


A model averaging approach for the ordered probit and nested logit models with applications

Longmei Chen, Alan T. K. Wan, Geoffrey Tso & Xinyu Zhang

To cite this article: Longmei Chen, Alan T. K. Wan, Geoffrey Tso & Xinyu Zhang (2018) A model averaging approach for the ordered probit and nested logit models with applications, Journal of Applied Statistics, 45:16, 3012-3052, DOI: [10.1080/02664763.2018.1450367](https://doi.org/10.1080/02664763.2018.1450367)

To link to this article: <https://doi.org/10.1080/02664763.2018.1450367>

 View supplementary material 

 Published online: 21 Mar 2018.

 Submit your article to this journal 

 Article views: 88

 View Crossmark data 



A model averaging approach for the ordered probit and nested logit models with applications

Longmei Chen^a, Alan T. K. Wan^a, Geoffrey Tso^a and Xinyu Zhang^b

^aDepartment of Management Sciences, City University of Hong Kong, Kowloon, Hong Kong; ^bAcademy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, People's Republic of China

ABSTRACT

This paper considers model averaging for the ordered probit and nested logit models, which are widely used in empirical research. Within the frameworks of these models, we examine a range of model averaging methods, including the jackknife method, which is proved to have an optimal asymptotic property in this paper. We conduct a large-scale simulation study to examine the behaviour of these model averaging estimators in finite samples, and draw comparisons with model selection estimators. Our results show that while neither averaging nor selection is a consistently better strategy, model selection results in the poorest estimates far more frequently than averaging, and more often than not, averaging yields superior estimates. Among the averaging methods considered, the one based on a smoothed version of the Bayesian Information criterion frequently produces the most accurate estimates. In three real data applications, we demonstrate the usefulness of model averaging in mitigating problems associated with the 'replication crisis' that commonly arises with model selection.

ARTICLE HISTORY

Received 23 September 2017
Accepted 4 March 2018


KEYWORDS

Hit rate; model averaging;
model selection; Monte
Carlo; nested logit; ordered
probit; screening

1. Introduction

Model selection lies at the heart of statistical modelling. It is of particular relevance to business, economics and social science research where the relevant theories often do not provide researchers with an apparatus that can guide them in formulating their statistical models with a high degree of certainty. For many researchers, the most common practice is to try many models, each containing a different combination of regressors, and eventually select the best of all models considered and report results based on this single 'champion' model. While expert judgment often helps establish the initial conceptual model and translate it into a quantitative model, the subsequent model selection is usually an automated data-driven exercise with minimal human intervention. Typically, model selection entails the cumbersome step of significance 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. However, due to sampling fluctuations it is highly unlikely for the best fitting model

CONTACT Alan T. K. Wan  msawan@cityu.edu.hk

 Supplemental data for this article can be accessed at <https://doi.org/10.1080/02664763.2018.1450367>

in one sample to be the preferred model across all samples. This points towards a ‘*replication crisis*’, a term used by scientists to refer to the inability to replicate results (and hence generalise findings) obtained in a single sample across situations.¹ The sensitivity of model selection techniques, even to slight variations in data [44], also contributes to the difficulty with replicating results. Moreover, when model selection is done in a single sample, subsequent inferences typically do not account for the fact that model selection is a random event; rather, inferences are contingent on the best fitting model, assuming that it is known in advance without accounting for the aforementioned uncertainty. This makes the analysis potentially vulnerable to over-confident inferences [9,14,27].

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.² 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 or supported by the data. 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. By pooling the findings from different models, model averaging can also circumvent the replication crisis. As has been noted, replication failures arise partly as a result of model selection uncertainty. Yuan and Yang [44] advocated the use of model selection diagnostics to examine if selection is stable across samples. Their numerical results suggest that model selection can often lead to very unstable results; on the other hand, averaging can substantially reduce this instability. Liu and Yang [30] concluded the same when interest focuses on panel data models.

Bayesian model averaging (BMA) has been widely applied in many disciplines. Although BMA provides a formal approach for incorporating prior knowledge, its estimates can be sensitive to the choice of the prior. 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. A common approach is to construct weights based on information criterion values obtained from the different models. This is the approach taken by Buckland *et al.* [3] and Hjort and Claeskens [21]. Other weight selection approaches that have been proposed include those based on the Mallows’ criterion [16,41], minimisation of MSE [28,40], cross-validation [2,18], and minimisation of Kullback–Leibler distance [48]. Some studies have shown that simple equally weighted average estimator can sometimes perform well [40]. Others have shown that screening out the very poor models prior to combining can often yield superior estimates [44]. FMA strategies have also been considered when data are missing from the sample [35,36]. The state of art of this rapidly expanding field is summarised in [5,8,33]. FMA has been successfully applied in many disciplines including biomedical sciences [37], climatology [11], ecology [25], health economics [23], growth economics [1], and tourism research [39]. Despite these advances, FMA has not come into usage in many other disciplines of social sciences such as political science and sociology, although arguments in favour of combining models over selecting a single model in sociological research have been put forward by Burnham and Anderson [4].

In this paper, we explore the promising approach of FMA within the context of two discrete choice models, the ordered probit and nested logit models, which are widely applied

in social science research. The ordered probit model is frequently preferred over its logit counterpart for ordinal data that contain many of the two extreme outcome values because the probit is tied to the normal distribution that has thicker tails than the logistic distribution. Well-known applications of the ordered probit model in social science research include [31], where the impacts of social background on educational attainment were investigated, and [15], who examined nurses' perception of the spiritual nature of their profession. The nested logit model has an advantage over the multinomial logit model in that it reconciles problems associated with the often unrealistic assumption of independence of irrelevant alternatives (IIA) embedded in the latter model. The nested logit model is the analytical platform of a seminal paper in sociology on the influence of personality on teenage premarital pregnancy by Plotnick [34]. The much-cited article by Knapp *et al.* [26] on household mobility in the US is also based on a nested logit model. Although FMA methods for the ordered probit and nested logit models are heretofore unexplored, there have been related studies of FMA for the simpler types of logit models. Claeskens *et al.* [6] considered the binary logit model and constructed model weights based on the scores of an information criterion. Wan *et al.* [40] 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.

We consider a wide range of FMA strategies including equal weighting (EW), weights based on a variety of information criterion scores, weights based on jackknife or leave-one-out cross validation, and weights based on two minimisation schemes related to the mean squared error of the FMA estimator. Our work is the first study that examines jackknife model averaging outside the framework of the linear model, and we prove that the jackknife estimator results in an optimal property within the context of the two model frameworks being examined. 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 [44]. Using data from the US General Social Survey of 2008, the AsiaBarometer survey of 2007, and a brand choice study, we demonstrate the difficulty of relying on a single best fitting model without acknowledging the anticipated replication problem, and show that model averaging can help mitigate this problem.

Our presentation proceeds as follows. In the next section, we describe the ordered probit and nested logit models. Section 3 discusses the various FMA model strategies. In Section 4, we conduct a comprehensive Monte Carlo study to compare the performance of these FMA methods with several traditional model selection methods. Section 5 contains applications to three large datasets, illustrates the replication crisis and further demonstrates the advantages of model averaging over selection. We offer our conclusions in Section 6. Proofs of theoretical results are contained in the appendix.

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 [20]. 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 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:

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

$$P(Y_i \leq J) = 1,$$

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 $\hat{\alpha}_1^{(s)}, \dots, \hat{\alpha}_{J-1}^{(s)}, \hat{\beta}^{(s)}$, 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 that 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

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

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 $(1, 2, \dots, J) = B_1 \cup B_2 \cup \dots \cup B_K$, 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

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

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 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 originated in the literature on econometrics and statistics. All methods have been shown to work well in other contexts. Several of these methods have been developed under the local misspecification framework (LMF). Readers are referred to Hjort and Claeskens [20] and Claeskens and Hjort [7] for a detailed description of this framework.

The FMA weight choice schemes we considered are as follows:³

- 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.* [3], and subsequently used in a number of FMA studies. Buckland *et al.* [3] justified this weighting scheme by noting that for the S-AIC, the ratio in Equation (8) is the relative penalised likelihood factor, and for the S-BIC, it is Schwarz's (1978) approximation to the Bayes factor. Hence the S-BIC is a BMA strategy. Hansen [16] also considered the S-BIC a simplified form of BMA. To the best of our knowledge, the S-AIC and S-BIC weights have not been proven to be asymptotically optimal, possibly because these weights are developed as variants of the AIC and BIC, but not as an optimal solution to any criterion.

- Smoothed-FIC (S-FIC) weights. The FIC, developed by Hjort and Claeskens [20] is a model selection criterion tailored to the parameter singled out for interest. Let μ be the parameter of interest. For the ordered Probit and nested Logit Models, μ can be one of p_{01}, \dots, p_{0j} . Hjort and Claeskens [20] showed that under the LMF, the FIC for minimising the MSE of the estimator of μ in the s th sub-model is

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

where $\hat{\omega}$ and $\hat{\delta}$ are the ML estimators of ω and δ (defined in the appendix) using the full model, $\hat{\mathcal{K}} \equiv (\mathcal{J}_{n,11} - \mathcal{J}_{n,10} \mathcal{J}_{n,00}^{-1} \mathcal{J}_{n,01})^{-1}$ 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 Equation (8) when using FIC_{MSE}^s as xIC_s , Hjort and Claeskens [20] suggested assigning the weight to the s th model by

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

where ϱ is an algorithmic parameter that bridges from the uniform weighting (ϱ near 0) to ‘hard’ FIC. Hjort and Claeskens [20] demonstrated that Equation (10) has an empirical Bayes justification. Following Zhang *et al.* [46], we set $\varrho = 1$. The S-FIC method has been used in various FMA studies including [6,45,46]. There is no result in the literature that shows the S-FIC-based FMA estimator leads to any asymptotic optimal property.

- Weight based on minimising the trace of an unbiased estimator of the FMA estimator’s mean square error [28] (LZWZ). Again, this has been developed under the LMF described in the appendix. Liang *et al.* [28] derived an unbiased estimator of the trace of the MSE of the FMA estimator, and suggested choosing $w = (w_1, \dots, w_{2^q})$ by minimising

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

subject to the constraints $0 \leq w_s \leq 1$ and $\sum_{s=1}^{2^q} w_s = 1$. This can be readily performed using routines in STATA, GAUSS or Matlab. The Appendix gives a description of the derivation of Equation (11) based on the LMF and explanations of the notations in Equation (11). Zhang *et al.* [49] proved that the LZWZ method minimises the asymptotic expected squared error of the resultant FMA estimator.

- Weight based on minimising a plug-in estimator of the asymptotic squared error risk of the FMA estimator [40] (A-opt). A-opt method chooses the weight vector $w = (w_1, \dots, w_{2^q})$ by minimising

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

subject to the constraints $0 \leq w_s \leq 1$ and $\sum_{s=1}^{2^q} w_s = 1$. Again, this is a standard minimisation problem that can be readily solved using STATA or other software.

A description of the derivation of Equation (12) based on the LMF is given in the Appendix. Wan *et al.* [40] showed that A-opt is asymptotically optimal in the sense that their weights converges to the infeasible optimal weights that minimise the asymptotic expected squared estimation error of the estimator. Wan *et al.* [40] considered the plug-in estimator of the asymptotic squared error risk because the plug-in estimator is a consistent estimator of the asymptotic squared error risk. Thus, by minimising the plug-in estimator, the resulting weights converge to the optimal weights.

- Weight based on Jackknife or leave-one-out cross validation criterion [18,47] (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 nonlinear. 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:

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

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 $Y_i = j$, while the best forecast of p_{ij} is 0 when $Y_i \neq j$. Hence we define the forecast error associated with the model average as $\sum_{s=1}^{2^q} w_s^{(-i)} \hat{p}_{ij}^{(s)} - \mathbb{I}(Y_i = j)$. 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 minimises $CV(w)$. Denote the weights minimising $CV(w)$ in Equation (13) as $\hat{w} = (\hat{w}_1, \dots, \hat{w}_{2^q})$. Let $\mathcal{W} = \{w : 0 \leq w_s \leq 1, \sum_{s=1}^{2^q} w_s = 1\}$ be the weight set. In the appendix, we prove that under some regular conditions,

$$\frac{\sum_{i=1}^n \sum_{j=1}^J \left(\sum_{s=1}^{2^q} \hat{w}_s \hat{p}_{ij}^{(s)} - p_{ij} \right)^2}{\inf_{w \in \mathcal{W}} \sum_{i=1}^n \sum_{j=1}^J \left(\sum_{s=1}^{2^q} w_s \hat{p}_{ij}^{(s)} - p_{ij} \right)^2} \rightarrow 1 \tag{14}$$

in probability, as $n \rightarrow \infty$. Equation (14) implies that the JMA estimator $\sum_{s=1}^{2^q} \hat{w}_s \hat{p}_{ij}^{(s)}$ has a squared error that is asymptotically identical to that of the infeasible best possible model average estimator. The JMA estimator is thus optimal.

The steps of the JMA strategy are as follows:

Step 1: Calculate $^{(-i)}\hat{p}_{ij}^{(s)}$, $i = 1, \dots, n, j = 1, \dots, J, s = 1, \dots, 2^q$ 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 $w = (w_1, \dots, w_{2^q})$ that minimises $CV(w)$ in Equation (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 Equation (7) to obtain the model averaged predictions \hat{p}_{0j}^w of p_{0j} , $j = 1, \dots, J$. The forecast of Y_0 is $\hat{Y}_0 = c$, the category that corresponds to $\hat{p}_{0c}^w = \max\{\hat{p}_{01}^w, \dots, \hat{p}_{0J}^w\}$.

- EW, by which $w_s = 1/2^q$.

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 [44] 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. An alternative method is backward elimination discussed in [6,46]. 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 simulation designs:

Design 1: The data are generated based on the ordered probit model in Equation (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 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 simulation design is based on the nested logit model in Equations (4) and (5). We let $\tau = l \times 0.25$, and set all other parameters to the same values as in the previous design.

Let n_1 and n_2 be the number of observations in the training and test samples. We consider the following combinations of (n_1, n_2) : (300, 100), (500, 200). Each part of our

simulation is based on 500 replications. We evaluate the performance of the various strategies 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:

$$\text{MSEF} = \frac{1}{500 \times n_2} \sum_{r=1}^{500} \sum_{i=1}^{n_2} \sum_{j=1}^J (p_{ij,r} - \hat{p}_{ij,r})^2, \tag{15}$$

$$\text{MAEF} = \frac{1}{500 \times n_2} \sum_{r=1}^{500} \sum_{i=1}^{n_2} \sum_{j=1}^J |p_{ij,r} - \hat{p}_{ij,r}|, \tag{16}$$

and

$$\text{HitRate} = \frac{1}{500 \times n_2} \sum_{r=1}^{500} \sum_{i=1}^{n_2} \mathbb{I}(Y_{i,r} = \hat{Y}_{i,r}), \tag{17}$$

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 $\hat{p}_{ic,r} = \max\{\hat{p}_{i1,r}, \dots, \hat{p}_{ij,r}\}$, $Y_{i,r}$ is the actual category of Y_i in the r th replication, and $\mathbb{I}(\cdot)$ is an indicator function that equals unity if the event inside the bracket occurs, and otherwise equals zero. Thus, $\mathbb{I}(Y_{i,r} = \hat{Y}_{i,r}) = 1$ if $\hat{Y}_{i,r}$ correctly predicts $Y_{i,r}$, and $\mathbb{I}(Y_{i,r} = \hat{Y}_{i,r}) = 0$ otherwise. The hit rate is therefore the percentage of observations in the test sample that is correctly predicted by the model.

