Statistica Sinica **26** (2016), 809-840
doi:http://dx.doi.org/10.5705/ss.202014.0067

MINIMUM MEAN SQUARED ERROR MODEL AVERAGING IN LIKELIHOOD MODELS

Ali Charkhi[1], Gerda Claeskens[1] and Bruce E. Hansen[2]

[1]*KU Leuven and* [2]*University of Wisconsin, Madison*

Abstract: A data-driven method for frequentist model averaging weight choice is developed for general likelihood models. We propose to estimate the weights which minimize an estimator of the mean squared error of a weighted estimator in a local misspecification framework. We find that in general there is not a unique set of such weights, meaning that predictions from multiple model averaging estimators might not be identical. This holds in both the univariate and multivariate case. However, we show that a unique set of empirical weights is obtained if the candidate models are appropriately restricted. In particular a suitable class of models are the so-called singleton models where each model only includes one parameter from the candidate set. This restriction results in a drastic reduction in the computational cost of model averaging weight selection relative to methods which include weights for all possible parameter subsets. We investigate the performance of our methods in both linear models and generalized linear models, and illustrate the methods in two empirical applications.

Key words and phrases: Frequentist model averaging, likelihood regression, local misspecification, mean squared error, weight choice.

1. Introduction

We study a focused version of frequentist model averaging where the mean squared error plays a central role. Suppose we have a collection of models $S \in \mathcal{S}$ to estimate a population quantity μ, this is the focus, leading to a set of estimators $\{\hat{\mu}_S : S \in \mathcal{S}\}$. The focus can be vector-valued. In this paper we study properties of the weight choice for constructing a combined, weighted, or aggregated, estimator

$$\hat{\mu}_w = \sum_{S \in \mathcal{S}} w_S \hat{\mu}_S. \tag{1.1}$$

Focused model selection (FIC, Claeskens and Hjort (2003)) assigns a single weight $\hat{w}_{S^*} = 1$ to the estimator for which the estimated mean squared error (MSE) is the smallest amongst all considered models, $\widehat{\mathrm{MSE}}(\hat{\mu}_{S^*}) = \min_{S \in \mathcal{S}} \widehat{\mathrm{MSE}}(\hat{\mu}_S)$, and $\hat{w}_S = 0$ for all other $S \in \mathcal{S}$. Due to the estimation of the MSE (the true model is unknown, hence unavailable for use in MSE computations), the collection of weights $\hat{w}_S$ is random. Similar random selection results from using any

other information criterion such as the Akaike information criterion (AIC, Akaike (1973)), the Bayesian information criterion (BIC, Schwarz (1978)) and Mallows' C_p (Mallows (1973)).

Small fluctuations in the data may cause the weights indicating the single best model to change from one to zero and vice versa. For this reason model averaging with weights outside the values $\{0, 1\}$ are considered as a more stable compromise. This paper concentrates on frequentist model averaging in a likelihood setting. For an overview of model averaging in a Bayesian framework see Hoeting et al. (1999).

Weight selection methods for regression models estimated via least squares include the Mallows' criterion for determining the weights to be used in model averaging for nested models (Hansen (2007)) and its extension to non-nested models (Wan, Zhang, and Zou (2010)). Hansen and Racine (2012) defined a jackknife model averaging estimator for heteroskedastic errors and showed the optimality of that model averaged estimator. Model averaging in econometrics is often used for improving forecast accuracy (Bates and Granger (1969); Granger and Ramanathan (1984); Hansen (2008)). For a further literature overview, see Cheng and Hansen (2015).

Liang et al. (2011) proposed to select the weights such that the estimated MSE of the weighted estimator $\hat{\mu}_w$ is minimal. In that paper, their 'optimal' set of weights for frequentist model-averaged estimators is, however, restricted to a specific *ad hoc* parametric form. They used their method for least squares estimation only, but explain that it could be extended to maximum likelihood estimation. For linear regression models with heteroscedastic errors Liu (2015) proposed a model averaging estimator in a local asymptotic framework and derived the asymptotic distribution of the so-called plug-in averaging estimator based on the asymptotic mean squared error expression. Multinomial regression was considered by Wan, Zhang, and Wang (2013) by minimizing a plug-in estimator of the asymptotic mean squared error for defining the weights.

In this paper we consider estimators obtained by maximum likelihood estimation in general. First, we propose an estimator of the mean squared error of $\widehat{\mu}_w$ under local misspecification, replacing the unknown localizing parameters by their plug-in estimators. We then propose selecting the weights that minimize this estimator of the MSE. This method can be considered as an extension of Liu (2015) to likelihood models. We also extend the approach of Liang et al. (2011) as we do not restrict the empirical weights to have a certain parametric form nor to lie in the unit simplex, although we impose that the sum of weights is equal to one as is necessary for consistency of the model averaging estimator (Hjort and Claeskens (2003)). In absence of imposing inequality restrictions, unlike other weight selection methods, no quadratic programming or nonlinear optimization

is required. When the aim of the model averaging is to improve estimation efficiency as compared to using a single models estimator, the interpretation of the separate weights is not of direct interest. By not restricting the weights to be between zero and one, more flexibility is allowed in the construction of the weighted estimator, and thus there is the possibility for reduced MSE.

A second part of this paper entails a study of the set of models $\mathcal{S}$ for which we can assign unique weights to the corresponding estimators. Perhaps surprisingly, it turns out that most of the studied weight selection methods result in a non-unique set of weights. This can be problematic when interpreting the weight values. We show that there are multiple weight vectors that yield equal model average predictions in linear regression using different sets of models. It is therefore sufficient to restrict attention to a subset of such models for which we can construct a unique MSE-minimizing weight vector. It turns out that one convenient choice is the class of singleton models, dramatically reducing the number of models to estimate. For example, if there are q candidate parameters for inclusion, then there are 2^q models consisting of all possible subsets of q parameters, but only $q+1$ singleton models (the baseline or narrow model which fixes all q parameters and the q models which each include only a single parameter) that suffice for model averaging.

Section 2 introduces notation, defines the local asymptotic framework and the asymptotic mean squared error. Estimators for the MSE are constructed and discussed in Section 3. Section 4 contains an extension of the weight selection method for vector-valued focuses. Simulations and data examples are given in Sections 5 and 6. Section 7 concludes.

2. Notation and Setting

Consider a likelihood regression model where it is uncertain which regression variables should best be included for the estimation of a population quantity μ. Different 'configurations' of covariates lead to define different models. A local misspecification setting avoids making the strong assumption that the true model is contained in the set of considered models. Take $\{Y_i; i = 1, \ldots, n\}$ independent with the ith density $f_n(y; x_i) = f(y; x_i, \theta_0, \gamma_0 + \delta/\sqrt{n})$, where the p-vector θ is included in every model and is not subject to selection. Components of the q-vector γ may or may not be relevant, these are subject to variable selection. The vectors θ and γ are non-overlapping. The true values of the parameter vector (θ, γ) are $(\theta_0, \gamma_0 + \delta/\sqrt{n})$ under the local misspecification setting, and (θ_0, γ_0) under a narrow model where the vector γ_0 is completely specified and known. For example, when γ represents regression coefficients, typically $\gamma_0 = 0$ in the narrow model, the smallest model one is willing to consider, indicating absence of the extra regression coefficients. The full model includes all q components of γ,

other models are indexed by a set $S \subset \{1, \ldots, q\}$. The narrow model corresponds to $S = \emptyset$. Since our method of finding weights is based on minimizing a mean squared error expression, see also Liang et al. (2011), this setting is justified since it balances the squared bias and the variance of the estimators in order for the mean squared error to be computable. Indeed, when not working under local misspecification, for a fixed true model not contained in the set of studied models asymptotically the bias would dominate, pointing toward always working with the most complicated model (Claeskens and Hjort (2008)).

In a regression setting the response values are typically not identically distributed due to the presence of the covariate vector x_i. We define the score vector, the vector of first derivatives of the log-likelihood,

$$\begin{pmatrix} U_\theta(y;x) \\ U_\gamma(y;x) \end{pmatrix} = \begin{pmatrix} \frac{\partial \log f(y;x,\theta_0,\gamma_0)}{\partial \theta} \\ \frac{\partial \log f(y;x,\theta_0,\gamma_0)}{\partial \gamma} \end{pmatrix},$$

and let the Fisher information matrix

$$J(x) = \mathrm{Var} \begin{pmatrix} U_\theta(Y;x) \\ U_\gamma(Y;x) \end{pmatrix} \text{ and } J_n = \frac{1}{n} \sum_{i=1}^{n} J(x_i),$$

be partitioned according to p and q, the lengths of θ and γ, as

$$J(x) = \begin{pmatrix} J_{00}(x) & J_{01}(x) \\ J_{10}(x) & J_{11}(x) \end{pmatrix}, \; J_n = \begin{pmatrix} J_{n,00} & J_{n,01} \\ J_{n,10} & J_{n,11} \end{pmatrix}, \; J_n^{-1} = \begin{pmatrix} J_n^{00} & J_n^{01} \\ J_n^{10} & J_n^{11} \end{pmatrix}.$$

We assume that J_n converges to an invertible matrix J when $n \to \infty$. Submatrices of J and J^{-1} are defined as above, though without using the subscript n.

The purpose of the model averaging procedure is to estimate a population focus $\mu = \mu(\theta, \gamma)$. Examples include a prediction of the response given covariate values (a forecast), a quantile of the response distribution, as well as a single coefficient of interest. Working with a population focus is more general than the commonly studied averaging of the regression coefficients. We assume that the first derivatives of μ with respect to θ and γ exist in a neighborhood of (θ_0, γ_0).

Maximum likelihood estimation is used in each submodel indexed by S. Under classical assumptions, each such submodel estimator of (θ, γ), and hence of the focus μ, has an asymptotically normal distribution with both the mean and the variance specific to the used model. Let $\widehat{\mu}_S = \mu(\widehat{\theta}_S, \widehat{\gamma}_S)$. Following the notation of Claeskens and Hjort (2008), it holds that

$$\sqrt{n}(\widehat{\mu}_S - \mu_{\text{true}}) \xrightarrow{d} \Lambda_S = \Lambda_0 + \nu^t(\delta - G_S D) \sim N(\text{mean}_S, \text{Var}_S), \tag{2.1}$$

where $\Lambda_0 \sim N(0, \tau_0^2)$ with the narrow model's variance $\tau_0^2 = (\partial\mu/\partial\theta)^t J_{00}^{-1} \partial\mu/\partial\theta$, the vector $\nu = J_{10}J_{00}^{-1}\partial\mu/\partial\theta - \partial\mu/\partial\gamma$, $D \sim N_q(\delta, Q)$ with $Q = J^{11}$. Further, define a $|S| \times q$ projection matrix π_S that selects those rows with an index belonging to S and let $Q_S = (\pi_S Q^{-1} \pi_S^t)^{-1}$, $Q_S^0 = \pi_S^t Q_S \pi_S$, and $G_S = Q_S^0 Q^{-1}$. We denote by I_q an identity matrix of dimension $q \times q$. By adding the squared bias and the variance, the asymptotic distribution in (2.1) implies that the mean squared error (MSE) of a single estimator $\widehat{\mu}_S$ converges to

$$\mathrm{MSE}(\widehat{\mu}_S, \delta) = \tau_0^2 + \nu^t Q_S^0 \nu + \nu^t (I_q - G_S)\delta\delta^t (I_q - G_S)^t \nu. \tag{2.2}$$

While selecting a model based on the estimator's estimated MSE value is the idea underlying the focused information criterion (Claeskens and Hjort (2003)), we here consider the choice of the weights via the mean squared error, similar as Liu (2015) and Liang et al. (2011), though for general likelihood estimation and a general choice of weights summing to one.

