Computational Economics (2023) 61:1637–1664
https://doi.org/10.1007/s10614-022-10255-5

Weighted-Average Least Squares (WALS): Confidence and Prediction Intervals

Giuseppe De Luca[1] · Jan R. Magnus[2] · Franco Peracchi[3]

Accepted: 21 March 2022 / Published online: 22 April 2022

Abstract

We consider inference for linear regression models estimated by weighted-average least squares (WALS), a frequentist model averaging approach with a Bayesian flavor. We propose a new simulation method that yields re-centered confidence and prediction intervals by exploiting the bias-corrected posterior mean as a frequentist estimator of a normal location parameter. We investigate the performance of WALS and several alternative estimators in an extensive set of Monte Carlo experiments that allow for increasing complexity of the model space and heteroskedastic, skewed, and thick-tailed regression errors. In addition to WALS, we include unrestricted and fully restricted least squares, two post-selection estimators based on classical information criteria, a penalization estimator, and Mallows and jackknife model averaging estimators. We show that, compared to the other approaches, WALS performs well in terms of the mean squared error of point estimates, and also in terms of coverage errors and lengths of confidence and prediction intervals.

Keywords Linear model · WALS · Confidence intervals · Prediction intervals · Monte Carlo simulations

JEL Classification C11 · C12 · C18 · C21 · C52

We thank Paolo Li Donni, Chu-An Liu, and Xinyu Zhang for useful discussions; and the referees for their constructive comments.

✉ Giuseppe De Luca
giuseppe.deluca@unipa.it

[1] University of Palermo, Palermo, Italy

[2] Vrije Universiteit Amsterdam and Tinbergen Institute, Amsterdam, The Netherlands

[3] University of Rome Tor Vergata and EIEF, Rome, Italy

1 Introduction

Many empirical studies in economics assume that the data are generated by a linear regression model where a distinction is made between 'focus regressors' and 'auxiliary regressors'. The focus regressors are included because we believe the model is not credible without them or because they are the subject of our investigation, while the number and the identity of the auxiliary regressors is less certain. The parameters of primary interest are the coefficients on the focus regressors (the 'focus parameters'), while the coefficients on the auxiliary regressors are treated as nuisance parameters. Instead of a single model for the data generating process (DGP), there is a 'model space' containing a finite but potentially large number of models, namely the unrestricted model that includes all auxiliary regressors, the fully restricted model that includes none, and all intermediate models. Adding auxiliary regressors tends to reduce omitted variable bias in estimating the focus parameters, but tends to increase sampling variability. Examples include studies concerning the determinants of economic growth (Sala-i-Martin et al. 2004; Magnus et al. 2010), risk premia (Sousa and Sousa 2019), product and labor market reforms (Duval et al. 2021), the impact of legalized abortion on crime (Donohue and Levitt 2001), and the relationship between body mass and income (Dardanoni et al. 2011).

Model uncertainty can be approached via 'model selection' or via 'model averaging'. In the model selection approach we attempt to find the 'best' model given the data, the model space, and a specific purpose (e.g., estimation of particular parameters or prediction of future outcomes). Given this best model, one then employs its estimates for the intended purpose. Like any other data-driven statistical decision, model selection is subject to sampling uncertainty which, if ignored, can lead to overestimate accuracy (Kabaila and Mainzer 2018). Typical examples are the classical pre-test estimator and post-selection estimators that select the model with the smallest value of some information criterion, such as the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC). More recently, considerable attention has been devoted to penalization estimators based on model sparsity and an absolute penalty criterion, such as the least absolute shrinkage and selection operator (LASSO), which address the sampling uncertainty problem by performing variable selection and regularization at the same time. These estimators typically require the choice of some 'tuning' parameters that control the trade-off between bias and variance. They also tend to be biased and to have nonstandard sampling distributions, so that inference based on the normal approximation can be misleading (Knight and Fu 2000; Claeskens and Hjort 2008).

The second approach is model averaging. In contrast to model selection, one is not concerned with finding a 'best' model but with finding a 'best' estimator of the focus parameters or a 'best' predictor of the outcome. The (well-established) terminology is a little confusing because we don't average over models but over estimators. In fact, one takes a weighted average of the estimators from all the available models, with data-dependent weights to account for the uncertainty associated with each model. There are many proposed model averaging estimators, typically obtained either from a Bayesian perspective (Bayesian model averaging: BMA) or from a frequentist perspective (frequentist model averaging: FMA). BMA weights can be interpreted as posterior model probabilities, while FMA weights are decreasing functions of some measure of predictive inaccuracy, such as Mallows' C_p (Hansen 2007) or leave-one-out cross-validation (Hansen and Racine 2012). There also exist Bayesian-frequentist 'fusions', such as weighted-average least squares (WALS), introduced by Magnus et al. (2010), which is frequentist but with a Bayesian flavor. We refer to Steel (2020) for an extensive survey of the various types of model averaging estimators and their use in economics. Like for model selection estimators, most of these estimators tend to be biased and their sampling distribution is not well approximated by the normal distribution. Furthermore, there is increasing evidence that, even after correcting for bias, inference for model averaging estimators can be misleading if based on the normal approximation (see, among others, Claeskens and Hjort 2008; Hansen 2014; Liu 2015; and DiTraglia 2016).

The finite-sample bias and variance of WALS have recently been analyzed by De Luca et al. (2021), who exploit results on the frequentist properties of the Bayesian posterior mean in a normal location model. The current paper extends their results to inference by proposing a simulation-based approach that yields re-centered confidence and prediction intervals using the bias-corrected posterior mean as a frequentist estimator of the normal location parameter. We assess its finite-sample performance by an extensive Monte Carlo experiment. To facilitate comparisons with the simulation study by Zhang and Liu (2019), we stay close to their framework and consider a finite model space that contains the true data-generating process (M-closed environment) but has little additional structure. Unlike Zhang and Liu (2019), who restrict attention to inference about a single auxiliary parameter, we consider inference about a single focus parameter, interpreted as the causal effect of a policy or intervention in the presence of a potentially large number of auxiliary parameters. This is likely to be the most interesting case for applied economists. We compare the performance of WALS point estimates and confidence intervals with the performance of several competing approaches, including least squares estimators for the unrestricted and fully restricted models, post-selection estimators based on AIC and BIC, Mallows and jackknife model averaging estimators, and one version of the LASSO (the adaptive LASSO). In addition, we discuss prediction intervals for the outcome of interest, which involves linear combinations of all focus and auxiliary parameters. The main conclusion of our Monte Carlo experiment is that, compared to other estimators, the coverage errors for WALS are small and confidence and prediction intervals are short, centered correctly, and allow for asymmetry. They are also easy and fast to compute.

The remainder of this paper is organized as follows. Section 2 introduces the framework and briefly describes the estimators that we consider. Section 3 discusses how to construct confidence intervals for a single parameter of interest. Section 4 describes the Monte Carlo experiment. Sections 5–7 contain the simulation results, separately for point estimates (Sect. 5), confidence intervals (Sect. 6), and prediction intervals (Sect. 7). Section 8 concludes. There are two appendices. Appendix A formalizes the nine estimators introduced in Sect. 2, while Appendix B describes the algorithm for simulation-based WALS confidence intervals.

2 Framework and Estimators

Our framework is the linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{y}$ $(n \times 1)$ is the vector of observations on the outcome of interest, $\boldsymbol{X}_1$ $(n \times k_1)$ and $\boldsymbol{X}_2$ $(n \times k_2)$ are matrices of nonrandom regressors, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are unknown parameter vectors, and $\boldsymbol{\epsilon}$ is a vector of random disturbances. The k_1 columns of $\boldsymbol{X}_1$ contain the 'focus regressors' which we want in the model on theoretical or other grounds, while the k_2 columns of $\boldsymbol{X}_2$ contain the 'auxiliary regressors' of which we are less certain. These auxiliary regressors could be controls that are added to avoid omitted-variable bias or transformations and interactions of the set of original regressors to allow for nonlinearities. We assume that $k_1 \geq 1$, $k_2 \geq 0$, and that $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ has full column-rank $k = k_1 + k_2 \leq n$. The disturbance vector $\boldsymbol{\epsilon}$ has zero mean and a positive definite variance matrix, diagonal but not necessarily scalar. The DGP thus allows for nonnormality and heteroskedasticity.