The results of the Monte Carlo study comparing the efficiency of various estimators are reported in Tables 1–6. 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 in terms of the performance yardsticks considered no one strategy uniformly dominates any of the others. Neither model selection nor model averaging is always a better strategy than the other. That being said, for $(n_1, n_2) = (300, 100), (500, 200)$, 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 FMA methods. Remarkably, the screened versions of all FMA methods are rarely among the three worst strategies. On the other hand, although selection can sometimes outperform averaging, it also delivers very poor estimates far more often than its averaging counterparts, especially when n_1 and n_2 are small. For example, in terms of the frequency in delivering the best estimates, the BIC selection is sometimes rated among the top three of all methods, but when $(n_1, n_2) = (300, 100)$, it also delivers one of the three worst estimates thrice with respect to MSEF, MAEF and hit rate. When $(n_1, n_2) = (500, 200)$, it again yields one of the three least accurate estimates thrice with respect to MSEF and MAEF, and once with respect to hit rate. On the other hand, when $(n_1, n_2) = (300, 100)$, the screened versions of the S-AIC, S-BIC, A-opt and JMA methods never result in the three worst estimates regardless of the performance

Table 1. MSEF results with $(n_1, n_2) = (300, 100)$.

l	S	Model selection			Model averaging without screening						
		AIC	BIC	FIC	S-AIC	S-BIC	S-FIC	LZWZ	A-opt	JMA	EW
0.5	1	0.00884	0.00987 ⁽⁻²⁾	0.00915 ⁽⁻³⁾	0.00795 ⁽¹⁾	0.00858	0.00830	0.00838	0.00834	0.00796 ⁽²⁾	0.01454 ⁽⁻¹⁾
		0.00835	0.00938 ⁽⁻²⁾	0.00905 ⁽⁻³⁾	0.00769 ⁽³⁾	0.00820	0.00822	0.00834	0.00831	0.00767 ⁽²⁾	0.01447 ⁽⁻¹⁾
		0.00820 ⁽⁻³⁾	0.00862 ⁽⁻¹⁾	0.00828 ⁽⁻²⁾	0.00715	0.00712	0.00704	0.00757	0.00791	0.00691 ⁽²⁾	0.00652 ⁽¹⁾
	2	0.00788	0.00786	0.01034 ⁽⁻²⁾	0.00725 ⁽³⁾	0.00738	0.00933	0.00974 ⁽⁻³⁾	0.00894	0.00766	0.03080 ⁽⁻¹⁾
		0.00688	0.00610 ⁽¹⁾	0.00915 ⁽⁻²⁾	0.00650	0.00617 ⁽³⁾	0.00909 ⁽⁻³⁾	0.00903	0.00857	0.00684	0.03051 ⁽⁻¹⁾
		0.00710	0.00576 ⁽³⁾	0.01172 ⁽⁻¹⁾	0.00647	0.00562 ⁽²⁾	0.00961	0.01114 ⁽⁻²⁾	0.00886	0.00665	0.01094 ⁽⁻³⁾
1	1	0.00710	0.00713	0.00779 ⁽⁻²⁾	0.00647 ⁽²⁾	0.00678	0.00719	0.00751 ⁽⁻³⁾	0.00734	0.00674	0.03304 ⁽⁻¹⁾
		0.00612	0.00578	0.00728 ⁽⁻³⁾	0.00577	0.00566 ⁽²⁾	0.00689	0.00728 ⁽⁻²⁾	0.00718	0.00590	0.03303 ⁽⁻¹⁾
		0.00663	0.00517 ⁽¹⁾	0.00747 ⁽⁻³⁾	0.00601	0.00539 ⁽³⁾	0.00699	0.00723	0.00749 ⁽⁻²⁾	0.00597	0.01044 ⁽⁻¹⁾
	2	0.00857	0.00869	0.01188 ⁽⁻²⁾	0.00778	0.00785	0.00960	0.01098 ⁽⁻³⁾	0.01056	0.00805	0.04649 ⁽⁻¹⁾
		0.00704	0.00609 ⁽³⁾	0.01044 ⁽⁻²⁾	0.00663	0.00606 ⁽²⁾	0.00884	0.01019 ⁽⁻³⁾	0.01001	0.00680	0.04574 ⁽⁻¹⁾
		0.00710	0.00560 ⁽³⁾	0.00868 ⁽⁻²⁾	0.00644	0.00554 ⁽²⁾	0.00831	0.00856 ⁽⁻³⁾	0.00814	0.00652	0.01457 ⁽⁻¹⁾
2	1	0.00726	0.00837	0.00880	0.00661 ⁽³⁾	0.00721	0.00748	0.00931 ⁽⁻²⁾	0.00896 ⁽⁻³⁾	0.00687	0.07015 ⁽⁻¹⁾
		0.00567	0.00477 ⁽³⁾	0.00743	0.00525	0.00468 ⁽²⁾	0.00661	0.00869 ⁽⁻²⁾	0.00858 ⁽⁻³⁾	0.00536	0.07029 ⁽⁻¹⁾
		0.00598	0.00452 ⁽³⁾	0.00636	0.00533	0.00444 ⁽²⁾	0.00582	0.00655 ⁽⁻³⁾	0.00674 ⁽⁻²⁾	0.00537	0.02116 ⁽⁻¹⁾
	2	0.00855	0.00944	0.01414 ⁽⁻²⁾	0.00779 ⁽³⁾	0.00818	0.01087	0.01335 ⁽⁻³⁾	0.01266	0.00805	0.06660 ⁽⁻¹⁾
		0.00685	0.00584 ⁽³⁾	0.01178	0.00638	0.00573 ⁽²⁾	0.00966	0.01216 ⁽⁻²⁾	0.01185 ⁽⁻³⁾	0.00657	0.06587 ⁽⁻¹⁾
		0.00697	0.00530 ⁽³⁾	0.00784 ⁽⁻²⁾	0.00623	0.00527 ⁽²⁾	0.00725	0.00770	0.00781 ⁽⁻³⁾	0.00629	0.01969 ⁽⁻¹⁾
(1)	0	2	0	1	0	0	0	0	0	1	
(2)	0	0	0	1	8	0	0	0	3	0	
(3)	0	7	0	4	2	0	0	0	0	0	
(-3)	1	0	4	0	0	1	7	4	0	1	
(-2)	0	2	9	0	0	0	5	2	0	0	
(-1)	0	1	1	0	0	0	0	0	0	16	

(continued).

Table 1. Continued.

I	S	Model averaging with screening							
		S-AIC	S-BIC	S-FIC	LZWZ	A-opt	JMA	EW	
0.5	1	0.00806	0.00860	0.00798↑	Design 1 0.00810↑	0.00804↑	0.00797 ⁽³⁾	0.00805↑	
		0.00773	0.00821	0.00780↑	0.00803↑	0.00797↑	0.00766 ⁽¹⁾ ↑	0.00786↑	
		0.00725	0.00725	0.00725	0.00743↑	0.00751↑	0.00692 ⁽³⁾	0.00697	
	2	0.00725	0.00737↑	0.00717 ⁽¹⁾ ↑	Design 2 0.00848↑	0.00805↑	0.00745↑	0.00718 ⁽²⁾ ↑	
		0.00648↑	0.00616 ⁽²⁾ ↑	0.00661↑	0.00768↑	0.00747↑	0.00663↑	0.00657↑	
		0.00634↑	0.00561 ⁽¹⁾ ↑	0.00599↑	0.00667↑	0.00688↑	0.00645↑	0.00592↑	
1	1	0.00647 ⁽¹⁾ ↑	0.00676↑	0.00652 ⁽³⁾ ↑	Design 1 0.00676↑	0.00675↑	0.00671↑	0.00678↑	
		0.00575 ⁽³⁾ ↑	0.00564 ⁽¹⁾ ↑	0.00608↑	0.00643↑	0.00651↑	0.00585↑	0.00623↑	
		0.00578↑	0.00533 ⁽²⁾ ↑	0.00585↑	0.00660↑	0.00690↑	0.00600	0.00592↑	
	2	0.00778 ⁽³⁾ ↑	0.00784↑	0.00757 ⁽¹⁾ ↑	Design 2 0.00895↑	0.00862↑	0.00793↑	0.00776 ⁽²⁾ ↑	
		0.00663↑	0.00606 ⁽¹⁾ ↑	0.00672↑	0.00797↑	0.00786↑	0.00668↑	0.00683↑	
		0.00629↑	0.00553 ⁽¹⁾ ↑	0.00595↑	0.00644↑	0.00683↑	0.00628↑	0.00578↑	
2	1	0.00661 ⁽²⁾ ↑	0.00721↑	0.00627 ⁽¹⁾ ↑	Design 1 0.00684↑	0.00666↑	0.00686↑	0.00770↑	
		0.00525↑	0.00468 ⁽¹⁾ ↑	0.00534↑	0.00606↑	0.00613↑	0.00551	0.00646↑	
		0.00519↑	0.00444 ⁽¹⁾ ↑	0.00499↑	0.00543↑	0.00584↑	0.00526↑	0.00466↑	
	2	0.00779 ⁽²⁾ ↑	0.00818↑	0.00763 ⁽¹⁾ ↑	Design 2 0.00944↑	0.00902↑	0.00797↑	0.00838↑	
		0.00638↑	0.00573 ⁽¹⁾ ↑	0.00652↑	0.00805↑	0.00794↑	0.00648↑	0.00722↑	
		0.00606↑	0.00526 ⁽¹⁾ ↑	0.00579↑	0.00629↑	0.00670↑	0.00611↑	0.00554↑	
		(1)	1	8	4	0	0	1	0
		(2)	2	2	0	0	0	0	2
		(3)	2	0	1	0	0	2	0
		(-3)	0	0	0	0	0	0	0
		(-2)	0	0	0	0	0	0	0
		(-1)	0	0	0	0	0	0	0

Note : (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 results with $(n_1, n_2) = (300, 100)$.

I	S	Model selection			Model averaging without screening						
		AIC	BIC	FIC	S-AIC	S-BIC	S-FIC	LZWZ	A-opt	JMA	EW
0.5	1	0.11749	0.12395 ⁽⁻²⁾	0.11952 ⁽⁻³⁾	0.11217	0.11604	0.11432	0.11486	0.11481	0.11200 ⁽¹⁾	0.15127 ⁽⁻¹⁾
		0.11325	0.12047 ⁽⁻²⁾	0.11827 ⁽⁻³⁾	0.10986 ⁽³⁾	0.11296	0.11333	0.11412	0.11416	0.10917 ⁽²⁾	0.15062 ⁽⁻¹⁾
		0.11142 ⁽⁻³⁾	0.11458 ⁽⁻¹⁾	0.11333 ⁽⁻²⁾	0.10561	0.10555	0.10524	0.10883	0.11131	0.10351 ⁽²⁾	0.10192 ⁽¹⁾
	2	0.11219	0.11162	0.12516 ⁽⁻²⁾	0.10748	0.10813	0.12338 ⁽⁻³⁾	0.12291	0.11953	0.11165	0.24541 ⁽⁻¹⁾
		0.10419	0.09780 ⁽¹⁾	0.11712	0.10170	0.09881 ⁽³⁾	0.12135 ⁽⁻²⁾	0.11934 ⁽⁻³⁾	0.11755	0.10535	0.24406 ⁽⁻¹⁾
		0.10990	0.09916 ⁽³⁾	0.13539 ⁽⁻²⁾	0.10539	0.09849 ⁽²⁾	0.12954	0.13430 ⁽⁻³⁾	0.12349	0.10726	0.14007 ⁽⁻¹⁾
1	1	0.10925	0.10842	0.11302 ⁽⁻²⁾	0.10416 ⁽²⁾	0.10601	0.10935	0.11166 ⁽⁻³⁾	0.11106	0.10635	0.24306 ⁽⁻¹⁾
		0.10070	0.09599 ⁽¹⁾	0.10826	0.09832	0.09643 ⁽³⁾	0.10693	0.10973 ⁽⁻²⁾	0.10971 ⁽⁻³⁾	0.09944	0.24243 ⁽⁻¹⁾
		0.10504	0.09216 ⁽¹⁾	0.11020 ⁽⁻³⁾	0.10084	0.09471 ⁽³⁾	0.10814	0.10960	0.11195 ⁽⁻²⁾	0.10018	0.13191 ⁽⁻¹⁾
	2	0.11523	0.11656	0.13159 ⁽⁻²⁾	0.10983 ⁽³⁾	0.11050	0.12178	0.12816 ⁽⁻³⁾	0.12606	0.11279	0.30754 ⁽⁻¹⁾
		0.10408	0.09697 ⁽¹⁾	0.12165	0.10142	0.09710 ⁽³⁾	0.11682	0.12313 ⁽⁻²⁾	0.12270 ⁽⁻³⁾	0.10352	0.30473 ⁽⁻¹⁾
		0.10965	0.09779 ⁽³⁾	0.11825	0.10505	0.09758 ⁽²⁾	0.11940 ⁽⁻³⁾	0.11947 ⁽⁻²⁾	0.11830	0.10592	0.16295 ⁽⁻¹⁾
2	1	0.09971	0.10848	0.10678	0.09518 ⁽³⁾	0.10034	0.10046	0.11058 ⁽⁻²⁾	0.10961 ⁽⁻³⁾	0.09867	0.38980 ⁽⁻¹⁾
		0.08802	0.08115 ⁽³⁾	0.09720	0.08520	0.08072 ⁽²⁾	0.09467	0.10671 ⁽⁻³⁾	0.10711 ⁽⁻²⁾	0.08801	0.39046 ⁽⁻¹⁾
		0.09961	0.08699 ⁽³⁾	0.10161	0.09471	0.08674 ⁽²⁾	0.09865	0.10429 ⁽⁻³⁾	0.10633 ⁽⁻²⁾	0.09507	0.19440 ⁽⁻¹⁾
	2	0.10828	0.11426	0.13141 ⁽⁻²⁾	0.10335 ⁽³⁾	0.10624	0.12043	0.12980 ⁽⁻³⁾	0.12708	0.10619	0.37255 ⁽⁻¹⁾
		0.09641	0.08953 ⁽³⁾	0.11801	0.09372	0.08908 ⁽²⁾	0.11338	0.12310 ⁽⁻²⁾	0.12258 ⁽⁻³⁾	0.09584	0.36996 ⁽⁻¹⁾
		0.10518	0.09208 ⁽¹⁾	0.11031	0.10024	0.09227 ⁽³⁾	0.10845	0.11116 ⁽⁻³⁾	0.11248 ⁽⁻²⁾	0.10092	0.18797 ⁽⁻¹⁾
(1)	0	5	0	0	0	0	0	0	1	1	
(2)	0	0	0	1	5	0	0	0	2	0	
(3)	0	5	0	4	5	0	0	0	0	0	
(-3)	1	0	3	0	0	2	8	4	0	0	
(-2)	0	2	6	0	0	1	5	4	0	0	
(-1)	0	1	0	0	0	0	0	0	0	17	

(continued).

Table 2. Continued.

