and Wang (2014) used a weight choice method based on the minimisation of a plug-in estimator of the asymptotic squared error risk of the FMA estimator for the multinomial and ordered logit models. To our knowledge, no studies have considered FMA for probit models or other more complex forms of logit models such as the nested logit, and the purpose of this paper is to take steps in this direction.

One problem with the multinomial logit model is that it imposes the assumption of independence of irrelevant alternatives (IIA), whereby the odds of one outcome versus another is assumed to be independent of all other alternatives. The IIA assumption is a direct consequence of the assumption that the errors in the log-odds equations in the multinomial logit model are independently distributed. Although the IIA assumption makes the likelihood function easy to compute, it is often unrealistic; for example, it assumes that the odds of voting for Bernie Sanders versus Donald Trump will not change if Michael Bloomberg is added or removed from the ballot. Indeed, the realism of the IIA assumption in sociological research has been questioned in a number of studies; e.g., Cheng and Long (2007). One way to avoid the problems associated with the IIA assumption is to use the nested logit model that partitions the choice set into nests. This model assumes that IIA is satisfied for choices at the same level, but not necessarily for choices at different levels. Within the discipline of sociology, the much-cited article by Plotnick (1992) on the influence of personality on teenage premarital pregnancy and its resolution is based on a nested logit model. Quinn and Rubb (2005) used the nested logit model to examine the effects of education-occupation matching on migration.

Another model that has found widespread applications in sociological research is the ordered probit model. For ordinal data that contain many of the two extreme outcome values, the ordered probit model is frequently preferred over its logit coun-

terpart because the probit is tied to the normal distribution that has thicker tails than the logistic distribution. Ordered probit was used in recent sociological studies by Lucas (2001) on the impacts of social background on educational attainment, and Grant, Morales and Sallaz (2009) on nurses' perception of the spiritual nature of their profession.

In this article, we extend the promising statistical method of FMA to the ordered probit and nested logit models, which are widely used as analytical platforms in empirical sociological research. We consider a range of FMA weight choice schemes including equal weighting, weights based on information criterion scores, weights based on jackknife or leave-one-out cross validation, and weights based on various minimisation schemes related to the mean squared error of the FMA estimator. In particular, to the best of our knowledge, our work is the first study that examines jackknife or leave-one-out cross validation model averaging outside the framework of the linear model. We also investigate the merit of a screening step that eliminates all but the very best subset of candidate models based on an information criterion (Yuan and Yang, 2005). We seek answers to the questions of whether FMA yields any improvement in estimator's efficiency compared to model selection in the contexts of the two model frameworks being examined, and if so, which FMA method is the most advantageous and whether pre-averaging model screening can lead to any gain in efficiency.

Our presentation proceeds as follows. In the next section, we describe the ordered probit and nested logit models. Section 3 discusses the various model FMA strategies. In Section 4, we conduct a Monte Carlo study to compare the performance of these FMA methods with several traditional model selection methods. Section 5 contains applications to two data sets in sociology. We offer our conclusions in Section 6.

2. The Ordered Probit and Nested Logit Models

2.1. Ordered Probit model

The ordered probit model is an appropriate analytical framework when the response categories have a natural ordering. Suppose there are J ordered alternatives indexed by the subscript j, and n independent observations indexed by the subscript i. Let Y_i be the choice made by the i^{th} individual. If individual i selects alternative j, then $Y_i = j$. We divide the independent variables into mandatory and optional variables according to the analytical framework of Hjort and Claeskens (2003). The mandatory variables are those that must be included in the model on theoretical or

other grounds, while the optional regressors can be excluded from any model. This framework is for convenience only and it poses no restriction to the model set-up because the mandatory variable vector can potentially be a null matrix when no variable is considered mandatory for inclusion. Let $X_i(p \times 1)$ be the mandatory regressors and $Z_i(q \times 1)$ be the optional regressors. The probability that individual i chooses a response category lower than or equal to j can be written as:

$$\begin{cases} P(Y_i \leq j|X_i, Z_i) = \int_{-\infty}^{\alpha_j + X_i'\beta + Z_i'\gamma} \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{t^2}{2}\right\} dt \quad \text{for } j = 1, ..., J-1, \\ P(Y_i \leq J) = 1, \end{cases} \tag{1}$$

where β and γ are slope coefficients of the regressor variables common for all categories, and α_j is an intercept coefficient that differs across the categories. The common method of estimating the unknown parameters is maximum likelihood (ML) in conjunction with an iterative procedure such as the Newton-Raphson algorithm. With q optional regressors in Z_i, there are 2^q sub-models to choose between. Let $\hat{\alpha}_1^{(s)}, ..., \hat{\alpha}_{J-1}^{(s)}$, $\hat{\beta}^{(s)}$, and $\hat{\gamma}^{(s)}$ be the ML estimators of the unknown parameters in the s^{th} sub-model; some elements of $\hat{\gamma}^{(s)}$ will be zero by default if the corresponding variables in Z_i are not included in the s^{th} sub-model.

Suppose there is a new observation 0 with an unknown response Y_0 and regressor variables (X_0, Z_0), the probability of selecting the j^{th} category based on the s^{th} sub-model can be written as

$$\begin{aligned} &\hat{p}_{0j}^{(s)} = \hat{P}(Y_0 \leq j|X_0, Z_0) - \hat{P}(Y_0 \leq j-1|X_0, Z_0) \\ &= \int_{-\infty}^{\hat{\alpha}_j^{(s)} + X_0'\hat{\beta}^{(s)} + Z_0'\hat{\gamma}^{(s)}} \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{t^2}{2}\right\} dt - \int_{-\infty}^{\hat{\alpha}_{j-1}^{(s)} + X_0'\hat{\beta}^{(s)} + Z_0'\hat{\gamma}^{(s)}} \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{t^2}{2}\right\} dt \end{aligned} \tag{2}$$

if $j < J$, or

$\hat{p}_{0j}^{(s)} = 1 - \hat{P}(Y_0 \leq J-1|X_0, Z_0) = 1 - \int_{-\infty}^{\hat{\alpha}_{J-1}^{(s)} + X_0'\hat{\beta}^{(s)} + Z_0'\hat{\gamma}^{(s)}} \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{t^2}{2}\right\} dt$

if $j = J$.

2.2. Nested Logit model

The nested logit model assumes that the J alternatives can be partitioned into K clusters, that is $(1, 2, ..., J) = B_1 \cup B_2 \cup ... \cup B_K$, such that each cluster consists of similar alternatives and each alternative belongs to exactly one nest. The probability that individual i chooses alternative j can be written as:

$$p_{ij} = P(Y_i \in B_k) \times P(Y_i = j|B_k), \tag{3}$$

where $P(Y_i \in B_k)$ is the marginal probability that individual i makes a choice within cluster B_k, and $P(Y_i = j|B_k)$ is the conditional probability that the j^{th} choice is selected within cluster B_k.

More explicitly, the conditional probability $P(Y_i = j|B_k)$ can be written as:

$$P(Y_i = j|B_k) = \frac{exp\left(\frac{\alpha_{j|B_k} + X'_{i,j|B_k}\beta + Z'_{i,j|B_k}\gamma}{\tau_k}\right)}{\sum_{j \in B_k} exp\left(\frac{\alpha_{j|B_k} + X'_{i,j|B_k}\beta + Z'_{i,j|B_k}\gamma}{\tau_k}\right)}, \tag{4}$$

where $X_{i,j|B_k}$ and $Z_{i,j|B_k}$ are respectively the mandatory and optional variables that determine the choice of alternative j within cluster B_k, β and γ are the slope coefficients of the regressor variables that are common for all the alternatives, $\alpha_{j|B_k}$ is an intercept coefficient that varies across the alternatives, and τ_k is an index of dissimilarity for the alternatives in B_k - a small τ_k indicates less dissimilarity, and vice versa. For ease of parameter identification and without loss of generality, we set the intercept coefficient corresponding to the last category in each cluster to zero.

The marginal probability $P(Y_i \in B_k)$ may be expressed as

$$P(Y_i \in B_k) = \frac{exp\{\tau_k IV_{i,B_k}\}}{\sum_{k=1}^{K} exp\{\tau_k IV_{i,B_k}\}}, \tag{5}$$

where

$$IV_{i,B_k} = ln\left[\sum_{j \in B_k} exp\left(\frac{\alpha_{j|B_k} + X'_{i,j|B_k}\beta + Z'_{i,j|B_k}\gamma}{\tau_k}\right)\right].$$

The quantity $\tau_k \times IV_{i,B_k}$ is the expected utility that individual i derives from choosing among the alternatives in cluster B_k.

Again, there are 2^q sub-models to choose between when $Z_{i,j|B_k}$ contains q variables. The probability of individual 0 selecting alternative j in cluster B_k based on the s^{th} sub-model is:

$$\hat{p}_{0j}^{(s)} = \frac{exp\{\hat{\tau}_k^{(s)} IV_{0,B_k}^{(s)}\}}{\sum_{k=1}^{K} exp\{\hat{\tau}_k^{(s)} IV_{0,B_k}^{(s)}\}} \times \frac{exp\left(\frac{\hat{\alpha}_{j|B_k}^{(s)} + X'_{0,j|B_k}\hat{\beta}^{(s)} + Z'_{0,j|B_k}\hat{\gamma}^{(s)}}{\hat{\tau}_k^{(s)}}\right)}{\sum_{j\in B_k} exp\left(\frac{\hat{\alpha}_{j|B_k}^{(s)} + X'_{0,j|B_k}\hat{\beta}^{(s)} + Z'_{0,j|B_k}\hat{\gamma}^{(s)}}{\hat{\tau}_k^{(s)}}\right)}, \tag{6}$$

where

$$IV_{0,B_k}^{(s)} = ln\left[\sum_{j\in B_k} exp\left(\frac{\hat{\alpha}_{j|B_k}^{(s)} + X'_{0,j|B_k}\hat{\beta}^{(s)} + Z'_{0,j|B_k}\hat{\gamma}^{(s)}}{\hat{\tau}_k^{(s)}}\right)\right],$$

and $\hat{\tau}_k^{(s)}$, $\hat{\alpha}_{j|B_k}^{(s)}$, $\hat{\beta}^{(s)}$ and $\hat{\gamma}^{(s)}$ are the ML estimators of their respective unknown parameters in the s^{th} sub-model. Again, some elements in $\hat{\gamma}^{(s)}$ will be zero by default if the corresponding variables in $Z_{0,j|B_k}$ are excluded from the s^{th} sub-model.