3. Estimation of the Mean Squared Error

3.1. Weight choice via minimum mean squared error

Rather than working with the estimator in a single model, we consider a finite set of M different models, this number not depending on the sample size. A weight is assigned to the estimator in each of the considered models to reach the model averaged estimator $\widehat{\mu}_w$ in (1.1).

Some common possibilities of sets of models to average over are (i) all possible subsets, these are $M = 2^q$ models, with q the length of the parameter vector γ; this is currently the most common construction for model averaging. (ii) A sequence of 'nested' models. Here one starts with a model with a single variable, then adds a second one, etc. Thus $S_1 = \{1\} \subset S_2 = \{1, 2\} \subset \ldots \subset S_q = \{1, \ldots, q\}$. When including also the narrow model, containing only θ and none of the components of γ, this leads to $M = q + 1$ models that depend on the order of inclusion of the variables. (iii) A collection of 'singleton' models. For singleton models, we only allow one variable γ_j to be present in the model, in addition to θ, implying that $S_j = \{j\}$. A major advantage is that only simple models need to be fit. Such a collection consists of $M = q + 1$ models when the narrow model is included.

When considering a non-random set of weights in $\mathcal{H} = \big\{(w_1, \ldots, w_M) : \sum_{j=1}^M w_j = 1\big\}$, then (see Hjort and Claeskens (2003)) for the weighted estimator it holds that

$$\sqrt{n}(\widehat{\mu}_w - \mu_{\mathrm{true}}) \xrightarrow{d} \sum_{j=1}^M w_j \Lambda_{S_j} = \sum_{j=1}^M w_j \{\Lambda_0 + \nu^t(\delta - G_{S_j} D)\},$$

from which, by (2.2), the limiting mean squared error is found to be $\mathrm{MSE}(\widehat{\mu}_w) = \tau_0^2 + R(\delta)$, where

$$\begin{aligned} R(\delta) &= \nu^t \Big\{ \Big(I_q - \sum_{j=1}^{M} w_j Q^0_{S_j} Q^{-1} \Big) \delta\delta^t \Big(I_q - \sum_{j=1}^{M} w_j Q^0_{S_j} Q^{-1} \Big)^t \\ &\quad + \Big(\sum_{j=1}^{M} w_j Q^0_{S_j} \Big) Q^{-1} \Big(\sum_{j=1}^{M} w_j Q^0_{S_j} \Big)^t \Big\} \nu. \end{aligned} \tag{3.1}$$

It is convenient to rewrite (3.1) as a quadratic function of the weights, $R(\delta) = w^t F(\delta) w$, where the (j,k)th entry of $F = F(\delta)$ is defined by (with $j,k = 1,\ldots,M$)

$$F_{jk}(\delta) = \nu^t \Big\{ \big(I_q - Q^0_{S_j} Q^{-1} \big)^t \delta\delta^t \big(I_q - Q^0_{S_k} Q^{-1} \big) + \big(Q^0_{S_j} Q^{-1} Q^0_{S_k} \big) \Big\} \nu. \tag{3.2}$$

The theoretical weights that minimize the MSE are

$$w_{\mathrm{mse}} = \operatorname*{argmin}_{w \in \mathcal{H}} w^t F w. \tag{3.3}$$

In practice, the MSE needs to be estimated in order to estimate the optimal weights.

3.2. Estimating the MSE and uniqueness of the weights

While almost all quantities in the MSE can be estimated by inserting consistent estimators for unknowns formed by plugging in estimators for (θ,γ) and using empirical Fisher information matrices, the situation is different for δ. With $\widehat{\delta} = \sqrt{n}(\widehat{\gamma}_{\mathrm{full}} - \gamma_0) \to_d D \sim N_q(\delta, Q)$, we cannot achieve the same accuracy as for the other estimators. To not overload the notation, we focus here on the estimation of δ, and leave the other quantities as they are, assumed to be known. However, all unknown quantities are consistently estimated for practical use.

Using the above defined unbiased estimator $\widehat{\delta}$ results in estimating the MSE by $\widehat{\mathrm{MSE}(\widehat{\mu}_w)} = \tau_0^2 + w^t \widehat{F} w$ where the $M \times M$ matrix $\widehat{F} = F(\hat{\delta})$, see (3.2). The minimum MSE weights are defined by $\widehat{w}_{\mathrm{mse}} = \mathrm{argmin}_{w \in \mathcal{H}}\, w^t \widehat{F} w$.

The estimated weights are unique if and only if the matrix $\widehat{F}$ is positive definite. By using (3.2) we can rewrite $\widehat{F} = \widehat{f}\widehat{f}^t + \widetilde{Q}$ where the $M \times 1$ vector $\widehat{f}$ has jth element equal to $\nu^t (I_q - Q^0_{S_j} Q^{-1})^t \widehat{\delta}$ and the $M \times M$ matrix $\widetilde{Q}$ has (j,k)th element $\nu^t Q^0_{S_j} Q^{-1} Q^0_{S_k} \nu$. Thus a sufficient condition for $\widehat{F}$ (and F as well) to be positive definite is that $\widetilde{Q}$ be positive definite. Lemma 1 presents a sufficient condition for this occurrence. All proofs are contained in the Appendix.

Lemma 1. *If Q is positive definite, ν is not equal to 0_M, and the matrices $Q^0_{S_j}$ $(j = 1,\ldots,M)$ are linearly independent, then $\widetilde{Q}$ is positive definite.*

Under the conditions of Lemma 1, the theoretical optimal weights that minimize (3.3) are unique and can be written as $w_{\text{mse}} = 1_M^t F^{-1}/(1_M^t F^{-1} 1_M)$; this is a well-known result for minimizing quadratic forms. The vector 1_M denotes a vector of all ones of length M. Our proposed model averaging weights are the values that minimize the MSE estimator $\widehat{\text{MSE}}(\widehat{\mu}_w)$. Given the conditions of Lemma 1, also these weights are unique and can be written as

$$\widehat{w}_{\widehat{\text{mse}}} = \underset{w \in \mathcal{H}}{\text{argmin}}\, w^t \widehat{F} w = \frac{1_M^t \widehat{F}^{-1}}{1_M^t \widehat{F}^{-1} 1_M}. \tag{3.4}$$

We call these the minimum MSE weights (mMSE).

Since Q has rank q, at most q linearly independent components $Q^0_{S_j}$ can be constructed, meaning that the rank of $\widetilde{Q}$ (and hence of $\widehat{F}$) is bounded by q. Together with the narrow model (the 'null' component), this means that the maximum number of models M needed to create a unique set of weights is $q+1$. This can be achieved by considering the class of nested models, or the class of singleton models. Each class has $M = q+1$ models and an estimate $\widehat{F}$ which is positive definite with rank equal to $q+1$, resulting in a unique set of weights. Note that several more situations lead to unique weights. For example, with $q = 7$, one model may contain the variables (γ_1, γ_2), a second model γ_3, and a third model $(\gamma_4, \ldots, \gamma_7)$. Here Lemma 1 guarantees uniqueness of the selected mMSE weights. Section 3.3 works out uniqueness properties of the *predictions* in the case of linear regression models.

Another conclusion of Lemma 1 is that we cannot find unique 'optimal' weights for the case of averaging over all 2^q subsets without considering more assumptions (Dostal (2009)). This may open a discussion about averaging over a set of not more than $q+1$ submodels only for which there are linearly independent Q^0_S matrices and for which we can find unique optimal weights versus the current common practice of averaging over all subsets resulting in a set of non-unique weights.

Remark 1. The above uniqueness property is tied to the estimation of δ. Alternatively, if we use $\widetilde{\delta}\widetilde{\delta}^t - Q$ in (3.3) as an unbiased estimator of $\delta\delta^t$ then, after removing some terms independent of the weight vector w,

$$\widehat{w}_{\widehat{\text{mse}},1} = \underset{w \in \mathcal{H}}{\text{argmin}}\{w^t \widehat{P} w + 2 w^t T\}$$

with, for $j, k = 1, \ldots, M$, $\widehat{P}_{jk} = \nu^t (I_q - Q^0_{S_j} Q^{-1}) \widehat{\delta}\widehat{\delta}^t (I_q - Q^0_{S_k} Q^{-1})^t \nu$ and $T_j = \nu^t Q^0_{S_j} \nu$. The matrix $\widehat{P}$ can be rewritten as a product $\widehat{a}\widehat{a}^t$ where the M-vector $\widehat{a}$ has jth component $\hat{\delta}^t (I_q - Q^0_{S_j} Q^{-1})^t \nu$. Obviously, $\widehat{P}$ is positive semi-definite only,

having rank one. An important consequence is that there is no unique solution for the weights, regardless of which models are averaged over. Nonetheless, we can find weights by using either generalized inverse matrices or by means of quadratic programming. With $\widehat{P}^-$ denoting such a (non-unique) generalized inverse of $\widehat{P}$, the (non-unique) weight is

$$\widehat{w}_{\widehat{\mathrm{mse}},1} = \widehat{P}^-[(-I_M + 1_M(1_M^t\widehat{P}^-1_M)^{-1}1_M^t\widehat{P}^-)T + 1_M(1_M^t\widehat{P}^-1_M)^{-1}].$$

Adding such constraints as having all weights positive and not larger than 1 is not a guarantee to get unique weights when $\widehat{P}$ is not positive definite (Propositions 2-10 and 2-20 Dostal (2009); Harville (2000)).

The mMSE weighted estimator with random, data-driven weights, as in (3.4), has different statistical properties as when using the unreachable theoretically optimal weights (3.3). Theorem 1 shows the limiting behavior of both the weights and of the mMSE weighted estimator.

It is straightforward to show that the estimator $\widehat{F} = F(\widehat{\delta})$ of F in (3.2) converges, for n tending to infinity, in distribution to F^* of which the (j,k)th element $(j,k = 1,\ldots,M)$ is

$$F^*_{j,k} = \nu^t\big(I_q - Q^0_{S_j}Q^{-1}\big)DD^t\big(I_q - Q^0_{S_k}Q^{-1}\big)^t\nu + \nu^t\big(Q^0_{S_j}Q^{-1}Q^0_{S_k}\big)\nu,$$

with $D \sim N(\delta, Q)$. Hence, it follows that $w^t\widehat{F}w \xrightarrow{d} w^tF^*w$ as $n \to \infty$. While the explicit form of the weights in (3.4) is useful for direct computation, it hints at a complicated limiting distribution.

Theorem 1. *Assume that $\widehat{F}$ and F^* are invertible. Let $\widehat{w}_{\widehat{\mathrm{mse}}} = \mathrm{argmin}_{w\in\mathcal{H}}\, w^t\widehat{F}w$ and $w^* = \mathrm{argmin}_{w\in\mathcal{H}}\, w^tF^*w$. Then* (i) $\widehat{w}_{\widehat{\mathrm{mse}}} \xrightarrow{d} w^*$; (ii) *the model averaging estimator satisfies*

$$\sqrt{n}(\widehat{\mu}_{\widehat{w}_{\widehat{\mathrm{mse}}}} - \mu_{\mathrm{true}}) \xrightarrow{d} \sum_{j=1}^{M} w_j^*\Lambda_{S_j}.$$

Since the weights w^* are random, the limiting distribution is not Gaussian, despite every Λ_{S_j} being Gaussian. For deterministic weights the limiting distribution is normal.

The randomness of the weights complicates inference for the model averaged estimator and, as a special case, also for the estimator post-selection when the uncertainty involved with the selection is taken into account. For more information about post-selection inference, see, e.g., Pötscher (1991), Kabaila (1995), Claeskens and Hjort (2003) and Danilov and Magnus (2004).

3.3. Uniqueness of predictions in linear regression models

While uniqueness of the weights is important, we investigate the uniqueness of the weighted predictions when different sets of models $\{S_1, \ldots, S_M\}$ are used to construct the weights. This discussion is restricted to linear models only.