I	S	Model averaging with screening							
		S-AIC	S-BIC	S-FIC	LZWZ	A-opt	JMA	EW	
0.5	1	0.11278	0.11617	0.11219↑	Design 1 0.11303↑	0.11277↑	0.11207 ⁽²⁾	0.11216 ⁽³⁾ ↑	
		0.10996	0.11301	0.11038↑	0.11208↑	0.11184↑	0.10910 ⁽¹⁾ ↑	0.11048↑	
		0.10626	0.10636	0.10639	0.10786↑	0.10857↑	0.10355 ⁽³⁾	0.10441	
	2	0.10745 ⁽³⁾ ↑	0.10805↑	0.10685 ⁽¹⁾ ↑	Design 2 0.11470↑	0.11251↑	0.10899↑	0.10732 ⁽²⁾ ↑	
		0.10158↑	0.09872 ⁽²⁾ ↑	0.10257↑	0.10911↑	0.10831↑	0.10263↑	0.10257↑	
		0.10433↑	0.09835 ⁽¹⁾ ↑	0.10170↑	0.10536↑	0.10799↑	0.10535↑	0.10127↑	
1	1	0.10412 ⁽¹⁾ ↑	0.10588↑	0.10436 ⁽³⁾ ↑	Design 1 0.10631↑	0.10622↑	0.10597↑	0.10580↑	
		0.09817↑	0.09631 ⁽²⁾ ↑	0.10059↑	0.10341↑	0.10430↑	0.09894↑	0.10130↑	
		0.09884↑	0.09421 ⁽²⁾ ↑	0.09926↑	0.10498↑	0.10773↑	0.10076	0.09987↑	
	2	0.10982 ⁽²⁾ ↑	0.11046↑	0.10825 ⁽¹⁾ ↑	Design 2 0.11640↑	0.11457↑	0.11104↑	0.11083↑	
		0.10139↑	0.09706 ⁽²⁾ ↑	0.10221↑	0.10949↑	0.10932↑	0.10177↑	0.10388↑	
		0.10381↑	0.09750 ⁽¹⁾ ↑	0.10124↑	0.10475↑	0.10810↑	0.10373↑	0.09998↑	
2	1	0.09518 ⁽²⁾ ↑	0.10034↑	0.09275 ⁽¹⁾ ↑	Design 1 0.09685↑	0.09580↑	0.09720↑	0.10665↑	
		0.08520↑	0.08072 ⁽¹⁾ ↑	0.08575↑	0.09131↑	0.09213↑	0.08642↑	0.09775↑	
		0.09344↑	0.08669 ⁽¹⁾ ↑	0.09168↑	0.09534↑	0.09905↑	0.09413↑	0.08915↑	
	2	0.10335 ⁽²⁾ ↑	0.10624↑	0.10216 ⁽¹⁾ ↑	Design 2 0.11128↑	0.10933↑	0.10464↑	0.10924↑	
		0.09372↑	0.08908 ⁽¹⁾ ↑	0.09465↑	0.10249↑	0.10256↑	0.09426↑	0.10129↑	
		0.09885↑	0.09220 ⁽²⁾ ↑	0.09685↑	0.10044↑	0.10395↑	0.09925↑	0.09502↑	
		(1)	1	5	4	0	0	1	0
		(2)	3	5	0	0	0	1	1
		(3)	1	0	1	0	0	1	1
		(-3)	0	0	0	0	0	0	0
		(-2)	0	0	0	0	0	0	0
		(-1)	0	0	0	0	0	0	0

Note : (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 results with $(n_1, n_2) = (300, 100)$.

<i>l</i>	<i>S</i>	Model selection			Model averaging without screening							
		AIC	BIC	FIC	S-AIC	S-BIC	S-FIC	LZWZ	A-opt	JMA	EW	
0.5	1	0.55466	0.55238 ⁽⁻²⁾	0.55306 ⁽⁻³⁾	0.55522	0.55350	Design 1					
		0.55206	0.55054	0.54960 ⁽⁻²⁾	0.55282 ⁽¹⁾	0.55176	0.55368	0.55322	0.55454	0.55528 ⁽³⁾	0.54616 ⁽⁻¹⁾	
		0.48450	0.47836 ⁽⁻¹⁾	0.48398	0.48496 ⁽²⁾	0.48262	0.55044 ⁽⁻³⁾	0.55070	0.55162	0.55166	0.54352 ⁽⁻¹⁾	
	2	0.55150	0.55248	0.55008 ⁽⁻³⁾	0.55250 ⁽³⁾	0.55288 ⁽²⁾	0.55132	Design 2				
		0.55192	0.55250 ⁽²⁾	0.55108	0.55164	0.55210	0.54970 ⁽⁻²⁾	0.54976 ⁽⁻²⁾	0.55022	0.55170	0.53750 ⁽⁻¹⁾	
		0.47990	0.48104 ⁽³⁾	0.47192 ⁽⁻¹⁾	0.47996	0.48154 ⁽²⁾	0.47694	0.54954 ⁽⁻³⁾	0.55002	0.55098	0.53540 ⁽⁻¹⁾	
1	1	0.61230	0.61148	0.61124	0.61246 ⁽¹⁾	0.61230	Design 1					
		0.60726	0.60736	0.60562	0.60766 ⁽²⁾	0.60768 ⁽¹⁾	0.61120	0.61082 ⁽⁻³⁾	0.61134	0.61212	0.60090 ⁽⁻¹⁾	
		0.49184	0.49396 ⁽³⁾	0.49106	0.49200	0.49428 ⁽¹⁾	0.60536 ⁽⁻²⁾	0.60536 ⁽⁻³⁾	0.60576	0.60756	0.59498 ⁽⁻¹⁾	
	2	0.63002	0.63030	0.62896	0.63066 ⁽³⁾	0.63030	0.62946	Design 2				
		0.63252	0.63366 ⁽¹⁾	0.63110	0.63256	0.63318 ⁽²⁾	0.63074	0.62802 ⁽⁻²⁾	0.62852 ⁽⁻³⁾	0.63002	0.61418 ⁽⁻¹⁾	
		0.53108 ⁽⁻³⁾	0.53278	0.53116	0.53250	0.53294 ⁽³⁾	0.53126	0.63014 ⁽⁻³⁾	0.63010 ⁽⁻²⁾	0.63204	0.61422 ⁽⁻¹⁾	
2	1	0.69274	0.69208 ⁽⁻²⁾	0.69244 ⁽⁻³⁾	0.69324	0.69322	Design 1					
		0.68898 ⁽³⁾	0.68966 ⁽¹⁾	0.68748	0.68862	0.68934 ⁽²⁾	0.68686	0.69286	0.69266	0.69250	0.69316	0.66842 ⁽⁻¹⁾
		0.54300 ⁽⁻²⁾	0.54444	0.54392	0.54418	0.54552 ⁽²⁾	0.54384	0.68594 ⁽⁻³⁾	0.68604 ⁽⁻²⁾	0.68892	0.66450 ⁽⁻¹⁾	
	2	0.71668	0.71626	0.71564 ⁽⁻²⁾	0.71810 ⁽¹⁾	0.71790 ⁽²⁾	0.71752	Design 2				
		0.72024 ⁽¹⁾	0.71996	0.71750 ⁽⁻³⁾	0.72000 ⁽³⁾	0.71994	0.71894	0.71662	0.71724	0.71776	0.69564 ⁽⁻¹⁾	
		0.61800	0.62020 ⁽¹⁾	0.61720 ⁽⁻²⁾	0.61866	0.61966 ⁽³⁾	0.61786	0.71724 ⁽⁻²⁾	0.71762	0.71962	0.69722 ⁽⁻¹⁾	
(1)	1	3	0	3	2	0	0	0	0	0		
(2)	0	1	0	3	5	0	0	0	0	0		
(3)	0	2	0	2	3	0	0	1	1	0		
(-3)	1	0	4	0	0	3	4	3	0	1		
(-2)	1	2	3	0	0	0	8	2	0	0		
(-1)	0	1	1	0	0	0	0	0	0	16		

(continued).

Table 3. Continued.

I	S	Model averaging with screening							
		S-AIC	S-BIC	S-FIC	LZWZ	A-opt	JMA	EW	
0.5	1	Design 1							
		0.55514	0.55330	0.55432↑	0.55458↑	0.55556 ⁽²⁾ ↑	0.55586 ⁽¹⁾ ↑	0.55416↑	
		0.55240 ⁽²⁾	0.55190↑	0.55238 ⁽³⁾ ↑	0.55136↑	0.55184↑	0.55204↑	0.55184↑	
	0.48318	0.48182	0.48164 ⁽⁻³⁾	0.48310	0.48546 ⁽¹⁾ ↑	0.48320↑	0.48112 ⁽⁻²⁾		
	2	Design 2							
		0.55250	0.55294 ⁽¹⁾ ↑	0.55180↑	0.55130↑	0.55202↑	0.55174↑	0.55166↑	
0.55164		0.55200	0.55230↑	0.55244 ⁽³⁾ ↑	0.55270 ⁽¹⁾ ↑	0.55138↑	0.55178↑		
0.48016↑	0.48158 ⁽¹⁾ ↑	0.48010↑	0.48026↑	0.47964↑	0.48072↑	0.48040↑			
1	1	Design 1							
		0.61244 ⁽²⁾	0.61236↑	0.61206↑	0.61164↑	0.61152↑	0.61176 ⁽³⁾	0.61066 ⁽⁻²⁾ ↑	
		0.60766 ⁽³⁾	0.60760	0.60726↑	0.60666↑	0.60658↑	0.60730	0.60654↑	
	0.49288↑	0.49408 ⁽²⁾	0.49310↑	0.49232↑	0.49126↑	0.49202↑	0.49372↑		
	2	Design 2							
		0.63066 ⁽³⁾	0.63034↑	0.63132 ⁽²⁾ ↑	0.62992↑	0.63024↑	0.63000	0.63134 ⁽¹⁾ ↑	
0.63254		0.63314 ⁽³⁾	0.63196↑	0.63072↑	0.63104↑	0.63248↑	0.63190↑		
0.53244	0.53308 ⁽²⁾ ↑	0.53274↑	0.53282↑	0.53196↑	0.53242↑	0.53336 ⁽¹⁾ ↑			
2	1	Design 1							
		0.69324	0.69322	0.69454 ⁽¹⁾ ↑	0.69354 ⁽³⁾ ↑	0.69406 ⁽²⁾ ↑	0.69332↑	0.69286↑	
		0.68862	0.68934 ⁽²⁾	0.68822↑	0.68764↑	0.68764↑	0.68856	0.68790↑	
	0.54452↑	0.54566 ⁽¹⁾ ↑	0.54500 ⁽³⁾ ↑	0.54472↑	0.54386↑	0.54412↑	0.54468↑		
	2	Design 2							
		0.71810 ⁽¹⁾	0.71790 ⁽²⁾	0.71778 ⁽³⁾ ↑	0.71626 ⁽⁻³⁾	0.71724	0.71756	0.71720↑	
		0.72000 ⁽³⁾	0.71994	0.71946↑	0.71840↑	0.71820↑	0.71994 ⁽²⁾ ↑	0.71896↑	
		0.61872↑	0.61960	0.61920↑	0.61826↑	0.61780↑	0.61854↑	0.61978 ⁽²⁾ ↑	
		(1)	1	3	1	0	2	1	2
		(2)	3	3	1	0	2	0	1
		(3)	2	3	2	2	0	0	0
		(-3)	0	0	1	1	0	0	0
(-2)		0	0	0	0	0	0	2	
(-1)	0	0	0	0	0	0	0		

Note : (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. MSEF results with $(n_1, n_2) = (500, 200)$.

I	S	Model selection			Model averaging without screening								
		AIC	BIC	FIC	S-AIC	S-BIC	S-FIC	LZWZ	A-opt	JMA	EW		
0.5	1	0.00534	0.00655 ⁽⁻²⁾	0.00551	0.00491 ⁽²⁾	0.00571	0.00514	0.00512	0.00503	0.00501	0.01292 ⁽⁻¹⁾		
		0.00480	0.00601 ⁽⁻²⁾	0.00539 ⁽⁻³⁾	0.00468 ⁽³⁾	0.00533	0.00505	0.00509	0.00499	0.00467 ⁽²⁾	0.01285 ⁽⁻¹⁾		
		0.00482	0.00553 ⁽⁻¹⁾	0.00518 ⁽⁻²⁾	0.00448 ⁽³⁾	0.00480	0.00462	0.00478	0.00490 ⁽⁻³⁾	0.00432 ⁽¹⁾	0.00463		
	2	Design 2											
		0.00481	0.00499	0.00647 ⁽⁻²⁾	0.00440 ⁽³⁾	0.00454	0.00526	0.00586 ⁽⁻³⁾	0.00554	0.00466	0.02932 ⁽⁻¹⁾		
		0.00413	0.00358 ⁽³⁾	0.00575 ⁽⁻²⁾	0.00386	0.00354 ⁽²⁾	0.00499	0.00559 ⁽⁻³⁾	0.00541	0.00399	0.02896 ⁽⁻¹⁾		
1	1	Design 1											
		0.00449	0.00454	0.00478 ⁽⁻²⁾	0.00396 ⁽²⁾	0.00410	0.00435	0.00471 ⁽⁻³⁾	0.00455	0.00418	0.03142 ⁽⁻¹⁾		
		0.00364	0.00308 ⁽¹⁾	0.00428	0.00342	0.00314 ⁽³⁾	0.00409	0.00457 ⁽⁻²⁾	0.00447 ⁽⁻³⁾	0.00348	0.03140 ⁽⁻¹⁾		
	2	Design 2											
		0.00503	0.00559	0.00801 ⁽⁻²⁾	0.00457 ⁽³⁾	0.00493	0.00615	0.00749 ⁽⁻³⁾	0.00716	0.00475	0.04447 ⁽⁻¹⁾		
		0.00418	0.00356 ⁽³⁾	0.00680	0.00393	0.00353 ⁽²⁾	0.00565	0.00711 ⁽⁻²⁾	0.00702 ⁽⁻³⁾	0.00397	0.04410 ⁽⁻¹⁾		
	2	1	Design 1										
			0.00427	0.00559	0.00568	0.00403 ⁽³⁾	0.00477	0.00480	0.00673 ⁽⁻²⁾	0.00634 ⁽⁻³⁾	0.00414	0.06852 ⁽⁻¹⁾	
			0.00318	0.00262 ⁽³⁾	0.00440	0.00295	0.00262 ⁽²⁾	0.00383	0.00588 ⁽⁻²⁾	0.00585 ⁽⁻³⁾	0.00301	0.06856 ⁽⁻¹⁾	
		2	Design 2										
			0.00330	0.00251 ⁽³⁾	0.00356	0.00298	0.00247 ⁽²⁾	0.00322	0.00377 ⁽⁻³⁾	0.00381 ⁽⁻²⁾	0.00302	0.01939 ⁽⁻¹⁾	
			0.00528	0.00652	0.01046 ⁽⁻²⁾	0.00494 ⁽²⁾	0.00558	0.00774	0.01029 ⁽⁻³⁾	0.00978	0.00496	0.06523 ⁽⁻¹⁾	
(1)		0	2	0	0	0	0	0	0	1	0		
		0	0	0	3	8	0	0	0	1	0		
		0	8	0	5	2	0	0	0	0	0		
	0	0	3	0	0	0	8	6	0	0			
	0	2	7	0	0	0	5	4	0	0			
	0	1	0	0	0	0	0	0	0	17			

(continued).

Table 4. Continued.