3. Model Averaging

Typically, an investigator would identify one single best "winning" model from the 2^q candidate models based on an information criterion such as the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), then proceed with this model to calculate $\hat{p}_{0j}$ for the new observation 0. As discussed in Section 1, model selection ignores the randomness embodied in the selection process, and reports the final results as if the selected model were not a random choice. On the other hand, model averaging combines forecasts obtained from the different models by the following weighted average:

$$\hat{p}_{0j}^{w} = \sum_{s=1}^{2^q} w_s \hat{p}_{0j}^{(s)}, \tag{7}$$

where $w_s (0 \leq w_s \leq 1)$ is the weight given to the s^{th} sub-model, and $\sum_{s=1}^{2^q} w_s = 1$. Thus, the model averaged predicted probability $\hat{p}_{0j}^{w}$ smoothes across the predicted probabilities from the 2^q candidate models.

3.1. Model averaging methods

A preponderance of the literature on model averaging emphasises the weight choice of the model average. Various methods of weight choice leading to model

average estimators with optimal properties have been proposed. Here, we consider a broad range of FMA methods which originated from the econometrics and statistics literatures. All methods have been shown to work well in other contexts. As several of these methods have been developed under the local misspecification framework (LMF) (Hjort and Claeskens, 2003; Claeskens and Hjort, 2003), we provide a summary encapsulating the essence of this framework in the Appendix.

The FMA weight choice schemes we considered are as follows:

- Smoothed-AIC (S-AIC) and Smoothed-BIC (S-BIC) weights, by which

$$[w_s = \frac{exp\{-xIC_s/2\}}{\sum_{s=1}^{2^q} exp\{-xIC_s/2\}}], \tag{8}$$

  where xIC$_s$ is the AIC or BIC score of the s^{th} model. The smoothed information criterion weighting scheme was proposed by Buckland et al. (1997), and subsequently used in a number of FMA studies. Buckland et al. (1997) justified this weighting scheme by noting that for the S-AIC, the ratio in (8) is the relative penalised likelihood factor, and for the S-BIC, it is Schwarz's (1978) approximation to the Bayes factor.

- Smoothed-FIC (S-FIC) weights. The FIC, developed by Hjort and Claeskens (2003) is a model selection criterion tailored to the parameter singled out for interest. Let μ be a vector of population unknown parameters to be estimated. Hjort and Claeskens (2003) showed that under the LMF, the FIC for minimising the MSE of the estimator of μ in the s^{th} sub-model is

$$FIC^s_{MSE} = \left(\hat{\omega}'(I_q - \varpi_s'\hat{\mathcal{K}}_s\varpi_s\hat{\mathcal{K}}^{-1})\hat{\delta}\right)^2 + 2\hat{\omega}'\varpi_s'\hat{\mathcal{K}}_s\varpi_s\hat{\omega}, \tag{9}$$

  where $\hat{\omega}$ and $\hat{\delta}$ are the ML estimators of ω and δ (defined in the Appendix) using the full model, $\hat{\mathcal{K}} \equiv \left(\mathcal{J}_{n,11} - \mathcal{J}_{n,10}\mathcal{J}^{-1}_{n,00}\mathcal{J}_{n,01}\right)^{-1}$ is a consistent estimator of $\mathcal{K}$ (defined in the Appendix), $\hat{\mathcal{K}}_s$ is obtained by replacing $\mathcal{K}$ with $\hat{\mathcal{K}}$ in $\mathcal{K}_s = (\varpi_s\mathcal{K}^{-1}\varpi_s')^{-1}$, and ϖ_s is a projection matrix that maps the vector δ to its subvector $\varpi_s\delta = \delta_s$ that contains the elements of δ in the s^{th} sub-model. As it is difficult to justify (8) when using FIC^s_{MSE} as xIC_s, Hjort and Claeskens (2003) suggested assigning the weight to the s^{th} model by

$$w_s = \frac{exp\left\{-FIC^s_{MSE}/2\varrho\hat{\omega}'\hat{\mathcal{K}}\hat{\omega}\right\}}{\sum_{s=1}^{2^q} exp\left\{-FIC^s_{MSE}/2\varrho\hat{\omega}'\hat{\mathcal{K}}\hat{\omega}\right\}}, \tag{10}$$

where ϱ is an algorithmic parameter that bridges from the uniform weighting (ϱ near 0) to "hard" FIC. Hjort and Claeskens (2003) demonstrated that (10) has an empirical Bayes justification. Following Zhang, Wan and Zhou (2012), we set $\varrho = 1$. The S-FIC method has been used in various FMA studies including Claeskens, Croux and Van Kerckhoven (2006), Zhang and Liang (2011) and Zhang, Wan and Zhou (2012).

- Weight based on minimising the trace of an unbiased estimator of the FMA estimator's mean square error (Liang, Zou, Wan and Zhang, 2011) (LZWZ). Again, this has been developed under the LMF described in the Appendix. Liang, Zou, Wan and Zhang (2011) derived an unbiased estimator of the trace of the MSE of the FMA estimator, and suggested choosing $w = (w_1, ..., w_{2^q})$ by minimising

$$\hat{R}(\hat{\mu}^w) = \left(\hat{\omega}'\hat{\mathcal{K}}^{1/2}\hat{\mathfrak{L}}(w)\hat{\mathcal{K}}^{-1/2}\hat{\delta}\right)^2 + 2\hat{\omega}'\hat{\mathcal{K}}^{1/2}\hat{H}(w)\hat{\mathcal{K}}^{1/2}\hat{\omega} \tag{11}$$

  subject to the constraints $0 \leq w_s \leq 1$ and $\sum_{s=1}^{2^q} = 1$. This can be readily performed using routines in STATA, GAUSS or Matlab. A description of the derivation of (11) based on the LMF is given in the Appendix.

- Weight based on minimizing a plug-in estimator of the asymptotic squared error risk of the FMA estimator (Wan, Zhang and Wang, 2014) (A-opt). A-opt method chooses the weight vector $w = (w_1, ..., w_{2^q})$ by minimising

$$\hat{R}_a(\hat{\mu}^w) = \left(\hat{\omega}'\hat{\mathcal{K}}^{1/2}\hat{\mathfrak{L}}(w)\hat{\mathcal{K}}^{-1/2}\hat{\delta}\right)^2 + \hat{\omega}'\hat{\mathcal{K}}^{1/2}\hat{H}^2(w)\hat{\mathcal{K}}^{1/2}\hat{\omega} \tag{12}$$

  subject to the constraints $0 \leq w_s \leq 1$ and $\sum_{s=1}^{2^q} = 1$. Again, this is a standard minimisation problem that can be readily solved using STATA or other software. A description of the derivation of (12) based on the LMF is given in the Appendix.

- Weight based on Jackknife or leave-one-out cross validation criterion (Hansen and Racine, 2012; Zhang, Wan and Zou, 2013) (JMA). It has been shown that in a linear model, the JMA estimator has squared errors that are asymptotically identical to those of the infeasible best possible model averaging estimator. Our present analysis is complicated by the fact that the relationship between the dependent and explanatory variable is non-linear. The known results therefore

do not directly apply. To the best of our knowledge, our work is the first analysis of the JMA method outside the linear model. To construct the JMA criterion, let $\hat{p}_{ij}^{(s)}$ be the forecast of p_{ij} based on the s^{th} model. The criterion is defined as:

$$CV(w) = \sum_{i=1}^{n} \sum_{j=1}^{J} \mathbb{I}(Y_i = j) \left(\sum_{s=1}^{2^q} w_s^{(-i)} \hat{p}_{ij}^{(s)} - 1 \right)^2, \tag{13}$$

where $\mathbb{I}(Y_i = j)$ is an indicator function that takes on the value of 1 if the i^{th} individual selects category j, and 0 otherwise, ${}^{(-i)}\hat{p}_{ij}^{(s)}$ is the estimator of p_{ij} based on the s^{th} sub-model with the i^{th} observation deleted from the sample, and $\sum_{s=1}^{2^q} w_s^{(-i)} \hat{p}_{ij}^{(s)}$ is the weighted average of ${}^{(-i)}\hat{p}_{ij}^{(s)}{}'s$ across the 2^q models. Clearly, the best forecast of p_{ij} is 1 when $Y_i = j$. Hence we define the forecast error associated with the model average as $\sum_{s=1}^{2^q} w_s^{(-i)} \hat{p}_{ij}^{(s)} - 1$. The overall accuracy of the model average is evaluated in terms of its squared forecast errors across all n observations in the sample. Our JMA weight selection strategy seeks a weight vector w that minimizes $CV(w)$. The steps of the JMA strategy are as follows:

Step 1: Calculate ${}^{(-i)}\hat{p}_{ij}^{(s)}$, $i = 1, ..., n$, $j = 1, ..., J$, $s = 1, ..., 2^q$ by "leaving out" the i^{th} observation from the sample. The calculations may be based on equation (2) in the case of the ordered probit model, and equation (6) in the case of the nested logit model.