In this subsection we consider the normal linear regression models $Y = X\theta + Z\gamma + \varepsilon$. The intercept is always present and is included in the vector θ that may also include coefficients of other fixed covariates, resulting in a design matrix X. In addition, there are q potential covariates $z_1, \ldots, z_q$ which are collected in the design matrix Z with corresponding coefficients γ. Take the ideal situation that $\sigma^2 = \text{var}(\varepsilon)$ is known to simplify the notation and calculations; this assumption is not necessary since we can include σ^2 in the vector θ and add one row and column to the Fisher information matrix J_n. For a linear model, the empirical Fisher information matrix is

$$J_n = \frac{1}{n}\begin{pmatrix} X^tX & X^tZ \\ Z^tX & Z^tZ \end{pmatrix} = \begin{pmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{pmatrix}.$$

The focus is on model-averaged prediction. We estimate the mean of Y at a given covariate vector (x, z), denoted by $\mu(x, z) = x^t\theta + z^t\gamma$ with $\gamma = (\gamma_1, \ldots, \gamma_q)^t$. We find minimal MSE weights for this purpose.

In Theorem 2 we show that model averaged predictions with mMSE weights for singleton and nested models result in identical predictions. We can best explain this phenomenon via the selection matrices.

Definition 1. The selection matrix (ζ) is an $M \times q$ matrix with $\{0, 1\}$ elements, constructed as

$$\zeta = \left(1_q^t \pi_{S_1}^t \pi_{S_1}, \ldots, 1_q^t \pi_{S_M}^t \pi_{S_M}\right)^t,$$

where each row represents a model such that elements equal to 1 correspond to auxiliary variables present in the model.

For example, the selection matrices for a set of singleton (ζ_s) and nested (ζ_n) models for $q = 3$ can be written as

$$\zeta_s = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \zeta_n = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}.$$

To facilitate the proof, we orthogonalize the design matrix $\mathcal{X} = (X, Z)$ so that the matrix $Q = J^{11}$ is diagonal. There is no loss of generality as, by the QR decomposition, for any matrix $\mathcal{X}$, there is an orthonormal matrix A and upper triangular matrix R for which $\mathcal{X} = AR$. The matrix R is the transformation matrix which can be used to convert $\mathcal{X}$ to the orthonormal matrix A. If one

uses the original matrix $\mathcal{X}$ and the estimated weights $\widehat{w}_{\text{mse}}$ via the proposed mMSE method, and calculates the prediction value for the new observation $\mathcal{X}_{\text{new}}$ resulting in $\widehat{\mu}_{\text{new}}$, or uses the orthonormal matrix A as the design matrix, uses the mMSE estimator of the weights $\widehat{w}^A_{\text{mse}}$, and calculates the prediction $\widehat{\mu}^A_{\text{new}}$ for the $\mathcal{X}_{\text{new}}R^{-1}$, then the prediction values are equal, $\widehat{\mu}^A_{\text{new}} = \widehat{\mu}_{\text{new}}$. This means that for prediction purposes, the orthogonalized version of $\mathcal{X}$ and the original $\mathcal{X}$ give the same results.

For any such diagonal matrix Q it follows that $\widehat{\gamma}_S = \pi_S^t \widehat{\gamma}_{\text{full}}$. This means that the estimators of γ_j in the full model and in each considered model that contains γ_j are identical. Thus in particular, $\widehat{\gamma}_j$ in the nested model is identical to $\widehat{\gamma}_j$ in the singleton models (for all $j = 1, \ldots, q$). In the case of a diagonal matrix Q, it is readily obtained that having linearly independent rows in a selection matrix ζ is equivalent to having linearly independent matrices $Q^0_{S_j}$, where the sets S_j are induced by the rows of ζ. Hence, finding models that satisfy the assumption of Lemma 1 is aided via the selection matrices.

The estimated weights for estimation of the value $\mu(x, z)$ are obtained via mMSE in (3.4). Consider the weighted prediction at a value (x, z) using a sequence of nested models with mMSE nested weights $\widehat{w}^{\text{nest}}$. Since the weights sum to 1,

$$\widehat{\mu}^{\text{nest}}_w = \sum_{i=1}^{q+1} \widehat{w}^{\text{nest}}_i \widehat{\mu}_i = x^t\widehat{\theta} + \sum_{i=1}^{q} \big(z_i\widehat{\gamma}_i \sum_{j=i+1}^{q+1} \widehat{w}^{\text{nest}}_j\big). \tag{3.5}$$

Next consider the weighted prediction at the same value (x, z) in the set of singleton models with mMSE singleton weights $\hat{w}^{\text{sing}}$,

$$\widehat{\mu}^{\text{sing}}_w = x^t\widehat{\theta} + \sum_{i=2}^{q+1} z_{i-1}\widehat{\gamma}_{i-1}\widehat{w}^{\text{sing}}_i. \tag{3.6}$$

In Theorem 2 we prove the equality of the mMSE averaged prediction values for singleton models and nested models when we use all q covariates.

Theorem 2. *Let $p \geq 1$ and $q \geq 2$. When using mMSE weights* (3.4) *for averaging predictions in linear regression models with least squares estimation, the weighted predictions are equal for averaging over singleton models and for averaging over nested models, $\widehat{\mu}^{\text{sing}}_w = \widehat{\mu}^{\text{nest}}_w$.*

Our calculations have illustrated that the result holds true more generally, the prediction values are equal for all sets of models that have the same number of linearly independent rows in the selection matrix and that use the same covariates to construct the models. The proof of such cases proceeds along the same lines as the proof of Theorem 2.

Another nice property of our method has appeared in the simulations for linear models. It turns out that the mMSE weighted predictions for the set

of models for which the corresponding matrices Q_S^0 form a basis of the matrix space of that dimension, are precisely the same as the mMSE weighted predictions formed by using these models and some extra ones. Hence, no information is lost when restricting to such a subset of models. Consequently, there is no need to consider all possible models anymore and the weight choice by mMSE can be considered as a screening method for which we do not lose any information regarding prediction. This suggests a simplification in the choice of models used for model averaging. This result allows a drastic simplification of the computational aspects. Indeed, one need not consider all 2^q submodels for model averaging, only q singleton models suffice and they yield for linear models identical predictions when the mMSE weights are used. Also determining the order of the variables in nested models becomes irrelevant with this choice of weights.

When the number of models is more than $q+1$, the matrix F is not positive definite, yet we get the same weighted prediction values for averaging over singleton models and over all possible models. This remarkable fact is explained by finding the weights via a quadratic programming application that searches the minimum of the estimated MSE. Since the matrix F is positive semi-definite when using all possible submodels, the solution for the weights is not unique but all solutions are global ones (Antoniou and Lu (2007, Chap. 13)), thus yielding the same prediction values for singleton models and for all possible subset models.

4. Weight Choice for Multiple Focuses

Here we consider a vector-valued focus. This is useful for example when considering predictions at more than one position or when considering a vector of regression coefficients for model averaged estimation.

Let $\vec{\mu} = (\mu_1, \ldots, \mu_r)^t$. Then $\boldsymbol{\tau}_0^2 = (\partial(\vec{\mu})/\partial\theta)^t J_{00}^{-1} \partial\vec{\mu}/\partial\theta$ is a $r \times r$ matrix and $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_r)$ is $q \times r$ with $\nu_i = J_{10}J_{00}^{-1}\partial\mu_i/\partial\theta - \partial\mu_i/\partial\gamma$. So, for a single submodel,

$$\sqrt{n}(\widehat{\vec{\mu}}_S - \mu_{\text{true}}) \xrightarrow{d} \vec{\Lambda}_S = \vec{\Lambda}_0 + \boldsymbol{\nu}^t(\delta - G_S D),$$

where $\vec{\Lambda}_0 \sim N_r(0, \boldsymbol{\tau}_0^2)$, and D and δ are as before. Here the MSE of $\widehat{\vec{\mu}}_S$ is defined as a matrix $E[(\widehat{\vec{\mu}}_S - \vec{\mu}_{\text{true}})(\widehat{\vec{\mu}}_S - \vec{\mu}_{\text{true}})^t]$ with diagonal elements the MSE of the individual focus component estimators. Similarly,

$$\sqrt{n}(\widehat{\vec{\mu}}_w - \mu_{\text{true}}) \xrightarrow{d} \sum_{j=1}^{M} w_j \vec{\Lambda}_{S_j} = \vec{\Lambda}_0 + \sum_{j=1}^{M} w_j \boldsymbol{\nu}^t(\delta - G_{S_j} D),$$

in which $\widehat{\mu}_w^i = \sum_{j=1}^{M} w_j \widehat{\mu}_j^i$. The $r \times r$ MSE matrix can be written as

$$\text{MSE}\,(\widehat{\vec{\mu}}_w, \delta) = \boldsymbol{\tau}_0^2 + \boldsymbol{R}(\delta), \tag{4.1}$$

where

$$\boldsymbol{R}(\delta) = \boldsymbol{\nu}^t \Big\{ \sum_{j=1}^{M} w_j (I_q - Q^0_{S_j} Q^{-1}) \delta\delta^t \sum_{j=1}^{M} w_j (I_q - Q^0_{S_j} Q^{-1})^t + \Big(\sum_{j=1}^{M} w_j Q^0_{S_j} \Big) Q^{-1} \Big(\sum_{j=1}^{M} w_j Q^0_{S_j} \Big)^t \Big\} \boldsymbol{\nu}.$$

As in the univariate case, all unknowns have consistent estimators except for δ. An additional issue with the multiple focuses case is deciding on the criterion for which we optimize the weight choice. Since the MSE is a matrix we consider both minimizing the trace and the determinant.

4.1. Minimizing the trace of the MSE matrix

The trace of the MSE matrix is the expected squared error loss function $E[\|\widehat{\vec{\mu}}_w - \vec{\mu}_{\text{true}}\|^2]$, the summation of the MSE values for the individual, univariate, focuses. Then

$$\text{tr}\{\text{MSE}\,(\widehat{\vec{\mu}}_w, \delta)\} = \text{tr}(\boldsymbol{\tau}_0^2) + \text{tr}\{\boldsymbol{R}(\delta)\} = \text{tr}(\boldsymbol{\tau}_0^2) + w^t \boldsymbol{F} w, \tag{4.2}$$

where $\boldsymbol{F}$ is a $M \times M$ matrix similar to the matrix F in (3.2) with (i,j)th entry

$$\boldsymbol{F}_{ij} = \text{tr}[\boldsymbol{\nu}^t \{(I_q - Q^0_{S_i} Q^{-1}) \delta\delta^t (I_q - Q^0_{S_j} Q^{-1})^t + Q^0_{S_i} Q^{-1} Q^0_{S_j}\} \boldsymbol{\nu}].$$

The optimal weights in (4.2) can be found as

$$w_{\text{mse}} = \underset{w \in \mathcal{H}}{\text{argmin}}\, w^t \boldsymbol{F} w. \tag{4.3}$$

If the unbiased estimator of δ, $\widehat{\delta} = \sqrt{n}(\widehat{\gamma} - \gamma_0)$ is plugged in to (4.3) using $\boldsymbol{R}(\widehat{\delta})$ and $\widehat{\boldsymbol{F}}$, minimizing the trace of the MSE matrix leads to

$$\widehat{w}_{\widehat{\text{mse}}} = \underset{w \in \mathcal{H}}{\text{argmin}}\, w^t \widehat{\boldsymbol{F}} w = \frac{1_M^t \widehat{\boldsymbol{F}}^{-1}}{(1_M^t \widehat{\boldsymbol{F}}^{-1} 1_M)}. \tag{4.4}$$

This results in a unique weight vector under the assumptions of Lemma 1. Also, Theorem 1 can be stated for a multivariate focus.