I	S	Model averaging with screening							
		S-AIC	S-BIC	S-FIC	LZWZ	A-opt	JMA	EW	
0.5	1	0.00498	0.00573 ⁽⁻³⁾	0.00499↑	Design 1 0.00491 ⁽³⁾ ↑	0.00483 ⁽¹⁾ ↑	0.00494↑	0.00514↑	
		0.00469	0.00534	0.00484↑	0.00486↑	0.00476↑	0.00459 ⁽¹⁾ ↑	0.00493↑	
		0.00453	0.00483	0.00469	0.00473↑	0.00475↑	0.00437 ⁽²⁾	0.00471	
	2	0.00440 ⁽²⁾ ↑	0.00454↑	0.00429 ⁽¹⁾ ↑	Design 2 0.00518↑	0.00494↑	0.00454↑	0.00449↑	
		0.00386↑	0.00354 ⁽¹⁾ ↑	0.00391↑	0.00473↑	0.00461↑	0.00390↑	0.00401↑	
		0.00333↑	0.00292 ⁽¹⁾ ↑	0.00314↑	0.00334↑	0.00356↑	0.00339↑	0.00308↑	
1	1	0.00396 ⁽³⁾	0.00410↑	0.00388 ⁽¹⁾ ↑	Design 1 0.00411↑	0.00405↑	0.00415↑	0.00432↑	
		0.00342↑	0.00313 ⁽²⁾ ↑	0.00352↑	0.00386↑	0.00392↑	0.00346↑	0.00380↑	
		0.00342↑	0.00290 ⁽²⁾ ↑	0.00332↑	0.00368↑	0.00393↑	0.00347↑	0.00322↑	
	2	0.00457 ⁽²⁾ ↑	0.00493↑	0.00448 ⁽¹⁾ ↑	Design 2 0.00550↑	0.00524↑	0.00473↑	0.00497↑	
		0.00393↑	0.00353 ⁽¹⁾ ↑	0.00399↑	0.00485↑	0.00478↑	0.00395↑	0.00436↑	
		0.00348↑	0.00300 ⁽¹⁾ ↑	0.00328↑	0.00354↑	0.00377↑	0.00350↑	0.00315↑	
2	1	0.00403 ⁽²⁾ ↑	0.00477↑	0.00391 ⁽¹⁾ ↑	Design 1 0.00426↑	0.00411↑	0.00412↑	0.00582↑	
		0.00295↑	0.00262 ⁽¹⁾ ↑	0.00298↑	0.00351↑	0.00352↑	0.00299↑	0.00451↑	
		0.00290↑	0.00247 ⁽¹⁾ ↑	0.00279↑	0.00303↑	0.00326↑	0.00294↑	0.00263↑	
	(1)	2	0.00494 ⁽¹⁾ ↑	0.00558↑	0.00495↑	Design 2 0.00639↑	0.00603↑	0.00494 ⁽³⁾ ↑	0.00615↑
			0.00405↑	0.00362 ⁽¹⁾ ↑	0.00414↑	0.00525↑	0.00514↑	0.00407↑	0.00501↑
			0.00335↑	0.00287 ⁽¹⁾ ↑	0.00320↑	0.00347↑	0.00372↑	0.00338↑	0.00305↑
		1	1	8	4	0	1	1	0
		(2)	3	2	0	0	0	1	0
		(3)	1	0	0	1	0	1	0
		(-3)	0	1	0	0	0	0	0
		(-2)	0	0	0	0	0	0	0
		(-1)	0	0	0	0	0	0	0

Note : (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 5. MAEF results with $(n_1, n_2) = (500, 200)$.

I	S	Model selection			Model averaging without screening						
		AIC	BIC	FIC	S-AIC	S-BIC	S-FIC	LZWZ	A-opt	JMA	EW
0.5	Design 1										
	1	0.09062	0.09866 ⁽⁻²⁾	0.09210	0.08720 ⁽³⁾	0.09342	0.08927	0.08913	0.08869	0.08776	0.14131 ⁽⁻¹⁾
	2	0.08430 ⁽³⁾	0.09331 ⁽⁻²⁾	0.09018 ⁽⁻³⁾	0.08435	0.08931	0.08796	0.08822	0.08781	0.08367 ⁽²⁾	0.14048 ⁽⁻¹⁾
	3	0.08371	0.08918 ⁽⁻¹⁾	0.08777 ⁽⁻²⁾	0.08232 ⁽³⁾	0.08482	0.08427	0.08530	0.08672 ⁽⁻³⁾	0.08057 ⁽¹⁾	0.08504
	Design 2										
	1	0.08792	0.08970	0.09858 ⁽⁻²⁾	0.08421 ⁽³⁾	0.08566	0.09177	0.09526 ⁽⁻³⁾	0.09329	0.08739	0.23966 ⁽⁻¹⁾
2	0.08130	0.07629 ⁽³⁾	0.09214 ⁽⁻³⁾	0.07913	0.07605 ⁽²⁾	0.08925	0.09279 ⁽⁻²⁾	0.09212	0.08076	0.23798 ⁽⁻¹⁾	
3	0.08005	0.07067 ⁽¹⁾	0.09510	0.07663	0.07094 ⁽³⁾	0.09691 ⁽⁻³⁾	0.09758 ⁽⁻²⁾	0.09154	0.07750	0.12433 ⁽⁻¹⁾	
1	Design 1										
	1	0.08605	0.08712	0.08809 ⁽⁻²⁾	0.08085 ⁽³⁾	0.08242	0.08438	0.08747 ⁽⁻³⁾	0.08668	0.08308	0.23530 ⁽⁻¹⁾
	2	0.07696	0.07093 ⁽¹⁾	0.08244	0.07487	0.07148 ⁽³⁾	0.08163	0.08587 ⁽⁻²⁾	0.08583 ⁽⁻³⁾	0.07546	0.23497 ⁽⁻¹⁾
	3	0.08011	0.07024 ⁽¹⁾	0.08345	0.07715	0.07066 ⁽³⁾	0.08219	0.08429 ⁽⁻³⁾	0.08599 ⁽⁻²⁾	0.07699	0.11886 ⁽⁻¹⁾
	Design 2										
	1	0.08831	0.09341	0.10522 ⁽⁻²⁾	0.08440 ⁽³⁾	0.08766	0.09627	0.10293 ⁽⁻³⁾	0.10116	0.08663	0.30129 ⁽⁻¹⁾
2	0.08043	0.07478 ⁽³⁾	0.09569	0.07838	0.07469 ⁽²⁾	0.09221	0.09978 ⁽⁻³⁾	0.09986 ⁽⁻²⁾	0.07933	0.29977 ⁽⁻¹⁾	
3	0.08180	0.07156 ⁽¹⁾	0.08488	0.07802	0.07180 ⁽³⁾	0.08426	0.08574 ⁽⁻³⁾	0.08663 ⁽⁻²⁾	0.07798	0.14904 ⁽⁻¹⁾	
2	Design 1										
	1	0.07633	0.08814	0.08439	0.07423 ⁽³⁾	0.08120	0.07986	0.09137 ⁽⁻³⁾	0.09017	0.07621	0.38534 ⁽⁻¹⁾
	2	0.06571	0.05994 ⁽¹⁾	0.07332	0.06364	0.06006 ⁽³⁾	0.07122	0.08469 ⁽⁻³⁾	0.08541 ⁽⁻²⁾	0.06523	0.38635 ⁽⁻¹⁾
	3	0.07315	0.06404 ⁽³⁾	0.07490	0.07010	0.06404 ⁽²⁾	0.07264	0.07792 ⁽⁻³⁾	0.07916 ⁽⁻²⁾	0.07038	0.18557 ⁽⁻¹⁾
	Design 2										
	1	0.08477	0.09455	0.10892 ⁽⁻³⁾	0.08214 ⁽³⁾	0.08761	0.09959	0.11009 ⁽⁻²⁾	0.10795	0.08306	0.36961 ⁽⁻¹⁾
	2	0.07725	0.07164 ⁽³⁾	0.09585	0.07507	0.07138 ⁽²⁾	0.09243	0.10385 ⁽⁻³⁾	0.10401 ⁽⁻²⁾	0.07557	0.36684 ⁽⁻¹⁾
	3	0.07837	0.06814 ⁽²⁾	0.08224	0.07444	0.06815 ⁽³⁾	0.07994	0.08327 ⁽⁻³⁾	0.08433 ⁽⁻²⁾	0.07467	0.17742 ⁽⁻¹⁾
	(1)	0	5	0	0	0	0	0	0	1	0
	(2)	0	1	0	0	4	0	0	0	1	0
	(3)	1	4	0	7	6	0	0	0	0	0
	(-3)	0	0	3	0	0	1	11	2	0	0
	(-2)	0	2	4	0	0	0	4	7	0	0
	(-1)	0	1	0	0	0	0	0	0	0	17

(continued)

Table 5. Continued.

I	S	Model averaging with screening							
		S-AIC	S-BIC	S-FIC	LZWZ	A-opt	JMA	EW	
0.5	1	Design 1							
		0.08763	0.09357 ⁽⁻³⁾	0.08788↑	0.08714 ⁽²⁾ ↑	0.08671 ⁽¹⁾ ↑	0.08720↑	0.08903↑	
		0.08432↑	0.08935	0.08586↑	0.08587↑	0.08552↑	0.08306 ⁽¹⁾ ↑	0.08652↑	
	3	0.08232	0.08501	0.08416↑	0.08494↑	0.08542↑	0.08088 ⁽²⁾	0.08405↑	
	2	Design 2							
		0.08420 ⁽²⁾ ↑	0.08565↑	0.08315 ⁽¹⁾ ↑	0.08959↑	0.08805↑	0.08569↑	0.08567↑	
0.07913↑		0.07604 ⁽¹⁾ ↑	0.07959↑	0.08530↑	0.08496↑	0.07942↑	0.08087↑		
3	0.07569↑	0.07090 ⁽²⁾ ↑	0.07361↑	0.07558↑	0.07808↑	0.07627↑	0.07304↑		
1	1	Design 1							
		0.08085 ⁽²⁾ ↑	0.08237↑	0.07980 ⁽¹⁾ ↑	0.08208↑	0.08146↑	0.08274↑	0.08378↑	
		0.07484↑	0.07145 ⁽²⁾ ↑	0.07586↑	0.07918↑	0.08001↑	0.07514↑	0.07833↑	
	3	0.07606↑	0.07056 ⁽²⁾ ↑	0.07515↑	0.07841↑	0.08111↑	0.07646↑	0.07408↑	
	2	Design 2							
		0.08440 ⁽²⁾ ↑	0.08766↑	0.08346 ⁽¹⁾ ↑	0.09053↑	0.08887↑	0.08595↑	0.08880↑	
0.07838↑		0.07469 ⁽¹⁾ ↑	0.07905↑	0.08487↑	0.08487↑	0.07862↑	0.08332↑		
3	0.07705↑	0.07177 ⁽²⁾ ↑	0.07508↑	0.07760↑	0.08018↑	0.07731↑	0.07385↑		
2	1	Design 1							
		0.07423 ⁽²⁾ ↑	0.08120↑	0.07311 ⁽¹⁾ ↑	0.07587↑	0.07499↑	0.07526↑	0.09278 ⁽⁻²⁾ ↑	
		0.06364↑	0.06006 ⁽²⁾ ↑	0.06397↑	0.06877↑	0.06944↑	0.06408↑	0.08182↑	
	3	0.06916↑	0.06401 ⁽¹⁾ ↑	0.06788↑	0.07054↑	0.07328↑	0.06948↑	0.06618↑	
	2	Design 2							
		0.08214 ⁽²⁾ ↑	0.08761↑	0.08210 ⁽¹⁾ ↑	0.09034↑	0.08845↑	0.08230↑	0.09392↑	
		0.07507↑	0.07138 ⁽¹⁾ ↑	0.07567↑	0.08199↑	0.08200↑	0.07497↑	0.08451↑	
		0.07338↑	0.06812 ⁽¹⁾ ↑	0.07197↑	0.07461↑	0.07742↑	0.07374↑	0.07061↑	
		(1)	0	5	5	0	1	1	0
		(2)	5	5	0	1	0	1	0
		(3)	0	0	0	0	0	0	0
		(-3)	0	1	0	0	0	0	0
(-2)		0	0	0	0	0	0	1	
(-1)	0	0	0	0	0	0	0		

Note : (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 6. HitRate results with $(n_1, n_2) = (500, 200)$.

I	S	Model selection			Model averaging without screening						
		AIC	BIC	FIC	S-AIC	S-BIC	S-FIC	LZWZ	A-opt	JMA	EW
0.5	1	0.55672	0.55547	0.55563	0.55762 ⁽¹⁾	0.55528 ⁽⁻²⁾	0.55660	0.55578	0.55568	0.55630	0.54970 ⁽⁻¹⁾
		0.55330 ⁽³⁾	0.55245	0.55127 ⁽⁻²⁾	0.55355 ⁽²⁾	0.55283	0.55203	0.55192 ⁽⁻³⁾	0.55227	0.55315	0.54532 ⁽⁻¹⁾
		0.48583	0.48172 ⁽⁻¹⁾	0.48612	0.48748 ⁽¹⁾	0.48420 ⁽⁻³⁾	0.48640	0.48658	0.48748 ⁽³⁾	0.48700	0.48535
	2	0.55320	0.55317	0.55087 ⁽⁻²⁾	0.55390	0.55385	0.55387	0.55280	0.55293	0.55407 ⁽³⁾	0.53895 ⁽⁻¹⁾
		0.55318	0.55325	0.55050 ⁽⁻²⁾	0.55307	0.55382 ⁽²⁾	0.55235	0.55153 ⁽⁻³⁾	0.55157	0.55340	0.53853 ⁽⁻¹⁾
		0.47972	0.48050 ⁽³⁾	0.47425 ⁽⁻²⁾	0.47972	0.48120 ⁽¹⁾	0.47667	0.47517 ⁽⁻³⁾	0.47808	0.47978	0.47217 ⁽⁻¹⁾
1	1	0.60788	0.60777	0.60782	0.60785	0.60817	0.60818	0.60790	0.60837 ⁽¹⁾	0.60753 ⁽⁻²⁾	0.59778 ⁽⁻¹⁾
		0.60338	0.60447 ⁽¹⁾	0.60248	0.60382	0.60433 ⁽²⁾	0.60278	0.60225 ⁽⁻²⁾	0.60238 ⁽⁻³⁾	0.60338	0.59470 ⁽⁻¹⁾
		0.49318	0.49412 ⁽³⁾	0.49240 ⁽⁻²⁾	0.49302	0.49435 ⁽¹⁾	0.49342	0.49243 ⁽⁻³⁾	0.49253	0.49327	0.48763 ⁽⁻¹⁾
	2	0.63073	0.63053	0.62978	0.63095	0.63110 ⁽³⁾	0.63085	0.62928 ⁽⁻²⁾	0.62945 ⁽⁻³⁾	0.63095	0.61632 ⁽⁻¹⁾
		0.63172	0.63303 ⁽³⁾	0.62995 ⁽⁻³⁾	0.63142	0.63313 ⁽¹⁾	0.63087	0.62997	0.62967 ⁽⁻²⁾	0.63177	0.61713 ⁽⁻¹⁾
		0.53293	0.53387 ⁽¹⁾	0.53255	0.53275	0.53375 ⁽²⁾	0.53248	0.53172 ⁽⁻³⁾	0.53157 ⁽⁻²⁾	0.53308	0.52478 ⁽⁻¹⁾
2	1	0.69380	0.69355	0.69335	0.69443	0.69483 ⁽²⁾	0.69428	0.69248 ⁽⁻³⁾	0.69220 ⁽⁻²⁾	0.69417	0.66623 ⁽⁻¹⁾
		0.68698	0.68793 ⁽³⁾	0.68690	0.68775	0.68808 ⁽¹⁾	0.68725	0.68613 ⁽⁻³⁾	0.68595 ⁽⁻²⁾	0.68717	0.66232 ⁽⁻¹⁾
		0.54225	0.54372	0.54288	0.54293	0.54402 ⁽¹⁾	0.54283	0.54212 ⁽⁻³⁾	0.54210 ⁽⁻²⁾	0.54353	0.53128 ⁽⁻¹⁾
	2	0.71842	0.71907	0.71655 ⁽⁻²⁾	0.71887	0.71905	0.71883	0.71667 ⁽⁻³⁾	0.71738	0.71932 ⁽²⁾	0.69750 ⁽⁻¹⁾
		0.71847	0.71923 ⁽¹⁾	0.71733	0.71922 ⁽²⁾	0.71918	0.71800	0.71692 ⁽⁻²⁾	0.71698 ⁽⁻³⁾	0.71908	0.69640 ⁽⁻¹⁾
		0.62043	0.62095	0.62020 ⁽⁻²⁾	0.62092	0.62093	0.62050	0.62038 ⁽⁻³⁾	0.62070	0.62097	0.60883 ⁽⁻¹⁾
(1)	0	3	0	2	5	0	0	1	0	0	
(2)	0	0	0	2	4	0	0	0	1	0	
(3)	1	4	0	0	1	0	0	1	1	0	
(-3)	0	0	1	0	1	0	10	3	0	0	
(-2)	0	0	7	0	1	0	3	5	1	0	
(-1)	0	1	0	0	0	0	0	0	0	17	

(continued).