Table 1 lists the nine estimators of $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$ that we consider in this paper. Except for LS-R and WALS, all other estimators also appear in Zhang and Liu (2019). In the remainder of this section, we describe briefly the various estimators with some emphasis on WALS. Appendix A provides a more detailed description of all estimators.

Our first two estimators are the least squares (LS) estimators of $\boldsymbol{\beta}$ in the unrestricted model that includes all auxiliary regressors and the fully restricted model that includes none. We shall refer to these two estimators as the unrestricted LS (LS-U) estimator and the fully restricted LS (LS-R) estimator, respectively. Under an $\mathcal{M}$-closed environment, the LS-U estimator is unbiased but is likely to have a large variance, especially when the sample size is small, the number of auxiliary variables is large, and the regressors are highly correlated. The LS-R estimator is subject to omitted variable bias when $\boldsymbol{X}_1'\boldsymbol{X}_2 \neq \boldsymbol{0}$ and $\boldsymbol{\beta}_2 \neq \boldsymbol{0}$, but has a smaller variance than the LS-U estimator under homoskedastic errors. These estimators require neither model selection nor model averaging as they rely on two *ad hoc* specifications of the unknown DGP.

When we account explicitly for uncertainty about the auxiliary regressors, the model space contains $J = 2^{k_2}$ possible models. Model selection tries to find a single 'best' model based on a specific criterion, while model averaging takes a weighted average of the estimators from all the models in the model space. For example, if $\widehat{\boldsymbol{\beta}}_{1j}$ and $\widehat{\boldsymbol{\beta}}_{2j}$ are the LS estimators of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ in model j, then a model averaging estimator takes the form

$$\widehat{\boldsymbol{\beta}}_1 = \sum_{j=1}^{J} \lambda_j \widehat{\boldsymbol{\beta}}_{1j}, \qquad \widehat{\boldsymbol{\beta}}_2 = \sum_{j=1}^{J} \lambda_j \widehat{\boldsymbol{\beta}}_{2j}, \tag{2}$$

Table 1 The estimators

Estimator	Description
Least squares (LS)	
LS-U	LS estimator of the unrestricted model with all auxiliary regressors
LS-R	LS estimator of the fully restricted model with no auxiliary regressors
Post-selection estimators based on information criteria	
IC-A	LS estimator of the model with smallest AIC
IC-B	LS estimator of the model with smallest BIC
Penalization methods	
ALASSO	Adaptive LASSO estimator (Zou 2006) with penalty parameter chosen by generalized cross-validation
Frequentist model averaging	
MMA	Mallows model averaging (Hansen 2007) with preordering ofthe auxiliary regressors
JMA	Jackknife model averaging (Hansen and Racine 2012)with preordering of the auxiliary regressors
JMA-M	Modified JMA estimator (Zhang and Liu 2019) with preordering of the auxiliary regressors and penalty parameter equal to $\log n$
Bayesian combination of frequentist estimators	
WALS	Weighted-average least squares (Magnus et al. 2010)

where the λ_j are nonnegative data-dependent model weights that add up to one. Even for moderate values of k_2 the computational burden of calculating 2^{k_2} estimates and the associated model weights can be substantial.

One possibility is to reduce the number of models by preordering, as suggested by Hansen (2007). If we can order the auxiliary regressors *a priori*, then we only need to consider $k_2 + 1$ nested models, with model p containing the focus regressors and the first p auxiliary regressors. Except for a few cases in which the auxiliary regressors admit a natural preordering (e.g., polynomial regression models), the question of how we should order the auxiliary regressors is difficult to answer, and if we use preliminary regressions to order the regressors then the statistical noise generated by these preliminary investigations should not be ignored.

Two common model selection strategies are based on information criteria such as AIC and BIC. AIC and BIC are known to represent two extreme strategies favoring, respectively, more and less complicated model structures. The IC-A and IC-B post-selection estimators are the LS estimators in the models with the smallest AIC and BIC respectively. As implemented in Zhang and Liu (2019), these estimators require preordering and the assumption of homoskedastic errors. There is no model averaging here, only model selection.

The adaptive LASSO (ALASSO) estimator, proposed by Zou (2006), does not rely on preordering. It solves a penalized LS problem with a penalty on the weighted sum of the absolute values of the estimated components of the full vector $\boldsymbol{\beta}$ and weights that depend on the LS-U estimates and a tuning parameter selected by generalized cross-validation. Following Zhang and Liu (2019), this version of the ALASSO estimator does not distinguish between focus and auxiliary regressors.

The Mallows model averaging (MMA) estimator was introduced by Hansen (2007). Although it can be applied to the full model space consisting of 2^{k_2} models, it is typically based on preordering in order to reduce the computational burden. This estimator is asymptotically efficient in the mean squared error (MSE) sense when the errors in (1) are homoskedastic (Hansen 2007; Wan et al. 2010; Zhang 2021).

The jackknife model averaging (JMA) estimator, introduced by Hansen and Racine (2012) , is also generally based on preordering but allows for heteroskedasticity. Under homoskedasticity it has the same (nonstandard) limiting distribution as MMA (Zhang and Liu 2019, p. 824) and it remains asymptotically efficient under heteroskedasticity (Hansen and Racine 2012; Zhang 2021). The modified JMA (JMA-M) estimator, introduced by Zhang and Liu (2019), is similar but is defined by weights that minimize a penalized cross-validation criterion.

The weighted-average least squares (WALS) estimator was introduced by Magnus et al. (2010) and reviewed by Magnus and De Luca (2016). Unlike other model averaging estimators, the WALS approach exploits a preliminary transformation of the auxiliary regressors that reduces the computational burden from order 2^{k_2} to order k_2 and leads to other important simplifications. In particular, after this transformation, model (1) may equivalently be written as

$$\boldsymbol{y} = \boldsymbol{Z}_1\boldsymbol{\gamma}_1 + \boldsymbol{Z}_2\boldsymbol{\gamma}_2 + \boldsymbol{\epsilon}, \tag{3}$$

where $\boldsymbol{Z}_2'\boldsymbol{M}_1\boldsymbol{Z}_2$ is equal to the identity matrix of order k_2. The WALS estimator $\widehat{\boldsymbol{\gamma}} = (\widehat{\boldsymbol{\gamma}}_1', \widehat{\boldsymbol{\gamma}}_2')$ of $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)$ is a weighted average of the LS estimators of $\boldsymbol{\gamma}$ over the J models in the model space.

From Theorem 2 of Magnus and Durbin (1999), the MSE of $\widehat{\boldsymbol{\gamma}}_1$ depends on the MSE of $\widehat{\boldsymbol{\gamma}}_2$. Thus, if we can choose the model weights optimally such that $\widehat{\boldsymbol{\gamma}}_2$ is a 'good' estimator of $\boldsymbol{\gamma}_2$ (in the MSE sense), the same weights will also provide a 'good' estimator of $\boldsymbol{\gamma}_1$. Moreover, the dependence of $\widehat{\boldsymbol{\gamma}}$ on the estimators from all possible models is completely captured by a random diagonal matrix $\boldsymbol{W}$, whose k_2 diagonal elements are partial sums of the model weights λ_j in (2). It follows that we can restrict attention to the WALS estimator of $\boldsymbol{\gamma}_2$, whose computational burden is of order k_2 as we need to determine only the diagonal elements of $\boldsymbol{W}$, not the full set of model weights.

The components of $\widehat{\boldsymbol{\gamma}}_2$ are shrinkage estimators of the components of $\boldsymbol{\gamma}_2$. Under the assumption of homoskedastic normal errors in (1) and the additional restriction that the hth diagonal element of the matrix $\boldsymbol{W}$ depends only on the hth component of the LS-U estimator of $\boldsymbol{\gamma}_2$, our shrinkage estimators are also independent. The initial k_2-dimensional problem then reduces to k_2 identical one-dimensional problems, namely: given a single observation x from the normal location model $\mathcal{N}(\eta, \sigma^2)$, what is the estimator $m(x)$ of η with minimum MSE? Since the risk properties of $m(x)$ are little affected by estimating the variance parameter (Danilov 2005), we assume that σ^2 is known.