The motivation for not using the unbiased estimator of $\delta\delta^t$, $\widehat{\delta}\widehat{\delta}^t - Q$, is as in the univariate case. Indeed, in that case the weight would equal $\text{argmin}_{w \in \mathcal{H}} \{w^t \boldsymbol{P} w + 2 w \boldsymbol{T}\}$, where $\boldsymbol{P}$ and $\boldsymbol{T}$ are similar to P and T in the univariate case, with

$$\boldsymbol{P}_{jk} = \text{tr}\{\boldsymbol{\nu}^t (I_q - Q^0_{S_j} Q^{-1}) \hat{\delta}\hat{\delta}^t (I_q - Q^0_{S_k} Q^{-1})^t \boldsymbol{\nu}\} \text{ and } \boldsymbol{T}_j = \text{tr}\left(\boldsymbol{\nu}^t Q^0_{S_j} \boldsymbol{\nu}\right).$$

Rewriting $\boldsymbol{P} = \vec{A}\vec{A}^t$ with $\vec{A}^t = (A_1, \ldots, A_M)$ and $\vec{A}_j = \widehat{\delta}^t (I_q - Q^0_{S_j} Q^{-1}) \boldsymbol{\nu}$, yields that $\text{rank}(\boldsymbol{P}) \leq \min(r, q, M)$. In order to have unique weights, a necessary

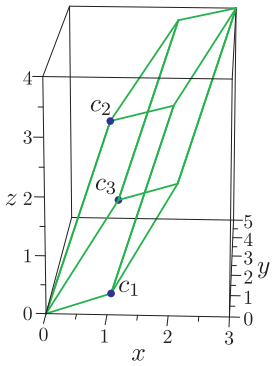

Figure 1. Geometrical representation of a matrix A with columns c_1, c_2 and c_3.

assumption is $\text{rank}(\boldsymbol{P}) = \min(r, q, M) = M$ which is true when ν is a full rank matrix and $M \leq \min(r, q)$. The necessary assumptions for the unicity of weights by using an unbiased estimator of $\delta\delta^t$ are more restrictive than the necessary assumptions for the unbiased estimation of δ. Moreover, the simulation results show that the biased estimator of the MSE performs better than the unbiased estimator with respect to out-of-sample mean squared prediction error.

4.2. Minimizing the determinant of the MSE matrix

The trace of the MSE matrix ignores the information stored in the off-diagonal elements. To use it, we consider the parallelepiped generated by the MSE column vectors, a geometric representation of the MSE matrix. For example, Figure 1 draws the parallelepiped produced by the three columns $c_1 = (1, 1, 0)$, $c_2 = (1, 1, 3)$ and $c_3 = (1, 3, 1)$ of a matrix A.

In Section 4.1., the weights were found by minimizing the trace of the MSE matrix which results in a parallelepiped with minimum sum of squares of the axes. In this section we assign weights to each model in such a way that the *volume* of the parallelepiped is minimized, equivalent to minimizing the determinant of the MSE matrix. The D-optimality criterion of experimental design studies seeks designs that minimize the covariance matrix of the parameter estimators (Atkinson, Denov, and Tobias (2007)). However, with the presence of δ, the estimators in our case are not unbiased, which motivates the use of the MSE matrix instead of the covariance matrix.

Our aim is to minimize the determinant of the MSE matrix in (4.1), hence

$$\widehat{w}_{|\widehat{\text{mse}}|} = \underset{w \in \mathcal{H}}{\text{argmin}} \det(\text{MSE}\,(\widehat{\bar{\mu}}_w, \widehat{\delta})). \tag{4.5}$$

It should be noted that this is a nonlinear optimization problem with a linear constraint for the weights and the results are no longer unique. The simulation

results in the next section show that the proposed method performs well in comparison with other methods.

Remark 2. The necessary assumption for starting the optimization in (4.5) is that the MSE matrix is non-singular with a non-zero determinant. In (4.1), $R(\delta)$ plays the crucial role, since τ_0^2 is zero when $p = 0$. If ν^t is a matrix with full column rank q, then the rank of the MSE matrix is equal to the rank of the middle part, B, of $R(\delta) = \nu^t B \nu$. In singleton models B is positive definite (rank=q) if all the weights are nonzero (starting values for optimization); moreover, the length of the focus vector cannot be more than q, else the MSE matrix is positive semi-definite, hence singular.

5. Simulation Studies

In this section, we consider linear models and generalized linear models. Since almost all previous studies insisted on linear regression, we first compare our method with other methods in this setting. Then, we present the results for generalized linear models and compare the proposed method with some other methods of model averaging.

5.1. Linear models

We investigate the finite sample performance of the minimum MSE estimator (mMSE) and compare the results with other methods of averaging, in particular by the plug-in estimator (Liu (2015)), the so-called optimal estimator (OPT, Liang et al. (2011)), Mallows model averaging (MMA, Hansen (2007)) and jackknife model averaging (JMA, Hansen and Racine (2012)). All of these methods are defined for averaging over all possible submodels. While the mMSE method is also applicable to using all submodels, we insist on unique weights (unique prediction) which entails using row linearly independent selection matrix models. By Theorem 2 and its discussion, we present the results only for singleton models for the mMSE weight choice method. Some methods such as the OPT estimator do not result in unique weights even for the singleton models because of using nonlinear optimization for constructing the weights. In Setting 1.2, averaging over singleton models and all possible models for different methods is compared.

General settings for the simulations in this section are summarized here. The data were generated from a finite order regression model of the form

$$Y_i = \sum_{j=1}^{p} \theta_j x_{ji} + \sum_{k=1}^{q} \frac{\delta_k}{\sqrt{n}} z_{ki} + e_i, \quad i = 1, \ldots, n. \tag{5.1}$$

We set $p = 3$, $q = 8$, $x_{1i} = 1$ as an intercept and $(x_{2i}, x_{3i}, z_{1i}, \cdots, z_{8i}) \sim N_{p-1+q}(0, \nu)$. The covariance matrix ν for the regressors contains a diagonal equal to 1 with off-diagonal entries ρ. The error term e_i was standard normal and independent of the regressors. We generated $n + 1$ observations from this model, the last observation acting as an out-of-sample value.

Setting 1. This setting compares different methods of model averaging for singleton models in different settings. The plug-in method is for linear models theoretically the same as the proposed method but with different assumptions for the weights (the weights are positive and sum to 1) and the implementation is based on least squares theory which causes some differences in the results. For that method the prediction values are unique, but not the weights since those authors used semi-definite quadratic programming for minimizing the MSE of the focus parameter, resulting in global minimizer weights that are not unique. Hence, if the goal is finding the most influential covariate or model, there is no unique answer given by those methods.

Three scenarios were considered,

$$\text{Scenario 1.1} : (\theta, \frac{\delta}{\sqrt{n}}) = \Big((5, -4, 3.5), \frac{c(3.9, 4.75, 4.2, 3.5, 4.95, -3.75, 4.4, -4)}{\sqrt{n}}\Big),$$
$$\text{Scenario 1.2} : (\theta, \frac{\delta}{\sqrt{n}}) = \Big((5, -4, 3.5), \frac{c(3.9, 4.75, 4.2, 0, 4.95, 0, 4.4, -4)}{\sqrt{n}}\Big),$$
$$\text{Scenario 1.3} : (\theta, \frac{\delta}{\sqrt{n}}) = \Big((5, -4, 3.5), \frac{c(0, 4.75, 4.2, 0, 4.95, 0, 4.4, 0)}{\sqrt{n}}\Big).$$

The effect of the importance of the δ values relative to θ is controlled by the constant c which varied in the set $\{0.5, 1, 2\}$. In Scenarios 1.2 and 1.3, the effect of true zero coefficients in the true model is studied. The sample sizes varied in $\{50, 100, 200, 500, 800,\ 1{,}500\}$ and the off-diagonal value ρ of ν varied in $\{0, 0.25, 0.5, 075\}$. All Monte-Carlo simulations were repeated 2,000 times and the numbers in Tables 1−3 are the mean squared prediction errors (MSPE) for the out-of-sample value (the $n + 1$st value that is not used in the estimation nor weight determination) over the simulation runs.

It is observed that the mMSE works better than the other methods for moderate and high values of c (Tables 2 and 3) in all scenarios. For small c (Table 1), in Scenarios 1.1 and 1.2 when the collinearity amongst the covariates is small ($\rho = 0$ and 0.25), the mMSE has the best performance, whereas higher collinearity cause better performance of the plug-in method but with small differences with mMSE. In Scenario 3, the plug-in method works well for low and moderate collinearity while MMA and JMA outperform for high collinearity. For high values of c, the difference between the proposed method and other methods is

Table 1. Simulation study for linear models. MSPE of singleton models for OPT, MMA, JMA, Plug-in and mMSE for $c = 0.5$.

		Scenario 1.1					Scenario 1.2					Scenario 1.3				
ρ	n	OPT	MMA	JMA	Plug-in	mMSE	OPT	MMA	JMA	Plug-in	mMSE	OPT	MMA	JMA	Plug-in	mMSE
0	50	0.615	0.608	0.607	0.325	**0.279**	0.506	0.471	0.471	0.284	**0.277**	0.385	0.335	0.334	**0.235**	0.274
	100	0.318	0.310	0.310	0.163	**0.118**	0.259	0.237	0.237	0.138	**0.117**	0.196	0.162	0.162	**0.114**	0.116
	200	0.149	0.143	0.143	0.077	**0.057**	0.128	0.115	0.115	0.067	**0.057**	0.099	0.080	0.080	**0.055**	0.057
	500	0.057	0.056	0.056	0.029	**0.021**	0.048	0.044	0.044	0.026	**0.021**	0.036	0.030	0.030	**0.021**	**0.021**
	1000	0.030	0.029	0.029	0.015	**0.010**	0.025	0.022	0.022	0.013	**0.010**	0.019	0.016	0.016	**0.010**	**0.010**
	1500	0.021	0.019	0.019	0.010	**0.007**	0.017	0.015	0.015	0.009	**0.007**	0.013	0.010	0.010	**0.007**	**0.007**
0.25	50	0.628	0.543	0.542	0.297	**0.268**	0.537	0.436	0.435	**0.265**	0.266	0.442	0.339	0.339	**0.239**	0.265
	100	0.294	0.254	0.254	0.138	**0.117**	0.253	0.205	0.205	0.121	**0.117**	0.207	0.154	0.154	**0.107**	0.117
	200	0.137	0.121	0.121	0.066	**0.055**	0.120	0.100	0.100	0.058	**0.055**	0.099	0.075	0.075	**0.051**	0.055
	500	0.060	0.051	0.051	0.026	**0.022**	0.052	0.042	0.042	0.023	**0.021**	0.043	0.032	0.032	**0.021**	**0.021**
	1000	0.029	0.025	0.025	0.013	**0.011**	0.025	0.020	0.020	0.011	**0.010**	0.020	0.015	0.015	**0.010**	0.011
	1500	0.020	0.017	0.017	0.009	**0.008**	0.017	0.014	0.014	**0.008**	**0.008**	0.014	0.010	0.010	**0.007**	0.008
0.5	50	0.464	0.407	0.407	**0.260**	0.290	0.402	0.339	0.339	**0.244**	0.289	0.335	0.271	0.271	**0.225**	0.288
	100	0.215	0.197	0.198	**0.126**	0.130	0.192	0.169	0.169	**0.116**	0.129	0.162	0.139	0.140	**0.109**	0.129
	200	0.113	0.098	0.098	**0.057**	**0.057**	0.101	0.084	0.084	**0.053**	0.057	0.085	0.067	0.068	**0.050**	0.057
	500	0.044	0.039	0.039	**0.023**	0.024	0.039	0.032	0.032	**0.022**	0.024	0.032	0.025	0.025	**0.020**	0.024
	1000	0.021	0.017	0.017	**0.010**	**0.010**	0.019	0.015	0.015	**0.010**	**0.010**	0.016	0.012	0.012	**0.009**	0.011
	1500	0.015	0.012	0.012	**0.007**	**0.007**	0.013	0.010	0.010	**0.007**	**0.007**	0.010	0.008	0.008	**0.006**	0.007
0.75	50	0.292	0.267	0.268	**0.235**	0.293	0.256	0.231	0.233	**0.224**	0.292	0.223	**0.199**	**0.199**	0.216	0.290
	100	0.132	0.119	0.120	**0.096**	0.125	0.117	0.104	0.104	**0.094**	0.124	0.102	**0.088**	**0.088**	0.091	0.124
	200	0.062	0.056	0.056	**0.047**	0.055	0.057	0.049	0.049	**0.046**	0.055	0.050	**0.043**	**0.043**	0.044	0.055
	500	0.026	0.023	0.023	**0.018**	0.021	0.024	0.020	0.020	**0.018**	0.021	0.021	**0.017**	**0.017**	**0.017**	0.021
	1000	0.012	0.011	0.011	**0.009**	0.011	0.011	0.010	0.010	**0.090**	0.010	0.009	**0.008**	**0.008**	**0.008**	0.010
	1500	0.008	0.008	0.008	**0.006**	0.007	0.008	0.007	0.007	**0.006**	0.007	0.007	**0.006**	**0.006**	**0.006**	0.007

Table 2. Simulation study for linear models. MSPE of singleton models for OPT, MMA, JMA, plug-in and mMSE for $c = 1$.