Step 2: Seek $w = (w_1, ..., w_{2^q})$ that minimizes $CV(w)$ in (13), subject to the constraints $0 \leq w_s \leq 1$ and $\sum_{s=1}^{2^q} w_s = 1$. This can be handled readily by software such as STATA, GAUSS or MATLAB.

Step 3: Substitute w obtained from *Step 2* in (7) to obtain the model averaged predictions $\hat{p}_{0j}^{w}$ of p_{0j}, $j = 1, ..., J$. The forecast of Y_0 is $\hat{Y}_0 = c$, the category that corresponds to $\hat{p}_{0c}^{w} = max\{\hat{p}_{01}^{w}, ..., \hat{p}_{0J}^{w}\}$.

- Equal weighting (EW), by which $w_s = \frac{1}{2^q}$.

4. A Monte Carlo Study

In this section, by means of a Monte Carlo study, we evaluate the finite sample performance of the various FMA strategies discussed in Section 3, and compare them with several common model selection strategies including the AIC, BIC and FIC. We also examine if anything can be gained by implementing a model screening step to remove the very poor models prior to combining. The screening procedure we

consider is the "top m" procedure (Yuan and Yang, 2005) that removes all but the $m(< 2^q)$ models corresponding to the m smallest values of an information criterion. In our analysis, we set m to 5 and choose the BIC as the information criterion. Another method we could have used was backward elimination discussed in Claeskens, Croux and Van Kerckhoven (2006) and Zhang, Wan and Zhou (2012). However, this method suffers from the deficiency that it always maintains one single model of each size in the final set of models. This means that a model that is not considered the best among the models of the same size will always be excluded even if it outperforms the best model of another size.

We consider the following two experimental designs:

Design 1: The data are generated based on the ordered probit model in (1), with $J = 3$, $p = 1$, $q = 4$, $(\alpha_1, \alpha_2, \beta) = l(-0.15, 0.5, 0.2)$, X_i, Z_{i1}, Z_{i2} and Z_{i4} each distributed as i.i.d $N(0, 1)$, Z_{i3} distributed as i.i.d Bernoulli(0.4), and γ set to one of the following scenarios:

S1: $\gamma = l(0.3, 0.7, 0.15, -0.04)$
S2: $\gamma = l(0.3, 0.7, 0, 0)$
S3: $\gamma = l(0.3, 0, 0, 0)$

The parameter l, which takes on 0.5, 1 or 2, has the purpose of controlling the magnitude of the coefficients. The three scenarios represent different sparsity levels of non-zero coefficients. Under S1, the true model contains no zero coefficients, and all sub-models except the full model are under-fitted. In contrast, under scenario S3, the majority of the coefficients are zero and consequently most sub-models are over-fitted. Scenario S2 with two zero coefficients is an intermediate scenario of the other two. As $q = 4$, there are $2^4 = 16$ sub-models within the model average. The number of sub-models reduces to $m = 5$ if screening is implemented prior to averaging.

Design 2: Our second experiment is based on the nested logit model in (4) and (5). We let $\tau = l \times 0.25$, and set all other parameters to the same values as in the previous design.

The number of observations in the training and test samples are set to 300 and 100 respectively. Each part of our experiment is based on 500 replications. We evaluate the various strategies' performance in terms of mean squared error of forecasts (MSEF), mean absolute error of forecasts (MAEF) and hit rate (HitRate). These

measures are defined as follows:

$$MSEF = \frac{1}{50000} \sum_{r=1}^{500} \sum_{i=1}^{100} \sum_{j=1}^{J} (p_{ij,r} - \hat{p}_{ij,r})^2, \tag{14}$$

$$MAEF = \frac{1}{50000} \sum_{r=1}^{500} \sum_{i=1}^{100} \sum_{j=1}^{J} |p_{ij,r} - \hat{p}_{ij,r}|, \tag{15}$$

and

$$HitRate = \frac{1}{50000} \sum_{r=1}^{500} \sum_{i=1}^{100} \mathbb{I}(Y_{i,r} = \hat{Y}_{i,r}), \tag{16}$$

where $p_{ij,r}$ is the probability of the i^{th} observation selecting category j in the r^{th} replication, $\hat{p}_{ij,r}$ is its forecast based on a given strategy, $\hat{Y}_{i,r} = c$ is the category that corresponds to $\hat{p}_{ic,r} = max\{\hat{p}_{i1,r}, ..., \hat{p}_{iJ,r}\}$, and $Y_{i,r}$ is the actual category of Y_i in the r^{th} replication.

The results of the Monte Carlo study comparing the efficiency of various estimators are reported in Tables 1 - 3. To facilitate comparisons, the best, second best, third best, third worst, second worst, and worst estimators in each case are flagged by (1), (2), (3), (-3), (-2) and (-1) respectively, and if the performance of a given averaging strategy is improved after screening out the poor models beforehand, it is flagged with a "↑".

The major conclusions of the Monte Carlo study may be summarised as follows:

It is clear from the results that no one strategy uniformly dominates any of the others in terms of the performance yardsticks considered. Neither model selection nor model averaging is always a better strategy than the other. That being said, the screened version of the S-BIC averaging strategy is frequently the best strategy and one of the best three strategies across all cases considered with respect to all three performance yardsticks. Among the three model selection methods, the BIC method generally performs the best, and can sometimes provide more accurate estimates than several of the FMA methods. Remarkably, the screened versions of all FMA methods are rarely among the three worst strategies. On the other hand, although model selection can sometimes outperform averaging, it also delivers very poor estimates far more often than its model averaging counterparts. For example, despite

being one of the top-rated methods with respect to the frequency in producing the best estimates, the BIC selection method also yields one of the three worst estimates thrice with respect to MSEF and MAEF, and four times with respect to hit rate. On the other hand, no matter the performance yardsticks, the screened versions of the S-AIC, S-BIC, S-FIC, LZWZ and A-opt methods never deliver the three worst estimates across all cases considered. Interestingly, the screened version of the EW method, arguably the worst of all screened versions of FMA methods considered, yields the poorest estimates far less frequently than BIC selection does, the latter being the best of the three selection methods. The AIC and FIC selection methods rarely perform at or near the top, but are frequently rated among the bottom three of all methods. These findings again highlight one major advantage of model averaging, that is, providing a kind of assurance against selecting a very poor model.

A close scrutiny of the performance of the three model selection methods reveals that they usually perform better when $l = 1$ or 2 than when $l = 0.5$. This result is not surprising because a small l makes it difficult for a selection criterion to differentiate the correct model from other models, especially those that include many zero coefficients. Other things being equal, a larger l makes it easier to identify the true model, which in turns makes model selection a more viable strategy. Wan, Zhang and Wang (2014) observed a similar finding in their study on comparing model selection to FMA in the contexts of the multinomial and ordered logit models.

Generally speaking, model averaging with screening is preferable to averaging without screening. Our results show that the various FMA methods commonly experience a deterioration in performance and deliver worse estimates far more frequently when the models that fail the screening are included in the model average. In particular, the non-screened version of the EW method frequently produces the worst of all seventeen estimators considered. The poor performance of EW may be explained by the fact that it assigns the same weight to all sub-models, including those with very poor explanatory power. In most cases, model screening improves the performance of FMA estimators; in the case of the EW estimator, the improvement is especially noticeable. Having said that, the non-screened versions of the S-AIC, S-BIC and S-FIC estimators are rarely the worst; as well, the non-screened S-BIC estimator is second only to its screened counterpart in terms of producing the highest HitRate and lowest MSEF most frequently. With few exceptions, the screened versions of the LZWZ, A-opt, JMA and EW methods are neither the best nor worst strategies among all of the seventeen strategies considered.

Lastly, we apply the Wilcoxon signed rank test (Wilcoxon, 1945) to test for pairwise performance equality of the methods. The non-rejection of the null hypothesis of this test implies that the pair performs equally well, and the observed difference in accuracy between the two methods in the pair is due to chance only. Table 4 reports the test results for the difference in all three accuracy measures between the screened version of the S-BIC method, generally the best of all FMA methods, and each of the three selection methods at the 1% level of significance. A cell entry of "1" indicates that for the pair in question we are able to reject the null hypothesis of no difference in accuracy; conversely, a "0" indicates that the difference in accuracy is not statistically significant. The results show that in the overwhelming majority of cases, the differences in MSEF and MAEF between the screened version of S-BIC and each of AIC, BIC, and FIC reported in Tables 1 and Tables 2 are statistically significant. On the other hand, the opposite is observed in terms of HitRate. This is because even if the methods produce different $\hat{p}_{0j}$'s, as long as that the category that corresponds to $max\{\hat{p}^w_{01}, ..., \hat{p}^w_{0J}\}$ is the same for the methods, they will yield the same forecast of Y_0. Table 5 reports the Wilcoxon signed rank test results for the difference in MSEF between the screened and non-screened versions of each FMA method. The results show that there are significant differences in MSEF between the screened and non-screened FMA methods in most cases. Similar results are observed in terms of MAEF, but the differences in the HitRates achieved as a result of screening are found to be insignificant in most cases due to the reason given above. To conserve space, these results are not shown here but are available on request from the authors.

5. Empirical Applications

In this section, we apply the various FMA methods to two real datasets used previously in sociological research. We evaluate the strategies in terms of the HitRate only and not with respect to MSEF and MAEF, because for each observation we can only observe the selected category and not the probability of selecting the different categories.