A Bayesian approach to the above problem requires two elements: a normal location model for the independently and identically distributed (i.i.d.) elements $\{x_h\}$ of the vector of t-ratios $\boldsymbol{x} = \widehat{\boldsymbol{\gamma}}_{2,u}/s_u$, where s_u^2 is the unbiased LS estimator of the error variance; and a prior distribution for the i.i.d. elements $\{\eta_h\}$ of the vector of 'theoretical' t-ratios $\boldsymbol{\eta} = \boldsymbol{\gamma}_2/\sigma_u$. For a proper treatment of admissibility, robustness, near-optimality in terms of minimax regret, and ignorance about η_h, we select a prior that is symmetric, leads to bounded risk, and satisfies the 'neutrality condition' $\mathbb{P}[|\eta_h| < 1] = 1/2$. The Bayesian approach to the normal location problem then yields the posterior mean $m_h = m(x_h)$ as an estimator of η_h, from which the WALS estimators of $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ and therefore of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are easily derived (see Appendix A for the details).

The mixture of Bayesian and frequentist approaches requires special attention when assessing the sampling properties of our model averaging estimator. First, for a prior which is symmetric around zero, the posterior mean m_h suffers from attenuation bias, so $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$ are in general biased estimators of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. Second, for any nonnegative bounded prior density, the posterior variance of η_h represents a first-order approximation to the sampling standard deviation (not the sampling variance) of the posterior mean m_h.

De Luca et al. (2021) presented Monte Carlo tabulations of the bias and variance of m_h under three neutral priors: Laplace, Weibull, and Subbotin. For each prior considered, they also compared two alternative plug-in estimators of these sampling moments of m_h: the frequentist maximum likelihood (ML) estimator and the Bayesian double shrinkage (DS) estimator. Based on these plug-in estimators, they derived new estimators for the sampling bias and variance of the WALS estimator. This paper investigates the implications of their findings for the construction of WALS confidence and prediction intervals.

3 Confidence Intervals

We concentrate on $(1-\alpha)$-level confidence intervals for the lth component β_l of $\boldsymbol{\beta}$, which could be either a focus or an auxiliary parameter. All confidence intervals take the form

$$\mathrm{CI}(\beta_l) = \left[\check{\beta}_l - \underline{c}_l, \check{\beta}_l + \overline{c}_l\right], \tag{4}$$

where $\check{\beta}_l$ is an estimator of β_l and the quantities $\underline{c}_l$ and $\overline{c}_l$ are chosen to attain the desired coverage level. If $\underline{c}_l = \overline{c}_l$, the interval is called symmetric. We consider sixteen types of confidence intervals — ten from Zhang and Liu (2019) and six based on WALS — that differ depending on the choice of $\check{\beta}_l$, $\underline{c}_l$, and $\overline{c}_l$.

LS-U and LS-R: $\check{\beta}_l$ is either the LS-U or the LS-R estimator and $\underline{c}_l = \overline{c}_l = z_{1-\alpha/2}\, s_l$, where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$th quantile of the standard normal distribution and s_l is the standard error of $\check{\beta}_l$.

IC-A and IC-B: $\check{\beta}_l = \widehat{\beta}_l(\widehat{p})$ and $\underline{c}_l = \overline{c}_l = z_{1-\alpha/2}\, s_l$, where $\widehat{\beta}_l(\widehat{p})$ is the LS estimator in the model with the smallest AIC or BIC, $\widehat{p}$ is the number of auxiliary regressors in the selected model, and s_l is the standard error of $\check{\beta}_l$. Zhang and Liu (2019) call these confidence intervals 'naive' because they ignore model selection noise.

ALASSO: $\check{\beta}_l$ is the ALASSO estimator and $\underline{c}_l = \overline{c}_l = n^{-1/2} q_l^*(\alpha)$, where $q_l^*(\alpha)$ is the αth quantile of the conditional distribution of $|\sqrt{n}(\check{\beta}_l^* - \check{\beta}_l)|$ given the data and $\check{\beta}_l^*$ is the ALASSO estimate from a bootstrap sample. These confidence intervals and rely on the asymptotic validity of the bootstrap for the ALASSO estimator (Chatterjee and Lahiri 2011; Camponovo 2015).

MMA: $\check{\beta}_l$ is the MMA estimator and we consider two alternative approaches to the choice of $\underline{c}_l$ and $\overline{c}_l$. In the bootstrap approach (MMA-B) we set $\underline{c}_l = \overline{c}_l = n^{-1/2} q_l^*(\alpha)$, where $q_l^*(\alpha)$ is the αth quantile of the bootstrap distribution of $|\sqrt{n}(\check{\beta}_l^* - \check{\beta}_l)|$ and $\check{\beta}_l^*$ is the MMA estimate from a bootstrap sample, while in the simulation-based approach (MMA-S) we set $\underline{c}_l = n^{-1/2} q_l(1-\alpha/2)$ and $\overline{c}_l = -n^{-1/2} q_l(\alpha/2)$, where $q_l(\alpha)$ is the αth quantile of the simulated asymptotic distribution of the estimator based on Zhang and Liu (2019, Theorem 2). The first interval is symmetric, the second is not.

JMA: $\check{\beta}_l$ is the JMA estimator and we again consider two alternative approaches to the choice of $\underline{c}_l$ and $\overline{c}_l$. In the first (JMA-B) we set $\underline{c}_l = \overline{c}_l = n^{-1/2} q_l^*(\alpha)$, where $q_l^*(\alpha)$ is the αth quantile of the bootstrap distribution of $|\sqrt{n}(\check{\beta}_l^* - \check{\beta}_l)|$ and $\check{\beta}_l^*$ is the JMA estimate from a bootstrap sample, while in the second (JMA-S) we set $\underline{c}_l = n^{-1/2} q_l(1-\alpha/2)$ and $\overline{c}_l = -n^{-1/2} q_l(\alpha/2)$, where $q_l(\alpha)$ is based on Zhang and Liu (2019, Theorem 4).

JMA-M: $\check{\beta}_l$ is the JMA-M estimator and $\underline{c}_l = \overline{c}_l = z_{1-\alpha/2}\, s_l^*$, where s_l^* is the standard error in the 'just-fitted' model, that is, the model obtained from the ordered sequence of models by deleting all redundant regressors at the end of the sequence. [1] Symmetry of these intervals is justified by the asymptotic normality of the JMA-M estimator (Zhang and Liu 2019, Theorem 5).

WALS: We consider three different methods for constructing confidence intervals, namely uncentered-and-naive (UN), centered-and-naive (CN), and simulation-based (S). The algorithm underlying the last two methods is presented in Appendix B.

In the UN method, $\check{\beta}_l$ is the WALS estimator $\widehat{\beta}_l$, $\underline{c}_l = \overline{c}_l = z_{1-\alpha/2}\, s_l$, and s_l is either the plug-in ML estimator or the plug-in DS estimator of the standard error of $\widehat{\beta}_l$. The resulting intervals take the classical normal approximation to the sampling distribution of $\widehat{\beta}_l$ at face value and neglect the bias of the WALS estimator.

In the CN method, $\check{\beta}_l$ is the bias-corrected WALS estimator, $\check{\beta}_l = \widehat{\beta}_l - b_l$, where b_l is either the plug-in ML estimator or the plug-in DS estimator of the bias of $\widehat{\beta}_l$. As in the UN method, $\underline{c}_l = \overline{c}_l = z_{1-\alpha/2}\, s_l$, but now s_l depends on the bias-corrected WALS estimator and is computed by the simulation-based algorithm discussed in Appendix B. The CN method again takes the classical normal approximation at face

[1] The just-fitted model is unknown in practice, so s_l^* is not a feasible estimator. In the simulations we follow Zhang and Liu (2019) and assume that the just-fitted model is known. As a consequence, the correct intervals will be larger than reported since some of the model selection noise has been ignored.

