香港城市大學
City University of Hong Kong

Research Centre for International Economics

Working Paper: 2013054

Title of Paper

Frequentist Model Averaging for Multinomial and Ordered Logit Models

Authors' List

Alan T.K. Wan, City University of Hong Kong
Xinyu Zhang and Shouyang Wang, Chinese Academy of Sciences

Abstract

Multinomial and ordered Logit models are quantitative techniques used in many disciplines nowadays. When applying these techniques, practitioners usually select a single model by information-based criteria or pretesting. In this paper, we consider the alternative strategy of combining models instead of selecting a single model. Our strategy of weight choice for the candidate models is based on the minimization of a plug-in estimator of the asymptotic squared error risk of the model average estimator. Theoretical justifications of this model averaging strategy are provided, and a Monte Carlo study shows that forecasts produced by the proposed strategy are often more accurate than those produced by other common model selection and model averaging strategies, especially when the regressors are only mildly to moderately correlated and the true model contains few zero coefficients. An empirical example based on credit rating data is used to illustrate the proposed method. To reduce the computational burden we also consider a model screening step that eliminates some of the very poor models before averaging.

Frequentist Model Averaging for Multinomial and Ordered Logit Models

ALAN T.K. WAN[1], XINYU ZHANG[2] AND SHOUYANG WANG[2]

[1] Department of Management Sciences, City University of Hong Kong, Kowloon, Hong Kong

[2] Center for Forecasting Science, Chinese Academy of Sciences, Beijing 100190, China

October 8, 2013

Abstract

Multinomial and ordered Logit models are quantitative techniques used in many disciplines nowadays. When applying these techniques, practitioners usually select a single model by information-based criteria or pretesting. In this paper, we consider the alternative strategy of combining models instead of selecting a single model. Our strategy of weight choice for the candidate models is based on the minimization of a plug-in estimator of the asymptotic squared error risk of the model average estimator. Theoretical justifications of this model averaging strategy are provided, and a Monte Carlo study shows that forecasts produced by the proposed strategy are often more accurate than those produced by other common model selection and model averaging strategies, especially when the regressors are only mildly to moderately correlated and the true model contains few zero coefficients. An empirical example based on credit rating data is used to illustrate the proposed method. To reduce the computational burden we also consider a model screening step that eliminates some of the very poor models before averaging.

Keywords and phrases: asymptotic squared error risk; local mis-specification; model screening; Monte Carlo; plug-in estimator

JEL classifications: C51; C52

1 Introduction

In the past two decades there has been a substantial deal of interest on modeling and forecasting using discrete choice models such as the Logit and Probit regression models. These models are now commonplace tools for studying brand choice and consumer satisfaction in marketing research. They are also used frequently in other fields including biomedicine, economics and sociology. A long-standing practice in much of regression analysis, whatever the functional form of the underlying model, is that a multitude of models, each involving a different combination of regressors, are tried until a model with all the favourable statistical measures of performance is found. Variable searching is considered necessary in practice because there is almost always a long list of variables to consider. Preliminary testing procedures, by which a regressor variable is either dropped or retained based on the outcome of a hypothesis test, and information criteria-based model selection strategies, whereby each candidate model is given a certain information score, are routinely applied for selecting regressor variables in practice. The popular "general-to-specific" econometric modeling methodology (Hendry and Richard, 1982) also involves extensive use of pretesting and model selection strategies. Typically, after arriving at the final model, the researcher would report standard errors of the estimates, construct confidence intervals of the unknowns, and conduct hypothesis tests on the basis of this final model as if it were known all along.

An often raised criticism of model selection is the lack of explicit recognition and understanding of the effects of the model uncertainty on any inferences made. That is, once a model is chosen it is used as if there were no randomization concerning the choice, and the results are treated as though they are unconditional as well. A disturbing effect is that reported variance estimates are smaller than what they really should be, resulting in over-optimistic confidence intervals of the unknowns. Furthermore, being discontinuous functions of the data, pretest and post-model selection estimators are well-known to have very poor sampling properties (Judge and Bock, 1984; Danilov and Magnus, 2004; Leeb and Pötscher, 2008). Instability is another major drawback of model selection. It is well-known that when ranking models by an information criterion, a small perturbation in the data can often alter the ranking, which in turn alters model selection. This problem is particularly serious when an ample sampling of observations is not available. Consequently, the variability of forecasts produced by this model selection strategy can often be very high.

An alternative procedure that offers promise for incorporating model selection uncertainty and reducing prediction errors is model averaging. This latter approach smoothes across a set of candidate models, and by so doing takes into account the uncertainty, and alleviates the instability associated with selecting a single model. The final estimator is a weighted combination of estimators from each model. Bayesian model averaging (BMA) has long proven to be a very successful tool, and has given rise to a large body of literature over the past two decades. Hoeting, Madigan, Raftery and Volinsky (1999) provided a review of BMA. For recent applications of BMA in conjunction with Logit or Probit regressions, see Viallefont, Raftery and Richardson (2001), Hobcraft and Sigle-Rushton (2005), and Burda, Harding and Hausman (2008). While we do not review the extensive collection of BMA literature, we draw attention here to the fact that BMA also has disadvantages. In particular, the necessity of assigning prior probabilities to individual models, which is often done in an ad hoc manner, holds the potential for generating too many conflicting prior probabilities when applying multiple models to a single parameter. This disadvantage is one factor that has led to the development of model averaging informed by frequentist considerations. A large part of this literature is concerned with ways of weighting models. Unlike BMA where models are usually weighted by their posterior model probabilities, the method of determining weights for frequentist model averaging (FMA) is a more intricate issue. Many of the FMA weighting strategies are formed using scores of information criteria. The studies of Buckland, Burnham and Augustin (1997), Claeskens, Croux and van Kerckhoven (2006), Zhang and Liang (2011), and Zhang, Wan and Zhou (2012) fall into this category. Other FMA strategies that have been developed include adaptive regression mixing by Yang (2001), Mallows model averaging (MMA) by Hansen (2007, 2008) and Wan, Zhang and Zou (2010), optimal mean square error averaging by Liang, Zou, Wan and Zhang (2011), and Jackknife model averaging (JMA) by Hansen and Racine (2012) and Zhang, Wan and Zou (2013). While the majority of this literature focuses on averaging estimators in the context of the linear regression model, FMA strategies have also been developed for the binary logit model (Claeskens, Croux and van Kerckhoven, 2006), the hazard regression model (Hjort and Claeskens, 2006), the partially linear semi-parametric model (Zhang and Liang, 2011; Wang, Zou and Wan, 2012), and the censored regression model (Zhang, Wan and Zhou, 2012).

This article develops an FMA strategy for the multinomial and ordered Logit models, with an eye to using this strategy in forecasting. The multinomial and ordered Logit models are widely used for marketing research data. Guadagni and Little (1983, reprinted in 2008) is arguably the best known study in brand choice using the multinomial Logit model, and Katahira's (1990) method of constructing perceptual map based on the ordered Logit model is widely considered to be seminal in the marketing literature. Recent papers in marketing research involving the multinomial and/or ordered Logit models include Brangule-Vlagsma, Pieters and Wedel (2002), Bodapati and Drolet (2005), Mantrala, Seetharaman, Kaul, Gopalakrishna and Stam (2006), Fiebig, Keane, Louviere, Wasi (2010), among others. We propose a FMA method that selects the model weights by minimizing a plug-in estimator of the asymptotic squared error risk of the model average estimator. This FMA method is similar in spirit to the method of Liang, Zou, Wan and Zhang (2011) (referred to as LZWZ hereafter), but there is one important technical difference being that the latter method selects weights by minimizing an approximately unbiased estimator of the asymptotic risk, whereas our method minimizes a plug-in estimator of the asymptotic risk. We show that our proposed FMA method is approximately optimal asymptotically under the local mis-specification set-up (Hjort and Claeskens, 2003). To date, the literature is virtually devoid of any optimality theorem of FMA methods applicable to discrete choice models; to the best of our knowledge, our results are the first theoretical results for model averaging with an explicit emphasis on Logit models. Model averaging for the related binary Logit model was considered by Claeskens, Croux and van Kerckhoven (2006), but their weight choice method was based on information criterion scores and they gave no theoretical justification for their method. Although LZWZ's method is applicable to Logit models, their proof of asymptotic optimality is limited to linear models. Similarly, the optimality theorems established for the MMA and JMA methods are valid only for linear estimators. These existing analyses might be extended to the Logit models, but the extent to which theoretical results may be forthcoming is likely to be limited in view of the fact that maximum likelihood (ML) estimators of Logit coefficients are non-linearly related to the response variable.

In addition to providing a theoretical justification for our proposed method, we demonstrate through a Monte Carlo study that gains in forecast efficiency in small samples can frequently result by adopting the proposed method over the LZWZ and other information criterion-based FMA and model selection methods for the two types of Logit models considered. In particular, our results show that the advantages offered by the proposed method are most pronounced when the regressors are mildly to moderately correlated and the true model contains few zero coefficients. While the method described here is illustrated in terms of the multinomial and ordered Logit models, the asymptotic theory of the method applies to any model subject to the regularity conditions within the local mis-specification set-up. To reduce the computational burden we also consider an information criterion-based model screening step that removes some of the very poor models prior to averaging.