Application 1: Analysis of Individual Self-realisation Data

The data used in this application are taken from the 2007 AsiaBarometer survey based on a sample of 990 ordinary residents of Indonesia[1]. The respondents were

[1]The data are available online at www.asiabarometer.org/. Inoguchi, Basànez, Tanaka and Dadabaev (2005) provided a detailed discussion of the survey.

asked to rate their self-assessed level of life accomplishment on a scale of 1 to 3, with 1=very little or none, 2=some, and 3=a great deal. The number of respondents selecting categories 1, 2 and 3 are 154, 585 and 251 respectively. The survey also provides information on the respondents' personal and demographic characteristics including gender, age, education, household income, employment status, and area of the residence.

The analysis attempts to model individuals' self-realisation evaluation level by the aforementioned personal and demographic characteristics as explanatory variables. As the responses are ordered, the ordered probit model is a meaningful framework for this analysis. We treat all six explanatory variables as non-mandatory, resulting in $2^6 = 64$ sub-models. We select a random sample of 600 observations from the full sample for parameter estimation, and use the remaining 390 observations for model evaluation. Applying the same top m model screening procedure and set $m = 5$ as in Section 3 reduces the number of sub-models within the model average from 64 to 5.

The top panel of Table 6 reports the hit rates produced by the various model selection and averaging methods based on the 390 observations in the test sample. Of the methods considered, the screened version of the S-BIC method has the best performance, followed closely by the non-screened version of the same method and the screened version of the S-FIC method, while the AIC and BIC model selection methods perform worst. Model screening improves the performance of two of the seven FMA estimators being considered; for the other five FMA estimators, the deterioration in performance due to screening is either none or negligible.

Application 2: Analysis of Travel Mode Choice Data

The data in this application, taken from Allison (1999b), contain information on the transportation mode of 210 individuals in Australia. The respondents were asked to indicate which of the following four modes of transportation they used on their most recent trip: 1=air, 2=train, 3=bus and 4=car. They were also asked to report on the cost of the trip, waiting time at the terminal, and the time spent in a vehicle at all stages during the trip. We treat the choice of transportation modes as the dependent variable, and the three other variables as optional explanatory variables, resulting in $2^3 = 8$ sub-models.

We begin our analysis by testing the IIA assumption. The Hausman and McFadden (1984) test rejects the IIA assumption at the 5% significance level. Hence we

adopt the nested logit model as our analytical framework. We split the four modes of transportation into two clusters, with air travel in the first, and train, bus and car in the second. We randomly select 160 observations for estimation and use the remaining 50 observations as a test sample for evaluation. Again, for the implementation of model screening, we set $m = 5$, which reduces the number of sub-models within the model average from 8 to 5.

The lower panel of Table 6 provides a comparison of forecasts based on the various methods in terms of hit rates. Of all methods considered, the non-screened version of the S-FIC method and the screened and non-screened versions of the JMA method are equally the best performer, while the screened version of the S-FIC and the non-screened version of the EW perform as close second. The FIC model selection method yields the worst hit rate. The non-screened and screened versions of various model FMA estimators typically yield very similar hit rates, with the exception of the LZWZ estimator, for which model screening results in a clear improvement in performance.

6. Conclusions

The ordered probit and nested logit models have received both theoretical and empirical support in the literature of quantitative sociology. The ordered probit model is useful for modelling responses that have a natural ordering. The nested logit model is an extension of the ordinary logit model to accommodate the unfulfillment of the IIA property. When applying these models, a sociological researcher would normally consider an array of models, each containing a different combination of regressors, select the best combination according to an off-the-shelf information criterion, and report results based on the final "best" model. In recent years, the practice of model selection has been criticised for ignoring the uncertainty embedded in the model selection process, with the risk associated with some very poor models being chosen. Model averaging, which smoothly interpolates estimates obtained across the different models, is a strategy to overcome the above-mentioned deficiencies of model selection. Model averaging within the frequentist paradigm has been widely applied in a number of disciplines, but has not come into usage in sociology.

In this study, by means of a Monte Carlo experiment, we compare a range of model averaging strategies with several common model selection methods for the ordered probit and nested logit models. We find that overall, model averaging is preferable to model selection, and averaging with screening generally compares favourably with

the strategy of pure averaging without removing the very poor models at the outset. One especially noteworthy aspect of our results is that model averaging with screening rarely if ever produces very poor results; by contrast, model selection can sometimes deliver very inaccurate estimates, especially in situations where the correct model does not "stand out" from the crowd. This finding reinforces a major advantage of averaging over selection, that is, assuring against the selection of a very poor model, which in turn leads to some very inaccurate results being reported in the end. Adding to this advantage is the bonus that some averaging strategies are found to frequently outperform the selection strategies, even in situations where selection is known to perform well. For example, our Monte Carlo results indicate that the S-BIC averaging methods frequently outperform all selection methods across all performance yardsticks considered.

In summary, the results in this paper indicate that model averaging generally produces more robust results and can be extremely useful for sociological researchers who make quantitative methods central to their study. We therefore put the case of the inclusion of this relatively new technique in the quantitative sociologists' repertoire. Clearly, a lot more remains to be done, but hopefully this paper has served to whet the appetite for further explorations of model averaging within the discipline of sociology.

7. Appendix

As mentioned in Section 4, the local misspecification framework forms the basis for the development of the S-FIC, LZWZ and A-opt averaging methods. This Appendix encapsulates the essence of this framework.

For notational convenience, let $\hbar$ be the vector of unknowns corresponding to the mandatory variables in a model. Thus, $\hbar = (\alpha_1, ..., \alpha_{J-1}, \beta')'$ for the ordered probit model and $\hbar = (\alpha_{j_1|B_1}, ..., \alpha_{j_K|B_K}, \tau_1, ..., \tau_K, \beta')'$ for the nested logit models. Let the true parameter vector of the model be $(\hbar'_{true}, \gamma'_0 + \delta'/n^{1/2})'$, where $\hbar_{true}$ is the vector containing the true values of the coefficients in $\hbar$, γ_0 is a vector that consists of values of γ in the narrow model that only contains the mandatory variables, and δ is a $q \times 1$ vector of parameters that signals the various degrees of departure from the narrow model. In our case, γ_0 is equal to a null vector. Together, there exist 2^q submodels obtained by setting different coefficients in δ to 0, leading to 2^q estimators of $\mu = \mu(\hbar, \gamma)$ to choose between or combine. Denote the FMA estimator of μ as $\hat{\mu}^w$.

Let $\mathcal{L}(\hbar,\gamma)$ be the likelihood function for the full model, and $\mathcal{J}_{n,full} = -\frac{1}{n}\frac{\partial^2 log\mathcal{L}(\hbar,\gamma)}{\partial(\hbar',\gamma')'\partial(\hbar',\gamma')}$ $= \begin{pmatrix} \mathcal{J}_{n,00} & \mathcal{J}_{n,01} \\ \mathcal{J}_{n,10} & \mathcal{J}_{n,11} \end{pmatrix}$ and $\mathcal{J}_{A,full} = \begin{pmatrix} \mathcal{J}_{00} & \mathcal{J}_{01} \\ \mathcal{J}_{10} & \mathcal{J}_{11} \end{pmatrix}$ be the corresponding information matrix and limiting information matrix respectively, where $|\hbar|$ is the length of $\hbar$ and $\mathcal{J}_{ij}$ $(i,j=0,1)$ is the limiting value of $\mathcal{J}_{n,ij}$ as n approaches infinity. Both $\mathcal{J}_{n,full}$ and $\mathcal{J}_{A,full}$ are of dimension $(|\hbar|+q)\times(|\hbar|+q)$. Let ϖ_s be the projection matrix that maps the vector $\delta = (\delta_1, ..., \delta_q)$ to the sub-vector $\varpi_s\delta = \delta_s$ that contains the coefficients of δ in the s^{th} sub-model. Write $\mathcal{K} = (\mathcal{J}_{11} - \mathcal{J}_{10}\mathcal{J}_{00}^{-1}\mathcal{J}_{01})^{-1}$, $\mathcal{K}_s = (\varpi_s\mathcal{K}^{-1}\varpi_s')^{-1}$, $H_s = \mathcal{K}^{-1/2}\varpi_s'\mathcal{K}_s\varpi_s\mathcal{K}^{-1/2}$, and $\omega = \mathcal{J}_{10}\mathcal{J}_{00}^{-1}\frac{\partial\mu}{\partial\hbar} - \frac{\partial\mu}{\partial\gamma}$, with the partial derivatives evaluated at $(\hbar_{true}, \gamma_0)$. Note that H_s is a $q\times q$ projection matrix that is orthogonal to $I_q - H_s$, and I_q is a $q\times q$ identity matrix. Hjort and Claeskens (2003) showed that

$$\sqrt{n}(\hat{\mu}^w - \mu) \xrightarrow{d} \Lambda \equiv \left(\frac{\partial\mu}{\partial\hbar}\right)' \mathcal{J}_{00}^{-1}M + \omega'\left\{\delta - \hat{\delta}(D)\right\}, \tag{17}$$

where $\xrightarrow{d}$ denotes convergence in distribution, $D \sim N_q(\delta, \mathcal{K})$, $M \sim N_{|\hbar|}(0, \mathcal{J}_{00})$ is independent of D, and $\hat{\delta}(D) = \mathcal{K}^{1/2}\left\{\sum_{s=1}^{2^q} w_s H_s\right\}\mathcal{K}^{-1/2}D \equiv \mathcal{K}^{1/2}H(w)\mathcal{K}^{-1/2}D$.