Table 6. Continued.

I	S	Model averaging with screening							
		S-AIC	S-BIC	S-FIC	LZWZ	A-opt	JMA	EW	
0.5	1	Design 1							
		0.55702 ⁽²⁾	0.55542 ^{(-3)↑}	0.55632	0.55638↑	0.55638↑	0.55653↑	0.55692 ^{(3)↑}	
		0.55387 ^{(1)↑}	0.55257	0.55307↑	0.55227↑	0.55285↑	0.55323↑	0.55267↑	
	0.48653	0.48398 ⁽⁻²⁾	0.48465	0.48633	0.48748 ⁽¹⁾	0.48712↑	0.48483		
	2	Design 2							
		0.55390	0.55387↑	0.55423 ^{(2)↑}	0.55265 ⁽⁻³⁾	0.55315↑	0.55368	0.55430 ^{(1)↑}	
0.55305		0.55382 ⁽²⁾	0.55302↑	0.55163↑	0.55183↑	0.55380↑	0.55393 ^{(1)↑}		
0.47970	0.48105 ⁽²⁾	0.48018↑	0.48005↑	0.47927↑	0.48030↑	0.47995↑			
1	1	Design 1							
		0.60782	0.60812	0.60832 ^{(2)↑}	0.60782	0.60820 ⁽³⁾	0.60765↑	0.60757 ^{(-3)↑}	
		0.60378	0.60433 ⁽²⁾	0.60410↑	0.60328↑	0.60328↑	0.60342↑	0.60308↑	
	0.49367↑	0.49418 ⁽²⁾	0.49397↑	0.49362↑	0.49347↑	0.49365↑	0.49403↑		
	2	Design 2							
		0.63097↑	0.63110 ⁽³⁾	0.63148 ^{(1)↑}	0.63080↑	0.63073↑	0.63115 ^{(2)↑}	0.63088↑	
		0.63142	0.63313 ⁽¹⁾	0.63162↑	0.63023↑	0.63003↑	0.63177	0.63155↑	
	0.53287↑	0.53370 ⁽³⁾	0.53363↑	0.53292↑	0.53213↑	0.53342↑	0.53357↑		
	2	1	Design 1						
			0.69443	0.69483 ⁽²⁾	0.69493 ^{(1)↑}	0.69438↑	0.69460↑	0.69405	0.69315↑
			0.68775	0.68808 ⁽¹⁾	0.68758↑	0.68688↑	0.68712↑	0.68715	0.68742↑
		0.54292	0.54397 ⁽²⁾	0.54342↑	0.54235↑	0.54252↑	0.54383↑	0.54383 ^{(3)↑}	
2		Design 2							
		0.71887	0.71905	0.71957 ^{(1)↑}	0.71802↑	0.71827↑	0.71915 ⁽³⁾	0.71883↑	
		0.71922 ⁽²⁾	0.71918	0.71910↑	0.71850↑	0.71868↑	0.71917↑	0.71830↑	
		0.62082	0.62100↑	0.62103 ^{(3)↑}	0.62048↑	0.62085↑	0.62142 ^{(1)↑}	0.62113 ^{(2)↑}	
		(1)	1	2	3	0	1	2	
		(2)	2	6	2	0	0	1	
		(3)	0	2	1	0	1	2	
		(-3)	0	1	0	1	0	1	
		(-2)	0	1	0	0	0	0	
(-1)		0	0	0	0	0	0		

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

yardsticks, while those of the S-FIC and LZWZ methods each yields the third worst estimate once with respect to hit rate but never results in the three worst estimates with respect to MSEF and MAEF; when $(n_1, n_2) = (500, 200)$, these methods yield poor estimates only very occasionally. 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, 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, shielding against the selection of 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 *et al.* [40] 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 17 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. Overall speaking, the non-screened S-BIC estimator is second only to its screened counterpart in terms of the frequency in producing the most precise estimates. 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 17 strategies considered.⁴

Lastly, we apply the Wilcoxon signed rank test [43] to test for pairwise performance equality of the methods. Our 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 2 are statistically significant. On the other hand, the same is not observed in terms of HitRate, because even if the methods produce different \hat{p}_{0j} 's, as long as the category that corresponds to $\max\{\hat{p}_{01}^w, \dots, \hat{p}_{0j}^w\}$ is the same for the methods, they will yield the same forecast of Y_0 . We also applied the Wilcoxon signed test to evaluate 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, the Wilcoxon tests results are not included in the paper but available upon request from the authors.

A referee has suggested that we conduct additional simulation experiments by increasing the number of explanatory variables and sample size. The corresponding results are contained in a supplementary file available at <http://personal.cb.cityu.edu.hk/msawan/researchprofile.htm>. We find no major qualitative difference in results under the additional experiments and those reported in the paper. Generally speaking, all the comments above apply to the additional results in broad terms.

5. Empirical applications

In this section, we apply the various FMA methods to three real datasets. 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 the General Social Survey Data, 2008

Our first application is based on 667 observations obtained from the US General Social Survey (GSS) of 2008. We use the data to investigate the relationship between belief in an afterlife and a range of demographic and social variables. The respondents were asked to indicate to what extent they believed in an afterlife (labelled as AFTERLIFE) on a 5-point scale that ranges from 1 = definitely believe to 5 = definitely not believe. They were also asked to indicate their gender (GENDER, 1 = male, 0 = female), and give answers measured on a rating scale 1 to U to the following questions, with 1 and U representing the most and least affirmative answers to the question, respectively: Do you consider yourself religious (RELIG, $U = 4$)? Do you think society trusts too much in science (TRUSTSCI, $U = 5$)? Do you consider the Christian Bible the word of God (BIBLE, $U = 3$)? To what extent do you believe in heaven (HEAVEN, $U = 4$)? To what extent do you believe in hell (HELL, $U = 4$)? Are you happy (HAPPY, $U = 3$)? Are you satisfied with your financial situation (SATFIN)? Do you consider yourself a liberal (POLVIEW, $U = 7$)? We treat AFTERLIFE as the dependent variable, and GENDER, RELIG, TRUSTSCI, BIBLE, HEAVEN, HELL, HAPPY, SATFIN, and POLVIEW as explanatory variables, resulting in $2^9 = 512$ sub-models. We index these sub-models as M_1, M_2, \dots, M_{512} , each containing a different set of explanatory variables. Given the ordinal nature of AFTERLIFE, we use the ordered probit model as the basis of our analysis.

The first goal of our study is to illustrate the replication failure that commonly arises with model selection. We randomly divide the data into two sub-samples containing 467 and 200 observations, to be used as training and test samples, respectively. This process is repeated five times, resulting in five cases of assessment, labelled as Cases 1–5. Clearly, there are overlapping observations across the training samples of the five cases. In each case, based on 467 training observations, we determine the best fitting ordered probit model by the AIC, BIC and FIC. Table 7 shows the model chosen by each of the three selection criteria for each of the five cases. The coefficient estimates and their standard errors produced by these three selection methods are also reported in the same table. We obtained the standard errors of the estimates through the output of the Hessian matrix produced by the procedure FSOLVE in Matlab. These standard errors are commonly reported in practice but they do not take into account the uncertainty arisen from model selection

Table 7. Coefficient estimates under the five assessment cases of Application 1.

Case	crit	model selected		$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$	GENDER	RELIG	TRUSTSCI	BIBLE	HEAVEN	HELL	HAPPY	SATFIN	POLVIEW
1	AIC	M13	coef.est.	1.56256	2.39644	2.91305	3.82427	-	-0.15967	-	-	-0.68391	-	-	-	-
			(s.e.)	(0.06120)	(0.05641)	(0.06604)	(0.12461)	-	(0.02277)	-	-	(0.02900)	-	-	-	-
	BIC	M5	coef.est.	1.33136	2.15697	2.67335	3.59300	-	-	-	-	-0.76949	-	-	-	-
			(s.e.)	(0.06073)	(0.05619)	(0.06617)	(0.12579)	-	-	-	-	(0.02898)	-	-	-	-
	FIC	M93	coef.est.	0.88166	1.53118	1.93473	2.76166	-	-	-0.29307	-	-	-	-0.07230	-	0.06918
			(s.e.)	(0.05241)	(0.04484)	(0.05435)	(0.11802)	-	-	(0.01656)	-	-	-	(0.03093)	-	(0.01315)
S-AIC	-	coef.est.	1.70198	2.53596	3.05262	3.96774	0.02794	-0.08827	-0.05887	-0.00799	-0.66520	-0.05166	-0.00326	-0.02243	-0.00639	
S-BIC	-	coef.est.	1.45590	2.28471	2.80091	3.71824	0.00459	-0.03717	-0.02079	-0.00045	-0.73058	-0.01367	-0.00098	-0.00322	-0.00048	
2	AIC	M275	coef.est.	2.26386	3.18417	3.59722	4.42705	-	-0.13546	-0.12018	-	-0.54051	-0.13571	-	-0.15206	-
			(s.e.)	(0.06342)	(0.05652)	(0.06320)	(0.12163)	-	(0.02327)	(0.01737)	-	(0.03047)	(0.02651)	-	(0.02914)	-
	BIC	M13	coef.est.	1.68938	2.59711	3.00867	3.84405	-	-0.20497	-	-	-0.66367	-	-	-	-
			(s.e.)	(0.06270)	(0.05617)	(0.06313)	(0.12254)	-	(0.02320)	-	-	(0.03044)	-	-	-	-
	FIC	M93	coef.est.	0.80063	1.52471	1.85534	2.62770	-	-	-0.30995	-	-	-	-0.08315	-	0.12369
			(s.e.)	(0.05499)	(0.04566)	(0.05290)	(0.11688)	-	-	(0.01682)	-	-	-	(0.03145)	-	(0.01346)
S-AIC	-	coef.est.	1.81533	2.72914	3.14271	3.97996	0.01643	-0.10277	-0.08243	-0.00839	-0.61765	-0.07148	-0.00098	-0.08675	-0.03244	
S-BIC	-	coef.est.	1.59353	2.49768	2.90980	3.75190	0.00300	-0.08362	-0.03530	-0.00430	-0.69182	-0.02285	-0.00118	-0.01906	0.01425	
3	AIC	M172	coef.est.	1.31788	2.19635	2.64680	3.40649	-0.32311	-0.13612	-	-	-0.65010	-	-	-	0.09249
			(s.e.)	(0.06288)	(0.05939)	(0.06809)	(0.12835)	(0.08744)	(0.02387)	-	-	(0.03320)	-	-	-	(0.01380)
	BIC	M5	coef.est.	1.46182	2.31169	2.75658	3.52727	-	-	-	-	-0.78550	-	-	-	-
			(s.e.)	(0.06148)	(0.05827)	(0.06780)	(0.13034)	-	-	-	-	(0.03309)	-	-	-	-
	FIC	M37	coef.est.	1.37023	2.15273	2.55682	3.28643	-0.45557	-	-	-	-	-0.48344	-	-	-
			(s.e.)	(0.05898)	(0.05355)	(0.06254)	(0.12451)	(0.08613)	-	-	-	-	(0.02781)	-	-	-
S-AIC	-	coef.est.	1.73460	2.61128	3.06105	3.82339	-0.26453	-0.06470	-0.3210	-0.06178	-0.66309	-0.00022	-0.07899	-0.03653	0.05517	
S-BIC	-	coef.est.	1.57705	2.43910	2.88615	3.65249	-0.13112	-0.04323	-0.01012	-0.02252	-0.72571	-0.00087	-0.02483	-0.01041	0.02162	

(continued).

Table 7. Continued.

Case	crit	model selected		$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$	GENDER	RELIG	TRUSTSCI	BIBLE	HEAVEN	HELL	HAPPY	SATFIN	POLVIEW	
4	AIC	M250	coef.est.	1.34319	2.11072	2.64651	3.44509	-0.39249	-	-	-	-0.71613	-	-	-0.12480	0.09296	
			(s.e.)	(0.06073)	(0.05909)	(0.07170)	(0.13176)	(0.08655)	-	-	-	(0.03402)	-	-	(0.02926)	(0.01359)	
	BIC	M33	coef.est.	1.52467	2.27943	2.80608	3.60544	-0.35618	-	-	-	-0.73687	-	-	-	-	
			(s.e.)	(0.06007)	(0.05819)	(0.07094)	(0.13215)	(0.08640)	-	-	-	(0.03397)	-	-	-	-	
	FIC	M123	coef.est.	1.58300	2.27409	2.75449	3.51988	-0.52482	-	-	-	-	-0.47154	-	-	-0.11518	-
			(s.e.)	(0.05720)	(0.05352)	(0.06595)	(0.12762)	(0.08559)	-	-	-	-	(0.02845)	-	-	(0.02908)	-
S-AIC	-	coef.est.	1.67994	2.44595	2.97632	3.77342	-0.36295	-0.00343	-0.03650	-0.04419	-0.67714	-0.01090	-0.06819	-0.07458	0.05497		
S-BIC	-	coef.est.	1.51720	2.27286	2.79957	3.59730	-0.27097	-0.00295	-0.00939	-0.01138	-0.73026	-0.00293	-0.02317	-0.02385	0.02410		
5	AIC	M226	coef.est.	2.10750	2.93258	3.43704	4.37213	-0.37614	-	-	-0.22567	-0.64215	-	-0.20400	-	-	
			(s.e.)	(0.06113)	(0.05823)	(0.07068)	(0.14649)	(0.08422)	-	-	(0.03050)	(0.03380)	-	(0.03204)	-	-	
	BIC	M33	coef.est.	1.47550	2.28504	2.78814	3.72405	-0.38925	-	-	-	-0.72872	-	-	-	-	
			(s.e.)	(0.06032)	(0.05769)	(0.07058)	(0.14600)	(0.08387)	-	-	-	(0.03373)	-	-	-	-	
	FIC	M43	coef.est.	0.54415	1.20480	1.62297	2.49166	-0.67299	-	-	-	-	-	-	0.00925	-	
			(s.e.)	(0.05320)	(0.04826)	(0.06071)	(0.14017)	(0.08133)	-	-	-	-	-	-	(0.02842)	-	
S-AIC	-	coef.est.	1.87007	2.69434	3.19820	4.13184	-0.38315	-0.02741	-0.04122	-0.13099	-0.65312	-0.02226	-0.14621	-0.00669	0.03476		
S-BIC	-	coef.est.	1.62379	2.43624	2.93788	3.87232	-0.32773	-0.01152	-0.01407	-0.05905	-0.69996	0.00215	-0.05052	-0.00285	0.01099		

and hence underestimate the true variability of the estimates. It is seen that the AIC, FIC and BIC yield 5, 4 and 3 different models, respectively, across the five assessment cases. More importantly, these models can be vastly different even when observations overlap across the five training samples and the models are selected based on the same criterion. For example, based on the AIC, while models for all cases contain HEAVEN as an explanatory variable, RELIG is included only for Cases 1, 2 and 3, but excluded for Cases 4 and 5; GENDER, in contrast, is excluded from the first three cases but included in the last two; each of SATFIN and POLVIEW is included twice, while each of the remaining variables, TRUSTSCI, BIBLE, HELL and HAPPY, is included once. In most cases, when a model includes a regressor that is excluded for another case, the coefficient estimate of this regressor often has a non-negligible magnitude and the coefficient is significantly different from zero. Clearly, these models all have very different conceptual interpretations. As expected, the BIC usually results in a more parsimonious model; for Cases 1 and 3, the BIC-based best fitting model includes only HEAVEN as an explanatory variable; for Cases 4 and 5, it includes GENDER in addition to HEAVEN, while for Case 3, it includes RELIG and HEAVEN. Interestingly, although HEAVEN is invariably included in the best fitted models determined by the AIC and BIC, it is never chosen when selection is based on the FIC; in contrast, the FIC selects each of HAPPY and HELL twice, but the BIC never selects either of these two variables. The lack of consistency produced by the different criteria, and by the same criterion across different samples, clearly demonstrates the replication problem. When results do not replicate, the conclusions are not generalisable. It may be argued that the relatively small sample size used in this study is one reason for non-replication. Nevertheless, sample sizes similar to the magnitude used in the present study frequently arise in practice.