The remainder of this article is structured in the following way. In Section 2, we introduce the notations and describe the local mis-specification set-up and the multinomial and ordered Logit models. In Section 3, we introduce the proposed FMA strategy and establish its asymptotic properties. Results of a Monte Carlo study that investigates the forecasting performance of the proposed method are reported in Section 4. This is followed by a real data application in Section 5. We offer our conclusions in Section 6, and provide the proof of the main theorem in an Appendix.

2 Notations, framework and the choice models

2.1 Notations and the local mis-specification framework

Let $Y_1, ..., Y_n$ be i.i.d. observations generated from the density f. The narrow and extended models take the form $f(y, \theta)$ and $f(y, \theta, \gamma)$ respectively, where θ and γ are unknown vectors of dimensions $p \times 1$ and $q \times 1$. When γ is known and equal to γ_0, $f(y, \theta) = f(y, \theta, \gamma_0)$. In the setting of local mis-specification (Hjort and Claeskens, 2003), the true density is specified to be

$$f_{\text{true}}(y) = f(y, \theta, \gamma) = f(y, \theta, \gamma_0 + \delta/\sqrt{n}), \tag{1}$$

where δ is a $q \times 1$ unknown vector that represents the extent to which a model deviates from the narrow model. For the models described in Subsections 2.2 and 2.3, γ_0 is equal to zero. The crucial assumption of the local mis-specification described in (1) is that γ, and hence the true model, depends on the sample size, and that the effects of γ decrease as n grows, eventually vanishing as n approaches infinity. Although there have been debates concerning the realism of the local mis-specification set-up (Hjort and Claeskens, 2003a; Raftery and Zheng, 2003), this set-up is nevertheless very plausible. Technically, it has the advantage of yielding exchangeable quantities of squared bias and variance, both of order $O(n^{-1})$. This latter property greatly facilitates the derivation of precise limiting distribution results (Hjort and Claeskens, 2003).

When considering the submodels for selection and averaging, we assume that all submodels contain θ, but each model can have some or all of the elements of γ restricted to 0. Thus, there are 2^q submodels to consider, each corresponding to the subset $S \subset \{1, ..., q\}$ such that $\delta_j = 0$ for $j \in S^c$, the complement of S. We let $\widehat{\theta}_s$ and $\widehat{\gamma}_s$ be the ML estimators of θ and γ in the s-th submodel that corresponds to the subset S. Note that some of the elements of $\widehat{\gamma}_s$ would be 0 by default if there were no corresponding elements of γ in the s-th submodel. We define the full model as the submodel that contains all q elements of γ in addition to θ. The narrow model that contains only θ is also known as the null model.

We will apply the above set-up to develop a FMA weighting strategy and establish an asymptotic theory for the FMA estimator resulting from this strategy in Section 3. In the remaining parts of the current section we will describe the multinominal and ordered Logit models.

2.2 Multinomial Logit model

Consider a general discrete choice model with n independent individuals, denoted by subscript i, and J nominal alternatives denoted by subscript j, numbered from 1 to J. Let Y_i be the choice made by individual i. Thus, $Y_i = j$ if individual i selects alternative j. The usual assumption leading to the multinomial Logit model is that the log odds of category j relative to the reference category is determined by a linear combination of regressor variables. Analogous to the set-up in Subsection 2.1, we categorize the regressor variables as either mandatory or optional, represented by X_i and Z_i respectively; the mandatory regressors by definition are those that must be included, while the optional regressors can be excluded in any given model. The choice probabilities for the

i-th individual may then be written as

$$\begin{cases} p_{ij} = P(Y_i = j|X_i, Z_i) = \frac{\exp(\alpha_j + X_i'\beta_j + Z_i'\gamma_j)}{1+\sum_{l=1}^{J-1}\exp(\alpha_l + X_i'\beta_l + Z_i'\gamma_l)} & for\ j = 1, \dots, J-1, \\ p_{iJ} = P(Y_i = J|X_i, Z_i) = \frac{1}{1+\sum_{l=1}^{J-1}\exp(\alpha_l + X_i'\beta_l + Z_i'\gamma_l)}, & \end{cases} \tag{2}$$

where $(\alpha_1, \beta_1', \dots, \alpha_{J-1}, \beta_{J-1}')'$ and $(\gamma_1', \dots, \gamma_{J-1}')'$ correspond to θ and γ of Subsection 2.1 respectively, and we set γ_0 to zero. Note that the division of regressors into X_i and Z_i induces no loss of generality as X_i can be an empty set. Here, the J-th category is designated as the reference category. The unknown parameters are usually estimated by ML, with the solution obtained by an iterative procedure such as the Newton-Raphson algorithm. It is straightforward to see that the sum of the p_{il}'s over the J alternatives is one for any individual i.

Let $\widehat{\alpha}_1^{(s)}, \dots, \widehat{\alpha}_{J-1}^{(s)}$, $\widehat{\beta}_1^{(s)}, \dots, \widehat{\beta}_{J-1}^{(s)}$, and $\widehat{\gamma}_1^{(s)}, \dots, \widehat{\gamma}_{J-1}^{(s)}$ be the ML estimators of the unknown parameters in the s-th submodel; some elements of $\widehat{\gamma}_1^{(s)}, \dots, \widehat{\gamma}_{J-1}^{(s)}$ would be zero by default if the corresponding variables in Z_i are excluded from the s-th submodel. Common approaches to model selection in the multinomial Logit model include model deviance and information criteria-based methods such as the AIC and BIC. Now, let (X_0, Z_0) be the regressor variables for a new individual with an unknown response Y_0. The predicted choice probabilities for this individual based on the s-th submodel are

$$\begin{cases} \widehat{p}_{0j}^{(s)} = \widehat{P}(Y_0 = j|X_0, Z_0) = \frac{\exp(\widehat{\alpha}_j^{(s)} + X_0'\widehat{\beta}_j^{(s)} + Z_0'\widehat{\gamma}_j^{(s)})}{1+\sum_{l=1}^{J-1}\exp(\widehat{\alpha}_l^{(s)} + X_0'\widehat{\beta}_l^{(s)} + Z_0'\widehat{\gamma}_l^{(s)})} & for\ j = 1, \dots, J-1, \\ \widehat{p}_{0J}^{(s)} = \widehat{P}(Y_0 = J|X_0, Z_0) = \frac{1}{1+\sum_{l=1}^{J-1}\exp(\widehat{\alpha}_l^{(s)} + X_0'\widehat{\beta}_l^{(s)} + Z_0'\widehat{\gamma}_l^{(s)})}. & \end{cases} \tag{3}$$

2.3 Ordered Logit model

The multinomial Logit model assumes no ordering in the response categories, and that the results are impervious to changes in their order. If the response categories are ranked, say, from "least favored" to "most favored", then a more appropriate model framework to adopt is the ordered Logit model which is usually described in terms of cumulative probabilities. Write $F_{ij} = \sum_{l=1}^{j} p_{il}$ as the cumulative probability that the individual i chooses a response category lower than or equal to j, and let the log odds be determined by $\log\left[F_{ij}/(1-F_{ij})\right] = \alpha_j + X_i'\beta + Z_i'\gamma$, $j = 1, \dots, J-1$. Then we have

$$\begin{cases} P(Y_i \le j|X_i, Z_i) = \frac{\exp(\alpha_j + X_i'\beta + Z_i'\gamma)}{1+\exp(\alpha_j + X_i'\beta + Z_i'\gamma)} & for\ j = 1, \dots, J-1, \\ P(Y_i \le J|X_i, Z_i) = 1, & \end{cases} \tag{4}$$

where the intercept coefficient α_j varies across the different equations, but the slope coefficients of the regressor variables are common for all equations. Note that $(\alpha_1, \dots, \alpha_{J-1}, \beta')'$ and γ in equation (4) correspond to θ and γ in Subsection 2.1 respectively.

The cumulative probabilities provide a basis for working out the probability of selecting a particular category; for example, for the individual with X_0 and Z_0 as regressor variables, based on the s-th submodel, the probability of selecting the j-th category is calculated to be

$$\begin{cases} \widehat{p}_{0j}^{(s)} = \widehat{P}(Y_0 \leq j|X_0, Z_0) - \widehat{P}(Y_0 \leq j-1|X_0, Z_0) \\ \quad = \frac{\exp(\widehat{\alpha}_j^{(s)} + X_0'\widehat{\beta}^{(s)} + Z_0'\widehat{\gamma}^{(s)})}{1+\exp(\widehat{\alpha}_j^{(s)} + X_0'\widehat{\beta}^{(s)} + Z_0'\widehat{\gamma}^{(s)})} - \frac{\exp(\widehat{\alpha}_{j-1}^{(s)} + X_0'\widehat{\beta}^{(s)} + Z_0'\widehat{\gamma}^{(s)})}{1+\exp(\widehat{\alpha}_{j-1}^{(s)} + X_0'\widehat{\beta}^{(s)} + Z_0'\widehat{\gamma}^{(s)})} \quad for\ j = 1, \ldots, J-1, \\ \widehat{p}_{0J}^{(s)} = 1 - \widehat{P}(Y_0 \leq J-1|X_0, Z_0) \\ \quad = 1 - \frac{\exp(\widehat{\alpha}_{J-1}^{(s)} + X_0'\widehat{\beta}^{(s)} + Z_0'\widehat{\gamma}^{(s)})}{1+\exp(\widehat{\alpha}_{J-1}^{(s)} + X_0'\widehat{\beta}^{(s)} + Z_0'\widehat{\gamma}^{(s)})}, \end{cases} \tag{5}$$

where $\widehat{\alpha}_1^{(s)}, \ldots, \widehat{\alpha}_{J-1}^{(s)}$, $\widehat{\beta}^{(s)}$, and $\widehat{\gamma}^{(s)}$ are be ML estimators of the unknown parameters under the s-th submodel. Again, some of the $\widehat{\gamma}^{(s)}$ may be zero by default as not every variable in Z is contained in all submodels.