It can be shown that the asymptotic squared error risk of $\hat{\mu}^w$ is

$$\begin{aligned} R(\hat{\mu}^w) = E(\Lambda^2) &= \varsigma_0^2 + E\left(\omega'\hat{\delta}(D) - \omega'\delta\right)^2 \\ &= \varsigma_0^2 + \omega'\mathcal{K}^{1/2}H^2(\omega)\mathcal{K}^{1/2}\omega + \left(\omega'\mathcal{K}^{1/2}\mathfrak{L}(w)\mathcal{K}^{-1/2}\delta\right)^2, \end{aligned} \tag{18}$$

where $\varsigma_0^2 = \left(\frac{\partial\mu}{\partial\hbar}\right)'\mathcal{J}_{00}^{-1}\left(\frac{\partial\mu}{\partial\hbar}\right)$ and $\mathfrak{L}(w) = I_q - H(w)$. Unfortunately, $R(\hat{\mu}^w)$ is of little practical utility for finding optimal values of w because ω, $\mathcal{K}$ and δ in $R(\hat{\mu}^w)$ are unknown. The LZWZ and A-opt methods are based on feasible variants of (18); LZWZ selects w by minimising an approximately unbiased estimator of $R(\hat{\mu}^w)$, while A-opt selects w by minimising a plug-in estimator of $R(\hat{\mu}^w)$.

Specifically, Liang, Zou, Wan and Zhang (2011) showed that

$$\tilde{R}(\hat{\mu}^w) = \varsigma_0^2 + \omega'\mathcal{K}\omega + (\omega'\mathcal{K}^{1/2}\mathfrak{L}(w)\mathcal{K}^{-1/2}D)^2 + 2\omega'\mathcal{K}^{1/2}H(w)\mathcal{K}^{1/2}\omega \tag{19}$$

is an unbiased estimator of $R(\hat{\mu}^w)$. The objective function $\hat{R}(\hat{\mu}^w)$ associated with LZWZ method given in (11) is obtained by deleting the first two terms that are unrelated to w on the r.h.s. of equation (19), and replacing ω, $\mathcal{K}$ $H(w)$, D and $\mathfrak{L}(w)$ in the last two terms of the same equation by their respective consistent estimators $\hat{\omega}$, $\hat{\mathcal{K}}$, $\hat{H}(w)$, $\hat{\delta}$ and $\hat{\mathfrak{L}}(w)$. Note that $\hat{H}(w)$ and $\hat{\mathfrak{L}}(w)$ in (11) have the same expressions

as $H(w)$ and $\mathfrak{L}(w)$ in (19), except that $\mathcal{K}$ contained in $H(w)$ and $\mathfrak{L}(w)$ are replaced by $\hat{\mathcal{K}}$ in the construction of $\hat{H}(w)$ and $\hat{\mathfrak{L}}(w)$.

For the A-opt method, the objective function (12) is obtained by removing ς_0^2 that is unrelated to w from the r.h.s. of (18), and replacing ω, $\mathcal{K}$, $H(w)$, $\mathfrak{L}(w)$ in (18) by δ with $\hat{\omega}$, $\hat{\mathcal{K}}$, $\hat{H}(w)$, $\hat{\mathfrak{L}}(w)$ and $\hat{\delta}$ respectively. See Wan, Zhang and Wang (2014) for details.

References

Allison, P.D., 1999a. Comparing logit and probit coefficients across groups. Sociological Methods & Research 28, 186–208.

Allison, P.D., 1999b. Logistic regression using the SAS system: theory and application. Cary, NC: SAS Institute Inc.

Amini, S.M., Parmeter, C.F., 2012. Comparison of model averaging techniques: Assessing growth determinants. Journal of Applied Econometrics 27, 870–876.

Breen, R., Karlson, K.B., Holm, A., 2013. Total, direct, and indirect effects in logit and probit models. Sociological Methods & Research 42, 164–191.

Buckland, S.T., Burnham, K.P., Augustin, N.H., 1997. Model selection: an integral part of inference. Biometrics 53, 603–618.

Burnham, K.P., Anderson, D.R., 2004. Multimodel inference understanding AIC and BIC in model selection. Sociological Methods & Research 33, 261–304.

Cheng, S., Long, J.S., 2007. Testing for IIA in the multinomial logit model. Sociological Methods & Research 35, 583–600.

Claeskens, G., Croux, C., Van Kerckhoven, J., 2006. Variable selection for logistic regression using a prediction-focused information criterion. Biometrics 62, 972–979.

Claeskens, G., Hjort, N.L., 2003. The focused information criterion. Journal of the American Statistical Association 98, 879–899.

Claeskens, G., Hjort, N.L., 2008. Model selection and model averaging. volume 330. Cambridge: Cambridge university Press.

Danilov, D., Magnus, J.R., 2004. On the harm that ignoring pretesting can cause. Journal of Econometrics 122, 27–46.

Deller, S., Amiel, L., Deller, M., 2011. Model uncertainty in ecological criminology: an application of Bayesian model averaging with rural crime data. International Journal of Criminology and Sociological Theory 4, 683–717.

Duan, K., Mei, Y., 2014. A comparison study of three statistical downscaling methods and their model-averaging ensemble for precipitation downscaling in China. Theoretical and Applied Climatology 116, 707–719.

Fullerton, A.S., 2009. A conceptual framework for ordered logistic regression models. Sociological Methods & Research 38, 306–347.

Gayle, V., Lambert, P.S., 2009. Logistic regression models in sociological research. Technical Report. DAMES Node.

Giles, J.A., Giles, D.E.A, 1993. Pre-test estimation and testing in econometrics: recent developments. Journal of Economic Surveys 7, 145–197.

Grant, D., Morales, A., Sallaz, J.J., 2009. Pathways to meaning: a new approach to studying emotions at work. American Journal of Sociology 115, 327–364.

Hansen, B.E., 2007. Least squares model averaging. Econometrica 75, 1175–1189.

Hansen, B.E, Racine, J.S., 2012. Jackknife model averaging. Journal of Econometrics 167, 38–46.

Hausman, J., McFadden, D., 1984. Specification tests for the multinomial logit model. Econometrica: Journal of the Econometric Society 52, 1219–1240.

Hjort, N.L., Claeskens, G., 2003. Frequentist model average estimators. Journal of the American Statistical Association 98, 879–899.

Hjort, N.L., Claeskens, G., 2006. Focused information criteria and model averaging for the Cox hazard regression model. Journal of the American Statistical Association 101, 1449–1464.

Inoguchi, T., Basànez, M., Tanaka, A., Dadabaev, T. (eds.), 2005. Values and life styles in urban Asia: a cross-cultural analysis and sourcebook based on the AsiaBarometer survey of 2003. volume 19. Siglo XXI.

Jackson, C. H., Thompson, S.G., Sharples, L.D., 2009. Accounting for uncertainty in health economic decision models by using model averaging. Journal of the Royal Statistical Society: Series A (Statistics in Society) 172, 383–404.

Johnson, J.B., Omland, K.S., 2004. Model selection in ecology and evolution. Trends in Ecology & Evolution 19, 101–108.

Karlson, K.B., Holm, A., Breen, R., 2012. Comparing regression coefficients between same-sample nested models using logit and probit: a new method. Sociological Methodology 42, 286–313.

Kuha, J., 2004. AIC and BIC: comparisons of assumptions and performance. Sociological Methods & Research 33, 188–229.

Leeb, H., Pötscher, B.M., 2005. Model selection and inference: Facts and fiction. Econometric Theory 21, 21–59.

Liang, H., Zou, G., Wan, A.T.K, Zhang, X., 2011. Optimal weight choice for Frequentist model average estimators. Journal of the American Statistical Association 106, 1053–1066.

Liu, C.A., 2015. Distribution theory of the least squares averaging estimator. Journal of Econometrics 186, 142–159.

Long, J.S., 2012. Regression models for nominal and ordinal outcomes. Technical Report. Indiana University.

Lucas, S.R., 2001. Effectively maintained inequality: education transitions, track mobility, and social background effects. American Journal of Sociology 106, 1642–1690.

Montgomery, J.M., Nyhan, B., 2010. Bayesian model averaging: theoretical developments and practical applications. Political Analysis 18, 245–270.

Moral-Benito, E., 2015. Model averaging in economics: an overview. Journal of Economic Surveys 29, 46–75.

Plotnick, R.D., 1992. The effects of attitudes on teenage premarital pregnancy and its resolution. American Sociological Review 57, 800–811.

Quinn, M.A., Rubb, S., 2005. The importance of education-occupation matching in migration decisions. Demography 42, 153–167.

Raftery, A.E., 1995. Bayesian model selection in social research. Sociological Methodology 25, 111–164.

Stine, R.A., 2004. Model selection using information theory and the MDL principle. Sociological Methods & Research 33, 230–260.

Wan, A.T.K., Zhang, X., 2009. On the use of model averaging in tourism research. Annals of Tourism Research 36, 525–532.

Wan, A.T.K., Zhang, X., Wang, S., 2014. Frequentist model averaging for multinomial and ordered logit models. International Journal of Forecasting 30, 118–128.

Wang, H., Zhang, X., Zou, G., 2009. Frequentist model averaging estimation: a review. Journal of Systems Science and Complexity 22, 732–748.

Weakliem, D.L., 2004. Introduction to the special issue on model selection. Sociological Methods & Research 33, 167–187.

Wilcoxon, F., 1945. Individual comparisons by ranking methods. Biometrics Bulletin 1, 80–83.

Williams, R., 2009. Using heterogeneous choice models to compare logit and probit coefficients across groups. Sociological Methods & Research 37, 531–559.

Yuan, Z., Yang, Y., 2005. Combining linear regression models: when and how? Journal of the American Statistical Association 100, 1202–1214.

Zhang, X., Liang, H., 2011. Focused information criterion and model averaging for generalized additive partial linear models. The Annals of Statistics 39, 174–200.