Table 2 Eight error distributions

Skedasticity	Distribution	u_i	σ_i
Homoskedastic	1	$\mathcal{N}(0, 1)$	2.5
	2	$t(5)$	$\sqrt{15/4}$
	3	$t^*(0, 1, 5, 0.5)$	2.5
	4	$t^*(0, 1, 5, 0.8)$	2.5
Heteroskedastic	5	$\mathcal{N}(0, 1)$	$2.5\,\tau_i$
	6	$t(5)$	$\sqrt{15/4}\,\tau_i$
	7	$t^*(0, 1, 5, 0.5)$	$2.5\,\tau_i$
	8	$t^*(0, 1, 5, 0.8)$	$2.5\,\tau_i$

value but re-centers to correct for estimation bias and accounts for randomness in the estimated bias.

The S method also yields re-centered confidence intervals by using the bias-corrected posterior mean as an estimator of the normal location parameter, and accounts for its randomness by exploiting a large set of pseudo-random Monte Carlo draws. However, since it does not require critical values from the normal distribution, its confidence intervals are not necessarily symmetric.

4 Monte Carlo Design

Our setup closely follows Zhang and Liu (2019) with some exceptions explained later in this section. We have $k_1 = 2$ focus regressors: $\boldsymbol{x}_{11}$ (the constant term) and $\boldsymbol{x}_{12}$; and k_2 auxiliary regressors: $\boldsymbol{x}_{21}, \dots, \boldsymbol{x}_{2k_2}$. Our parameter of interest is the coefficient β_{12} on $\boldsymbol{x}_{12}$, which may be interpreted as the causal effect of $\boldsymbol{x}_{12}$ on $\boldsymbol{y}$.

The $k_2 + 1$ regressors $\boldsymbol{x}_{12}, \boldsymbol{x}_{21}, \dots, \boldsymbol{x}_{2k_2}$ are drawn from a multivariate normal distribution with mean zero and variance $\sigma_x^2\, \boldsymbol{\Sigma}_x(\rho)$, where

$$\boldsymbol{\Sigma}_x(\rho) = \begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & & \vdots \\ \rho & \rho & \dots & 1 \end{pmatrix},$$

with $-1/k_2 < \rho < 1$. We set $\sigma_x^2 = \rho = 0.7$.

The error term is generated by $\epsilon_i = \sigma_i u_i$, where the u_i are independently distributed following either a standard normal distribution or a skewed t^*-distribution $t^*(\mu, \sigma, d, \lambda)$ with mean μ, variance σ^2, d degrees of freedom and skewness parameter $|\lambda| < 1$ (defined for $d > 3$). In addition to the standard normal distribution, we consider three skewed t^*-distributions with $\mu = 0$ and $d = 5$: (i) the standard $t(5)$-distribution, which is obtained by setting $\sigma = \sqrt{5/3}$ and $\lambda = 0$, (ii) a distribution with moderate positive skewness ($\sigma = 1$ and $\lambda = 0.5$), and (iii) a distribution with large positive skewness ($\sigma = 1$ and $\lambda = 0.8$). We also consider four homoskedastic and four heteroskedastic error distributions, as shown in Table 2. In the homoskedastic cases

Table 3 Four configurations of the $k_2 = 8$ auxiliary parameters

Conf.	$\boldsymbol{\beta}_2$
(a)	$(\xi, \xi^2, \xi^3, \xi^4, 0, 0, 0, 0)'$
(b)	$(\xi^4, \xi^3, \xi^2, \xi, 0, 0, 0, 0)'$
(c)	$(\xi, \xi^2, 0, 0, \xi^3, \xi^4, 0, 0)'$
(d)	$(0, 0, 0, 0, \xi^4, \xi^3, \xi^2, \xi)'$

we take $\sigma_i = 2.5$ when the distribution of u_i has variance one. For the standard $t(5)$-distribution the variance is $5/3$, so we need the correction factor $2.5/\sqrt{5/3} = \sqrt{15/4}$. In the heteroskedastic cases we define

$$\tau_i = \frac{1 + 2|x_{12}^{(i)}| + 4|x_{21}^{(i)}|}{1 + 6\sigma_x\sqrt{2/\pi}},$$

where $x_{12}^{(i)}$ and $x_{21}^{(i)}$ respectively denote observation i on the second focus regressor and the first auxiliary regressor, and the scaling is chosen such that $\mathbb{E}[\tau_i] = 1$ for all i.

Setting $k_2 = 8$, we have $2^{k_2} = 256$ possible models that include the two focus regressors and a subset of the eight auxiliary regressors. We fix $\boldsymbol{\beta}_1 = (1, 1)'$ and consider four configurations of the eight auxiliary parameters, as shown in Table 3.

Our setup is intentionally similar to that in Zhang and Liu (2019) with three important exceptions:

- Our parameter of interest is one of the focus parameters, not one of the auxiliary parameters, because it is focus parameters that we are primarily interested in.
- Zhang and Liu (2019) ignore the possibility of skewness in the error distribution. In fact, of the eight cases in Table 2 they only consider two: homoskedastic under normality (case 1) and heteroskedastic under a t-distribution (case 6). In the heteroskedastic setup we take 5 rather than 4 degrees of freedom, so as to ensure the existence of both skewness and kurtosis. In addition, our scaling in design 6 gives $\mathbb{E}[\sigma_i] = 2.5/\sqrt{\mathbb{V}[t(5)]} \approx 1.94$ thus ensuring comparability with the other designs, whereas in the case considered by Zhang and Liu (2019) we would have $\mathbb{E}[\sigma_i] \approx 3.23$. Finally, we let τ_i depend on one focus and one auxiliary regressor (instead of two auxiliary regressors).
- To the three cases (a)–(c) in Table 3, we have added case (d) to show what can happen when the preliminary ordering is poor. As in case (b), the auxiliary regressors with nonzero coefficients enter with a decreasing order of importance as measured by the magnitude of their coefficients (since we set $|\xi| < 1$). In addition, case (d) implies that all submodels in the preordered sequence of $k_2 + 1$ nested models (except for the unrestricted model) are subject to omitted-variable bias.

We set $\xi = 0.5$ and consider sample sizes of $n = 100$ and $n = 400$. By combining the eight specifications of the regression error in Table 2 with the four configurations of the auxiliary parameters in Table 3, we obtain 32 simulation designs for $n = 100$ and 32 simulation designs for $n = 400$. Using 5,000 Monte Carlo replications for each design (instead of 500 replications as in Zhang and Liu 2019), we compute the bias, variance, and MSE of the nine estimators discussed in Sect. 2: LS-U, LS-R, IC-A, IC-B, ALASSO, MMA, JMA, JMA-M, and WALS. The LS-U, LS-R and WALS estimators are implemented in Stata, the other estimators in MATLAB. [2] Since WALS has been shown to be robust to different choices of the prior (De Luca et al. 2018; De Luca et al. 2021), we focus on the Laplace prior to exploit its computational advantages in computing the posterior mean.

5 Monte Carlo Results: Point Estimates

In this and the next two sections we present the results of the Monte Carlo experiment in a number of graphs. The current section discusses point estimates; confidence intervals and prediction intervals are discussed in Sects. 6 and 7 , respectively.

In Figs. 1 and 2 we present the first two sampling moments of the nine estimators for $n = 100$. The sixteen plots in Fig. 1 represent the homoskedastic designs, the sixteen plots in Fig. 2 the heteroskedastic designs. Each plot contains the squared bias–variance decomposition of the MSE of the nine estimators and, in addition, two 'iso-MSE' lines, which consist of all points with the same MSE as the LS-U estimator (red dash-dotted line) and the WALS estimator (blue dashed line). Design 1a refers to distribution 1 (normal, homoskedastic) and configuration (a), and so on, as described in Tables 2 and 3 .