3 A theory of model averaging

3.1 A general strategy for parametric models

Here, we develop a general model averaging theory that is applicable to any parametric model setting that assumes local mis-specification, then illustrate this FMA strategy in the contexts of the multinomial and ordered Logit models. Now, in the setting of Subsection 2.1, let $\mathcal{L}(\theta, \gamma)$ be the likelihood function under the full model, $J_{n,\text{full}} = -\frac{1}{n}\frac{\partial^2 \log \mathcal{L}(\theta,\gamma)}{\partial(\theta',\gamma')'\partial(\theta',\gamma')} = \begin{pmatrix} J_{n,00} & J_{n,01} \\ J_{n,10} & J_{n,11} \end{pmatrix}$ be the corresponding $(p+q) \times (p+q)$ information matrix, $\begin{pmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{pmatrix}$ be the limiting information matrix, and J_{ij} the limiting value of $J_{n,ij}$, $i, j = 0, 1$. Unless otherwise stated, all limiting processes are with respect to $n \to \infty$. Denote π_s as the projection matrix mapping the vector $v = (v_1, ..., v_q)'$ to its subvector $\pi_s v = v_s$ that consists of v_j with $j \in S$.

Let $\mu = \mu(\theta, \gamma) = \mu(\theta, \gamma_0 + \delta/\sqrt{n})$ be the estimand of interest. The FMA estimator of μ is $\widehat{\mu}(w) = \sum_{s=1}^{2^q} w_s \widehat{\mu}_s$, where w_s's are the weights, $w = (w_1, \ldots, w_{2^q})'$, and $\widehat{\mu}_s$ is the ML estimator of μ in the s-th submodel. Write $K = (J_{11} - J_{10}J_{00}^{-1}J_{01})^{-1}$, $K_s = (\pi_s K^{-1} \pi_s')^{-1}$, $H_s = K^{-1/2}\pi_s' K_s \pi_s K^{-1/2}$, and $\omega = J_{10}J_{00}^{-1}\partial\mu/\partial\theta - \partial\mu/\partial\gamma$, with the partial derivatives evaluated at the null point (θ, γ_0). Define H_1 as the null matrix of size $q \times q$, and $\widehat{\delta}_{\text{full}} = \sqrt{n}(\widehat{\gamma}_{\text{full}} - \gamma_0)$. The following results are obtained using results of Hjort and Claeskens (2003):

$$\widehat{\delta}_{\text{full}} \xrightarrow{d} D \sim N_q(\delta, K), \tag{6}$$

$$\sqrt{n}(\widehat{\mu}_s - \mu) \xrightarrow{d} \Lambda_s \equiv \left(\frac{\partial \mu}{\partial \theta}\right)' J_{00}^{-1} M + \omega'(\delta - K^{1/2} H_s K^{-1/2} D), \tag{7}$$

and

$$\sqrt{n}(\widehat{\mu}(w) - \mu) \xrightarrow{d} \Lambda \equiv \left(\frac{\partial \mu}{\partial \theta}\right)' J_{00}^{-1} M + \omega'\{\delta - \widehat{\delta}(D)\}, \tag{8}$$

where "$\xrightarrow{d}$" denotes convergence in distribution, $M \sim N_p(0, J_{00})$ is independent of D, and $\widehat{\delta}(D) = K^{1/2}\left\{\sum_{s=1}^{2^q} w_s H_s\right\} K^{-1/2} D \equiv K^{1/2} H(w) K^{-1/2} D$. Thus, the asymptotic risk of $\widehat{\mu}(w)$ under squared errors is given by

$$\begin{aligned} R_a(\widehat{\mu}(w)) &= E(\Lambda^2) = \tau_0^2 + E\left(\omega'\widehat{\delta}(D) - \omega'\delta\right)^2 \\ &= \tau_0^2 + \omega' K^{1/2} H^2(w) K^{1/2}\omega + (\omega' K^{1/2} L(w) K^{-1/2}\delta)^2, \end{aligned} \tag{9}$$

where $\tau_0^2 = \left(\frac{\partial \mu}{\partial \theta}\right)' J_{00}^{-1} \left(\frac{\partial \mu}{\partial \theta}\right)$ and $L(w) = I_q - H(w)$. Our goal here is to seek w that minimizes $R_a(\widehat{\mu}(w))$, the asymptotic risk of $\widehat{\mu}(w)$.

Write $\mathcal{W} = \left\{w \in [0,1]^{2^q} : \sum_s w_s = 1\right\}$, a general weight set. We define the optimal weight vector as

$$w^{\text{opt}} = \underset{w \in \mathcal{W}}{\text{argmin}} R_a(\widehat{\mu}(w)). \tag{10}$$

Thus, the estimator $\widehat{\mu}(w^{\text{opt}})$ has the minimum asymptotic risk under squared errors among the class of estimators defined by $\widehat{\mu}(w)$. The problem with $\widehat{\mu}(w^{\text{opt}})$ is that it is infeasible because ω and K in $R_a(\widehat{\mu}(w))$ are unknown. A feasible version of $\widehat{\mu}(w^{\text{opt}})$ may be obtained by replacing these unknowns by their consistent estimators. A consistent estimator of K is $\widehat{K} \equiv (J_{n,11} - J_{n,10} J_{n,00}^{-1} J_{n,01})^{-1}$. Also, since the ML estimators $\widehat{\theta}_{\text{full}}$ and $\widehat{\gamma}_{\text{full}}$ based on the full model are consistent estimators of their respective unknowns, we can use $\widehat{\omega} = \omega \mid_{J_{\text{full}} = J_{n,\text{full}}, \theta = \widehat{\theta}_{\text{full}}, \gamma = \widehat{\gamma}_{\text{full}}}$ to estimate ω consistently. Let

$$\delta_n = E(\widehat{\delta}_{\text{full}}). \tag{11}$$

When $\widehat{\delta}_{\text{full}}$ is absolutely integrable, we have, from (6),

$$\delta_n \to \delta. \tag{12}$$

Now, only the second and third terms in the expression of $R_a(\widehat{\mu}(w))$ in (9) are related to w, and these two terms may be estimated by

$$A(w) = \widehat{\omega}' \widehat{K}^{1/2} \widehat{H}^2(w) \widehat{K}^{1/2} \widehat{\omega} + (\widehat{\omega}' \widehat{K}^{1/2} \widehat{L}(w) \widehat{K}^{-1/2} \delta_n)^2, \tag{13}$$

where $\widehat{H}(w)$ and $\widehat{L}(w)$ have the same expressions as $H(w)$ and $L(w)$ respectively, except that K is replaced by $\widehat{K}$ everywhere. Similarly, $\widehat{H}_s$ has the same expression as H_s, except that K is replaced by $\widehat{K}$ everywhere. Denote

$$\widehat{w}^{\text{opt}} = \underset{w \in \mathcal{W}}{\text{argmin}} A(w). \tag{14}$$

The following theorem shows that under some regularity conditions, $\widehat{w}^{\text{opt}}$ converges to the optimal weight vector w^{opt} in probability.

THEOREM 3.1 *When $n \to \infty$, provided that (12) holds and $R_a(\widehat{\mu}(w))$ has an identifiable unique*[1] *minimizer w^{opt} on $\mathcal{W}$, then*

$$\widehat{w}^{\text{opt}} \xrightarrow{p} w^{\text{opt}}. \tag{15}$$

Proof: See the Appendix.