Zhang, X., Wan, A.T.K., Zhou, S.Z., 2012. Focused information criteria, model selection, and model averaging in a Tobit model with a nonzero threshold. Journal of Business & Economic Statistics 30, 132–142.

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

Table 1: MSEF

		model selection			model averaging without screening							model averaging with screening						
l	S	AIC	BIC	FIC	S-AIC	S-BIC	S-FIC	LZWZ	A-opt	JMA	EW	S-AIC	S-BIC	S-FIC	LZWZ	A-opt	JMA	EW
									Design 1									
0.5	1	0.00884	$0.00987^{(-2)}$	$0.00915^{(-3)}$	$0.00795^{(1)}$	0.00858	0.00830	0.00838	0.00834	0.00824	$0.01454^{(-1)}$	0.00806	0.00860	$0.00798^{(2)}\uparrow$	$0.00810\uparrow$	$0.00804^{(3)}\uparrow$	$0.00824\uparrow$	$0.00805\uparrow$
	2	0.00835	$0.00938^{(-2)}$	$0.00905^{(-3)}$	$0.00769^{(1)}$	0.00820	0.00822	0.00834	0.00831	0.00793	$0.01447^{(-1)}$	$0.00773^{(2)}$	0.00821	$0.00780^{(3)}\uparrow$	$0.00803\uparrow$	$0.00797\uparrow$	$0.00792\uparrow$	$0.00786\uparrow$
	3	$0.00820^{(-3)}$	$0.00862^{(-1)}$	$0.00828^{(-2)}$	0.00715	0.00712	$0.00704^{(3)}$	0.00757	0.00791	0.00769	$0.00652^{(1)}$	0.00725	0.00725	0.00725	$0.00743\uparrow$	$0.00751\uparrow$	$0.00755\uparrow$	$0.00697^{(2)}$
									Design 2									
	1	0.00788	0.00786	$0.01034^{(-2)}$	$0.00725^{(3)}$	0.00738	0.00933	$0.00974^{(-3)}$	0.00894	0.00769	$0.03080^{(-1)}$	0.00725	$0.00737\uparrow$	$0.00717^{(1)}\uparrow$	$0.00848\uparrow$	$0.00805\uparrow$	$0.00769\uparrow$	$0.00718^{(2)}\uparrow$
	2	0.00688	$0.00610^{(1)}$	$0.00915^{(-2)}$	0.00650	$0.00617^{(3)}$	$0.00909^{(-3)}$	0.00903	0.00857	0.00686	$0.03051^{(-1)}$	$0.00648\uparrow$	$0.00616^{(2)}\uparrow$	$0.00661\uparrow$	$0.00768\uparrow$	$0.00747\uparrow$	$0.00686\uparrow$	$0.00657\uparrow$
	3	0.00710	$0.00576^{(3)}$	$0.01172^{(-1)}$	0.00647	$0.00562^{(2)}$	0.00961	$0.01114^{(-2)}$	0.00886	0.00720	$0.01094^{(-3)}$	$0.00634\uparrow$	$0.00561^{(1)}\uparrow$	$0.00599\uparrow$	$0.00667\uparrow$	$0.00688\uparrow$	$0.00715\uparrow$	$0.00592\uparrow$
									Design 1									
1	1	0.00710	0.00713	$0.00779^{(-2)}$	$0.00647^{(2)}$	0.00678	0.00719	$0.00751^{(-3)}$	0.00734	0.00681	$0.03304^{(-1)}$	$0.00647^{(1)}\uparrow$	$0.00676\uparrow$	$0.00652^{(3)}\uparrow$	$0.00676\uparrow$	$0.00675^{(2)}\uparrow$	$0.00679\uparrow$	$0.00678\uparrow$
	2	0.00612	0.00578	$0.00728^{(-3)}$	0.00577	$0.00566^{(2)}$	0.00689	$0.00728^{(-2)}$	0.00718	0.00600	$0.03303^{(-1)}$	$0.00575^{(3)}\uparrow$	$0.00564^{(1)}\uparrow$	$0.00608\uparrow$	$0.00643\uparrow$	$0.00651\uparrow$	$0.00598\uparrow$	$0.00623\uparrow$
	3	0.00663	$0.00517^{(1)}$	$0.00747^{(-3)}$	0.00601	$0.00539^{(3)}$	0.00699	0.00723	$0.00749^{(-2)}$	0.00664	$0.01044^{(-1)}$	$0.00578\uparrow$	$0.00533^{(2)}\uparrow$	$0.00585\uparrow$	$0.00660\uparrow$	$0.00690\uparrow$	$0.00656\uparrow$	$0.00592\uparrow$
									Design 2									
	1	0.00857	0.00869	$0.01188^{(-2)}$	$0.00778^{(3)}$	0.00785	0.00960	$0.01098^{(-3)}$	0.01056	0.00824	$0.04649^{(-1)}$	$0.00778\uparrow$	$0.00784\uparrow$	$0.00757^{(1)}\uparrow$	$0.00895\uparrow$	$0.00862\uparrow$	$0.00824\uparrow$	$0.00776^{(2)}\uparrow$
	2	0.00704	$0.00609^{(3)}$	$0.01044^{(-2)}$	0.00663	$0.00606^{(2)}$	0.00884	$0.01019^{(-3)}$	0.01001	0.00697	$0.04574^{(-1)}$	$0.00663\uparrow$	$0.00606^{(1)}\uparrow$	$0.00672\uparrow$	$0.00797\uparrow$	$0.00786\uparrow$	$0.00697\uparrow$	$0.00683\uparrow$
	3	0.00710	$0.00560^{(3)}$	$0.00868^{(-2)}$	0.00644	$0.00554^{(2)}$	0.00831	$0.00856^{(-3)}$	0.00814	0.00711	$0.01457^{(-1)}$	$0.00629\uparrow$	$0.00553^{(1)}\uparrow$	$0.00595\uparrow$	$0.00644\uparrow$	$0.00683\uparrow$	$0.00708\uparrow$	$0.00578\uparrow$
									Design 1									
2	1	0.00726	0.00837	0.00880	$0.00661^{(3)}$	0.00721	0.00748	$0.00931^{(-2)}$	$0.00896^{(-3)}$	0.00692	$0.07015^{(-1)}$	$0.00661^{(2)}\uparrow$	$0.00721\uparrow$	$0.00627^{(1)}\uparrow$	$0.00684\uparrow$	$0.00666\uparrow$	$0.00692\uparrow$	$0.00770\uparrow$
	2	0.00567	$0.00477^{(3)}$	0.00743	0.00525	$0.00468^{(2)}$	0.00661	$0.00869^{(-2)}$	$0.00858^{(-3)}$	0.00551	$0.07029^{(-1)}$	$0.00525\uparrow$	$0.00468^{(1)}\uparrow$	$0.00530\uparrow$	$0.00606\uparrow$	$0.00613\uparrow$	0.00551	$0.00646\uparrow$
	3	0.00598	$0.00452^{(3)}$	0.00636	0.00533	$0.00444^{(2)}$	0.00582	$0.00655^{(-3)}$	$0.00674^{(-2)}$	0.00598	$0.02116^{(-1)}$	$0.00519\uparrow$	$0.00444^{(1)}\uparrow$	$0.00499\uparrow$	$0.00543\uparrow$	$0.00584\uparrow$	$0.00596\uparrow$	$0.00466\uparrow$
									Design 2									
	1	0.00855	0.00944	$0.01414^{(-2)}$	$0.00779^{(3)}$	0.00818	0.01087	$0.01335^{(-3)}$	0.01266	0.00818	$0.06660^{(-1)}$	$0.00779^{(2)}\uparrow$	$0.00818\uparrow$	$0.00763^{(1)}\uparrow$	$0.00944\uparrow$	$0.00902\uparrow$	$0.00817\uparrow$	$0.00838\uparrow$
	2	0.00685	$0.00584^{(3)}$	0.01178	0.00638	$0.00573^{(2)}$	0.00966	$0.01216^{(-2)}$	$0.01185^{(-3)}$	0.00658	$0.06587^{(-1)}$	$0.00638\uparrow$	$0.00573^{(1)}\uparrow$	$0.00652\uparrow$	$0.00805\uparrow$	$0.00794\uparrow$	$0.00657\uparrow$	$0.00722\uparrow$
	3	0.00697	$0.00530^{(3)}$	$0.00784^{(-2)}$	0.00623	$0.00527^{(2)}$	0.00725	0.00770	$0.00781^{(-3)}$	0.00676	$0.01969^{(-1)}$	$0.00606\uparrow$	$0.00526^{(1)}\uparrow$	$0.00579\uparrow$	$0.00629\uparrow$	$0.00670\uparrow$	$0.00672\uparrow$	$0.00554\uparrow$
	(1)	0	2	0	2	0	0	0	0	0	1	1	8	4	0	0	0	0
	(2)	0	0	0	1	8	0	0	0	0	0	3	2	1	0	0	0	3
	(3)	0	7	0	4	2	1	0	0	0	0	1	0	2	1	1	0	0
	(-3)	1	0	4	0	0	1	7	4	0	1	0	0	0	0	0	0	0
	(-2)	0	2	9	0	0	0	5	2	0	0	0	0	0	0	0	0	0
	(-1)	0	1	1	0	0	0	0	0	0	16	0	0	0	0	0	0	0

[1] : (1), (2), (3), (-3), (-2), (-1) = Number of cases yielding the best, second best, third best, third worst, second worst and worst estimates respectively.