The similarity of the sixteen plots in Fig. 1 is remarkable. The LS-U, LS-R, ALASSO, and WALS estimators are not affected by preordering, hence their moments and MSEs are the same across configurations. This is not the case for the other five estimators, IC-A, IC-B, MMA, JMA, and JMA-M, for which the effect of preordering can be substantial (comparing across rows), but the effect of nonnormality (skewness and excess kurtosis) appears to be small (comparing across columns). The LS-R estimator has a large bias which dominates the small variance, and hence its MSE is large. ALASSO has a small bias but a large variance, hence a large MSE. The MSE is also large for IC-B based on the BIC criterion because of its large bias, especially in configurations (b) and (d) where the ordering is unfavorable. The IC-A estimator based on the AIC criterion behaves about the same as the LS-U estimator in configurations (a) and (c), but considerably worse in configurations (b) and (d). As predicted by the asymptotic theory, MMA (Mallows) and JMA (jackknife) perform essentially the same under homoskedasticity and are indistinguishable in the figure, but again their performance deteriorates when the preordering is unfavorable. Unlike Zhang and Liu (2019), we find that JMA is 7–14% more efficient relative to JMA-M (in MSE sense) in the sixteen designs of Fig. 1.

The dominating estimator is WALS, whose bias is more than offset by a much smaller variance, thus capturing the essence of model averaging. The efficiency of WALS relative to the next-best JMA estimator is about 12% in configurations (a) and (c), 23% in configuration (b) and 31% in configuration (d). The MSE of WALS is 0.23–0.24 depending on the error distribution, hence showing considerable robustness

[2] The MATLAB routines were kindly provided by Xinyu Zhang and Chu-An Liu. All Stata routines are available from the authors upon request.

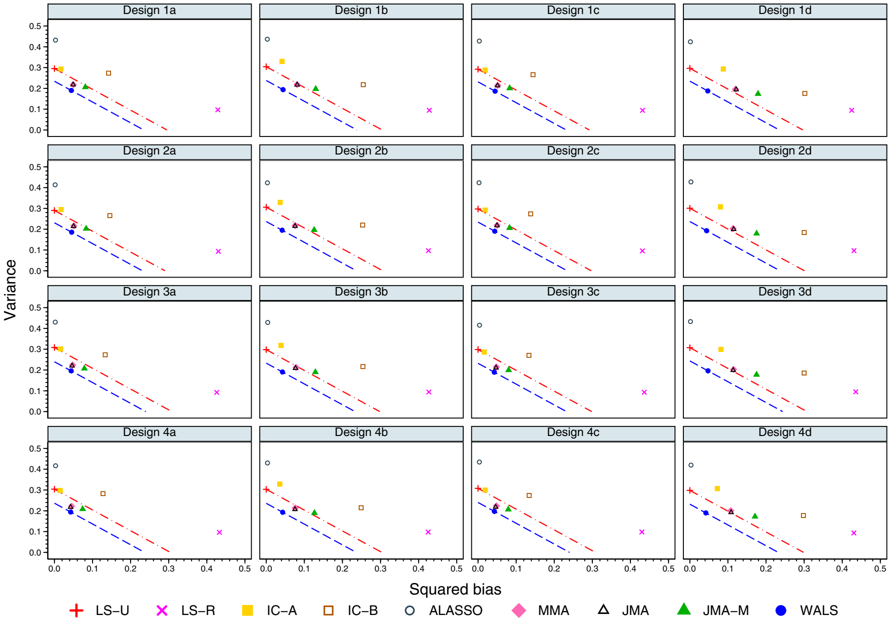

Fig. 1 Squared bias and variance of the estimators of the focus parameter β_{12} in the simulation designs with $k_2 = 8$, $n = 100$, and homoskedastic errors. *Notes*. The sixteen plots represent different specifications of the DGP as indicated in Tables 2 and 3 . The nine estimators considered are described in Table 1. The two ‘iso-MSE’ lines represent all points (squared bias and variance) with the same MSE as the LS-U estimator (red dash-dotted line) and the WALS estimator (blue dashed line)

to violations of the normality assumption, probably because $n = 100$ is already large enough to justify asymptotic approximations.

Now consider the case of heteroskedastic errors, still for $n = 100$, as plotted in Fig. 2. Averaging over all estimators and all designs, this leads to a deterioration of the MSE by about 30% but does not change the ordering of estimators. Contrary to what the asymptotic theory predicts, MMA is 2% more efficient than JMA under heteroskedasticity. WALS remains the preferred estimator in terms of MSE.

When the sample size increases, things change. Since we work in an M-closed environment and the number of models remains fixed, the LS-U estimator remains unbiased and its variance and MSE decrease at the rate of $1/n$. So eventually it dominates all other estimators unless we also let k_2 increase.

Fig. 3 only presents designs 1 and 5 because the t- and skewed t^*-distributions produce moments that are almost identical. For example, in the homoskedastic case the MSE ranges from 0.066 to 0.070 for LS-U and from 0.083 to 0.087 for WALS, while in the heteroskedastic case it ranges from 0.095 to 0.101 for LS-U and from 0.110 to 0.115 for WALS. When n increases from 100 to 400, one would expect the variance to decrease by about 75%, and this is more or less what happens. Averaged over all estimators, the variance decreases by about 73% in both the homoskedastic and the heteroskedastic cases. The (absolute) bias also decreases but at a lower speed. The LS-U estimator is unbiased, while the bias of the LS-R estimator does not change

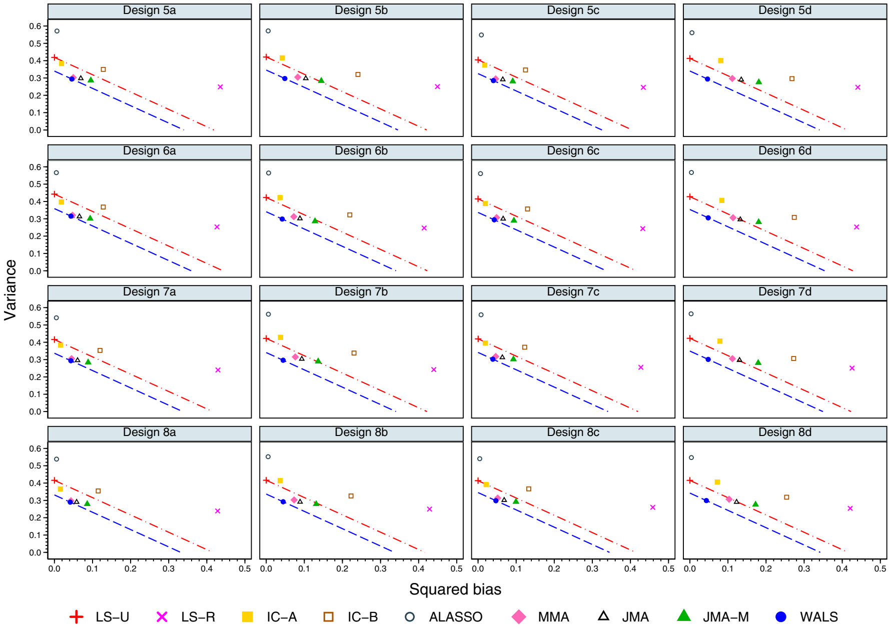

Fig. 2 Squared bias and variance of the estimators of the focus parameter β_{12} in the simulation designs with $k_2 = 8$, $n = 100$, and heteroskedastic errors. *Notes*. See Fig. 1