Note that the computation of $A(w)$ requires knowledge of the unknown quantity δ_n. If we replace δ_n in $A(w)$ by an estimator $\widehat{\delta}_n$ such that $\widehat{\delta}_n - \delta_n \xrightarrow{p} 0$, then the weight vector that results from (14) still converges to w^{opt} in probability. However, it is difficult if not impossible to find the estimator $\widehat{\delta}_n$ under the assumption of local mis-specification. In view of (11), we suggest to estimate δ_n by its unbiased estimator $\widehat{\delta}_{\text{full}}$ obtained based on the full model. Let

$$\tilde{A}(w) = \widehat{\omega}'\widehat{K}^{1/2}\widehat{H}^2(w)\widehat{K}^{1/2}\widehat{\omega} + (\widehat{\omega}'\widehat{K}^{1/2}\widehat{L}(w)\widehat{K}^{-1/2}\widehat{\delta}_{\text{full}})^2 \tag{16}$$

be the objective function that results after replacing δ_n by $\widehat{\delta}_{\text{full}}$. The weight vector that minimizes $\tilde{A}(w)$ is

$$\tilde{w}^{\text{opt}} = \underset{w \in \mathcal{W}}{\text{argmin}} \tilde{A}(w). \tag{17}$$

This weight vector is a feasible version of $\widehat{w}^{\text{opt}}$. We propose to construct FMA estimators based on $\tilde{w}^{\text{opt}}$, and we call it the "approximately optimal" (A-opt) weight choice.

It is worth noting that this FMA strategy is similar in spirit to that proposed by LZWZ (2011), but there is one important technical difference: LZWZ (2011) selects w by minimizing an approximately unbiased estimator of the asymptotic risk (see formula (33) in their paper), whereas in the present paper we selects w by minimizing a plug-in estimator of the asymptotic risk (9).

Define Ψ as a $2^q \times 2^q$ matrix with $\Psi_{sr} = \widehat{\omega}'\widehat{K}^{1/2}\widehat{H}_s\widehat{H}_r\widehat{K}^{1/2}\widehat{\omega} + \widehat{\omega}'\widehat{K}^{1/2}(I_q - \widehat{H}_s)\widehat{K}^{-1/2}\widehat{\delta}_{\text{full}}\widehat{\omega}'\widehat{K}^{1/2}(I_q - \widehat{H}_r)\widehat{K}^{-1/2}\widehat{\delta}_{\text{full}}$ as its *sr*-th element. It is readily seen that $\tilde{A}(w) = w'\Psi w$. Thus, the minimization of $\tilde{A}(w)$ with respect to w is a quadratic programming problem. Computational routines available from various software packages (e.g., Matlab and SAS) can be used to obtain solutions to this problem, and they generally work effectively and efficiently even when 2^q is large.

[1] Readers may refer to Definition 3.3 of White (1994) for the definition of identifiable uniqueness.

3.2 Specialization to multinomial and ordered Logit models

For the multinomial Logit model (2), $\theta = (\alpha_1, \beta_1', \ldots, \alpha_{J-1}, \beta_{J-1}')'$ and $\gamma = (\gamma_1', \ldots, \gamma_{J-1}')'$. Let $\eta = (\eta_1', \ldots, \eta_{J-1}')' = (\alpha_1, \beta_1', \gamma_1', \ldots, \alpha_{J-1}, \beta_{J-1}', \gamma_{J-1}')'$ and Π be a project matrix such as $(\theta', \gamma')' = \Pi\eta$. Straightforward calculations show that for model (2),

$$\frac{\partial^2 \log \mathcal{L}(\theta,\gamma)}{\partial \eta_{j_1} \partial \eta_{j_2}'} = -\sum_{i=1}^{n} p_{ij_1} \left[I(j_1 = j_2) - p_{ij_2} \right] (1, X_i', Z_i')'(1, X_i', Z_i') \equiv \Xi_{j_1 j_2}, \tag{18}$$

where $I(\cdot)$ is the usual indicator function. Let Ξ be a matrix with $\Xi_{j_1 j_2}$ as its $j_1 j_2$-th block. Thus, for the multinomial Logit model (2), $J_{n,\text{full}} = -\frac{1}{n}\Pi\Xi\Pi'$. The unknowns p_{ij} in Ξ are estimated by the full model. Given $J_{n,\text{full}}$, Ψ_{sr} can be calculated directly using the procedure described in Subsection 3.1. Thus, $\tilde{w}^{\text{opt}}$ can be obtained by minimizing $w'\Psi w$. The *quadprog* function of Matlab can be utilized to solve this minimization problem.

The calculations and the steps involved are largely similar for the ordered Logit model (4), except that for this model,

$$\begin{aligned} J_{n,\text{full}} &= -\frac{1}{n}\frac{\partial^2 \log \mathcal{L}(\theta,\gamma)}{\partial \eta \partial \eta'} \\ &= -\frac{1}{n}\sum_{i=1}^{n} \left\{ p_{iy_i}^{-1} \left[(1 - 2C_{iy_i})\xi_{iy_i} h_{iy_i}' h_{iy_i} - (1 - 2C_{iy_i-1})\xi_{iy_i-1} h_{iy_i-1}' h_{iy_i-1} \right] \right. \\ &\quad \left. - p_{iy_i}^{-2} (\xi_{iy_i} h_{iy_i}' - \xi_{iy_i-1} h_{iy_i-1}')(\xi_{iy_i} h_{iy_i}' - \xi_{iy_i-1} h_{iy_i-1}')' \right\}, \end{aligned} \tag{19}$$

where $(y_1, \ldots, y_n)$ are realizations of $(Y_1, \ldots, Y_n)$, $p_{ij} = P(Y_i = j)$, $C_{ij} = P(Y_i \le j)$, $\xi_{ij} = C_{ij} - C_{ij}^2$, and $h_{ij} = (I(j = 1), \ldots, I(j = J-1), (X_i', Z_i')I(1 \le j < J))$. As in the case of the multinomial model, the unknowns p_{ij} in $J_{n,\text{full}}$ are estimated based on the full model.

4 A Monte Carlo study

In this section, an examination of the finite sample performance of the proposed model averaging strategy is undertaken in a number of Monte Carlo experiments with designs that include both the multinomial and ordered Logit models. Our study has the following specific objectives: i) compare the proposed A-opt weight choice model averaging scheme with some alternative FMA and common model selection schemes, and ii) examine the effects of the changing magnitude and sparsity level of non-zero coefficients on the various strategies' performance.

Included for comparison are post-model selection estimators based on AIC and BIC, and model average estimators based on the smoothed-Focused Information criterion (S-FIC) (Claeskens, Croux and van Kerckhoven, 2006), the optimal mean square error (o-MSE) criterion (LZWZ, 2011) and the equal weight criterion. The AIC and BIC are penalized versions of the attained log likelihood, and arguably the most widely applied model selection criteria in practice. The S-FIC strategy assigns the weight

$$\exp\{-FIC_s/(2\widehat{\omega}'\widehat{K}\widehat{\omega})\}/(\sum_{s^*}[\exp\{-FIC_{s^*}/(2\widehat{\omega}'\widehat{K}\widehat{\omega})\}])$$

to the s-th submodel, where FIC_s is the FIC score achieved by the s-th submodel. The FIC, introduced by Claeskens and Hjort (2003), is an approximately unbiased estimator of the asymptotic squared error risk of the unknown coefficient vector in the s-th submodel[2]. The o-MSE criterion, developed by LZWZ (2011), is based on the minimization of an approximately unbiased estimator of the asymptotic squared error risk of the FMA estimator [3]. The equal weighted model average simply assigns to each model a weight that equals the reciprocal of the total number of models contained in the average.

We consider two schemes for computing a model average. The first one combines all 2^q candidate models, whereas the second one combines only the subset of models that survive an initial screening step. Model screening has the advantage of narrowing down the array of models before combining and thus saving computing cost. Here, we adopt the "top m model screening procedure" (Yuan and Yang, 2009) that selects $m(< 2^q)$ leading models using a model selection criterion; specifically, it eliminates all but the m models with the smallest values of an information criterion. In our simulations, we use the BIC as the criterion for model elimination and inclusion and choose $m = 5$. Another model screening procedure that could be adopted is backward elimination (Claeskens, Croux and van Kerckhoven, 2006). One drawback of this latter procedure is that it always includes exactly one model of each size in the final set of models for averaging. This means even if the best model of a given size may be worse than the second best model of another size, the procedure will include the former model but exclude the latter.

The experimental designs of our Monte Carlo experiments can be summarized as follows:

Design 1: The responses are generated based on the set-up of a multinomial Logit model (2) with the following specifications: $J = 3$, $X_i = 0$ (i.e., no mandatory regressors), each of $(Z_{i1}, \ldots, Z_{i8})' \sim N(0, \Omega)$, where $\Omega = (\Omega_{ij})$ and $\Omega_{ij} = \rho^{|i-j|}$ for $i \neq j$ and ρ = 0, 0.3, and 0.6, $(\alpha_1, \alpha_2) = \kappa(0.3, 0.5)$, and γ_1 and γ_2 chosen according to the following scenarios:

`Scenario 1 :` $\gamma_1 = \kappa(1.4, 0.9, 1.3, 1.5, 1.5, 1.2, 0.9, 0)$; $\gamma_2 = \kappa(1.0, 1.2, 1.1, 0.9, 0.7, 1.1, 1.0, 0)$

`Scenario 2 :` $\gamma_1 = \kappa(1.4, 0.9, 1.3, 1.5, 1.5, 0, 0, 0)$; $\gamma_2 = \kappa(1.0, 1.2, 1.1, 0.9, 0.7, 0, 0, 0)$

[2]See equation (3.3) of Claeskens and Hjort (2003).