		Scenario 1.1					Scenario 1.2					Scenario 1.3				
ρ	n	OPT	MMA	JMA	Plug-in	mMSE	OPT	MMA	JMA	Plug-in	mMSE	OPT	MMA	JMA	Plug-in	mMSE
0	50	2.291	2.229	2.227	0.897	**0.286**	1.843	1.696	1.692	0.720	**0.285**	1.347	1.072	1.069	0.499	**0.284**
	100	1.168	1.123	1.124	0.461	**0.120**	0.934	0.831	0.831	0.351	**0.120**	0.683	0.522	0.522	0.243	**0.119**
	200	0.547	0.525	0.524	0.215	**0.059**	0.464	0.406	0.406	0.175	**0.058**	0.345	0.255	0.255	0.118	**0.058**
	500	0.214	0.208	0.208	0.087	**0.022**	0.178	0.161	0.161	0.069	**0.022**	0.130	0.099	0.099	0.045	**0.022**
	1000	0.111	0.106	0.106	0.045	**0.011**	0.090	0.080	0.080	0.034	**0.011**	0.068	0.052	0.052	0.022	**0.011**
	1500	0.076	0.072	0.072	0.031	**0.007**	0.063	0.054	0.054	0.024	**0.007**	0.047	0.034	0.034	0.016	**0.007**
0.25	50	2.218	1.931	1.935	0.719	**0.273**	1.836	1.508	1.507	0.585	**0.273**	1.440	1.067	1.068	0.443	**0.272**
	100	1.068	0.923	0.924	0.336	**0.120**	0.904	0.723	0.724	0.264	**0.119**	0.712	0.498	0.498	0.193	**0.119**
	200	0.500	0.436	0.436	0.154	**0.056**	0.435	0.347	0.347	0.121	**0.056**	0.349	0.240	0.240	0.091	**0.056**
	500	0.220	0.184	0.184	0.066	**0.022**	0.193	0.149	0.149	0.053	**0.023**	0.153	0.103	0.103	0.039	**0.022**
	1000	0.107	0.089	0.089	0.032	**0.011**	0.090	0.069	0.069	0.025	**0.011**	0.071	0.047	0.047	0.018	**0.011**
	1500	0.074	0.062	0.062	0.023	**0.008**	0.063	0.049	0.049	0.018	**0.008**	0.050	0.033	0.033	0.013	**0.008**
0.5	50	1.617	1.409	1.410	0.501	**0.298**	1.373	1.109	1.110	0.433	**0.297**	1.093	0.798	0.799	0.351	**0.297**
	100	0.736	0.625	0.626	0.225	**0.133**	0.646	0.512	0.512	0.190	**0.132**	0.527	0.380	0.381	0.158	**0.132**
	200	0.399	0.332	0.332	0.113	**0.058**	0.350	0.269	0.270	0.095	**0.058**	0.283	0.194	0.194	0.079	**0.058**
	500	0.156	0.131	0.131	0.045	**0.024**	0.134	0.103	0.103	0.038	**0.022**	0.106	0.070	0.070	0.030	**0.024**
	1000	0.075	0.060	0.060	0.021	**0.011**	0.066	0.049	0.049	0.017	**0.011**	0.054	0.035	0.035	0.014	**0.011**
	1500	0.052	0.042	0.042	0.014	**0.007**	0.045	0.034	0.034	0.012	**0.007**	0.035	0.023	0.023	0.009	**0.007**
0.75	50	0.924	0.786	0.787	0.379	**0.304**	0.788	0.625	0.627	0.337	**0.303**	0.653	0.474	0.475	**0.301**	0.302
	100	0.436	0.358	0.358	0.153	**0.129**	0.370	0.286	0.286	0.138	**0.129**	0.315	0.216	0.216	**0.124**	0.129
	200	0.197	0.162	0.162	0.072	**0.057**	0.174	0.132	0.132	0.067	**0.057**	0.144	0.101	0.101	0.059	**0.057**
	500	0.085	0.069	0.069	0.029	**0.022**	0.074	0.055	0.055	0.026	**0.023**	0.062	0.042	0.042	0.023	**0.022**
	1000	0.039	0.032	0.032	0.014	**0.011**	0.034	0.027	0.027	0.013	**0.011**	0.027	0.020	0.020	**0.011**	**0.011**
	1500	0.027	0.022	0.022	0.010	**0.007**	0.023	0.018	0.018	0.009	**0.007**	0.019	0.014	0.014	0.008	**0.007**

Table 3. Simulation study for linear models. MSPE of singleton models for OPT, MMA, JMA, plug-in and mMSE for $c = 2$.

		Scenario 1.1					Scenario 1.2					Scenario 1.3				
ρ	n	OPT	MMA	JMA	Plug-in	mMSE	OPT	MMA	JMA	Plug-in	mMSE	OPT	MMA	JMA	Plug-in	mMSE
0	50	8.821	8.761	8.759	3.236	**0.288**	6.839	6.640	6.640	2.446	**0.287**	4.595	4.063	4.055	1.535	**0.287**
	100	4.480	4.368	4.371	1.670	**0.121**	3.399	3.207	3.208	1.197	**0.121**	2.231	1.979	1.979	0.752	**0.120**
	200	2.128	2.070	2.069	0.769	**0.059**	1.731	1.586	1.585	0.607	**0.059**	1.141	0.970	0.969	0.364	**0.059**
	500	0.847	0.818	0.818	0.325	**0.022**	0.698	0.629	0.628	0.244	**0.022**	0.479	0.376	0.376	0.141	**0.022**
	1000	0.432	0.416	0.416	0.165	**0.011**	0.351	0.309	0.309	0.118	**0.011**	0.262	0.194	0.194	0.070	**0.011**
	1500	0.297	0.283	0.283	0.115	**0.007**	0.244	0.213	0.213	0.084	**0.007**	0.180	0.131	0.131	0.051	**0.007**
0.25	50	7.776	7.533	7.545	2.362	**0.275**	6.170	5.805	5.808	1.809	**0.274**	4.485	3.981	3.983	1.193	**0.274**
	100	3.763	3.635	3.637	1.123	**0.120**	3.001	2.807	2.809	0.829	**0.120**	2.129	1.886	1.887	0.526	**0.120**
	200	1.760	1.682	1.682	0.503	**0.057**	1.435	1.332	1.333	0.369	**0.057**	1.019	0.900	0.899	0.246	**0.057**
	500	0.791	0.719	0.719	0.222	**0.022**	0.650	0.578	0.578	0.170	**0.022**	0.459	0.389	0.389	0.109	**0.022**
	1000	0.403	0.349	0.349	0.106	**0.011**	0.324	0.266	0.266	0.077	**0.011**	0.233	0.176	0.176	0.049	**0.011**
	1500	0.286	0.242	0.242	0.077	**0.008**	0.236	0.186	0.186	0.057	**0.008**	0.180	0.123	0.122	0.037	**0.008**
0.5	50	5.601	5.431	5.438	1.416	**0.300**	4.557	4.223	4.226	1.129	**0.300**	3.366	2.897	2.901	0.768	**0.300**
	100	2.428	2.306	2.307	0.595	**0.133**	2.044	1.860	1.859	0.461	**0.133**	1.530	1.317	1.318	0.326	**0.133**
	200	1.353	1.263	1.263	0.326	**0.059**	1.128	1.014	1.014	0.254	**0.059**	0.834	0.695	0.695	0.181	**0.059**
	500	0.547	0.495	0.495	0.129	**0.024**	0.449	0.382	0.382	0.098	**0.024**	0.328	0.250	0.250	0.067	**0.024**
	1000	0.280	0.233	0.233	0.062	**0.011**	0.242	0.187	0.187	0.047	**0.011**	0.191	0.128	0.128	0.032	**0.011**
	1500	0.201	0.164	0.164	0.042	**0.007**	0.171	0.129	0.129	0.033	**0.007**	0.132	0.084	0.084	0.021	**0.007**
0.75	50	3.078	2.857	2.856	0.851	**0.307**	2.537	2.180	2.183	0.688	**0.307**	1.973	1.550	1.552	0.520	**0.307**
	100	1.491	1.319	1.320	0.354	**0.131**	1.228	1.008	1.009	0.275	**0.131**	0.983	0.716	0.717	0.216	**0.130**
	200	0.690	0.586	0.586	0.160	**0.057**	0.596	0.466	0.467	0.138	**0.057**	0.470	0.329	0.329	0.100	**0.057**
	500	0.316	0.248	0.249	0.067	**0.022**	0.271	0.195	0.195	0.055	**0.022**	0.223	0.139	0.139	0.041	**0.022**
	1000	0.147	0.116	0.116	0.031	**0.011**	0.128	0.096	0.096	0.026	**0.011**	0.100	0.064	0.064	0.018	**0.011**
	1500	0.101	0.081	0.081	0.021	**0.007**	0.086	0.063	0.063	0.017	**0.007**	0.070	0.044	0.044	0.013	**0.007**

remarkable, even with the plug-in method which is the theoretically closest method to the mMSE, the difference arising from the fewer restrictions for the weights in the mMSE. If we remove this additional constraint for the weights for the plug-in method, it performs similarly to mMSE in linear regression. Explicitly allowing for heteroscedasticity in order to improve prediction accuracy, the plug-in method of Liu (2015) results in different estimated Fisher information matrices as compared to mMSE, explaining the slightly different results between the two methods. Tables 1−3 reveal the stability of the mMSE values for different choices of c and n for the three scenarios, whereas other methods are sensitive to the particular setting.

To investigate potential information loss by other methods when using singleton models as compared to all subsets, we ran a simulation with all possible models for Scenario 1.2 with $c = 1$. To ease the comparison, the mMSE method is also shown in Table 4. This simulation using all subsets took around 45 hours on a supercomputer. As Table 4 shows, all competing methods improved significantly by using all possible models, especially the OPT method which performs the best in all settings. The plug-in method with the additional constraint for the weights to belong to $[0, 1]$ using all available models performs equally well as mMSE with singleton models. An important drawback, however, is that the computational intensity grows exponentially by adding extra auxiliary covariates; adding one covariate doubles the number of models. The mMSE performance for singleton model averaging is almost as good as all subsets model averaging for other methods, and Table 4 shows that except for the smallest sample size the difference is in most cases negligible.

Table 5 gives the computation time in seconds, using the same computer, for the OPT and mMSE method for the average time over five runs in the simulation when $\rho = 0.5$. The other values of ρ give similar computation times. With practically the same accuracy as OPT, mMSE benefits from a much shorter computation time, regardless of the sample size.

5.2. Generalized linear models: Poisson regression

Here we explore the performance of the mMSE method in Poisson regression by using five out-of-sample observations of which we estimate the mean.