Table 2: MAEF

		model selection			model averaging without screening							model averaging with screening						
l	S	AIC	BIC	FIC	S-AIC	S-BIC	S-FIC	LZWZ	A-opt	JMA	EW	S-AIC	S-BIC	S-FIC	LZWZ	A-opt	JMA	EW
								Design 1										
0.5	1	0.11749	$0.12395^{(-2)}$	$0.11952^{(-3)}$	$0.11217^{(2)}$	0.11604	0.11432	0.11486	0.11481	0.11404	$0.15127^{(-1)}$	0.11278	0.11617	$0.11219^{(3)}\uparrow$	0.11303↑	0.11277↑	0.11395↑	$0.11216^{(1)}\uparrow$
	2	0.11325	$0.12047^{(-2)}$	$0.11827^{(-3)}$	$0.10986^{(1)}$	0.11296	0.11333	0.11412	0.11416	0.11113	$0.15062^{(-1)}$	$0.10996^{(2)}$	0.11301	$0.11038^{(3)}\uparrow$	0.11208↑	0.11184↑	0.11099↑	0.11048↑
	3	$0.11142^{(-3)}$	$0.11458^{(-1)}$	$0.11333^{(-2)}$	0.10561	0.10555	$0.10524^{(3)}$	0.10883	0.11131	0.10872	$0.10192^{(1)}$	0.10626	0.10636	0.10639	0.10786↑	0.10857↑	0.10779↑	$0.10441^{(2)}$
								Design 2										
	1	0.11219	0.11162	$0.12516^{(-2)}$	0.10748	0.10813	$0.12338^{(-3)}$	0.12291	0.11953	0.11085	$0.24541^{(-1)}$	$0.10745^{(3)}\uparrow$	0.10805↑	$0.10685^{(1)}\uparrow$	0.11470↑	0.11251↑	0.11081↑	$0.10732^{(2)}\uparrow$
	2	0.10419	$0.09780^{(1)}$	0.11712	0.10170	$0.09881^{(3)}$	$0.12135^{(-2)}$	$0.11934^{(-3)}$	0.11755	0.10430	$0.24406^{(-1)}$	0.10158↑	$0.09872^{(2)}\uparrow$	0.10257↑	0.10911↑	0.10831↑	0.10429↑	0.10257↑
	3	0.10990	$0.09916^{(3)}$	$0.13539^{(-2)}$	0.10539	$0.09849^{(2)}$	0.12954	$0.13430^{(-3)}$	0.12349	0.11086	$0.14007^{(-1)}$	0.10433↑	$0.09835^{(1)}\uparrow$	0.10170↑	0.10536↑	0.10799↑	0.11047↑	0.10127↑
								Design 1										
1	1	0.10925	0.10842	$0.11302^{(-2)}$	$0.10416^{(2)}$	0.10601	0.10935	$0.11166^{(-3)}$	0.11106	0.10693	$0.24306^{(-1)}$	$0.10412^{(1)}\uparrow$	0.10588↑	$0.10436^{(3)}\uparrow$	0.10631↑	0.10622↑	0.10684↑	0.10580↑
	2	0.10070	$0.09599^{(1)}$	0.10826	0.09832	$0.09643^{(3)}$	0.10693	$0.10973^{(-2)}$	$0.10971^{(-3)}$	0.10003	$0.24243^{(-1)}$	0.09817↑	$0.09631^{(2)}\uparrow$	0.10059↑	0.10341↑	0.10430↑	0.09996↑	0.10130↑
	3	0.10504	$0.09216^{(1)}$	$0.11020^{(-3)}$	0.10084	$0.09471^{(3)}$	0.10814	0.10960	$0.11195^{(-2)}$	0.10538	$0.13191^{(-1)}$	0.09884↑	$0.09421^{(2)}\uparrow$	0.09926↑	0.10498↑	0.10773↑	0.10491↑	0.09987↑
								Design 2										
	1	0.11523	0.11656	$0.13159^{(-2)}$	$0.10983^{(3)}$	0.11050	0.12178	$0.12816^{(-3)}$	0.12606	0.11294	$0.30754^{(-1)}$	$0.10982^{(2)}\uparrow$	0.11046↑	$0.10825^{(1)}\uparrow$	0.11640↑	0.11457↑	0.11294	0.11083↑
	2	0.10408	$0.09697^{(1)}$	0.12165	0.10142	$0.09710^{(3)}$	0.11682	$0.12313^{(-2)}$	$0.12270^{(-3)}$	0.10362	$0.30473^{(-1)}$	0.10139↑	$0.09706^{(2)}\uparrow$	0.10221↑	0.10949↑	0.10932↑	0.10362↑	0.10388↑
	3	0.10965	$0.09779^{(3)}$	0.11825	0.10505	$0.09758^{(2)}$	$0.11940^{(-3)}$	$0.11947^{(-2)}$	0.11830	0.10990	$0.16295^{(-1)}$	0.10381↑	$0.09750^{(1)}\uparrow$	0.10124↑	0.10475↑	0.10810↑	0.10970↑	0.0999↑8
								Design 1										
2	1	0.09971	0.10848	0.10678	$0.09518^{(3)}$	0.10034	0.10046	$0.11058^{(-2)}$	$0.10961^{(-3)}$	0.09711	$0.38980^{(-1)}$	$0.09518^{(2)}\uparrow$	0.10034↑	$0.09275^{(1)}\uparrow$	0.09685↑	0.09580↑	0.09711↑	0.10665↑
	2	0.08802	$0.08115^{(3)}$	0.09720	0.08520	$0.08072^{(2)}$	0.09467	$0.10671^{(-3)}$	$0.10711^{(-2)}$	0.08711	$0.39046^{(-1)}$	0.08520↑	$0.08072^{(1)}\uparrow$	0.08575↑	0.09131↑	0.09213↑	0.08711	0.09775↑
	3	0.09961	$0.08699^{(3)}$	0.10161	0.09471	$0.08674^{(2)}$	0.09865	$0.10429^{(-3)}$	$0.10633^{(-2)}$	0.09997	$0.19440^{(-1)}$	0.09344↑	$0.08669^{(1)}\uparrow$	0.09168↑	0.09534↑	0.09905↑	0.09980↑	0.08915↑
								Design 2										
	1	0.10828	0.11426	$0.13141^{(-2)}$	$0.10335^{(3)}$	0.10624	0.12043	$0.12980^{(-3)}$	0.12708	0.10589	$0.37255^{(-1)}$	$0.10335^{(2)}\uparrow$	0.10624↑	$0.10216^{(1)}\uparrow$	0.11128↑	0.10933↑	0.10583↑	0.10924↑
	2	0.09641	$0.08953^{(3)}$	0.11801	0.09372	$0.08908^{(2)}$	0.11338	$0.12310^{(-2)}$	$0.12258^{(-3)}$	0.09487	$0.36996^{(-1)}$	0.09372↑	$0.08908^{(1)}\uparrow$	0.09465↑	0.10249↑	0.10256↑	0.09485↑	0.10129↑
	3	0.10518	$0.09208^{(1)}$	0.11031	0.10024	$0.09227^{(3)}$	0.10845	$0.11116^{(-3)}$	$0.11248^{(-2)}$	0.10398	$0.18797^{(-1)}$	0.09885↑	$0.09220^{(2)}\uparrow$	0.09685↑	0.10044↑	0.10395↑	0.10373↑	0.09502↑
	(1)	0	5	0	1	0	0	0	0	0	1	1	5	4	0	0	0	1
	(2)	0	0	0	2	5	0	0	0	0	0	4	5	0	0	0	0	2
	(3)	0	5	0	3	5	1	0	0	0	0	1	0	3	0	0	0	0
	(-3)	1	0	3	0	0	2	8	4	0	0	0	0	0	0	0	0	0
	(-2)	0	2	6	0	0	1	5	4	0	0	0	0	0	0	0	0	0
	(-1)	0	1	0	0	0	0	0	0	0	17	0	0	0	0	0	0	0

[1] : (1), (2), (3), (-3), (-2), (-1) = Number of cases yielding the best, second best, third best, third worst, second worst and worst estimates respectively.