[3]See equation (33) of LZWZ (2011). This criterion allows for both fixed and random weights. In our computation, we assume that the weights are fixed.

$$\texttt{Scenario 3:} \qquad \gamma_1 = \kappa(1.4, 0.9, 0, 0, 0, 0, 0); \qquad \gamma_2 = \kappa(1.0, 1.2, 0, 0, 0, 0, 0)$$

The parameter κ is used to control the magnitude of the coefficients, and we let it vary in the set $\{0.5, 1, 2\}$. The three scenarios also represent different sparsity levels of non-zero coefficients. Under Scenario 1, the true model is almost the full model, and thus the majority of models in the model average are under-fitted. Scenario 3 contains many zero coefficients resulting in a large number of over-fitted models in the model average. Scenario 2 represents an intermediate scenario. With $q = 8$, there are $2^8 = 256$ submodels to combine. On the other hand, if the above mentioned top m model screening procedure is applied, then the model average only combines the $m = 5$ submodels that attain the smallest BIC values.

Design 2: This experimental design has the same specifications as the previous design, except that here we generate the p_{ij}s' based on the ordered Logit model in (4), and the following scenarios determine the choice of γ:

$$\begin{aligned}
\texttt{Scenario I:} &\quad \gamma = \kappa(1.0, 1.2, 0.9, 1.4, 1.1, 0.8, 0.9, 0) \\
\texttt{Scenario II:} &\quad \gamma = \kappa(1.0, 1.2, 0.9, 1.4, 1.1, 0, 0, 0) \\
\texttt{Scenario III:} &\quad \gamma = \kappa(1.0, 1.2, 0, 0, 0, 0, 0, 0)
\end{aligned}$$

All of our Monte Carlo simulations are based on 1000 replications. We generate 100 observations as training data and 10 observations as test data. Our objectives are to evaluate the accuracy of the out-of-sample forecasts produced by the coefficient estimates. We assess the accuracy of forecasts based on the mean squared error forecast error (MSFE):

$$\text{MSFE} = \frac{1}{10000} \sum_{r=1}^{1000} \sum_{t=1}^{10} \sum_{j=1}^{J} (\widehat{p}_{tj}^{[r]} - p_{tj}^{[r]})^2 \tag{20}$$

and the mean absolute forecast error (MAFE):

$$\text{MAFE} = \frac{1}{10000} \sum_{r=1}^{1000} \sum_{t=1}^{10} \sum_{j=1}^{J} |\widehat{p}_{tj}^{[r]} - p_{tj}^{[r]}|, \tag{21}$$

where $\widehat{p}_{tj}^{[r]}$ is the forecast of $p_{tj}^{[r]}$, the probability of the t-th test observation resulting in choice j for the r-th replication. Some representative results are shown in Tables 1 - 3. As the results based on the screening and non-screening versions of the model averages are quite similar, to conserve space, we choose to report only those based on the screening version. In the tables, AIC and BIC denote the two post-model selection estimators, and S-FIC, LZWZ, EW, A-opt and opt denote the FMA estimators based on S-FIC weighting, o-MSE weighting, equal weighting, our proposed $\tilde{w}^{\text{opt}}$ weight choice, and the (infeasible) optimal weight $\widehat{w}^{\text{opt}}$ respectively. The opt estimator is of no practical utility and used only as a benchmark for assessing other estimators' efficiency. To facilitate readability, the opt estimators' forecast errors are shown in brackets in all cases, and the best and worst estimators (excluding the infeasible opt estimator) in each case are flagged by a ”†” and a ”⋆” respectively. At the bottom of each table a summary is provided for the percentages of cases of the various estimators producing the best and worst forecasts, and the A-opt estimator yielding superior forecasts to the other strategies. We also apply the Morgan-Granger-Newbold (MGN) test (Granger and Newbold, 1977) to test for equal accuracy in forecasts between the A-opt and other methods; a forecast accuracy figure is highlighted in bold if it differs significantly from the corresponding A-opt figure at the 10% level.

The following conclusions may be drawn from the Monte Carlo results:

First, it can be seen that in the majority of cases forecasts produced by the four model averaging strategies are superior to forecasts obtained by model selection. Other things being equal, model averaging appears to work better when κ is small or moderate (Tables 1 and 2) than when it is large (Table 3). This result is not unexpected, because when κ is small, the non-zero coefficients in the true model are all close to zero, making it difficult to distinguish the truth from a false model that contains many zeros. As model selection criterion scores can be quite similar for different models, the choice of models becomes unstable. On the other hand, when κ is large, the absolute values of the non-zero coefficients are also large, and a model selection criterion can more readily identify a non-zero coefficient. This reduces the forecast variability of the post-model selection estimator. For example, when $\kappa = 0.5$ (Table 1), the worst forecast is invariably produced by either one of the two model selection strategies, but when $\kappa = 2$ (Table 3), the two model selection strategies together produce over half of the best forecasts across all cases in both MSFE and MAFE terms. These results reinforces the intuition that model averaging is more credible when the uncertainty in finding the best model is high, but less suitable when there is little instability in selection.

Second, for small to moderate κ (Tables 1 and 2), the proposed A-opt estimator most frequently delivers the best forecast. In most cases the A-opt estimator is preferred to the S-FIC and LZWZ averaging strategies. This is an encouraging finding given the merits of S-FIC and LZWZ demonstrated in other contexts. The bold figures in the tables indicate that in the majority of cases the differences in forecast performance between the A-opt and other methods are statistically significant at the 10% level. It also appears that the value of ρ which controls the degrees of regressor collinearity has some bearings on the performance of the A-opt forecast. There is generally a higher frequency of the A-opt estimators yielding better (worse) forecasts when ρ is small (large) than when it is large (small), suggesting that the advantages of the A-opt estimators may be stronger when the collinearity of regressors is small to moderate than when it is large. The advantages of the A-opt estimator are also more pronounced under Scenarios 1 and 2 where the true model contains few zero coefficients than Scenario 3 where the true model contains only a small number of non-zero coefficients. Under the latter scenario the BIC estimator which favors parsimony exhibits frequent empirical superiority over other strategies, especially when κ is large.

Third, the results show that the simple equal weighted model average is often a strong competitor that out-compete other strategies. This strong showing of the EW averaging scheme is a surprising feature of our results. It is usually under Scenario 3 that EW outperforms the proposed A-opt method. Interestingly, even for $\kappa = 2$ (Table 3) where selection is arguably the preferred strategy, EW has the ability to produce the best forecast and outperforms the A-opt estimator in a large number of cases, although the latter estimator has an overall advantage when other values of κ are also considered.

Fourth, all of the above comments regarding the relative merits and shortcomings of the various strategies apply to both the multinomial and ordered Logit models. We do not observe any major difference in the pattern of the Monte Carlo results under Designs 1 and 2. Although we do not report the results here, the non-screening versions of the various model average estimators generally exhibit behavior very similar to their screening counterparts.

5 An empirical application

In this section, we consider an application of the proposed model averaging strategy to real data. The dataset, taken from Compustat, and used by Ashbaugh-Skaife, Collins and LaFond (2006) and Verbeek (2007), contains observations of Standard and Poor's credit ratings of 921 U.S. firms in 2005. The ratings range from AAA (highest rating) to D (lowest rating). We analyze this dataset by the binary and ordered Logit models. The binary model is a special case of both the multinomial and ordered Logit models.

The dependent variable used in our binary Logit analysis has the value of 1 if the firm's rating is above BB+ (investment-grade rating), and 0 otherwise (speculative-grade rating). The following explanatory variables are available: working capital of the firm (wc), which proxies the firm's short-term liquidity; retained earnings (re) and earnings before interest and taxes (ebit), which proxy historical profitability and current profitability respectively; book leverage (bl), the ratio of the firm's debt to assets; and log sales volume (ls) which proxies the firm's size. We scale the first three of these variables by total assets (ta) in our analysis. All of the explanatory variables are treated as optional. This results in $2^5 = 32$ binary Logit submodels. As in the Monte Carlo study, we set m, the number of models to be retained after model screening, to 5. We estimate the models using the first 460 observations according to the sequence listed in Verbeek (2007), and use the remaining 461 observations for forecast evaluation purpose. The predicted value of an observation is 1 (0) if the predicted probability score of the observation taking on 1 is greater (smaller) than 0.5. We evaluate the forecasts by the hit-rate, obtained by dividing the number of correct predictions by the the size of the evaluation sample.

Panel I of Table 4 presents the coefficient estimates produced by the AIC and BIC model selection and the screening versions of the various model averaging methods. The AIC and BIC methods yield the same coefficient estimates as both methods select the full model that contains all five explanatory variables. The four model averaging methods also produce very similar estimates between themselves and to those obtained by model selection. Panel II of Table 4 shows that all methods perform well in terms of out-of-sample hit-rates, with the EW method having a slight edge over its competitors.