Setting 2. The response values Y_i have a Poisson distribution with mean $\mu_i = \exp(x_i^t\delta/\sqrt{n})$, with $p = 0$ (no core regressors), $q = 8$ with $(x_{1i}, \ldots, x_{8i}) \sim N_{p+q}(0, \nu)$ in which $\nu_{ij} = 1$ for $i = j$ and $\nu_{ij} = \rho$ for $i \neq j$. The value of ρ varied in the set $\{0, 0.25, 0.5, 0.75\}$. The value of γ_0 was set to zero and δ values were considered according to Scenario 2.1: $\delta = (-1, -4, 3, -4, 0.6, 4, 2, 5)^t$, Scenario 2.2: $\delta = (-1, -4, 3, -4, 0, 4, 0, 5)^t$, and Scenario 2.3: $\delta = (0, -4, 3, -4,$

Table 4. Simulation study for linear models. MSPE for averaging over all possible models for OPT, MMA, JMA, plug-in and singleton averaging for the mMSE weighted estimator for Scenario 1.2 and $c = 1$.

		Method				
ρ	n	OPT	MMA	JMA	Plug-in	mMSE
0	50	0.273	0.291	0.297	0.285	0.285
	100	0.120	0.126	0.127	0.120	0.120
	500	0.021	0.022	0.022	0.022	0.022
	1,000	0.011	0.011	0.011	0.011	0.011
0.25	50	0.259	0.294	0.295	0.272	0.273
	100	0.115	0.123	0.124	0.120	0.119
	500	0.022	0.023	0.023	0.023	0.023
	1,000	0.011	0.011	0.011	0.011	0.011
0.5	50	0.272	0.323	0.327	0.296	0.297
	100	0.125	0.147	0.147	0.132	0.132
	500	0.021	0.024	0.024	0.022	0.022
	1,000	0.011	0.012	0.012	0.011	0.011
0.75	50	0.260	0.309	0.307	0.301	0.303
	100	0.112	0.136	0.136	0.127	0.129
	500	0.020	0.025	0.024	0.022	0.023
	1,000	0.010	0.012	0.012	0.011	0.011

Table 5. Simulation time in seconds for the average time over five repetitions of the simulation for $\rho = 0.5$ in Table 4.

	n			
method	50	100	500	1,000
OPT	1.87	2.63	17.04	60.21
mMSE	0.04	0.03	0.03	0.04

$0, 4, 0, 0)^t$. The considered sample sizes were $\{50, 100, 200, 500, 1{,}000\}$. All Monte Carlo simulations were based on 2,000 replications. The multivariate focus had length five, fixed for each setting, and was generated randomly for each setting; the focus was the same for all $n_{\text{sim}} = 2{,}000$ simulation runs in each setting. Each method was evaluated based on the *empirical* mean squared error matrix of dimension 5×5, $\text{MSE}_{\text{emp}} = (1/n_{\text{sim}}) \sum_{i=1}^{n_{\text{sim}}} (\widehat{\mu}_{\widehat{w},i} - \mu_{\text{true}})(\widehat{\mu}_{\widehat{w},i} - \mu_{\text{true}})^t$, where for each test data set with values x, $\widehat{\mu}_{\widehat{w},i} = \exp\{\sum_{j=1}^{M} \widehat{w}_j x^t \widehat{\delta}_j / \sqrt{n}\}$.

In this simulation, we estimated the weight vector w by minimizing separately the trace and the determinant of the MSE matrix. In order to distinguish the multivariate mMSE methods, we used mtrMSE and mdetMSE. Although, by minimizing the trace of the MSE matrix, we lose some precision in comparison with estimating a separate weight vector for each out-of-sample value, the comparison with other methods that estimate one weight vector for all test data is fairer; moreover, the computations were faster in the multivariate case than

when performing separate univariate optimizations. Table 6 reports the ratios of the trace of the empirical MSE matrix for the mtrMSE method divided by those for each other method. For the weight choice minimizing the determinant we report the ratio of the generalized standard deviations $\{\det(MSE)\}^{1/5}$ of the empirical MSE matrix over the simulation runs when using the weight choice by mdetMSE and that resulting by the other methods; if the number is bigger than 1, that other method performs better than the proposed method mMSE, and vice versa.

For singleton models, the AIC and BIC values are identical (the penalty does not have an effect in singleton models), hence, we show the results for AIC, SAIC and the mMSE methods. The AIC selects a single model, assigns weight 1 to that model and weight zero to all other models. The smoothed AIC, SAIC, gives weights proportional to the value of the AIC, the better the AIC value, the larger the weight. All weights in SAIC were rescaled to be in the interval $[0, 1]$.

Table 6 shows that in almost all settings and scenarios the mtrMSE method performs well, in some cases ten times better than other methods. The averaging method, SAIC, works better than AIC. In two settings, the SAIC method outperforms the mtrMSE method slightly, while in most of the other settings the mtrMSE method performs at least two times better than the other averaging method. In all settings, the mdetMSE method results in a relative low determinant value in comparison with the AIC and SAIC methods. The AIC and SAIC methods always perform worse than mdetMSE, while SAIC performs relatively better than the AIC selection method.

6. Data Analysis

6.1. Growth model, linear regression

We employed the proposed method of model averaging in a dataset used in several studies, including Liu (2015) and Magnus, Powell, and Prufer (2010). The economic growth measured as the gross domestic product per capita (GDPc), is modeled as a function of several covariates. A rise in GDP per capita shows growth in the economy and usually results in an increase in productivity. In the dataset there are 74 observations for average growth rate of GDP per capita between 1960 and 1996.

We compared the application of mMSE by the frequentist model averaging approach of Liu (2015). We adopted the model setup A in their study and fit a linear model as

$$GDPc_i = X_i\theta + Z_i\gamma + \epsilon_i,$$

with the same fixed regressors (X), a constant for the intercept, GDP60 which is the logarithm of the GPD in 1960, 'equipinv' the investment part of the GDP

Table 6. Simulation study. Top part of the table: ratio of the trace of the empirical MSE matrix for the mtrMSE method to the trace of the empirical MSE matrices resulting from the AIC and SAIC methods. Lower part of the table: ratio of the generalized standard deviation, that is, the determinant of the empirical MSE matrix over all simulation runs for the mdetMSE method to the determinant corresponding to the AIC and SAIC methods, raised to the power 1/5.

		Scenario 2.1				Scenario 2.2				Scenario 2.3			
		ρ				ρ				ρ			
method	n	0	0.25	0.5	0.75	0	0.25	0.5	0.75	0	0.25	0.5	0.75
tr(MSE) vs. AIC	50	0.507	0.226	0.795	0.428	0.487	0.241	0.613	0.659	0.577	0.131	0.668	0.891
	100	0.224	0.222	0.158	0.511	0.241	0.215	0.194	0.964	0.238	0.229	0.311	0.484
	200	0.088	0.180	0.390	0.470	0.125	0.187	0.521	0.446	0.185	0.258	0.461	0.736
	500	0.147	0.128	0.151	0.531	0.135	0.158	0.181	0.540	0.230	0.208	0.251	0.622
	1000	0.141	0.203	0.256	0.750	0.134	0.264	0.275	0.714	0.146	0.285	0.363	0.748
tr(MSE) vs. SAIC	50	0.506	0.232	0.794	0.475	0.485	0.253	0.613	0.676	0.572	0.147	0.664	0.990
	100	0.243	0.223	0.170	0.539	0.262	0.219	0.207	0.968	0.276	0.230	0.327	0.527
	200	0.100	0.189	0.413	0.552	0.143	0.192	0.543	0.441	0.221	0.267	0.532	1.050
	500	0.150	0.128	0.152	0.566	0.138	0.157	0.179	0.503	0.253	0.211	0.244	0.686
	1000	0.148	0.224	0.274	1.101	0.142	0.292	0.285	0.854	0.154	0.309	0.391	0.789
det(MSE) vs. AIC	50	0.155	0.129	0.217	0.065	0.146	0.163	0.164	0.126	0.190	0.152	0.191	0.231
	100	0.138	0.128	0.275	0.147	0.134	0.171	0.272	0.169	0.236	0.169	0.125	0.149
	200	0.148	0.046	0.547	0.121	0.106	0.059	0.709	0.151	0.307	0.128	0.378	0.163
	500	0.067	0.195	0.138	0.085	0.117	0.413	0.164	0.125	0.394	0.110	0.214	0.095
	1000	0.139	0.306	0.098	0.314	0.138	0.274	0.145	0.310	0.394	0.184	0.139	0.198
det(MSE) vs. SAIC	50	0.159	0.140	0.253	0.081	0.154	0.174	0.197	0.194	0.206	0.165	0.251	0.432
	100	0.151	0.149	0.333	0.195	0.148	0.203	0.326	0.241	0.249	0.206	0.167	0.293
	200	0.171	0.055	0.663	0.160	0.123	0.066	0.912	0.215	0.329	0.164	0.564	0.348
	500	0.083	0.244	0.174	0.120	0.142	0.354	0.225	0.174	0.425	0.150	0.290	0.231
	1000	0.168	0.384	0.115	0.434	0.170	0.326	0.182	0.483	0.417	0.218	0.209	0.456

during 1960–1985, 'school60' primary school enrollment rate in the year 1960, 'life60' the life expectancy at birth in 1960 and 'dpop' the population growth between 1960 and 1990. As potential regressors (Z) we took 'law' referred to as a rule of a law index, 'tropics' the fraction of tropical area of the country, 'avelf' which is an average index of ethnolinguistic fragmentation and 'confuc', the fraction of Confucian population. For more details see Magnus, Powell, and Prufer (2010).

Liu (2015) performed model averaging for all possible submodels and presented the estimated coefficients and weights for each model for the plug-in method and other methods of averaging, including OPT, MMA, and JMA. Those estimated coefficients and weights are not unique and one can find another estimate for the coefficients by changing the optimization routine. Table 7 presents two such sets of estimators by changing the optimization method (using *fmincon* instead of *quadprog* in Matlab, or changing the starting point). For the plug-in method we observed a large difference for the estimate of 'dpop'. The JMA estimates did not change that much, while MMA resulted in different estimates for 'dpop' and for the four potential z-variables. The method OPT showed the largest changes in values for nearly all of the variables, except for 'life60' and 'avelf'.

Liu (2015) used GDP60 as a focus, while here the goal was estimation of the coefficients for all regressors. We used the results on a multivariate focus of Section 4 defining $\mu^t = (\theta^t, \gamma^t)^t$, the full parameter vector, and we minimized the trace and next the determinant of the MSE matrix for this focus vector. Singleton models were used to ensure a unique solution for the trace method; for the determinant method the weights are not unique. Table 8 gives the estimated weighted coefficients when averaging over singleton models.

As Table 8 illustrates, some of the coefficients were estimated to be equal to zero. This is not surprising since this indicates that those variables are not correlated with growth per capita (see Magnus, Powell, and Prufer (2010)). The sign of 'dpop' that is found for the plug-in and mtrMSE methods is in line with Solow's model, see Solow (1956) and Durlauf, Kourtellos, and Tan (2008). The results for the mdetMSE method are close to those of the OPT method.

We further used 10-fold cross-validation and calculated the mean of the average squared prediction error (ASPE) for both averaging over singleton models and over all possible submodels. The chosen focuses are now the out-of-sample mean values. Table 9 reports the relative out-of-sample prediction errors. Entries larger than one indicate an inferior performance of that method relative to mtrMSE method. It should be noted that the focuses are out-of-sample observations not the coefficients. While all methods perform close to the mtrMSE method, especially using all subsets, the mtrMSE weight choice method outperforms all others.