Table 7 also reports the coefficient estimates that resulted from S-AIC and S-BIC model averaging, obtained by combining estimates across 512 sub-models. Relative to the selection-based estimates, the model average estimates are more stable across the five assessment cases. The fact that model averaging does not automatically translate any coefficient estimate to zero, also means that it is less probable for estimates to experience abrupt changes across samples; a phenomenon commonly observed with model selection. Like the S-AIC and S-BIC-based estimates, the coefficient estimates obtained using the S-FIC, LZWZ, A-opt, JMA and EW averaging methods also do not differ substantially across the five cases. They are not shown here to conserve space.

The hit rates obtained using the various methods in the test sample relevant to this example are reported in the top panel of Table 10. The integer in the bracket next to a hit rate represents the rank of the estimation method in a given case. We omit the EW averaging estimator from the comparison because of the poor performance of this estimator shown in the simulation study. With few exceptions, the S-AIC, S-BIC and S-FIC estimators deliver estimates that are at least as good as their model selection counterparts; in most cases, an improvement in hit rate is observed when model averaging is implemented. Although there are exceptions, the ordinal rankings of the S-AIC, S-BIC and S-FIC model averaging estimators, with and without screening, generally follow the same pattern as the ordinal rankings of their model selection counterparts. In nearly all cases, the FIC selection estimator yields the poorest estimates. This also results in the S-FIC estimator performing poorly relative to the S-AIC and S-BIC methods. While the performance of the non-screened versions of the LZWZ, A-opt and JMA estimators is unremarkable, a noticeable

improvement in their performance is observed after model screening is implemented; the screened versions of the LZWZ, A-opt and JMA estimators habitually provide better hit rates than the three model selection estimators. The poor showing of the non-screened versions of these three model average estimators may be explained by the fact that a relatively small number of observations (467) is used to estimate a large number of sub-models (512). Because all three strategies involve the substitution of plug-in estimators in the objective equations, the errors associated with the plug-in estimators accumulate as the number of sub-models increases and impact the final results. This also explains the marked improvement observed for these three estimators after model screening reduces the number of sub-models from 512 to 5. We find the good performance of the screened version of the S-BIC estimator particularly encouraging.

Application 2: 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.⁵ The respondents were 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, which include gender (GENDER, 1 = male, 0 = female), age (AGE, 1 = 20–29 years old, 2 = 30–39 years old, 3 = 40–49 years old, 4 = 50–59 years old, 5 = 60–69 years old), the highest level of education attainment (EDU, 1 = no final education/elementary school/junior high school/middle school, 2 = high school, 3 = professional school/technical college/university), household annual income (INC, 1 = less than 7.2 million rupiah, 2 = 7.2 to 12 million rupiah, 3 = over 12 million rupiah), employment status (EMP, 1 = employed, 0 = unemployed), and area of residence (RES, 1 = urban, 0 = rural).

The intent of the analysis is to evaluate the levels of self-realisation among the survey participants, with 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 index them as M_1, M_2, \dots, M_{64} . Applying the same top m model screening procedure, and setting $m = 5$ as in Section 4, reduces the number of sub-models within the model average from 64 to 5. We randomly select 600 observations from the full sample for model estimation, and use the remaining 390 observations for model evaluation. As in the preceding example, we repeat this process five times, yielding 5 cases of assessment with overlapping observations.

Table 8 presents the coefficient estimates and their (unadjusted) standard errors of the models selected by AIC, BIC and FIC under the five cases, in a format similar to Table 7 reported for the last example. Table 8 shows that there is a lack of consistency in results produced by a given model selection criterion across the different cases. The AIC, in particular, yields five different models for the five cases considered. By the AIC, the regressor AGE is excluded for Cases 1 and 2 but included from all other cases. Similarly, for Cases 1 and 5, the AIC produces a zero coefficient for EMP, but for Cases 2, 3 and 4, it leads to estimated effects of EMP that are not only non-zero but also quite large. Under Cases 1 and 2, the lowest BIC occurs for a model that includes EDU, INC and RES as regressors, but under Cases 3 and 4, the model selected by the BIC includes neither EDU nor RES. Clearly, the

Table 8. Coefficient estimates under the five assessment cases of Application 2.

Case	criterion	model selected		$\hat{\alpha}_1$	$\hat{\alpha}_2$	GENDER	AGE	EDU	INC	EMP	RES	
1	AIC	M40	coef.est.	-0.10450	1.525258	-	-	-0.20228	-0.35729	-	0.41009	
			(s.e.)	(0.0606)	(0.053652)	-	-	(0.029099)	(0.020292)	-	(0.074329)	
	BIC	M40	coef.est.	-0.10450	1.525258	-	-	-0.20228	-0.35729	-	0.41009	
			(s.e.)	(0.0606)	(0.053652)	-	-	(0.029099)	(0.020292)	-	(0.074329)	
	FIC	M26	coef.est.	-1.2292	0.334712	-0.09405	0.092114	-	-	-	-	0.0228779
			(s.e.)	(0.058895)	(0.052573)	(0.072643)	(0.018784)	-	-	-	-	(0.073541)
S-BIC	-	coef.est.	-0.22222	1.401587	-0.00299	0.009511	-0.09829	-0.37884	0.00391	0.0393412		
JMA	-	coef.est.	-0.45813	1.150262	-0.06861	0.026234	-0.06176	-0.29918	0.073946	0.313859		
2	AIC	M57	coef.est.	-0.24517	1.41514	-	-	-0.29317	-0.34306	0.148795	0.492503	
			(s.e.)	(0.062982)	(0.053262)	-	-	(0.030582)	(0.020884)	(0.061325)	(0.07581)	
	BIC	M40	coef.est.	-0.17999	1.477136	-	-	-0.28095	-0.33634	-	0.469174	
			(s.e.)	(0.062932)	(0.053187)	-	-	(0.030559)	(0.020871)	-	(0.075766)	
	FIC	M26	coef.est.	-1.27084	0.304245	-0.06882	0.074603	-	-	-	-	0.0224613
			(s.e.)	(0.060723)	(0.05191)	(0.071465)	(0.018805)	-	-	-	-	(0.074785)
S-BIC	-	coef.est.	-0.21392	1.441774	-0.00321	0.002272	-0.25731	-0.34209	0.018597	0.467407		
JMA	-	coef.est.	-0.42336	1.214645	-0.11727	0.0	-0.18633	-0.2879	0.171208	0.390863		
3	AIC	M46	coef.est.	-0.68773	1.114487	-0.20072	0.087902	-	-0.37062	0.189244	-	
			(s.e.)	(0.068845)	(0.053613)	(0.07215)	(0.01875)	-	(0.021229)	(0.060008)	-	
	BIC	M5	coef.est.	-0.44499	1.338675	-	-	-	-0.37327	-	-	
			(s.e.)	(0.066431)	(0.053308)	-	-	-	(0.021174)	-	-	
	FIC	M8	coef.est.	-1.37483	0.358116	-0.12845	0.106534	-	-	-	-	
			(s.e.)	(0.065039)	(0.05261)	(0.0715566)	(0.0188566)	-	-	-	-	
S-BIC	-	coef.est.	-0.52415	1.264646	-0.00992	0.034211	-0.0083	-0.37141	0.012865	0.012095		
JMA	-	coef.est.	-0.73113	1.052285	-0.15705	0.048083	-0.05086	-0.28606	0.162969	0.101394		

(continued).

Table 8. Continued.

Case	criterion	model selected		$\hat{\alpha}_1$	$\hat{\alpha}_2$	GENDER	AGE	EDU	INC	EMP	RES
4	AIC	M61	coef.est.	-0.89723	1.08790	-0.27794	0.08689	-	-0.31054	0.213804	0.172353
			(s.e.)	(0.068327)	(0.055702)	(0.071982)	(0.019367)	-	(0.021527)	(0.060379)	(0.074044)
	BIC	M5	coef.est.	-0.60911	1.3460	-	-	-	-0.30966	-	-
			(s.e.)	(0.067806)	(0.05518)	-	-	-	(0.021441)	-	-
	FIC	M26	coef.est.	-1.45539	0.47383	-0.17307	0.10935	-	-	-	0.12726
		(s.e.)	(0.066965)	(0.054948)	(0.071593)	(0.019214)	-	-	-	(0.073442)	
	S-BIC	-	coef.est.	-0.68849	1.27350	-0.03498	0.034317	-0.00652	-0.3073	0.01479	0.02659
	JMA	-	coef.est.	-0.87662	1.098186	-0.25123	0.0598	-0.02543	-0.26898	0.18491	0.17229
5	AIC	M54	coef.est.	-0.3917	1.40875	-	0.063832	-0.15446	-0.32234	-	0.395732
			(s.e.)	(0.060019)	(0.057176)	-	(0.018437)	(0.030454)	(0.020904)	-	(0.070221)
	BIC	M21	coef.est.	-0.35707	1.42860	-	-	-	-0.36711	-	0.376983
			(s.e.)	(0.059708)	(0.056934)	-	-	-	(0.020865)	-	(0.070031)
	FIC	M26	coef.est.	-1.24718	0.48619	-0.02483	0.097902	-	-	-	0.236482
		(s.e.)	(0.058531)	(0.056095)	(0.069963)	(0.018269)	-	-	-	(0.069471)	
	S-BIC	-	coef.est.	-0.35517	1.43434	-0.00013	0.015139	-0.04931	-0.35283	0.002136	0.0378645
	JMA	-	coef.est.	-0.49814	1.27743	0.0	0.037696	-0.095	-0.25362	0.0	0.28190

difficulty in replicating results is an issue here. For this data set, the FIC yields two models and hence the most replicable results across the five assessment cases. Table 8 also provides the estimates based on the S-BIC and JMA strategies. Compared to the estimates obtained from model selection, the averaging-based estimates are more stable. For a given averaging method, the estimate of a given coefficient is about the same across the five cases. The middle panel of Table 10 reports the hit rates based on the 390 observations in the test sample for each of the five cases. Again, the S-AIC, S-BIC and S-FIC estimators generally yield improved hit rates over their model selection counterparts. The LZWZ, A-opt and JMA methods rarely yield very inferior hit rates and are often ranked among the most accurate of all. Generally speaking, the results of this data example are consistent with those observed under Application 1.

Application 3: Analysis of saltine cracker purchase data

The data in this application, taken from Jain *et al.* [24] and Franses and Paap [12], contain 3292 observations of saltine cracker purchases in Rome, Georgia. The data provide information on consumers' purchase decisions among four brands of saltine crackers: Nabisco, Sunshine, Keebler, and Private labels, and factors affecting their decisions including the price of the brand (PRICE), and whether the brand was on aisle display only (DISP, 1 = yes, 0 = no), featured only (FEATU, 1 = yes, 0 = no), or jointly on display and featured (DF, 1 = yes, 0 = no). In our application, we treat the intercept and the price variable as mandatory explanatory variables, and the three dummy variables as optional explanatory variables, resulting in $2^3 = 8$ sub-models. We index these models as M_1, M_2, \dots, M_8 . Franses and Paap [12] used the same explanatory variables in their analysis.

We begin our analysis by testing the IIA assumption. The Hausman and McFadden [19] test rejects the IIA assumption at the 5% significance level. This is consistent with the result of Franses and Paap [12], who applied the same test to a subset of the data. Hence we adopt the nested logit model as our analytical framework. Following Franses and Paap [12], we split the four brands into two clusters, with Private label in the first, and the other three brands in the second. When implemented without screening, model averaging combines the forecasts obtained from all of the eight sub-models; with screening, the number of sub-models is reduced to five. We randomly select 2492 observations for estimation and use the remaining 800 observations as a test sample for evaluation. We repeat this process 5 times, resulting in 5 cases of assessment with overlapping observations.

Table 9 presents the results. It is found that the AIC and FIC each delivers three, whereas the BIC yields two different models for the five cases considered. The AIC and FIC each select the full model (M_8) twice but the BIC never selects it. Model M_4 that contains DF but excludes the DISPLAY and FEATURE is also frequently selected. With the exception of Case 2 in which the three criteria each select a different model, there is more uniformity in results produced by model selection in the current example than in the previous two. This is likely attributed to the much reduced number of candidate models as a result of a smaller number of regressors. That said, the estimates of a given coefficient can still experience considerable variations across the cases when different models are selected by a given model selection criterion. On the other hand, estimates obtained by model averaging are generally more stable. In terms of hit-rate comparisons, while no one strategy is uniformly the best, model averaging generally has an edge over model selection, as is evidenced from the lower panel of Table 10.

Table 9. Coefficient estimates under the five assessment cases of Application 3.

Case	criterion	model selected		$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\tau}$	PRICE	DISP	FEATU	DF
1	AIC	M7	coef.est.	-1.83782	-1.71443	-1.21627	0.62683	-2.65907	-	0.23612	0.29393
			(s.e.)	(0.04495)	(0.05072)	(0.04908)	(0.01826)	(0.09028)	(0.12337)	(0.09782)	
	BIC	M1	coef.est.	-1.83512	-1.53760	-1.04877	0.54136	-2.58974	-	-	-
			(s.e.)	(0.04482)	(0.04394)	(0.04247)	(0.01597)	(0.08639)	-	-	
	FIC	M1	coef.est.	-1.83512	-1.53760	-1.04877	0.54136	-2.58974	-	-	-
(s.e.)	(0.04482)	(0.04394)	(0.04247)	(0.01597)	(0.08639)	-	-				
S-AIC	-	coef.est.	-1.83513	-1.69188	-1.19458	0.61672	-2.64888	0.01176	0.16154	0.27317	
S-BIC	-	coef.est.	-1.83703	-1.58470	-1.09359	0.56401	-2.60855	0.00017	0.01593	0.10124	
2	AIC	M7	coef.est.	-1.56761	-1.45750	-1.03269	0.51441	-2.10320	-	0.19353	0.29425
			(s.e.)	(0.04411)	(0.04392)	(0.04163)	(0.01513)	(0.08506)	-	(0.11070)	(0.08681)
	BIC	M4	coef.est.	-1.57411	-1.41146	-0.98948	0.49273	-2.09578	-	-	0.26848
			(s.e.)	(0.04409)	(0.04208)	(0.03988)	(0.01453)	(0.08383)	-	-	(0.08410)
	FIC	M1	coef.est.	-1.55425	-1.26207	-0.85568	0.42735	-2.01034	-	-	-
(s.e.)			(0.04399)	(0.03656)	(0.03462)	(0.01271)	(0.07923)	-	-	-	
S-AIC	-	coef.est.	-1.56110	-1.44399	-1.02002	0.51011	-2.09401	0.02413	0.12850	0.28987	
S-BIC	-	coef.est.	-1.56558	-1.35539	-0.93934	0.46836	-2.06241	0.00107	0.01175	0.16389	
3	AIC	M4	coef.est.	-1.56000	-1.41361	-1.03405	0.53576	-2.30458	-	-	0.32078
			(s.e.)	(0.04370)	(0.04225)	(0.04280)	(0.01619)	(0.08476)	-	-	(0.08998)
	BIC	M4	coef.est.	-1.56000	-1.41361	-1.03405	0.53576	-2.30458	-	-	0.32078
			(s.e.)	(0.04370)	(0.04225)	(0.04280)	(0.01619)	(0.08476)	-	-	(0.08998)
	FIC	M7	coef.est.	-1.55617	-1.44223	-1.06204	0.55109	-2.31393	-	0.15304	0.33855
(s.e.)			(0.04372)	(0.04344)	(0.04400)	(0.01662)	(0.08556)	-	(0.11851)	(0.09179)	
S-AIC	-	coef.est.	-1.54918	-1.43036	-1.05003	0.54693	-2.30333	0.02298	0.07644	0.33719	
S-BIC	-	coef.est.	-1.55820	-1.39688	-1.01720	0.52756	-2.29613	0.00146	0.00603	0.27262	

(continued).