For the ordered Logit analysis, we use a dependent variable with seven categories indexed by integer values ranging from 1 (lowest credit rating) to 7 (highest credit rating), and the same independent variables as in the binary Logit analysis. Again, we treat all explanatory variables as optional, estimate the model based on the first 460 observations and evaluate forecasts using the last 461 observations. An observation has a predicted value of j ($1 \leq j \leq 7$) if the predicted probability of the observation taking on j is the highest among the seven predicted probability values. Table 5 reports the estimates of coefficients and the hit-rates, where intercept-j's, $j = 1, .., 6$, denote the intercepts of the first six equations as described in (5). All six methods yield similar estimates, but the A-opt and LZWZ methods produce the most accurate out-of-sample hit-rates. The deterioration in hit-rates for all methods relative to those observed under binary Logit analysis is expected due to the much larger grouping of choice categories in the ordered Logit analysis.

6 Conclusions

Model averaging has advantages over model selection in that it guards against the selection of a very poor model. These advantages hold the potential to produce estimates and forecasts that improve those obtained by model selection. In this paper we have developed a FMA weight choice criterion by minimizing a plug-in estimator of the FMA estimator's asymptotic risk. Our proposed method shares the spirit of a similar method devised by Liang, Zou, Wan and Zhang (2011), but there are also important differences as discussed in Section 3. Although this paper focuses on the multinomial and ordered Logit model, the proposed method can be applied to any parametric model. We have proved that the proposed method has an approximate asymptotic optimality property under the local mis-specification set-up. Our Monte Carlo results demonstrate that the method frequently delivers more accurate forecasts than other model model selection and averaging methods; the superiority of the proposed method is most marked when there is small to moderate collinearity among regressors and high uncertainty in identifying the best model.

One surprising feature of our Monte Carlo results is the strong showing of the simple equal weighted average estimator. Bates and Granger (1969) showed that when all forecasts are uncorrelated and have identical variances, the equal weighting method has an optimal property. While this result is not directly applicable to the present context, the frequent empirical superiority exhibited by the equal weighted estimator should perhaps reinvigorate thinking about how best to combine estimators in general. This remains for future research.

Also, the bulk of this paper only addresses issues relating to the efficiency of estimators and forecasts obtained from model averaging. By comparison, little attention has been paid to matters of inference. Our work in progress explores the inferential aspects of model averaging in the context of the types of Logit models being analyzed in this paper, as well as extending the analysis to other discrete choice models including the nested Logit and ordered Probit models.

Appendix: Proof of Theorem 3.1

Let $A^o(w) = \omega' K^{1/2} H^2(w) K^{1/2}\omega + (\omega' K^{1/2} L(w) K^{-1/2}\delta)^2$. By (10) and recognising that τ_0^2 is unrelated to w, we have

$$w^{\text{opt}} = \underset{w\in\mathcal{W}}{\text{argmin}} A^o(w). \tag{A.1}$$

Let Φ^o and Φ_n be $2^q \times 2^q$ matrices with their respective *sr*-th elements being

$$\Phi^o_{sr} = \omega' K^{1/2} H_s H_r K^{1/2}\omega + \omega' K^{1/2}(I_q - H_s)K^{-1/2}\delta\omega' K^{1/2}(I_q - H_r)K^{-1/2}\delta$$

and

$$\Phi_{n,sr} = \widehat{\omega}'\widehat{K}^{1/2}\widehat{H}_s\widehat{H}_r\widehat{K}^{1/2}\widehat{\omega} + \widehat{\omega}'\widehat{K}^{1/2}(I_q - \widehat{H}_s)\widehat{K}^{-1/2}\delta_n\widehat{\omega}'\widehat{K}^{1/2}(I_q - \widehat{H}_r)\widehat{K}^{-1/2}\delta_n.$$

Recognising that $\theta_{\text{full}} \xrightarrow{p} \theta$, $\gamma_{\text{full}} \xrightarrow{p} \gamma$, $J_{n,\text{full}} \to J_{\text{full}}$, and $\delta_n \to \delta$, we obtain

$$\begin{aligned}\sup_{w\in\mathcal{W}}(A(w) - A^o(w)) &= \sup_{w\in\mathcal{W}}\sum_{s=1}^{2^q}\sum_{r=1}^{2^q} w_s w_r(\Phi_{n,sr} - \Phi^o_{sr}) \\ &\le \sup_{w\in\mathcal{W}}\sum_{r=1}^{2^q}|\Phi_{n,sr} - \Phi^o_{sr}| = \sum_{r=1}^{2^q}|\Phi_{n,sr} - \Phi^o_{sr}| = o_p(1).\end{aligned} \tag{A.2}$$

Hence

$$A(w) \xrightarrow{p} A^o(w) \tag{A.3}$$

uniformly for $w \in \mathcal{W}$. Now, from (14), (A.1), (A.3), the identifiable uniqueness stated in Theorem 3.1, and Theorem 3.4 of White (1994), we obtain $\widehat{w} \xrightarrow{p} w^{\text{opt}}$.

Acknowledgements

The authors thank the editor Graham Elliot, the associate editor and two referees for very helpful comments and suggestions. The usual disclaimer applies.

References

Ashbaugh-Skaife, H., Collins, D. W. and LaFond, R. (2006). The effects of corporate governance on firms' credit ratings. *Journal of Accounting and Economics* 42, 203-243.

Bates, J. and Granger, C. (1969). The combination of forecasts. *Operations Research Quarterly* 20, 451-468.

Bodapati, A.V. and Drolet, A. (2005). A hybrid choice model that uses actual and ordered attribute value information. *Journal of Marketing Research* 42, 256-265.

Brangule-Vlagsma, K., Pieters, R.G.M. and Wedel, M. (2002). The dynamics of value segments: modeling framework and empirical illustration. *International Journal of Marketing* 19, 267-285.

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

Burda, M., Harding, M. and Hausman, J. (2008). A Bayesian mixed logit-probit model for multinomial choice. *Journal of Econometrics* 147, 232-246.

Claeskens, G., Croux, C. and Van Kerckhoven, J. (2006). Variable selection for logit regression using a prediction-focused information criterion. *Biometrics* 62, 972-979.

Claeskens, G. and Hjort, N.L. (2003). The focused information criterion. *Journal of the American Statistical Association* 98, 900-916.

Danilov, D. and Magnus, J.R. (2004). On the harm that ignoring pretesting can cause. *Journal of Econometrics* 122, 27-46.

Diebold, F.X. and Mariano, R.S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics* 13, 253-263.

Fiebig, D.G., Keane, M.P., Louviere, J. and Wasi, N. (2010). The generalized multinomial logit model: accounting for scale and coefficient heterogeneity. *Marketing Science* 29, 393-421.

Granger, C.W.J. and Newbold, P. (1977). *Forecasting Economic Time Series*. Academic Press: Orlando, U.S.A.

Guadagni, P.M. and Little, J.D.C. (1983). A logit model of brand choice calibrated on scanner data. *Marketing Science* 2, 203-238 (reprinted in 2008 in *Marketing Science* 27, 29-48).

Hansen, B.E. (2007). Least squares model averaging. *Econometrica* 75, 1175-1189.

Hansen, B.E. (2008). Least squares forecast averaging. *Journal of Econometrics* 146, 342-350.

Hansen, B.E. and Racine, J.S. (2012). Jackknife model averaging. *Journal of Econometrics* 167, 38-46.

Harvey, D. and Leybourne, S. and Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting* 13, 281-291.

Hendry, D.F. and Richard, J.F. (1982). On the formulation of empirical models in dynamic econometrics. *Journal of Econometrics* 20, 3-33.

Hjort, N.L. and Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association* 98, 879-899.

Hjort, N.L. and Claeskens, G. (2003a). Rejoinder to ”Frequentist model average estimators” and ”The focused information criterion”, *Journal of the American Statistical Association* 98, 938-945.

Hjort, N.L. and Claeskens, G. (2006). Focused information criteria and model averaging for the Cox hazard regression model. *Journal of the American Statistical Association* 110, 1449-1464.

Hobcraft, J. and Sigle-Rushton, W. (2005). An exploration of childhood antecedents of female adult malaise in two British birth cohorts: combining Bayesian model averaging and recursive partitioning. CASE paper: 95, Centre for Analysis of Social Exclusion, London School of Economics and Political Science.

Hoeting, J.A., Madigan, D., Raftery, A.E. and Volinsky, C.T. (1999). Bayesian model averaging: a tutorial. *Statistical Science* 14, 382-417.

Judge, G.G. and Bock, M.E. (1978). *The Statistical Implications of Pre-Test and Stein-Rule Estimators in Econometrics.* North Holland: Amsterdam.

Katahira, H. (1990). Perceptual mapping using ordered logit analysis. *Marketing Science* 9, 1-17.

Leeb, H. and Pötscher, B.M. (2008). Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory* 24, 338-376.

Liang H, Zou, G., Wan A.T.K. and Zhang X. (2011). Optimal weight choice for frequentist model average estimators. *Journal of the American Statistical Association* 106, 1053-1066.

Mantrala, M., Seetharaman, P.B., Kaul, R., Gopalakrishna, S. and Stam, A. (2006). Optimal pricing strategies for an automotive aftermarket retailer. *Journal of Marketing Research* 43, 588-604.

Raftery, A.E. and Zheng, Y. (2003). Discussion of ”Frequentist model average estimators” and ”Focused information criterion”, *Journal of the American Statistical Association* 98, 931-938.