Table 7. GDP data. Two different coefficient estimates show the non-uniqueness for averaging over all possible submodels for different weight choice methods.

method	constant	GDP60	equipinv	school60	life60	dpop	law	tropics	avelf	confuc
OPT	0.0596	-0.0156	0.1807	0.0172	0.0009	0.1785	0.0098	-0.0041	-0.0048	0.0345
	0.0560	-0.0149	0.1600	0.0165	0.0008	0.2267	0.0126	-0.0049	-0.0049	0.0527
MMA	0.0558	-0.0150	0.1526	0.0173	0.0008	0.2596	0.0144	-0.0057	-0.0039	0.0521
	0.0559	-0.0156	0.1511	0.0181	0.0009	0.2465	0.0166	-0.0043	-0.0026	0.0430
JMA	0.0559	-0.0156	0.1511	0.0181	0.0009	0.2465	0.0166	-0.0043	-0.0026	0.0430
	0.0558	-0.0155	0.1518	0.0180	0.0009	0.2443	0.0164	-0.0043	-0.0026	0.0431
Plug-in	0.0641	-0.0156	0.2263	0.0137	0.0010	0.0055	0.0000	-0.0000	-0.0104	0.0251
	0.0648	-0.0156	0.2241	0.0144	0.0010	0.0400	0.0000	-0.0016	-0.0099	0.0256

Table 8. GDP data. Estimated coefficients for the growth regression for singleton model averaging with different methods.

method	constant	GDP60	equipinv	school60	life60	dpop	law	tropics	avelf	confuc
OPT	0.0589	-0.0158	0.2146	0.0177	0.0010	0.0551	0.0041	-0.0017	-0.0022	0.0161
MMA	0.0467	-0.0138	0.1932	0.0157	0.0009	0.0275	0.0062	-	-	0.0489
JMA	0.0492	-0.0145	0.1882	0.0165	0.0009	0.0543	0.0086	-	-	0.0403
Plug-in	0.0434	-0.0124	0.2086	0.0136	0.0009	-0.0343	-	-	-0.0011	0.0637
mtrMSE	0.0414	-0.0121	0.2073	0.0137	0.0009	-0.0388	-	-	-	0.0667
mdetMSE	0.0596	-0.0159	0.2157	0.0178	0.0010	0.0565	0.0041	-0.0017	-0.0023	0.0140

Table 9. GDP data. Relative average out-of-sample squared prediction errors of competing methods relative to the mtrMSE method.

	Method				
models	OPT	MMA	JMA	Plug-in	mMSE
singleton models	1.157	1.086	1.087	1.060	1.000
all submodels	1.013	1.021	1.042	1.015	1.000

6.2. The automobile dataset, Gamma regression

Insurance companies wish to predict the losses that a car incurs based on its characteristics. The automobile dataset contains the normalized losses in use as compared to other cars. This dataset was obtained from the Machine Learning Repository at UCI (Bache and Lichman (2013)). There are 14 variables that can be used to model the losses. A description of all the variables is presented in Table 12 in the Appendix.

The response variable is 'nloss', positive and skewed to the right, motivating one to use gamma regression with a logistic link function. Out of 205 observations, we used the subset of 160 observations with no missing records. None of the variables was forced to be in a model. As a focus we took the vector of the regression coefficients. In this multivariate setting we minimized the trace of the estimated MSE matrix to determine the weights.

Table 10 presents the estimated coefficients for smoothed AIC and mMSE for singleton models, the corresponding weight for the singleton model is given between parentheses. The estimated coefficients are different even in the sign of each covariate. The highest weight is related to the biggest coefficient, width. This variable has been selected by the AIC method as the best model in singleton models.

Because of the large differences between SAIC and mMSE averaging methods in estimated coefficients, we assessed these methods by 5-fold crossvalidation. Four parts of the data set were used as training data and one part as a test data set. This procedure was repeated five times to consider each part in turn as a training data set. The focus vector was the expected response for the covariates included in the test data, $\mu = \exp(z\gamma)$. We estimated the mean of the response in gamma regression with a logistic link function and calculated the average squared prediction error (ASPE) for the test data sets,

$$\text{ASPE} = \frac{1}{5}\sum_{i=1}^{5}\frac{1}{n_i}\sum_{j=1}^{n_i}(y_{ij} - \widehat{\mu}_{ij})^2.$$

Table 10. Automobile data. The estimated coefficients and between parenthesis the weights assigned to the singleton models.

Variable	AIC	SAIC	mMSE	
wheelb	0	0.0055 (0.1118)	-0.0038	(-0.0773)
length	0	0.0029 (0.1038)	-0.0003	(-0.0089)
width	0.0732	0.0097 (0.1324)	0.0862	(1.1780)
height	0	0.0094 (0.1049)	-0.0331	(-0.3690)
cwei	0	0.0001 (0.0549)	-0.0001	(-0.0264)
engi	0	0.0020 (0.0465)	0.0029	(0.0675)
bore	0	0.1372 (0.0934)	-0.0571	(-0.0389)
stroke	0	0.1167 (0.0779)	-0.0524	(-0.0350)
compr	0	0.0205 (0.0395)	-0.0032	(-0.0061)
hpower	0	0.0019 (0.0346)	-0.0009	(-0.0154)
peak	0	0.0001 (0.0944)	0.0002	(0.2436)
cmpg	0	0.0078 (0.0395)	-0.0472	(-0.2403)
hmpg	0	0.0071 (0.0447)	0.0522	(0.3278)
price	0	0.0000 (0.0217)	0.0000	(0.0003)

Table 11. Automobile data. Crossvalidation results of relative AMSPE for singleton and all possible models as compared to that of the mMSE using only singleton models. For values smaller than one mMSE is best.

models	AIC	SAIC	BIC	SBIC
Singleton models	0.8396	0.2924	0.8396	0.2924
All submodels	1.1531	1.0457	1.0224	0.9932

Table 11 shows the relative risk of AIC, BIC, smooth AIC, and smooth BIC as compared to that of mMSE for weight selection. The average squared prediction error of the mMSE method in singleton models was divided by each other method's ASPE value. We calculated the crossvalidation results for singleton models with the mMSE method, but for other methods both the singleton models and all possible submodels were fitted. It can be seen that the mMSE method performs best in singleton models. The SAIC performs worse than the AIC resulting in some of the coefficients being weakly estimated. For singleton models there is no difference between AIC and BIC. The results for all possible submodels show that the singleton model averaging by the mMSE method performs comparable, and even better for SBIC to an all subsets averaging. The mMSE method here uses only 14 models instead of 2^{14} for the other methods. This small difference in performance often does not outweigh the increased computational cost.

7. Discussion

Minimum mean squared error weight choice is studied in a general setup, not restricting to linear normal models and not restricting the weights to belong to certain parametric classes. The broadness of the model scope avoids a case by case treatment for model averaging. All likelihood-based models can be averaged in this way. Existing studies of the mean squared error expression under local misspecification in settings such as generalized additive partially linear models (Zhang and Liang (2011)), Cox proportional hazard models (Hjort and Claeskens (2006)), and for quantile regression (Behl, Claeskens and Dette (2014)) open the way to construct similar extensions of the proposed mMSE weight choice method to those settings.

To the best of our knowledge we have not found other work dealing with the issue of non-unique estimators. Our work in linear models shows that even though the weights might not be unique, there are occasions where the predictions using mMSE weights are unique. This is a welcome relief when considering high-dimensional models. A reduction to singleton models, resulting in the same predictions as when using all subsets may be an interesting alternative to screening methods that first try to reduce the set of potential variables and then perform averaging. This topic is currently under investigation.

An interesting extension is the investigation of the limiting distribution for inference on the weighted estimator and on the post-selection estimator. The explicit form of the estimator for the mMSE method may aid the construction of confidence intervals and tests. This research is beyond the scope of this paper and deserves further attention.

Acknowledgements

We thank the reviewers for their helpful comments. We acknowledge the support of the Research Foundation Flanders, KU Leuven grant GOA/12/14 and of the IAP Research Network P7/06 of the Belgian Science Policy. Hansen thanks the National Science Foundation for research support. The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Hercules Foundation and the Flemish Government - department EWI. The authors wish to thank Chu-An Liu and Xinyu Zhang for providing their code for the plug-in and OPT methods.

Appendix. Proofs

Proof of Lemma 1. For all $x \in \mathbb{R}^M \setminus 0_M$,

$$\begin{aligned} x^t \widetilde{Q} x &= \sum_{j=1}^{M}\sum_{k=1}^{M} x_j x_k \mathrm{Tr}\big(Q^0_{S_j} Q^{-1} Q^0_{S_k} \nu\nu^t\big) \\ &= \mathrm{Tr}\Big(\big(\sum_{j=1}^{M} x_j Q^0_{S_j}\big) Q^{-1} \big(\sum_{k=1}^{M} x_k Q^0_{S_k}\big)\nu\nu^t\Big) > 0. \end{aligned}$$

The last inequality results from the fact that Q^{-1} is a positive definite matrix. By the linear independence assumption for the $Q^0_{S_j}$, $\sum_{j=1}^{M} x_j Q^0_{S_j} \neq 0$. Also, $\nu\nu^t$ is positive semi-definite. This proves the Lemma.

Proof of Theorem 1. (i) To connect with the notation of Theorem 2.7 of Kim and Pollard (1990), take $Z_n(w) = w^t \widehat{F} w$ and let $t_n = \widehat{w}_{\widehat{\mathrm{mse}}}$, then since $w^t \widehat{F} w \to_d w^t F^* w$ and $\widehat{w}_{\widehat{\mathrm{mse}}} = O_p(1)$, we have that $Z_n(t_n) \leq \inf_w Z_n(w) + a_n$ for random variables a_n of order $o_p(1)$. Hence all conditions of that theorem hold and we may conclude that $\widehat{w}_{\widehat{\mathrm{mse}}} \to_d w^*$. (ii) is proven by the joint convergence of $\widehat{w}_j$ and $\widehat{\mu}_j$ $(j = 1, \ldots, M)$.

Proof of Theorem 2. Without loss of generality, we present the proof for $Q = I_q$, $p = 1$, and $\sigma = 1$; it can be generalized, but with more calculations. We consider sets of singleton models and of nested models and denote the matrix F for nested and singleton models by F_{nest} and F_{sing}, respectively. Moreover, we use the exact values for ν and δ instead of their estimation for the ease of presentation.

First, consider nested models with partitioned matrices

$$F_{\mathrm{nest}} = \begin{pmatrix} T_{\mathrm{nest},1\times 1} & U^t_{\mathrm{nest},1\times q} \\ U_{\mathrm{nest},q\times 1} & W_{\mathrm{nest},q\times q} \end{pmatrix} \quad \text{and} \quad F^{-1}_{\mathrm{nest}} = \begin{pmatrix} T^{\mathrm{nest},1\times 1} & (U^{\mathrm{nest},1\times q})^t \\ U^{\mathrm{nest},q\times 1} & W^{\mathrm{nest},q\times q} \end{pmatrix},$$

where $T_{\mathrm{nest}} = \sum_{k=1}^{q}\sum_{l=1}^{q} z_k z_l \delta_k \delta_l$, $U_{\mathrm{nest},i} = \sum_{k=i+1}^{q}\sum_{l=1}^{q} z_k z_l \delta_k \delta_l$ for $i = 1, \ldots, q-1$, $U_{\mathrm{nest},q} = 0$ and W_{nest} is a symmetric matrix with entries

$$W_{\mathrm{nest},ij} = \begin{cases} \displaystyle\sum_{k=i+1}^{q}\sum_{l=j+1}^{q} z_k z_l \delta_k \delta_l + \sum_{k=1}^{i} z_k^2 & \text{if} \quad j \leq i, \quad i = 1, \ldots, q-1, \\ \displaystyle\sum_{k=1}^{i} z_k^2 & \text{if} \quad j = q \quad \text{and} \quad i = 1, \ldots, q. \end{cases}$$

We follow Harville (2000, Thm. 8.5.11) to calculate the inverse matrix F^{-1}_{nest}; this requires the invertibility of the symmetric matrix $N_{\mathrm{nest},q\times q} = W_{\mathrm{nest}}$