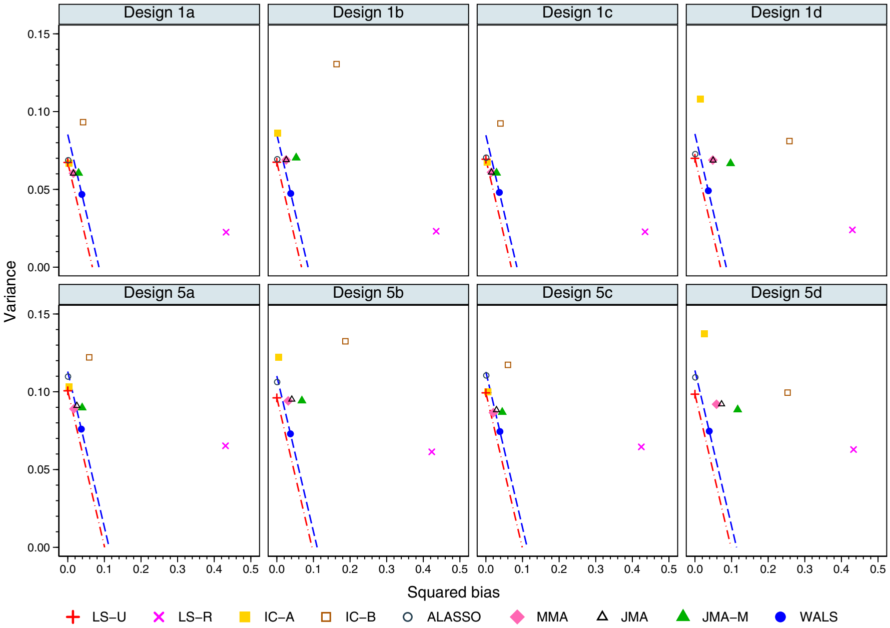

Fig. 3 Squared bias and variance of the estimators of the focus parameter β_{12} in the simulation designs with $k_2 = 8$, $n = 400$, and normal errors. *Notes*. See Fig. 1

with n. Averaging over the remaining estimators we find a decrease of the absolute bias of about 35% in the homoskedastic case and 29% in the heteroskedastic case. The decrease in absolute bias of the WALS estimator is particularly slow. The resulting MSE decreases by about 60% averaged over all estimators, both under homoskedasticity and heteroskedasticity. The preferred estimator is now the LS-U estimator, with ALASSO as second-best and WALS as third-best. These three estimators are not influenced by the order of the auxiliary variables. For the other estimators (except LS-R which clearly performs badly) a poor choice of preordering may lead to poor behavior of the estimator.

Let us now extend our design in four directions. First, we consider not only $n = 100$ and $n = 400$ but also two intermediate values 200 and 300. Second, we extend the number of auxiliary variables from $k_2 = 8$ to $16, 24, 32, \dots, 64$ by setting $\boldsymbol{\beta}_2 = (\xi, \xi^2, \dots, \xi^{k_2/2}, 0, 0, \dots, 0)'$. Third, we consider not only $\xi = 0.5$ but also $\xi = -0.5$, so that we allow for both positive and negative influences or, what is the same, for positive and negative correlations between the regressors. Fourth, in addition to $\sigma_x^2 = \rho = 0.7$ (high correlation), we also consider $\sigma_x^2 = \rho = 0.3$ (low correlation). In total, our second Monte Carlo experiment includes 128 simulation designs for the different combinations of n, k_2, ξ, and ρ, each with 5,000 Monte Carlo replications. To simplify the presentation, we restrict ourselves to homoskedastic normal errors and two estimators: LS-U and WALS.

Fig. 4 considers the efficiency of the WALS estimator relative to the LS-U estimator, i.e. the ratio of the MSE of LS-U and WALS. Theory predicts that, in every setup, WALS will dominate LS-U when n is 'small' and LS-U will dominate WALS when n is 'large'. The question is where to draw the line between small and large. We see that LS-U dominates when n is larger than about 250. But when ξ is negative or the correlation is small, WALS also dominates LS-U for large values of n, certainly larger than 400. As expected, we also see that an increase in k_2 increases the efficiency of WALS relative to LS-U.

6 Monte Carlo Results: Confidence Intervals

We now consider confidence intervals for β_{12} of the form (4) with nominal coverage probability of (at least) $1 - \alpha$. We compare the sixteen methods discussed in Sect. 3. The confidence intervals for ALASSO, MMA-B, and JMA-B are based on 499 bootstrap replications, those for MMA-S and JMA-S are based on 499 Monte Carlo replications, those for WALS (DS-S, ML-S, DS-CN, and ML-CN) on 5,000 Monte Carlo replications. For given α, we calculate $\check{\beta}_{12}$, $\underline{c}_{12}(\alpha)$, and $\overline{c}_{12}(\alpha)$ for each method and each replication of the 32 simulation designs. We then obtain the coverage probability and the length of the interval by averaging over the 5,000 Monte Carlo replications for each simulation design.

Figs. 5 and 6 summarize the simulation results for $n = 100$ and $n = 400$, respectively. Both figures contain 16 panels, one for each method. On the horizontal axis we plot the coverage probabilities for the three values of α: 10% (red long-dashed line), 5% (green dashed line), and 1% (blue dash-dotted line). The lengths of the intervals are plotted on the vertical axis. Since there are 32 designs (labeled 1a–8d), there are

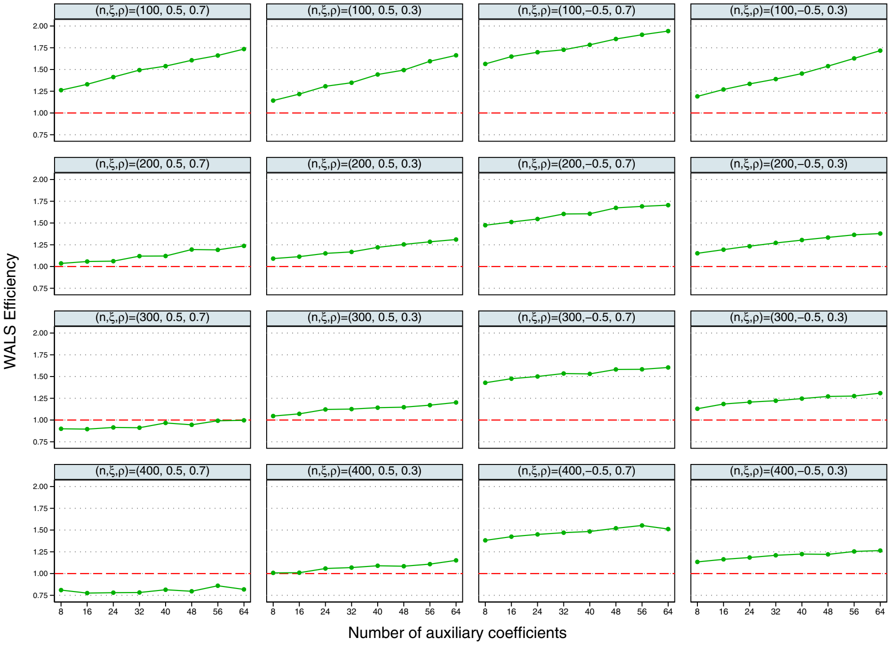

Fig. 4 Efficiency of WALS relative to LS-U of the estimator of β_{12} in the simulation designs with homoskedastic normal errors under alternative values of n, k_2, ξ, and ρ. *Notes*. The sixteen plots represent different specifications of the DGP obtained by varying the sample size ($n = 100$, $n = 200$, $n = 300$ or $n = 400$), the values of the auxiliary parameters ($\xi = 0.5$ or $\xi = -0.5$), and the correlation coefficient among regressors ($\rho = 0.7$ or $\rho = 0.3$). We also allow the number of auxiliary parameters to range from $k_2 = 8$ to 16, 24, 32, ..., 64 by setting $\beta_2 = (\xi, \xi^2, \dots, \xi^{k_2/2}, 0, 0, \dots, 0)'$. For each of the 128 simulation designs, on the vertical axis we plot the efficiency of WALS relative to LS-U (i.e., the ratio between the MSE of LS-U and WALS)

32 points in each panel for each level of α (marked as triangles for $\alpha = 10\%$, squares for $\alpha = 5\%$, and circles for $\alpha = 1\%$). The markers are full for the homoskedastic designs and empty for the heteroskedastic designs. Not all points are visible because many overlap, but what really matters is how much the coverage probabilities differ from their nominal levels and how short the confidence intervals are.