Verbeek, M. (2007). *A Guide to Modern Econometrics*, 3rd edition. Wiley: Chichester, U.K.

Viallefont, V., Raftery, A.E. and Richardson, S. (2001). Variable selection and Bayesian model averaging in epidemiological case-control studies. *Statistics in Medicine* 20, 3215-3230.

Wan, A.T.K., Zhang, X. and Zou, G. (2010). Least squares model averaging by Mallows criterion. *Journal of Econometrics* 156, 277-283.

Wang, H., Zou, G. and Wan, A.T.K. (2012). Model averaging for varying-coefficient partially linear measurement errors models. *Electronic Journal of Statistics* 6, 1017-1039.

White, H. (1994). *Estimation, Inference and Specification Analysis*. Cambridge University Press: Cambridge, U.K.

Yang, Y. (2001). Adaptive regression by mixing. *Journal of the American Statistical Association* 96, 574-588.

Yuan, Z. and Yang, Y. (2005). Combining linear regression models: when and how? *Journal of the American Statistical Association* 100, 1202-1204.

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

Zhang, X., Wan, A.T.K. and Zhou, Z. (2012). Focused information criteria, model selection and model averaging in a Tobit model with a non-zero threshold. *Journal of Business and Economic Statistics* 30, 132-142.

Zhang, X., Wan, A.T.K. and Zou, G. (2013). Model averaging by Jackknife criterion in models with dependent data. *Journal of Econometrics* 174, 82-94.

Table 1: $\kappa = 0.5$

		MSFE							MAFE						
ρ	Scenario	AIC	BIC	S-FIC	LZWZ	EW	A-opt	(opt)	AIC	BIC	S-FIC	LZWZ	EW	A-opt	(opt)
		Design 1													
0	1	**0.061**	**0.075⋆**	**0.061**	0.055	**0.069**	0.055†	(**0.037**)	**0.316**	**0.344⋆**	**0.313**	**0.299**	**0.339**	0.297†	(**0.237**)
	2	**0.058**	**0.063⋆**	**0.048**	**0.047**	**0.052**	0.046†	(**0.029**)	**0.309**	**0.323⋆**	**0.285**	**0.283**	**0.300**	0.280†	(**0.213**)
	3	**0.045⋆**	0.030	**0.029**	0.037	**0.027†**	0.037	(**0.017**)	**0.271⋆**	0.224	**0.223**	0.251	**0.217†**	0.251	(**0.168**)
		Design 2													
	1	**0.057**	**0.085⋆**	**0.056**	**0.045**	**0.065**	0.043†	(**0.033**)	**0.262**	**0.336⋆**	**0.257**	**0.230**	**0.282**	0.225†	(**0.197**)
	2	0.052	**0.069⋆**	**0.051**	**0.044**	**0.058**	0.043†	(**0.034**)	0.253	**0.303⋆**	**0.256**	**0.237**	**0.276**	0.232†	(**0.200**)
	3	**0.030**	**0.033⋆**	0.027†	0.028	0.028	0.028	(**0.017**)	**0.199**	**0.212⋆**	0.197	**0.196**	**0.203**	0.195†	(**0.152**)
		Design 1													
0.3	1	**0.073**	**0.089⋆**	**0.066**	**0.065**	**0.072**	0.063†	(**0.042**)	**0.339**	**0.381⋆**	**0.327**	**0.324**	**0.345**	0.318†	(**0.250**)
	2	**0.062**	**0.065⋆**	**0.048†**	**0.051**	0.049	0.050	(**0.030**)	**0.314**	**0.325⋆**	**0.278†**	**0.289**	0.283	0.285	(**0.214**)
	3	**0.043⋆**	**0.032**	**0.028**	0.037	**0.026†**	0.037	(**0.017**)	**0.268⋆**	**0.234**	**0.218**	0.253	**0.211†**	0.253	(**0.169**)
		Design 2													
	1	**0.040**	**0.058⋆**	**0.044**	**0.035**	**0.052**	0.034†	(**0.027**)	**0.215**	**0.269⋆**	**0.227**	**0.202**	**0.249**	0.197†	(**0.175**)
	2	**0.041**	**0.059⋆**	**0.041**	**0.037**	**0.045**	0.036†	(**0.026**)	**0.226**	**0.275⋆**	**0.227**	**0.214**	**0.241**	0.209†	(**0.181**)
	3	**0.027**	**0.032⋆**	**0.024**	0.026	**0.023†**	0.026	(**0.014**)	0.186	**0.208⋆**	**0.181†**	0.185	0.181	0.185	(**0.140**)
		Design 1													
0.6	1	**0.080**	**0.082⋆**	**0.061†**	**0.070**	**0.063**	0.067	(**0.044**)	**0.355**	**0.364⋆**	**0.315†**	**0.335**	**0.322**	0.327	(**0.259**)
	2	**0.060⋆**	**0.059**	**0.042**	**0.052**	**0.041†**	0.050	(**0.029**)	**0.317**	**0.318⋆**	**0.268**	**0.298**	**0.264†**	0.292	(**0.220**)
	3	**0.048⋆**	**0.037**	**0.030**	**0.041**	**0.027†**	0.041	(**0.017**)	**0.292⋆**	**0.250**	**0.228**	**0.269**	**0.219†**	0.271	(**0.173**)
		Design 2													
	1	**0.038**	**0.046⋆**	**0.041**	**0.037**	**0.045**	0.036†	(**0.024**)	**0.207**	**0.234⋆**	**0.214**	**0.202**	**0.225**	0.197†	(**0.159**)
	2	**0.038**	**0.041⋆**	**0.036**	**0.035**	**0.038**	0.034†	(**0.021**)	**0.206**	**0.215⋆**	**0.199**	**0.197**	**0.204**	0.191†	(**0.152**)
	3	**0.038⋆**	**0.032**	**0.030**	0.035	**0.030†**	0.035	(**0.017**)	**0.214⋆**	**0.204**	**0.200**	**0.213**	**0.199†**	0.211	(**0.152**)
	[1]	0	0	17	0	33	50	N.A.	0	0	17	0	28	56	N.A.
	[2]	28	72	0	0	0	0	N.A.	22	78	0	0	0	0	N.A.
	[3]	100	78	50	78	56	N.A.	N.A.	100	78	56	83	56	N.A.	N.A.

Notes:
[1]: Percentage of cases with the best forecast
[2]: Percentage of cases with the worst forecast
[3]: Percentage of cases with an inferior forecast to A-opt

Table 2: $\kappa = 1$

		MSFE							MAFE						
ρ	Scenario	AIC	BIC	S-FIC	LZWZ	EW	A-opt	(opt)	AIC	BIC	S-FIC	LZWZ	EW	A-opt	(opt)
		Design 1													
0	1	**0.075**	**0.100**⋆	**0.077**	**0.077**	**0.085**	0.072†	(**0.054**)	**0.308**	**0.363**⋆	**0.316**	**0.315**	**0.338**	0.304†	(**0.256**)
	2	**0.062**	0.083⋆	**0.057**†	**0.065**	0.060	0.062	(**0.043**)	**0.272**	0.323⋆	**0.268**†	**0.289**	0.277	0.281	(**0.228**)
	3	0.044⋆	**0.041**	**0.030**	**0.039**	**0.028**†	0.040	(**0.017**)	**0.259**⋆	**0.244**	**0.221**	**0.254**	**0.217**†	0.259	(**0.172**)
		Design 2													
	1	**0.037**	**0.049**	**0.043**	**0.037**	**0.049**⋆	0.035†	(**0.029**)	**0.191**	**0.221**	**0.209**	**0.193**	**0.228**⋆	0.188†	(**0.172**)
	2	0.028†	0.032⋆	0.029	**0.030**	0.030	0.029	(**0.019**)	**0.163**†	0.173⋆	0.166	0.167	**0.172**	0.166	(**0.140**)
	3	0.024⋆	**0.015**†	**0.017**	**0.022**	**0.016**	0.024	(**0.013**)	0.168	**0.143**	**0.143**	**0.164**	**0.140**†	0.171⋆	(**0.132**)
		Design 1													
0.3	1	**0.077**	**0.096**⋆	**0.071**	**0.073**	**0.077**	0.069†	(**0.052**)	**0.300**	**0.346**⋆	**0.299**	**0.302**	**0.314**	0.293†	(**0.243**)
	2	**0.057**†	**0.082**⋆	**0.057**	**0.065**	0.059	0.062	(**0.044**)	**0.260**†	**0.316**⋆	**0.267**	**0.283**	0.274	0.276	(**0.224**)
	3	0.038	**0.040**⋆	**0.028**	**0.039**	**0.026**†	0.039	(**0.017**)	**0.245**	**0.243**	**0.215**	**0.253**	**0.210**†	0.256⋆	(**0.170**)
		Design 2													
	1	0.037	**0.063**⋆	**0.033**	**0.030**	**0.037**	0.028†	(**0.023**)	**0.171**	**0.236**⋆	**0.174**	**0.163**	**0.194**	0.156†	(**0.141**)
	2	0.038†	**0.045**⋆	**0.041**	0.040	**0.044**	0.039	(**0.026**)	0.178	**0.194**	**0.186**	**0.178**	**0.196**⋆	0.175†	(**0.147**)
	3	0.028⋆	**0.020**	**0.019**	**0.024**	**0.018**†	0.025	(**0.015**)	0.178⋆	**0.147**†	**0.152**	**0.171**	**0.149**	0.176	(**0.135**)
		Design 1													
0.6	1	**0.099**	**0.113**⋆	0.080†	**0.085**	0.083	0.082	(**0.054**)	0.352	**0.373**⋆	0.319†	**0.329**	0.329	0.323	(**0.257**)
	2	**0.052**†	**0.069**⋆	**0.053**	**0.060**	**0.053**	0.058	(**0.037**)	**0.262**†	**0.303**⋆	**0.266**	**0.284**	**0.269**	0.278	(**0.216**)
	3	**0.049**⋆	**0.030**†	**0.035**	0.047	**0.032**	0.047	(**0.022**)	0.278	**0.216**†	**0.239**	0.281⋆	**0.232**	0.280	(**0.193**)
		Design 2													
	1	**0.039**	**0.049**⋆	**0.036**	**0.033**	**0.039**	0.031†	(**0.026**)	**0.157**	**0.188**⋆	**0.165**	**0.152**	**0.177**	0.147†	(**0.132**)
	2	0.034†	**0.042**⋆	0.037	**0.038**	0.039	0.036	(**0.023**)	0.153†	**0.181**⋆	**0.161**	**0.158**	**0.167**	0.154	(**0.130**)
	3	**0.021**	**0.016**†	**0.019**	0.023	**0.018**	0.023⋆	(**0.010**)	**0.157**	**0.133**†	**0.151**	**0.164**	**0.150**	0.167⋆	(**0.111**)
	[1]	28	17	11	0	17	28	N.A.	22	17	11	0	17	33	N.A.
	[2]	22	67	0	0	6	6	N.A.	11	56	0	6	11	17	N.A.
	[3]	61	78	39	72	50	N.A.	N.A.	50	67	39	72	50	N.A.	N.A.