Table 9. Continued.

Case	criterion	model selected		$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\tau}$	PRICE	DISP	FEATU	DF
4	AIC	M8	coef.est.	-2.10795	-2.35639	-1.55600	0.82855	-3.69735	0.14047	0.34099	0.47031
			(s.e.)	(0.04483)	(0.07029)	(0.06493)	(0.02408)	(0.09391)	(0.06402)	(0.14809)	(0.11851)
	BIC	M4	coef.est.	-2.17544	-2.28843	-1.49772	0.78339	-3.73642	-	-	0.39414
			(s.e.)	(0.04478)	(0.06647)	(0.06152)	(0.02293)	(0.09301)	-	-	(0.11493)
	FIC	M8	coef.est.	-2.10795	-2.35639	-1.55600	0.82855	-3.69735	0.14047	0.34099	0.47031
			(s.e.)	(0.04483)	(0.07029)	(0.06493)	(0.02408)	(0.09391)	(0.06402)	(0.14809)	(0.11851)
S-AIC	-	coef.est.	-2.12876	-2.33858	-1.54092	0.81604	-3.71024	0.09513	0.25520	0.44607	
S-BIC	-	coef.est.	-2.17633	-2.28376	-1.49030	0.78096	-3.74394	0.00881	0.03356	0.33455	
5	AIC	M8	(0.044951198251968)	-1.63508	-1.71526	-1.13393	0.61263	-2.42427	0.10832	0.25770	0.45524
			(s.e.)	(0.04428)	(0.05278)	(0.04720)	(0.01806)	(0.08898)	(0.05290)	(0.12964)	(0.09931)
	BIC	M4	coef.est.	-1.69372	-1.68202	-1.10389	0.58552	-2.47578	-	-	0.39286
			(s.e.)	(0.04426)	(0.05040)	(0.04512)	(0.01731)	(0.08787)	-	-	(0.09627)
	FIC	M8	coef.est.	-1.63508	-1.71526	-1.13393	0.61263	-2.42427	0.10832	0.25770	0.45524
			(s.e.)	(0.04428)	(0.05278)	(0.04720)	(0.01806)	(0.08898)	(0.05290)	(0.12964)	(0.09931)
S-AIC	-	coef.est.	-1.69038	-1.68100	-1.10281	0.58573	-2.47252	0.00661	0.01748	0.38610	
S-BIC	-	coef.est.	-1.65721	-1.70322	-1.12306	0.60261	-2.44378	0.06688	0.16417	0.43165	

Table 10. HitRate comparisons for Applications 1, 2 and 3.

Case	Model selection			Model averaging without screening						Model averaging with screening					
	AIC	BIC	FIC	S-AIC	S-BIC	S-FIC	LZWZ	A-opt	JMA	S-AIC	S-BIC	S-FIC	LZWZ	A-opt	JMA
Application 1															
1	0.67416(2)	0.67041(7)	0.57678(15)	0.67790(1)	0.67041(7)	0.58427(14)	0.64794(11)	0.64794(11)	0.64794(11)	0.67416(2)	0.67041(7)	0.67041(7)	0.67416(2)	0.67416(2)	0.67416(2)
2	0.62547(1)	0.61049(8)	0.52434(14)	0.62547(1)	0.61423(5)	0.52434(14)	0.59925(11)	0.59925(11)	0.59925(11)	0.61423(5)	0.61423(5)	0.61049(8)	0.62172(3)	0.61798(4)	0.61049(8)
3	0.54682(6)	0.53933(9)	0.48315(15)	0.55056(3)	0.54682(6)	0.49813(14)	0.53558(11)	0.53558(11)	0.53558(11)	0.55056(3)	0.55056(3)	0.53933(9)	0.55805(1)	0.55805(1)	0.54682(6)
4	0.57678(2)	0.57303(3)	0.50936(15)	0.56929(7)	0.57303(3)	0.51685(14)	0.58427(1)	0.55805(11)	0.56554(9)	0.57303(3)	0.57303(3)	0.55056(13)	0.56554(9)	0.56929(7)	0.55805(11)
5	0.59176(10)	0.61423(1)	0.52434(15)	0.61423(1)	0.61049(5)	0.53558(14)	0.56929(11)	0.56929(11)	0.56929(11)	0.61049(5)	0.61423(1)	0.61423(1)	0.61049(5)	0.61049(5)	0.61049(5)
Application 2															
1	0.63846(14)	0.63846(14)	0.65128(6)	0.64872(10)	0.65385(3)	0.65128(6)	0.65128(6)	0.65128(6)	0.65385(3)	0.64615(13)	0.65641(1)	0.65385(3)	0.64872(10)	0.64872(10)	0.65641(1)
2	0.62308(11)	0.62821(4)	0.65641(1)	0.62308(11)	0.62821(4)	0.65641(1)	0.62564(8)	0.62564(8)	0.63333(3)	0.62308(11)	0.62821(4)	0.62821(4)	0.62308(11)	0.62308(11)	0.62564(8)
3	0.59487(15)	0.60000(2)	0.60000(2)	0.60256(1)	0.60000(2)	0.60000(2)	0.60000(2)	0.60000(2)	0.60000(2)	0.60000(2)	0.60000(2)	0.60000(2)	0.60000(2)	0.60000(2)	0.60000(2)
4	0.51026(13)	0.51282(1)	0.51282(1)	0.51282(1)	0.51282(1)	0.51282(1)	0.51026(13)	0.51026(13)	0.51282(1)	0.51282(1)	0.51282(1)	0.51282(1)	0.51282(1)	0.51282(1)	0.51282(1)
5	0.57179(1)	0.56667(6)	0.56667(6)	0.56667(6)	0.56667(6)	0.56667(6)	0.57179(1)	0.57179(1)	0.56667(6)	0.56667(6)	0.56667(6)	0.56667(6)	0.57179(1)	0.57179(1)	0.56667(6)
Application 3															
1	0.53875(2)	0.53500(11)	0.53500(11)	0.53875(2)	0.53375(14)	0.53750(6)	0.53750(6)	0.53625(9)	0.53500(11)	0.53875(2)	0.53375(14)	0.54000(1)	0.53750(6)	0.53625(9)	0.53875(2)
2	0.54000(9)	0.54125(4)	0.54000(9)	0.54125(4)	0.53750(12)	0.54250(1)	0.53750(12)	0.54250(1)	0.53000(15)	0.54125(4)	0.53750(12)	0.54125(4)	0.54250(1)	0.54125(4)	0.54000(9)
3	0.59250(12)	0.59250(12)	0.59500(2)	0.59500(2)	0.59250(12)	0.59500(2)	0.59500(2)	0.59500(2)	0.59500(2)	0.59625(1)	0.59250(12)	0.59500(2)	0.59500(2)	0.59500(2)	0.59500(2)
4	0.51000(14)	0.51875(3)	0.50875(15)	0.51750(8)	0.51875(3)	0.51500(13)	0.51875(3)	0.52000(1)	0.51750(8)	0.51750(8)	0.51875(3)	0.51750(8)	0.51875(3)	0.52000(1)	0.51750(8)
5	0.54625(3)	0.54375(12)	0.54125(15)	0.54625(3)	0.54375(12)	0.54625(3)	0.54625(3)	0.54500(10)	0.54750(1)	0.54625(3)	0.54375(12)	0.54625(3)	0.54625(3)	0.54500(10)	0.54750(1)

Note : The figure in bracket gives the rank of the estimator with respect to hit rate.

6. Conclusions

The ordered probit and nested logit models have received both theoretical and empirical support in the literature. 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, needed to accommodate the unfulfillment of the IIA property. When applying these models, a researcher normally considers an array of models, each containing a different combination of regressors, selecting 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 many areas of social science.

In this study, 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 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 and unstable 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, which is, assuring against the selection of a very poor model that may not withstand replications, and thereby mitigating the replication crisis that commonly arises in empirical research. Adding to this advantage is the fact that some averaging strategies 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. Our analysis is also the first that considers the jackknife averaging strategy outside the framework of the linear model. We prove that the jackknife estimator achieves an asymptotic optimality. We consider this result a novelty and an important theoretical advance.

As has been apparent from our preceding discussion, the emphasis of the model averaging literature has been on the efficiency of point estimators of the unknowns. Relatively little is known about model diagnosis and post-model averaging inference. To address this lack of understanding, we need information on the full distributions on model average estimators. The recent work of Hansen [17] and Liu [29] serves as a useful guide in this regard. Also, in recent years, stability selection [32,38], which involves applying a variable selection method to random sub-samples of data and choosing the variables that are most frequently selected, is gaining popularity. This approach has the attractive advantage of error control via an upper bound on the falsely selected variables. It remains for future research to compare this enhanced approach to model selection with model averaging. Clearly, more remains to be done, but hopefully this paper serves to pique an interest for further explorations of model averaging in statistical applications.

Notes

1. The subject of replication crisis has attracted enormous attention among scientists in recent years. See [13] for a high-level introduction. Andrew Gelman's blog (<http://andrewgelman.com/>) provides links to many interesting articles written on this subject.
2. Model averaging is a fundamentally different approach from boosting used extensively in machine learning. In contrast to model averaging, boosting adds new models to the model ensemble sequentially, creating a new model space that is more complex than the original. Davidson and Fan [10] showed that when there exists considerable uncertainty in the original model space, model averaging is often preferred to boosting which is perceived as building an overly complex model out of insufficient data.
3. The Matlab codes for computing the FMA estimates are available for download from the corresponding author's website: <http://personal.cb.cityu.edu.hk/msawan/researchprofile.htm>.
4. With the exception of the Jackknife methods, the FMA methods considered in this paper are not computationally demanding. To give an idea, in our simulations, under the nested logit model with $(n_1, n_2) = (300, 100)$, it takes 7–9 s to complete one round of replication if the JMA estimators are excluded from the set; however, if the JMA estimators are also included, then the corresponding computing time increases to 120–140 s. It is also observed that the time required for computing the JMA estimates increases with the sample size. The scalability of the model averaging methods in relation to computing time is not an issue except for the JMA methods. For model averaging in a high-dimensionality setup, see [2].
5. The data are available online at www.asiabarometer.org/. Inoguchi *et al.* [22] provided a detailed discussion of the survey.

Acknowledgements

The authors thank the associate editor and two referees for thoughtful review of the manuscript. The usual disclaimer applies.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

Wan's work was supported by a strategic grant from the City University of Hong Kong (Project no. 7004985). Zhang's research was supported by the National Science Foundation of China (Project nos. 71522004, 71463012, 71631008 and 11471324) and a grant from the Ministry of Education of China (Project no. 17YJC910011).

References

- [1] S.M. Amini and C.F. Parmeter, *Comparison of model averaging techniques: Assessing growth determinants*, *J. Appl. Econometrics* 27 (2012), pp. 870–876.
- [2] T. Ando and K.C. Li, *A model-averaging approach for high-dimensional regression*, *J. Amer. Statist. Assoc.* 109 (2014), pp. 254–265.
- [3] S.T. Buckland, K.P. Burnham, and N.H. Augustin, *Model selection: An integral part of inference*, *Biometrics* 53 (1997), pp. 603–618.
- [4] K.P. Burnham and D.R. Anderson, *Multimodel inference understanding AIC and BIC in model selection*, *Sociol. Methods Res.* 33 (2004), pp. 261–304.
- [5] G. Claeskens, *Statistical model choice*, *Annu. Rev. Stat. Appl.* 3 (2016), pp. 233–256.
- [6] G. Claeskens, C. Croux, and J. Van Kerckhoven, *Variable selection for logistic regression using a prediction-focused information criterion*, *Biometrics* 62 (2006), pp. 972–979.

- [7] G. Claeskens and N.L. Hjort, *The focused information criterion*, J. Amer. Statist. Assoc. 98 (2003), pp. 879–899.
- [8] G. Claeskens and N.L. Hjort, *Model Selection and Model Averaging*, Cambridge, New York, 2008.
- [9] D. Danilov and J.R. Magnus, *On the harm that ignoring pre-testing can cause*, J. Econometrics 122 (2004), pp. 27–46.
- [10] I. Davidson and W. Fan, *When efficient model averaging outperforms boosting and bagging*, in *Knowledge Discovery in Databases: PKDD 2006, LNAI 4213*, J. Fürnkranz, T. Scheffer, and M. Spiliopoulou, eds., Springer, Heidelberg, 2006, pp. 478–486.
- [11] K. Duan and Y. Mei, *A comparison study of three statistical downscaling methods and their model-averaging ensemble for precipitation downscaling in China*, Theor. Appl. Climatol. 116 (2014), pp. 707–719.
- [12] P.H. Franses and R. Paap, *Quantitative Models in Marketing Research*, Cambridge, New York, 2001.
- [13] A. Gelman, *Statistics and the crisis of scientific replication*, Significance 12 (2015), pp. 33–35.
- [14] J.A. Giles and D.E.A. Giles, *Pre-test estimation and testing in econometrics: Recent developments*, J. Econ. Surv. 7 (1993), pp. 145–197.
- [15] D. Grant, A. Morales, and J.J. Sallaz, *Pathways to meaning: A new approach to studying emotions at work*, Am. J. Sociol. 115 (2009), pp. 327–364.
- [16] B.E. Hansen, *Least squares model averaging*, Econometrica 75 (2007), pp. 1175–1189.
- [17] B.E. Hansen, *Model averaging, asymptotic risk and regressor groups*, Quant. Econom. 5 (2014), pp. 495–530.
- [18] B.E. Hansen and J.S. Racine, *Jackknife model averaging*, J. Econom. 167 (2012), pp. 38–46.
- [19] J. Hausman and D. McFadden, *Specification tests for the multinomial logit model*, Econometrica 52 (1984), pp. 1219–1240.
- [20] N.L. Hjort and G. Claeskens, *Frequentist model average estimators*, J. Amer. Statist. Assoc. 98 (2003), pp. 879–899.
- [21] N.L. Hjort and G. Claeskens, *Focused information criteria and model averaging for the Cox hazard regression model*, J. Amer. Statist. Assoc. 101 (2006), pp. 1449–1464.
- [22] T. Inoguchi, M. Basànez, A. Tanaka, and T. Dadabaev (eds.), *Values and Life Styles in Urban Asia: A Cross-Cultural Analysis and Sourcebook based on the AsiaBarometer Survey of 2003, Vol. 19*, Tokyo, Institute of Oriental Culture, University of Tokyo, 2005.
- [23] C.H. Jackson, S.G. Thompson, and L.D. Sharples, *Accounting for uncertainty in health economic decision models by using model averaging*, J. R. Stat. Soc. Ser. A 172 (2009), pp. 383–404.
- [24] D.C. Jain, N.J. Vilcassim, and P.K. Chintagunta, *A random-coefficients logit brand-choice model applied to panel data*, J. Bus. Econ. Stat. 12 (1994), pp. 317–328.
- [25] J.B. Johnson and K.S. Omland, *Model selection in ecology and evolution*, Trends Ecol. Evol. 19 (2004), pp. 101–108.
- [26] T.A. Knapp, N.E. White, and D.E. Clark, *A nested logit approach to household mobility*, J. Reg. Sci. 41 (2001), pp. 1–22.
- [27] H. Leeb and B.M. Pötscher, *Model selection and inference: Facts and fiction*, Econ. Theory. 21 (2005), pp. 21–59.
- [28] H. Liang, G. Zou, A.T.K. Wan, and X. Zhang, *Optimal weight choice for frequentist model average estimators*, J. Amer. Statist. Assoc. 106 (2011), pp. 1053–1066.
- [29] C.A. Liu, *Distribution theory of the least squares averaging estimator*, J. Econometrics 186 (2015), pp. 142–159.
- [30] S. Liu and Y. Yang, *Combining models in longitudinal data analysis*, Ann. Inst. Stat. Math. 64 (2012), pp. 233–254.
- [31] S.R. Lucas, *Effectively maintained inequality: Education transitions, track mobility, and social background effects*, Am. J. Sociol. 106 (2001), pp. 1642–1690.
- [32] N. Meinshausen and P. Blmann, *Stability selection*, J. R. Stat. Soc. Ser. B 72 (2010), pp. 417–473. 260.
- [33] E. Moral-Benito, *Model averaging in economics: An overview*, J. Econ. Surv. 29 (2015), pp. 46–75.