Regarding the coverage probabilities we see that there are five methods that produce accurate coverage probabilities, namely LS-U and the four centered versions of WALS: centered-and-naive (WALS-DS-CN and WALS-ML-CN) and simulation-based (WALS-DS-S and WALS-ML-S). The other eleven methods are much less accurate. In particular, the naive confidence intervals for IC-A and IC-B lead to large undercoverage errors because they ignore model selection noise. The MMA-B and JMA-B confidence intervals are more accurate than the simulation-based algorithms proposed by Zhang and Liu (2019) , but the underlying undercoverage errors are still sizeable and increase with the sample size. [3] The JMA-M confidence intervals also have nonnegligible undercoverage errors which tend to increase with the sample size.

[3] With $n = 100$ the undercoverage errors of MMA-B and JMA-B are -0.03 for $\alpha = 10\%$ and -0.02 for $\alpha = 5\%$, while with $n = 400$ the undercoverage errors become -0.07 for $\alpha = 10\%$ and -0.05 for $\alpha = 5\%$.

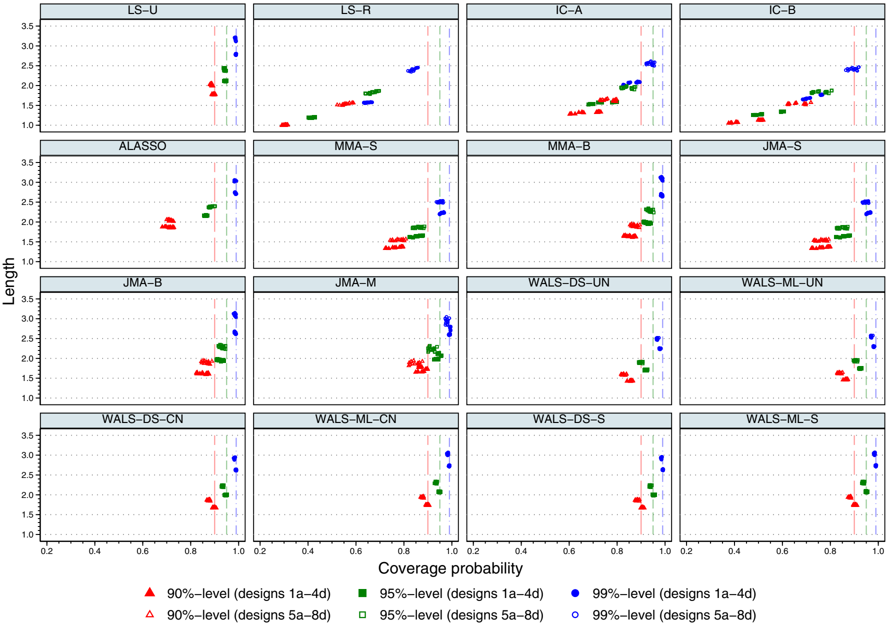

Fig. 5 Coverage probability and length of the confidence intervals for the focus parameter β_{12} in the simulation designs with $k_2 = 8$ and $n = 100$. *Notes*. The sixteen plots refer to different types of confidence intervals for β_{12}, with coverage probabilities on the horizontal axis and lengths of the intervals on the vertical axis. The vertical lines represent three values of the nominal confidence level $1 - \alpha$: 90% (red long-dashed line), 95% (green dashed line), and 99% (blue dash-dotted line). For each level of α we plot 32 points corresponding to the 32 simulation designs in Tables 2 and 3 . The points are marked as triangles for $\alpha = 10\%$, squares for $\alpha = 5\%$, and circles for $\alpha = 1\%$. The markers are full for the homoskedastic designs and empty for the heteroskedastic designs

ALASSO performs well for $n = 400$, but the undercoverage errors of its 90% and 95% confidence intervals for $n = 100$ are rather large (-0.19 for $\alpha = 10\%$ and -0.08 for $\alpha = 5\%$). The UN confidence intervals for WALS do not perform well because they use critical values from the normal distribution and ignore estimation bias. Ignoring estimation bias is much more important than naively using critical values from the normal distribution, as shown by first comparing UN and CN intervals (large difference) and then CN and S intervals (small difference). Obviously to use the correct critical values is better, but the improvement is very small. Similar conclusions are obtained when looking at higher moments of the bias-corrected WALS estimator of the focus parameter β_{12}, computed via the simulation-based algorithm discussed in Appendix B. We find that this estimator is left-skewed and exhibits positive excess kurtosis, but the deviations from zero are in general very small.

Regarding the interval lengths for our five favourite methods we see that for $n = 100$ the interval lengths in the homoskedastic designs are about 1.7 when $\alpha = 10\%$, 2.1 when $\alpha = 5\%$, and 2.7 when $\alpha = 1\%$; about 12% higher in the heteroskedastic designs. For $n = 400$ the interval lengths decrease by about 50%. WALS performs slightly better than LS-U, but the differences are small and require further investigation in an extended design.

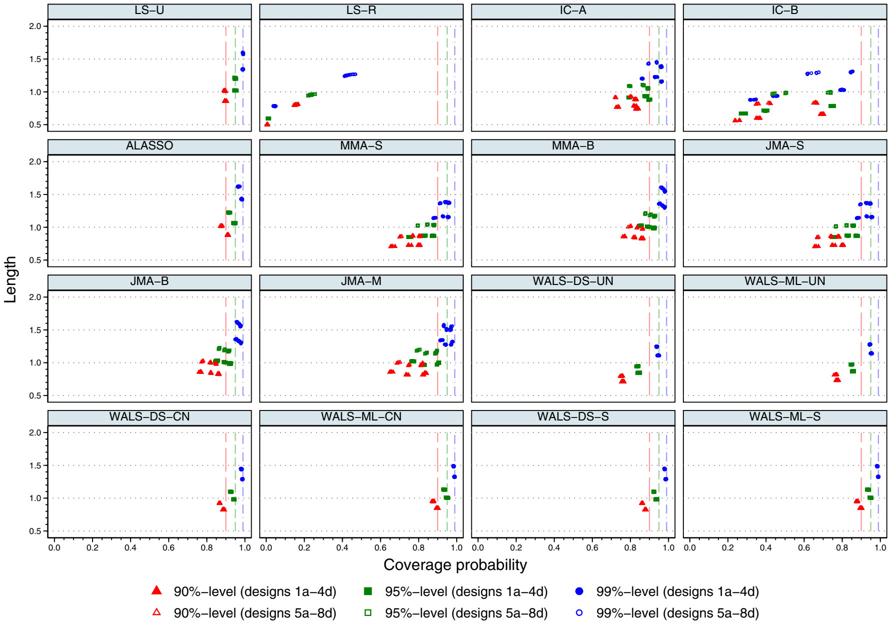

Fig. 6 Coverage probability and length of the confidence intervals for the focus parameter β_{12} in the simulation designs with $k_2 = 8$ and $n = 400$. *Notes*. See Fig. 5

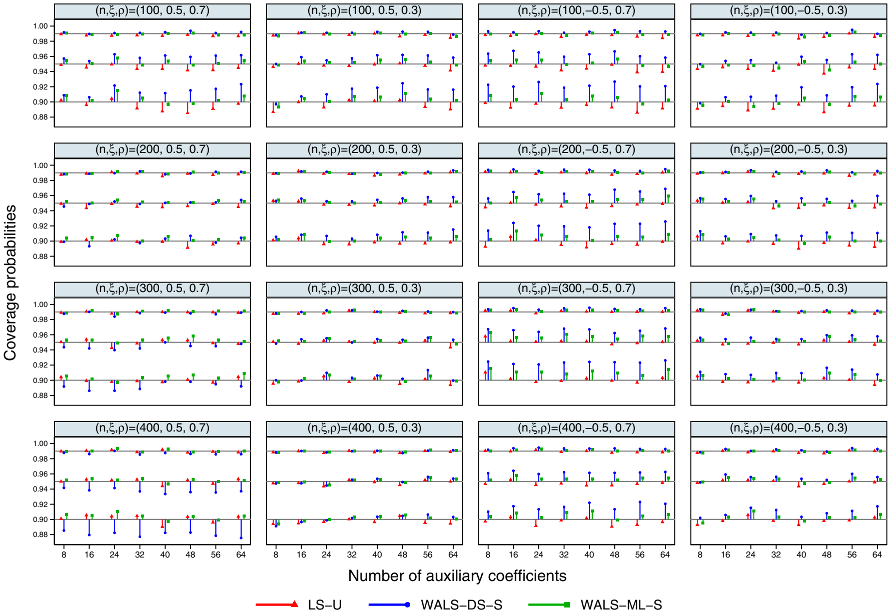

Fig. 7 Coverage probabilities of confidence interval of β_{12} in the simulation designs with homoskedastic normal errors and alternative values of n, k_2, ξ, and ρ. *Notes*. Same as Fig. 4, but on the vertical axis we now plot the coverage probabilities of the LS-U (red line with triangle), WALS-DS-S (blue line with circle), and WALS-ML-S (green lines with square) confidence intervals of β_{12} for the 90%, 95% and 99% confidence levels