Notes:

[1]: Percentage of cases with the best forecast

[2]: Percentage of cases with the worst forecast

[3]: Percentage of cases with an inferior forecast to A-opt

Table 3: $\kappa = 2$

		MSFE							MAFE						
ρ	Scenario	AIC	BIC	S-FIC	LZWZ	EW	A-opt	(opt)	AIC	BIC	S-FIC	LZWZ	EW	A-opt	(opt)
		Design 1													
0	1	**0.090**	**0.114⋆**	**0.085†**	**0.105**	0.090	0.093	(**0.088**)	**0.268†**	**0.317⋆**	**0.273**	**0.301**	0.289	0.284	(**0.276**)
	2	**0.070**	**0.066**	**0.064†**	0.079⋆	**0.064**	0.078	(**0.050**)	**0.250**	**0.246**	**0.239†**	0.264	**0.240**	0.265⋆	(**0.215**)
	3	**0.039⋆**	**0.023**	**0.022**	**0.032**	**0.020†**	0.034	(**0.015**)	**0.226⋆**	**0.166**	**0.170**	**0.203**	**0.164†**	0.211	(**0.151**)
		Design 2													
	1	**0.040†**	**0.040†**	0.041	**0.044**	**0.046⋆**	0.042	(**0.030**)	**0.155**	**0.152†**	0.167	**0.169**	**0.189⋆**	0.166	(**0.144**)
	2	0.030⋆	**0.025**	**0.024**	**0.027**	**0.023†**	0.028	(**0.022**)	0.143⋆	**0.132**	**0.130†**	**0.137**	**0.130**	0.140	(**0.125**)
	3	0.032⋆	**0.018†**	**0.020**	**0.025**	**0.018**	0.028	(**0.016**)	0.171⋆	**0.129†**	**0.139**	**0.154**	**0.135**	0.162	(**0.124**)
		Design 1													
0.3	1	0.090	**0.113⋆**	**0.085†**	**0.100**	0.088	0.091	(**0.079**)	**0.268†**	**0.317⋆**	**0.270**	**0.290**	0.279	0.277	(**0.257**)
	2	0.078	**0.076**	**0.073**	**0.089⋆**	**0.072†**	0.087	(**0.056**)	**0.263**	**0.261**	**0.253†**	**0.280⋆**	**0.254**	0.277	(**0.222**)
	3	**0.039⋆**	**0.024**	**0.023**	**0.033**	**0.021†**	0.035	(**0.017**)	**0.222⋆**	**0.169**	**0.171**	**0.204**	**0.164†**	0.213	(**0.156**)
		Design 2													
	1	0.033†	**0.050⋆**	**0.042**	**0.041**	**0.046**	0.039	(**0.034**)	**0.136†**	**0.157**	**0.157**	**0.153**	**0.172⋆**	0.149	(**0.140**)
	2	0.028	**0.023†**	**0.025**	0.029	**0.026**	0.030⋆	(**0.022**)	0.127	**0.116†**	**0.118**	0.127	**0.122**	0.129⋆	(**0.115**)
	3	0.024⋆	**0.014†**	**0.016**	**0.021**	**0.015**	0.022	(**0.011**)	0.145⋆	**0.116†**	**0.121**	**0.136**	**0.117**	0.141	(**0.106**)
		Design 1													
0.6	1	0.083†	**0.105⋆**	0.087	**0.095**	0.087	0.090	(**0.075**)	0.252†	**0.292⋆**	0.274	**0.283**	0.276	0.275	(**0.246**)
	2	**0.080**	**0.066†**	**0.070**	**0.085⋆**	**0.071**	0.079	(**0.049**)	**0.258**	**0.232†**	**0.252**	**0.275⋆**	0.258	0.264	(**0.207**)
	3	**0.038⋆**	**0.020†**	**0.025**	0.037	**0.023**	0.037	(**0.016**)	**0.219**	**0.162†**	**0.183**	**0.217**	**0.181**	0.220⋆	(**0.150**)
		Design 2													
	1	0.052	**0.070⋆**	**0.051**	**0.044**	**0.055**	0.041†	(**0.038**)	0.142	**0.183⋆**	**0.160**	**0.140**	**0.174**	0.135†	(0.132)
	2	0.027†	**0.030**	0.029	**0.031**	**0.031⋆**	0.029	(**0.020**)	**0.105†**	**0.115**	**0.118**	**0.117**	**0.128⋆**	0.113	(**0.097**)
	3	**0.020⋆**	**0.013**	**0.013**	**0.016**	**0.013†**	0.017	(**0.010**)	**0.129⋆**	**0.105†**	**0.108**	**0.116**	**0.106**	0.121	(**0.096**)
	[1]	22	33	17	0	28	6	N.A.	28	39	17	0	11	6	N.A.
	[2]	39	28	0	17	11	6	N.A.	33	22	0	11	17	17	N.A.
	[3]	50	33	11	56	22	N.A.	N.A.	39	33	22	50	39	N.A.	N.A.

Notes:
[1]: Percentage of cases with the best forecast
[2]: Percentage of cases with the worst forecast
[3]: Percentage of cases with an inferior forecast to A-opt

Table 4: Binary Logit analysis for credit rating application

	AIC	BIC	S-FIC	LZWZ	EW	A-opt
	Panel I: estimates of coefficients					
intercept	-6.924	-6.924	-7.726	-7.284	-8.378	-7.265
bl	-4.497	-4.497	-4.290	-4.177	-2.384	-4.198
ebit/ta	6.778	6.778	6.705	6.177	5.352	6.226
ls	0.939	0.939	1.019	0.946	1.023	0.946
re/ta	3.863	3.863	4.018	4.104	4.000	4.081
wc/ta	-4.375	-4.375	-3.971	-4.143	-2.297	-4.155
	Panel II: Out-of-sample hit-rates					
	0.824	0.824	0.824	0.833	0.839	0.829

Table 5: Ordered Logit models for credit rating application

	AIC	BIC	S-FIC	LZWZ	EW	A-opt
	Panel I: estimates of coefficients					
intercept-1	-1.814	-1.814	-1.614	-1.693	-0.194	-1.700
intercept-2	3.655	3.655	3.780	3.779	4.891	3.768
intercept-3	6.127	6.127	6.326	6.284	7.273	6.271
intercept-4	8.449	8.449	8.742	8.637	9.537	8.624
intercept-5	11.549	11.549	11.947	11.785	12.620	11.772
intercept-6	13.374	13.374	13.820	13.619	14.419	13.614
bl	3.226	3.226	3.184	3.247	1.727	3.258
ebit/ta	-6.322	-6.322	-6.125	-6.020	-4.892	-6.034
ls	-0.793	-0.793	-0.852	-0.798	-0.868	-0.798
re/ta	-3.746	-3.746	-3.874	-3.859	-3.908	-3.852
wc/ta	3.279	3.279	3.051	3.144	1.738	3.150
	Panel II: out-of-sample hit-rates					
	0.479	0.479	0.484	0.503	0.484	0.503