Table 3: HitRate

		model selection			**model averaging without screening**							**model averaging with screening**						
l	S	AIC	BIC	FIC	S-AIC	S-BIC	S-FIC	LZWZ	A-opt	JMA	EW	S-AIC	S-BIC	S-FIC	LZWZ	A-opt	JMA	EW
										Design 1								
.5	1	0.55466	$0.55238^{(-2)}$	$0.55306^{(-3)}$	$0.55522^{(3)}$	0.55350	0.55368	0.55322	0.55454	0.55522	$0.54616^{(-1)}$	0.55514	0.55330	0.55432↑	0.55458↑	$0.55556^{(1)}$↑	$0.55556^{(2)}$↑	0.55416↑
	2	0.55206	0.55054	$0.54960^{(-2)}$	$0.55282^{(1)}$	0.55176	$0.55044^{(-3)}$	0.55070	0.55162	0.55236	$0.54352^{(-1)}$	$0.55240^{(3)}$	0.55190↑	0.55238↑	0.55136↑	0.55184↑	$0.55260^{(2)}$↑	0.55184↑
	3	0.48450	$0.47836^{(-1)}$	0.48398	$0.48496^{(2)}$	0.48262	0.48266	0.48334	0.48484	0.48482	0.48232	0.48318	0.48182	$0.48164^{(-3)}$	0.48310	$0.48546^{(1)}$↑	$0.48490^{(3)}$↑	$0.48112^{(-2)}$
										Design 2								
	1	0.55150	0.55248	$0.55008^{(-3)}$	$0.55250^{(3)}$	$0.55288^{(2)}$	0.55132	$0.54976^{(-2)}$	0.55022	0.55172	$0.53750^{(-1)}$	0.55250	$0.55294^{(1)}$↑	0.55180↑	0.55130↑	0.55202↑	0.55172	0.55166↑
	2	0.55192	$0.55250^{(2)}$	0.55108	0.55164	0.55210	$0.54970^{(-2)}$	$0.54954^{(-1)}$	0.55002	0.55186	$0.53540^{(-3)}$	0.55164	0.55200	0.55230↑	$0.55244^{(3)}$↑	$0.55270^{(1)}$↑	0.55180	0.55178↑
	3	0.47990	$0.48104^{(3)}$	$0.47192^{(-1)}$	0.47996	$0.48154^{(2)}$	0.47694	$0.47382^{(-2)}$	0.47740	0.47988	$0.47514^{(-3)}$	0.48016↑	$0.48158^{(1)}$↑	0.48010↑	0.48026↑	0.47964↑	0.47990↑	0.48040↑
										Design 1								
1	1	0.61230	0.61148	0.61124	$0.61246^{(1)}$	0.61230	0.61120	$0.61082^{(-3)}$	0.61134	0.61234	$0.60090^{(-1)}$	$0.61244^{(2)}$	0.61236↑	0.61206↑	0.61164↑	0.61152↑	$0.61244^{(3)}$↑	$0.61066^{(-2)}$↑
	2	0.60726	0.60736	0.60562	$0.60766^{(2)}$	$0.60768^{(1)}$	$0.60536^{(-2)}$	$0.60536^{(-3)}$	0.60576	0.60740	$0.59498^{(-1)}$	$0.60766^{(3)}$	0.60760	0.60726↑	0.60666↑	0.60658↑	0.60758↑	0.60654↑
	3	0.49184	$0.49396^{(3)}$	0.49106	0.49200	$0.49428^{(1)}$	0.49102	$0.49012^{(-2)}$	$0.49020^{(-3)}$	0.49126	$0.48514^{(-1)}$	0.49288↑	$0.49408^{(2)}$	0.49310↑	0.49232↑	0.49126↑	0.49148↑	0.49372↑
										Design 2								
	1	0.63002	0.63030	0.62896	$0.63066^{(3)}$	0.63030	0.62946	$0.62802^{(-2)}$	$0.62852^{(-3)}$	0.62996	$0.61418^{(-1)}$	$0.63066^{(3)}$	0.63034↑	$0.63132^{(2)}$↑	0.62992↑	0.63024↑	0.62998↑	$0.63134^{(1)}$↑
	2	0.63252	$0.63366^{(1)}$	0.63110	0.63256	$0.63318^{(2)}$	0.63074	$0.63014^{(-3)}$	$0.63010^{(-2)}$	0.63268	$0.61422^{(-1)}$	0.63254	$0.63314^{(3)}$	0.63196↑	0.63072↑	0.63104↑	0.63268	0.63190↑
	3	$0.53108^{(-3)}$	0.53278	0.53116	0.53250	$0.53294^{(3)}$	0.53126	0.53136	$0.53106^{(-2)}$	0.53132	$0.52352^{(-1)}$	0.53244	$0.53308^{(2)}$↑	0.53274↑	0.53282↑	0.53196↑	0.53126	$0.53336^{(1)}$↑
										Design 1								
2	1	0.69274	$0.69208^{(-2)}$	$0.69244^{(-3)}$	0.69324	0.69322	0.69286	0.69266	0.69250	0.69310	$0.66842^{(-1)}$	0.69324	0.69322	$0.69454^{(1)}$↑	$0.69354^{(3)}$↑	$0.69406^{(2)}$↑	0.69310	0.69286↑
	2	$0.68898^{(3)}$	$0.68966^{(1)}$	0.68748	0.68862	$0.68934^{(2)}$	0.68686	$0.68594^{(-3)}$	$0.68604^{(-2)}$	0.68888	$0.66450^{(-1)}$	0.68862	$0.68934^{(2)}$	0.68822↑	0.68764↑	0.68764↑	0.68888	0.68790↑
	3	$0.54300^{(-2)}$	0.54444	0.54392	0.54418	$0.54552^{(2)}$	0.54384	$0.54328^{(-3)}$	0.54372	0.54332	$0.53108^{(-1)}$	0.54452↑	$0.54566^{(1)}$↑	$0.54500^{(3)}$↑	0.54472↑	0.54386↑	0.54330	0.54468↑
										Design 2								
	1	0.71668	0.71626	$0.71564^{(-2)}$	$0.71810^{(1)}$	$0.71790^{(2)}$	0.71752	0.71662	0.71724	0.71768	$0.69564^{(-1)}$	$0.71810^{(1)}$	$0.71790^{(2)}$	$0.71778^{(3)}$↑	$0.71626^{(-3)}$	0.71724	0.71764	0.71720↑
	2	$0.72024^{(1)}$	0.71996	$0.71750^{(-3)}$	$0.72000^{(3)}$	0.71994	0.71894	$0.71724^{(-2)}$	0.71762	$0.72002^{(2)}$	$0.69722^{(-1)}$	$0.72000^{(3)}$	0.71994	0.71946↑	0.71840↑	0.71820↑	$0.72002^{(2)}$	0.71896↑
	3	0.61800	$0.62020^{(1)}$	$0.61720^{(-2)}$	0.61866	$0.61966^{(3)}$	0.61786	$0.61752^{(-3)}$	0.61760	0.61806	$0.60770^{(-1)}$	0.61872↑	0.61960	0.61920↑	0.61826↑	0.61780↑	0.61816↑	$0.61978^{(2)}$↑
(1)		1	3	0	3	2	0	0	0	0	0	1	3	1	0	3	0	2
(2)		0	1	0	2	6	0	0	0	1	0	1	4	1	0	1	3	1
(3)		1	2	0	4	2	0	0	0	0	0	4	1	2	2	0	2	0
(-3)		1	0	4	0	0	1	6	2	0	2	0	0	1	1	0	0	0
(-2)		1	2	3	0	0	2	5	3	0	0	0	0	0	0	0	0	2
(-1)		0	1	1	0	0	0	1	0	0	15	0	0	0	0	0	0	0

[1] : (1), (2), (3), (-3), (-2), (-1) = Number of cases yielding the best, second best, third best, third worst, second worst and worst estimates respectively.

Table 4: Results of Wilcoxon's sign rank test for equal performance between the screened S-BIC and model selection methods

		MSEF			MAEF			HitRate		
l	S	AIC	BIC	FIC	AIC	BIC	FIC	AIC	BIC	FIC
					Design 1					
0.5	1	1	1	1	0	1	1	0	0	0
	2	0	1	1	0	1	1	0	0	0
	3	1	1	1	1	1	1	0	1	0
					Design 2					
	1	1	1	1	1	1	1	0	0	0
	2	1	1	1	1	1	1	0	0	0
	3	1	1	1	1	1	1	0	0	1
					Design 1					
1	1	1	1	1	1	1	1	0	0	0
	2	0	1	1	0	1	1	0	0	1
	3	1	1	1	1	1	1	1	0	1
					Design 2					
	1	1	1	1	1	1	1	0	0	0
	2	1	1	1	1	1	1	0	0	1
	3	1	1	1	1	1	1	1	0	0
					Design 1					
2	1	0	1	1	0	1	1	0	0	0
	2	1	1	1	1	1	1	0	0	1
	3	1	1	1	1	1	1	1	0	0
					Design 2					
	1	1	1	1	1	1	1	0	1	0
	2	1	1	1	1	1	1	0	0	1
	3	1	1	1	1	1	1	0	0	1

[1] A cell entry of 1 indicates rejection of the null of equal accuracy between the S-BIC method and the model selection method, whereas 0 indicates non-rejection. All tests are conducted at the 1% level of significance.

Table 5: Results of Wilcoxon's sign rank test for equal performance between the screened and non-screened versions of each FMA method

l	S	S-AIC	S-BIC	S-FIC	EW	LZWZ	A-opt	JMA
				Design 1				
0.5	1	1	1	1	1	1	1	0
	2	1	1	1	1	1	1	0
	3	1	1	1	1	1	1	0
				Design 2				
	1	0	0	1	1	1	1	0
	2	1	1	1	1	1	1	1
	3	1	1	1	1	1	1	1
				Design 1				
1	1	0	1	1	1	1	1	0
	2	1	1	1	1	1	1	1
	3	1	1	1	1	1	1	1
				Design 2				
	1	1	1	1	1	1	1	0
	2	1	1	1	1	1	1	1
	3	1	1	1	1	1	1	1
				Design 1				
2	1	1	1	1	1	1	1	1
	2	1	1	1	1	1	1	1
	3	1	1	1	1	1	1	1
				Design 2				
	1	1	1	1	1	1	1	1
	2	1	1	1	1	1	1	1
	3	1	1	1	1	1	1	1

[1] A cell entry of 1 indicates rejection of the null of equal accuracy between the screened and non-screened versions of a given FMA method, whereas 0 indicates non-rejection. All tests are conducted at the 1% level of significance.

Table 6: Results of empirical applications

	model selection			**model averaging without screening**							**model averaging with screening**						
	AIC	BIC	FIC	S-AIC	S-BIC	S-FIC	LZWZ	A-opt	JMA	EW	S-AIC	S-BIC	S-FIC	LZWZ	A-opt	JMA	EW
	Application I																
Hit Rate	0.63846*	0.63846*	0.65128	0.64872	0.65385	0.65128	0.65128	0.65128	0.64872	0.65128	0.64615	0.65641‡↑	0.65385↑	0.64872	0.64872	0.64615	0.64615
	Application II																
Hit Rate	0.74000	0.74000	0.60000*	0.74000	0.74000	0.78000‡	0.68000	0.72000	0.78000‡	0.76000	0.74000	0.74000	0.76000	0.70000↑	0.72000	0.78000‡	0.74000

[1] :‡= Best estimator
[2] :* = Worst estimator