[34] R.D. Plotnick, *The effects of attitudes on teenage premarital pregnancy and its resolution*, Am. Sociol. Rev. 57 (1992), pp. 800–811.

[35] M. Schomaker and C. Heumann, *Model selection and model averaging after multiple imputation*, Comput. Statist. Data Anal. 71 (2014), pp. 758–770.

[36] M. Schomaker, A.T.K. Wan, and C. Heumann, *Frequentist model averaging with missing observations*, Comput. Statist. Data Anal. 54 (2010), pp. 3336–3347.

[37] K. Schorning, B. Bornkamp, F. Bretz, and H. Dette, *Model selection versus model averaging in dose finding studies*, Stat. Med. 35 (2016), pp. 4021–404.

[38] R.D. Shah and R.J. Samworth, *Variable selection with error control: Another look at stability selection*, J. R. Stat. Soc. Ser. B 75 (2013), pp. 55–80.

[39] A.T.K. Wan and X. Zhang, *On the use of model averaging in tourism research*, Ann. Tour. Res. 36 (2009), pp. 525–532.

[40] A.T.K. Wan, X. Zhang, and S. Wang, *Frequentist model averaging for multinomial and ordered logit models*, Int. J. Forecast. 30 (2014), pp. 118–128.

[41] A.T.K. Wan, X. Zhang, and G. Zou, *Least squares model averaging by Mallows criterion*, J. Econometrics 156 (2010), pp. 277–283.

[42] H. White, *Maximum likelihood estimation of misspecified models*, Econometrica 50 (1982), pp. 1–25.

[43] F. Wilcoxon, *Individual comparisons by ranking methods*, Biometrics 1 (1945), pp. 80–83.

[44] Z. Yuan and Y. Yang, *Combining linear regression models: When and how?* J. Amer. Statist. Assoc. 100 (2005), pp. 1202–1214.

[45] X. Zhang and H. Liang, *Focused information criterion and model averaging for generalized additive partial linear models*, Ann. Statist. 39 (2011), pp. 174–200.

[46] X. Zhang, A.T.K. Wan, and S.Z. Zhou, *Focused information criteria, model selection, and model averaging in a Tobit model with a nonzero threshold*, J. Bus. Econ. Stat. 30 (2012), pp. 132–142.

[47] X. Zhang, A.T.K. Wan, and G. Zou, *Model averaging by jackknife criterion in models with dependent data*, J. Econometrics 174 (2013), pp. 82–94.

[48] X. Zhang, G. Zou, and R.J. Carroll, *Model averaging based on Kullback–Leibler distance*, Stat. Sinica 25 (2015), pp. 1583–1598.

[49] X. Zhang, G. Zou, and H. Liang, *Model averaging and weight choice in linear mixed-effects models*, Biometrika 101 (2014), pp. 205–218.

Appendix

A description of the LMF

As mentioned in Section 3, the LMF 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 \tilde{h} be the vector of unknowns corresponding to the mandatory variables in a model. Thus, $\tilde{h} = (\alpha_1, \dots, \alpha_{j-1}, \beta')'$ for the ordered probit model and $\tilde{h} = (\alpha_{j_1 | B_1}, \dots, \alpha_{j_K | B_K}, \tau_1, \dots, \tau_K, \beta')'$ for the nested logit models. Let the true parameter vector of the model be $(\tilde{h}'_{true}, \gamma'_0 + \delta'/n^{1/2})'$, where \tilde{h}_{true} is the vector containing the true values of the coefficients in \tilde{h} , 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 sub-models obtained by setting different coefficients in δ to 0, leading to 2^q estimators of $\mu = \mu(\tilde{h}, \gamma)$ to choose between or combine. Denote the FMA estimator of μ as $\hat{\mu}^w$.

Let $\mathcal{L}(\tilde{h}, \gamma)$ be the likelihood function for the full model, and $\mathcal{J}_{n,full} = -(1/n)(\partial^2 \log \mathcal{L}(\tilde{h}, \gamma) / \partial(\tilde{h}', \gamma')' \partial(\tilde{h}', \gamma')) = \begin{pmatrix} \mathcal{J}_{n,00} & \mathcal{J}_{n,01} \\ \mathcal{J}_{n,10} & \mathcal{J}_{n,11} \end{pmatrix}$ 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 $|\tilde{h}|$ is the length of \tilde{h} 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 $(|\tilde{h}| + q) \times (|\tilde{h}| + q)$. Let ϖ_s be the projection matrix that maps the vector $\delta = (\delta_1, \dots, \delta_q)$ 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 $\omega = \mathcal{J}_{10} \mathcal{J}_{00}^{-1} (\partial \mu / \partial \tilde{h}) -$

$\partial\mu/\partial\gamma$, with the partial derivatives evaluated at $(\hat{h}_{true}, \gamma_0)$. 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 [20] showed that

$$\sqrt{n}(\hat{\mu}^w - \mu) \xrightarrow{d} \Lambda \equiv \left(\frac{\partial\mu}{\partial\hat{h}} \right)' \mathcal{J}_{00}^{-1}M + \omega' \left\{ \delta - \hat{\delta}(D) \right\}, \tag{A1}$$

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

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

$$\begin{aligned} R(\hat{\mu}^w) &= E(\Lambda^2) = \zeta_0^2 + E \left(\omega' \hat{\delta}(D) - \omega' \delta \right)^2 \\ &= \zeta_0^2 + \omega' \mathcal{K}^{1/2} H^2(w) \mathcal{K}^{1/2} \omega + \left(\omega' \mathcal{K}^{1/2} \mathcal{L}(w) \mathcal{K}^{-1/2} \delta \right)^2, \end{aligned} \tag{A2}$$

where $\zeta_0^2 = (\partial\mu/\partial\hat{h})' \mathcal{J}_{00}^{-1} (\partial\mu/\partial\hat{h})$ and $\mathcal{L}(w) = I_q - H(w)$. 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 (A.2); 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 *et al.* [28] showed that

$$\tilde{R}(\hat{\mu}^w) = \zeta_0^2 + \omega' \mathcal{K} \omega + (\omega' \mathcal{K}^{1/2} \mathcal{L}(w) \mathcal{K}^{-1/2} D)^2 + 2\omega' \mathcal{K}^{1/2} H(w) \mathcal{K}^{1/2} \omega \tag{A3}$$

is an unbiased estimator of $R(\hat{\mu}^w)$. The objective function $\hat{R}(\hat{\mu}^w)$ associated with LZWZ method given in Equation (11) is obtained by deleting the first two terms that are unrelated to w on the r.h.s. of Equation (A.3), and replacing $\omega, \mathcal{K}, H(w), D$ and $\mathcal{L}(w)$ 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 $\hat{\mathcal{L}}(w)$. Note that $\hat{H}(w)$ and $\hat{\mathcal{L}}(w)$ in Equation (11) have the same expressions as $H(w)$ and $\mathcal{L}(w)$ in Equation (A.3), except that \mathcal{K} contained in $H(w)$ and $\mathcal{L}(w)$ are replaced by $\hat{\mathcal{K}}$ in the construction of $\hat{H}(w)$ and $\hat{\mathcal{L}}(w)$.

For the A-opt method, the objective function (12) is obtained by removing ζ_0^2 that is unrelated to w from the r.h.s. of Equation (A.2), and replacing $\omega, \mathcal{K}, H(w), \mathcal{L}(w)$ in Equation (A.2) by δ with $\hat{\omega}, \hat{\mathcal{K}}, \hat{H}(w), \hat{\mathcal{L}}(w)$ and $\hat{\delta}$, respectively. See [40] for details.

Proof of asymptotic optimality of the JMA estimator

Here, we show the proof of Equation (14). It is assumed that q and J are fixed. Let $\hat{\theta}^{(s)} = \text{stack}(\hat{\alpha}_{j-1}^{(s)}, \dots, \hat{\alpha}_{j-1}^{(s)}, \hat{\beta}^{(s)}, \hat{\gamma}^{(s)})$, where the function $\text{stack}(\cdot)$ stacks the vectors inside the brackets on top of one another in the order given. It can be seen from Assumptions A1–A3 of [42] that for the s^{th} candidate model, there exists a limit $\theta^{(s)*}$ such that

$$\hat{\theta}^{(s)} - \theta^{(s)*} = O_p(n^{-1/2}). \tag{A4}$$

Let $p_{ij}^{(s)*} = \hat{p}_{ij}^{(s)} |_{\hat{\theta}^{(s)} = \theta^{(s)*}}$ and $\xi_n = \inf_{w \in \mathcal{W}} \sum_{i=1}^n \sum_{j=1}^J \left(\sum_{s=1}^{2q} w_s p_{ij}^{(s)*} - p_{ij} \right)^2$.

The proof of Equation (14) requires the following technical conditions:

Condition 1: $\inf \xi_n^{-1} n^{1/2} = o(1)$.

Condition 2: For any s , $\partial \hat{p}_{ij}^{(s)} / \partial \hat{\theta}^{(s)} |_{\hat{\theta}^{(s)} = \hat{\theta}_i^{(s)}} = O_p(1)$ uniformly for $i = 1, \dots, n$ and any $\tilde{\theta}_i^{(s)}$ that lies between $\hat{\theta}^{(s)}$ and its limit.

Condition 1 requires that as the sample size n increases, the minimum limiting squared error expands at a faster rate than $n^{-1/2}$. Condition 2 requires $\hat{p}_{ij}^{(s)}$ to have uniformly bounded derivatives. In layman’s terms, under Condition 1, all candidate models are misspecified and at best approximations to the data generating process; under Condition 2, $\hat{p}_{ij}^{(s)}$ are smooth with respect to the parameters.

Let $\hat{p}_{ij}(w) = \sum_{s=1}^{2^q} w_s \hat{p}_{ij}^{(s)}$, ${}^{(-i)}\hat{p}_{ij}(w) = \sum_{s=1}^{2^q} w_s {}^{(-i)}\hat{p}_{ij}^{(s)}$, $p_{ij}^*(w) = \sum_{s=1}^{2^q} w_s p_{ij}^{(s*)}$, $\hat{p}(w) = (\hat{p}_{11}(w), \dots, \hat{p}_{nJ}(w))'$, $\tilde{p}(w) = ({}^{(-1)}\hat{p}_{11}(w), \dots, {}^{(-n)}\hat{p}_{nJ}(w))'$, $p^*(w) = (p_{11}^*(w), \dots, p_{nJ}^*(w))'$, $p = (p_{11}, \dots, p_{nJ})'$, and $\mathbb{I} = (\mathbb{I}(Y_1 = 1), \mathbb{I}(Y_1 = 2), \dots, \mathbb{I}(Y_n = J))'$. We have

$$\begin{aligned} CV_J(w) &= \|\tilde{p}(w) - \mathbb{I}\|^2 \\ &= \|\hat{p}(w) - p + \tilde{p}(w) - p^*(w) - (\hat{p}(w) - p^*(w)) + p - \mathbb{I}\|^2 \\ &\leq \|\hat{p}(w) - p\|^2 + \|\tilde{p}(w) - p^*(w)\|^2 + \|\hat{p}(w) - p^*(w)\|^2 \\ &\quad + \|\hat{p}(w) - p^*(w)\| \|\tilde{p}(w) - p^*(w)\| + \|p^*(w) - p\| \|\tilde{p}(w) - p^*(w)\| \\ &\quad + \|\hat{p}(w) - p^*(w)\| \|\hat{p}(w) - p^*(w)\| + \|p^*(w) - p\| \|\hat{p}(w) - p^*(w)\| \\ &\quad + \|\hat{p}(w) - p^*(w)\| \|p - \mathbb{I}\| + |(p^*(w) - p)'(p - \mathbb{I})| + \|\tilde{p}(w) - p^*(w)\| \|\hat{p}(w) - p^*(w)\| \\ &\quad + \|\tilde{p}(w) - p^*(w)\| \|p - \mathbb{I}\| + \|\hat{p}(w) - p^*(w)\| \|p - \mathbb{I}\| + \|p - \mathbb{I}\|^2 \\ &\equiv \|\hat{p}(w) - p\|^2 + \Pi_n(w) + \|p - \mathbb{I}\|^2 \end{aligned}$$

and

$$\begin{aligned} \|\hat{p}(w) - p\|^2 &= \|\hat{p}(w) - p^*(w) + p^*(w) - p\|^2 \\ &= \|p^*(w) - p\|^2 + \|\hat{p}(w) - p^*(w)\|^2 + 2(\hat{p}(w) - p^*(w))'(p^*(w) - p) \\ &\equiv \|p^*(w) - p\|^2 + \Xi_n(w). \end{aligned}$$

To prove Equation (14), we need only to verify that

$$\sup_{w \in \mathcal{W}} \frac{\Xi_n(w)}{\|p^*(w) - p\|^2} = o_p(1) \quad \text{and} \quad \sup_{w \in \mathcal{W}} \frac{\Pi_n(w)}{\|p^*(w) - p\|^2} = o_p(1). \tag{A5}$$

Now, for any $\delta > 0$,

$$\begin{aligned} \Pr \left\{ \sup_{w \in \mathcal{W}} \xi_n^{-1} |(p^*(w) - p)'(p - \mathbb{I})| > \delta \right\} &\leq \Pr \left\{ \sup_{w \in \mathcal{W}} \xi_n^{-1} \sum_{s=1}^{2^q} w_s |(p_m^* - p)'(p - \mathbb{I})| > \delta \right\} \\ &= \Pr \left\{ \max_s |(p_s^* - p)'(p - \mathbb{I})| > \xi_n \delta \right\} \leq \sum_{s=1}^{2^q} \Pr \{ |(p_s^* - p)'(p - \mathbb{I})| > \xi_n \delta \} \\ &\leq \xi_n^{-2} \delta^{-2} \sum_{s=1}^{2^q} E\{(p_s^* - p)'(p - \mathbb{I})\}^2. \end{aligned}$$

Together with Condition 1 and recognising that $0 \leq p_{ij} \leq 1$, $0 \leq p_{ij}^{(s)*} \leq 1$ and J is fixed, this implies that

$$\sup_{w \in \mathcal{W}} \frac{|(p^*(w) - p)'(p - \mathbb{I})|}{\|p^*(w) - p\|^2} = o_p(1). \tag{A6}$$

Now, by Equation (A.4) and Condition 2, we have

$$\sup_{w \in \mathcal{W}} \|\hat{p}(w) - p^*(w)\|^2 = O_p(1) \quad \text{and} \quad \sup_{w \in \mathcal{W}} \|\tilde{p}(w) - p^*(w)\|^2 = O_p(1). \tag{A7}$$

In addition,

$$\begin{aligned} \|p - \mathbb{I}\|^2 &= O_p(n) \quad \text{and} \quad \sup_{w \in \mathcal{W}} \frac{\|p^*(w) - p\| \|\hat{p}(w) - p^*(w)\|}{\|p^*(w) - p\|^2} \\ &= \sup_{w \in \mathcal{W}} \frac{\|\hat{p}(w) - p^*(w)\|}{\|p^*(w) - p\|}. \end{aligned} \tag{A8}$$

Combining Equations (A.6)–(A.8) and Condition 1, we obtain Equation (A.5), which leads to Equation (14).