$-U_{\text{nest}}T_{\text{nest}}^{-1}U_{\text{nest}}^{t}$, guaranteed by Lemma 1, where $N_{\text{nest},ij} = \sum_{k=1}^{\min(i,j)} z_k^2 \quad i,j = 1,\ldots,q$; N_{nest} is a symmetric matrix and its inverse matrix N_{nest}^{-1} contains the entries

$$(N_{\text{nest}}^{-1})_{ij} = \begin{cases} \frac{z_i^2+z_{i+1}^2}{z_i^2 z_{i+1}^2}, \text{if} \quad i=j \quad \text{and} \quad i,j=1,\ldots,q-1, \\ \frac{-1}{z_{i+1}^2}, \quad \text{if} \quad i=j+1 \quad \text{and} \quad j=1,\ldots,q-1, \\ \frac{1}{z_q^2}, \quad \text{if} \quad i=j=q, \\ 0, \quad \text{otherwise.} \end{cases}$$

Using some matrix calculation,

$$T_{\text{nest}} = \frac{z_1^2(1+\sum_{k=2}^{q}\delta_k^2) + (\sum_{k=2}^{q}\sum_{l=2}^{q} z_k z_l \delta_k \delta_l)}{z_1^2(\sum_{k=1}^{q}\sum_{l=1}^{q} z_k z_l \delta_k \delta_l)},$$

$$U_{\text{nest},i} = \begin{cases} -\dfrac{\sum_{k=2}^{q} z_k\delta_k z_2 + z_1^2 d_2}{\sum_{k=1}^{q} z_k \delta_k z_1^2 z_2}, \text{if} \quad i=1, \\ \dfrac{z_{i+1}\delta_i - z_i\delta_{i+1}}{\sum_{k=1}^{q} z_k\delta_k z_i z_{i+1}}, \quad \text{if} \quad i=2,\ldots,q-1, \\ \dfrac{\delta_q}{\sum_{k=1}^{q} z_k\delta_k z_q}, \quad \text{if} \quad i=q \end{cases}$$

and $W_{\text{nest}} = N_{\text{nest}}^{-1}$. From (3.4), $w_{\text{mse}}^{\text{nest}}$ follows after computing

$$1_M^t F_{\text{nest}}^{-1} 1_M = \frac{1+\sum_{k=1}^{q}\delta_k^2}{\sum_{k=1}^{q}\sum_{l=1}^{q} z_k z_l \delta_k \delta_l}. \tag{A.1}$$

Considering the singleton models, we again partition the matrices F_{sing} and F_{sing}^{-1} as

$$F_{\text{sing}} = \begin{pmatrix} T_{\text{sing},1\times 1} & U_{\text{sing},1\times q}^t \\ U_{\text{sing},q\times 1} & W_{\text{sing},q\times q} \end{pmatrix} \quad \text{and} \quad F_{\text{sing}}^{-1} = \begin{pmatrix} T^{\text{sing},1\times 1} & (U^{\text{sing},1\times q})^t \\ U^{\text{sing},q\times 1} & W^{\text{sing},q\times q} \end{pmatrix},$$

where $T_{\text{sing}} = T_{\text{nest}}$, $U_{\text{sing},i} = (\sum_{k=1}^{q} z_k\delta_k - z_i\delta_i)(\sum_{k=1}^{q} z_k\delta_k)$, $\quad i = 1\ldots\ q$, and

W_{sing} is a symmetric matrix with (i,j)th entry equal to

$$W_{\text{sing},ij} = \begin{cases} (\sum_{k=1}^{q} z_k\delta_k - z_i\delta_i)(\sum_{k=1}^{q} z_k\delta_k - z_i\delta_i) + z_i^2 \text{ if } i = j \text{ and} \quad i = 1,\ldots,q, \\ (\sum_{k=1}^{q} z_k\delta_k - z_i\delta_i)(\sum_{k=1}^{q} z_k\delta_k - z_j\delta_j) \qquad \text{if } i \neq j \text{ and} \quad i,j = 1,\ldots,q. \end{cases}$$

Here $N_{\text{sing}} = \text{diag}(z_i^2)$ for $i = 1,\ldots,q$.

By some cumbersome calculations it can be shown that, with $S_i = \{i_k; k = 1,\ldots,q-1\}$,

$$T_{\text{sing}} = \frac{\sum_{i_1=1}^{2}\sum_{i_2=i_1+1}^{3}\cdots\sum_{i_j=i_{j-1}+1}^{j+1}\cdots\sum_{i_{q-1}=i_{q-2}+1}^{q}\Big(\prod_{i_k\in S_i} z_{i_k}^2\Big)\Big(\sum_{i_k\in S_i} z_{i_k}\delta_{i_k}\Big)^2 + \prod_{i=1}^{q} z_i^2}{\prod_{i=1}^{q} z_i^2 \sum_{k=1}^{q}\sum_{l=1}^{q} z_k z_l \delta_k\delta_l},$$

$$U_i^{\text{sing}} = -\frac{\sum_{k=1}^{q} z_k\delta_k) - z_i d_i}{z_i^2\sum_{k=1}^{q} z_k\delta_k}, \quad i = 1,\ldots,q,$$

and $W_{\text{sing}} = N_{\text{sing}}^{-1} = \text{diag}(-1/z_i^2)$. It follows that $1_M^t F_{\text{sing}}^{-1} 1_M = 1_M^t F_{\text{nest}}^{-1} 1_M$, see (A.1).

From (3.5) and (3.6), showing $\hat{\mu}_w^{\text{nest}} = \hat{\mu}_w^{\text{sing}}$ is equivalent to showing that

$$\begin{cases} w_2^{\text{nest}} + \ldots + w_{q+1}^{\text{nest}} = w_2^{\text{sing}}, \\ w_3^{\text{nest}} + \ldots + w_{q+1}^{\text{nest}} = w_3^{\text{sing}}, \\ \quad \vdots \\ w_{q+1}^{\text{nest}} = w_{q+1}^{\text{sing}}. \end{cases} \tag{A.2}$$

Since the denominators of the weights in nested and singleton models are equal, we need to consider the numerators which are equal to the sum of the elements in each column of F_{nest}^{-1} and F_{sing}^{-1}. The first weight w_0, which is related to the narrow model, does not have an effect in (A.2). By using the structure of F^{-1} in nested and singleton models, we can show that

$$(1_M^t F_{\text{nest}}^{-1} 1_M) w_i^{\text{nest}} = \begin{cases} \dfrac{z_{i+1}\delta_i - z_i\delta_{i+1}}{\sum_{k=1}^{q} z_k\delta_k z_i z_{i-1}}, \text{ if} \quad i = 1,\ldots,q-1, \\ \dfrac{\delta_q}{\sum_{k=1}^{q} z_k\delta_k z_q}, \qquad \text{if} \quad i = q, \end{cases}$$

and $(1_M^t F_{\text{sing}}^{-1} 1_M) w_i^{\text{sing}} = \delta_i/(\sum_{k=1}^{q} z_k\delta_k z_i), \quad i = 1,\ldots,q$. It is not difficult to show that (A.2) is satisfied and this completes the proof.

Description of the variables of the automobile data

Table 12. Description of the variables in the automobile dataset.

Variable	Description	Range
nloss	Normalized loss	65–256
wheelb	Wheel base of the car	86.6–120.9
length	Length of the car	141.1–208.1
width	Width of the car	60.3–72.3
height	Height of the car	47.8–59.8
cwei	Curb weight in thousands	1.488–4.066
engi	Engine size in hundreds	0.61–3.26
bore	Bore quantity	2.54–3.94
stroke	Stroke of the car	2.07–4.17
compr	Compression ratio	7–23
hpower	Horsepower of the car	48–288
peak	Peak revolutions per minute in hundreds	41.50–66.00
cmpg	City miles per gallon	13–49
hmpg	Highway miles per gallon	16–54
price	Price of the car in thousands	5.118–45.400

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. 2*nd International Symposium on Information Theory*, 267-281.

Antoniou, A. and Lu, W.-S. (2007). *Practical Optimization: Algorithms and Engineering Applications*. Springer, New York.

Atkinson, A., Denov, A. and Tobias, R. (2007). *Optimum Experimental Designs with SAS*. Oxford Univesity Press, Oxford.

Bache, K. and Lichman, M. (2013). UCI machine learning repository.

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

Behl, P., Claeskens, G. and Dette, H. (2014). Focused model selection in quantile regression. *Statist. Sinica* **24**, 601-624.

Cheng, X. and Hansen, B. E. (2015). Forecasting with factor-augmented regression: A frequentist model averaging approach. *J. Econom.* **186**, 280-293.

Claeskens, G. and Hjort, N. L. (2003). The focused information criterion. *J. Amer. Statist. Assoc.* **98**, 900-916.

Claeskens, G. and Hjort, N. L. (2008). *Model Selection and Model Averaging*. Cambridge University Press, Cambridge.

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

Dostal, Z. (2009). *Optimal Quadratic Programming Algorithms, with Applications to Variational Inequalities*. Springer, New York.

Durlauf, S., Kourtellos, A. and Tan, C. (2008). Are any growth theories robust? *Economic J.* **118**, 329-346.

Granger, C. and Ramanathan, R. (1984). Improved methods of combining forecast accuracy. *J. Forecast.* **19** 197-204.

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

Hansen, B. E. (2008). Least-squares forecast averaging. *J. Econom.* **146**, 342-350.

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

Harville, D. A. (2000). *Matrix Algebra From a Statisticians Perspective*. Springer, New York.

Hjort, N. L. and Claeskens, G. (2003). Frequentist model average estimators. *J. Amer. Statist. Assoc.* **98**, 879-899.

Hjort, N. L. and Claeskens, G. (2006). Focused information criteria and model averaging for the Cox hazard regression model. *J. Amer. Statist. Assoc.* **101**, 1449-1464.

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

Kabaila, P. (1995). The effect of model selection on confidence regions and prediction regions. *Econom. Theory* **11**, 537-549.

Kim, J. and Pollard, D. (1990). Cube root asymptotics. *Ann. Statist.* **18**, 191-219.

Liang, H., Zou, G., Wan, A. T. K. and Zhang, X. (2011). Optimal weight choice for frequentist model average estimators. *J. Amer. Statist. Assoc.* **106**, 1053-1066.

Liu, C.-A. (2015). Distribution theory of the least square averaging estimator. *J. Econom.* **186**, 142-159.

Magnus, J., Powell, O. and Prufer, P. (2010). A comparison of two model averaging techniques with an application to growth empirics. *J. Econom.* **154**, 139-153.

Mallows, C. L. (1973). Some comments on Cp. *Technometrics* **15**, 661-675.

Pötscher, B. (1991). Effects of model selection on inference. *Econom. Theory* **7**, 163-185.

Schwarz, G. (1978). Estimating the dimension of a models. *Ann. Statist.* **6**, 461-464.

Solow, R. (1956). A contribution to the theory of economic growth. *Quarterly J. Econom.* **70**, 65-94.

Wan, A., Zhang, X. and Zou, G. (2010). Least squares model averaging by Mallows criterion. *J. Econom.* **156**, 277-283.

Wan, A. T. K., Zhang, X. and Wang, S. (2013). Frequentist model averaging for multinomial and ordered logit models. *Internat. J. Forecast.* **30**, 118-128.

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

ORSTAT and Leuven Statistics Research Center, KU Leuven, Naamsestraat 69, 3000 Leuven, Belgium.

E-mail: ali.charkhi@kuleuven.be

ORSTAT and Leuven Statistics Research Center, KU Leuven, Naamsestraat 69, 3000 Leuven, Belgium.

E-mail: gerda.claeskens@kuleuven.be

Department of Economics, University of Wisconsin, 1180 Observatory Drive, Madison, Wisconsin 53706-1393, U.S.A.

E-mail: behansen@wisc.edu

(Received October 2014; accepted July 2015)