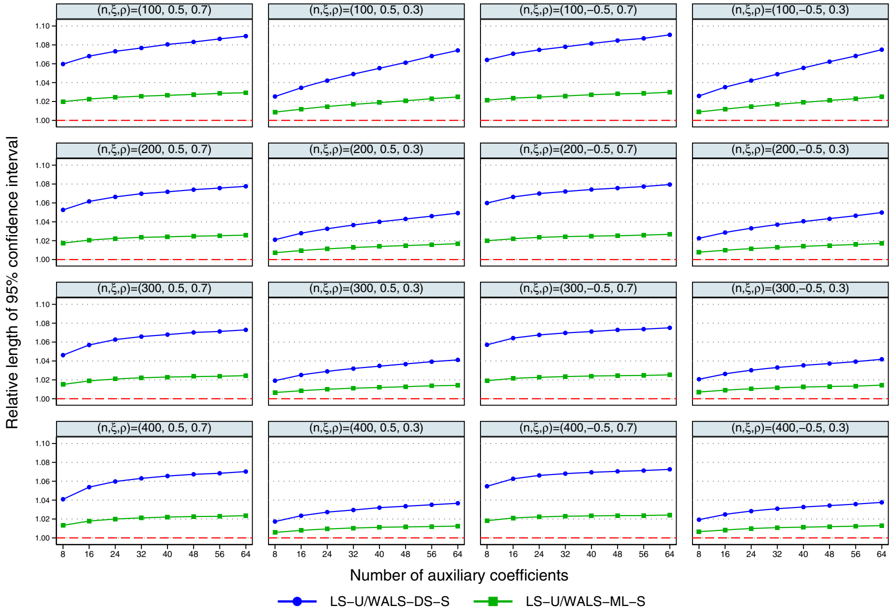

Fig. 8 Relative lengths of the 95% confidence interval of β_{12} in the simulation designs with homoskedastic normal errors and alternative values of n, k_2, ξ, and ρ. *Notes*. Same as Fig. 4, but on the vertical axis we now plot the relative lengths of the 95% WALS-DS-S and WALS-ML-S confidence intervals of β_{12} (i.e. LS-U divided by WALS-DS-S and LS-U divided by WALS-ML-S)

In the extended design defined in the Sect. 5 we consider only the classical LS-U confidence interval and the two simulation-based WALS confidence intervals, WALS-DS-S and WALS-ML-S, based on the plug-in DS and ML estimators of the bias of the posterior mean in the normal location model. [4] The coverage probabilities of the three methods (LS-U, WALS-DS-S, WALS-ML-S) are compared in Fig. 7 for the 90%, 95% and 99% confidence levels. The coverage errors of the three methods are in general small. In the 128 simulation designs considered in our second Monte Carlo experiment they are always smaller than 0.03 in absolute value and they are more often positive (overcoverage) than negative (undercoverage). The fact that WALS-DS-S yields slightly larger coverage errors than WALS-ML-S is consistent with the finite-sample properties of the underlying plug-in estimators of the bias of the posterior mean in the normal location model. Specifically, under the Laplace prior, the plug-in DS estimator of the bias of the posterior mean has always a larger bias than the plug-in ML estimator. We also find that the absolute value of the coverage errors for WALS-DS-S increases with α, reaching a maximum of 0.006 when $\alpha = 1\%$, 0.018 when $\alpha = 5\%$, and 0.028 when $\alpha = 10\%$.

Fig. 8 shows the relative length of the confidence intervals: LS-U divided by WALS-DS-S and LS-U divided by WALS-ML-S. We only present results for the 95% level since they are indistinguishable from those for the 90% and 99% levels. For all cases

[4] The centered-and-naive results for WALS-DS-CN and WALS-ML-CN are again very close to those obtained with the simulation-based approach.

we find that the simulation-based WALS confidence intervals are always smaller than the classical LS-U confidence intervals, even when the LS-U estimator dominates the WALS estimator in terms of MSE. The average length reduction with respect to classical LS-U confidence intervals is about 1.8% for WALS-ML-S and about 5.4% for WALS-DS-S. This result agrees with the fact that, although more biased, the plug-in DS estimator of the bias of the posterior mean has better MSE performance than the plug-in ML estimator, at least for small or moderate values of the location parameter.

The relative gains of WALS on LS-U in terms of confidence interval length are much smaller than those in terms of MSE obtained for the point estimators, which agrees with the findings of Kabaila and Leeb (2006) and Wang and Zhou (2013) for other model averaging approaches to inference. A possible explanation is the randomness of the estimated bias. We have seen that re-centering based on the bias-corrected estimator is important to obtain small coverage errors. However, correcting for bias increases sampling variability, which is reflected in the length of the confidence interval.

7 Monte Carlo Results: Prediction Intervals

Finally we consider the problem of predicting a single observation y_f from model (1) and covariate vector $\boldsymbol{x}_f = (\boldsymbol{x}'_{1f}, \boldsymbol{x}'_{2f})'$, that is,

$$y_f = \boldsymbol{x}'_f \boldsymbol{\beta} + \epsilon_f = \boldsymbol{x}'_{1f}\boldsymbol{\beta}_1 + \boldsymbol{x}'_{2f}\boldsymbol{\beta}_2 + \epsilon_f,$$

where $\boldsymbol{\epsilon}$ and ϵ_f are independent of each other and jointly normally distributed with zero means and variances $\mathbb{V}[\boldsymbol{\epsilon}] = \sigma^2 \boldsymbol{I}_n$ and $\mathbb{V}[\epsilon_f] = \sigma^2$. If $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$ denote the WALS estimators of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, then the WALS predictor of y_f is defined as

$$\widehat{y}_f = \boldsymbol{x}'_{1f}\widehat{\boldsymbol{\beta}}_1 + \boldsymbol{x}'_{2f}\widehat{\boldsymbol{\beta}}_2,$$

and its prediction error is

$$\widehat{y}_f - y_f = \boldsymbol{x}'_{1f}(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) + \boldsymbol{x}'_{2f}(\widehat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2) - \epsilon_f.$$

Because of (9), the WALS predictor of y_f may be viewed as a weighted average of the predictors from all 2^{k_2} models in the model space. We are interested in constructing prediction intervals for $\mathbb{E}[y_f] = \boldsymbol{x}'_{1f}\boldsymbol{\beta}_1 + \boldsymbol{x}'_{2f}\boldsymbol{\beta}_2$. Unlike the confidence intervals described in Sect. 3, these prediction intervals require dealing with the sampling uncertainty on all model parameters, focus and auxiliary.

We consider two variants of WALS prediction intervals. The first, which we call the naive approach, starts from the bias-corrected WALS estimator $\check{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}} - b(\widehat{\boldsymbol{\beta}})$ and then constructs a symmetric prediction interval with nominal coverage probability $1 - \alpha$:

$$\boldsymbol{x}'_f \check{\boldsymbol{\beta}} - z_{1-\alpha/2}\sqrt{\boldsymbol{x}'_f \check{\boldsymbol{V}} \boldsymbol{x}_f} < \mathbb{E}[y_f] < \boldsymbol{x}'_f \check{\boldsymbol{\beta}} + z_{1-\alpha/2}\sqrt{\boldsymbol{x}'_f \check{\boldsymbol{V}} \boldsymbol{x}_f},$$

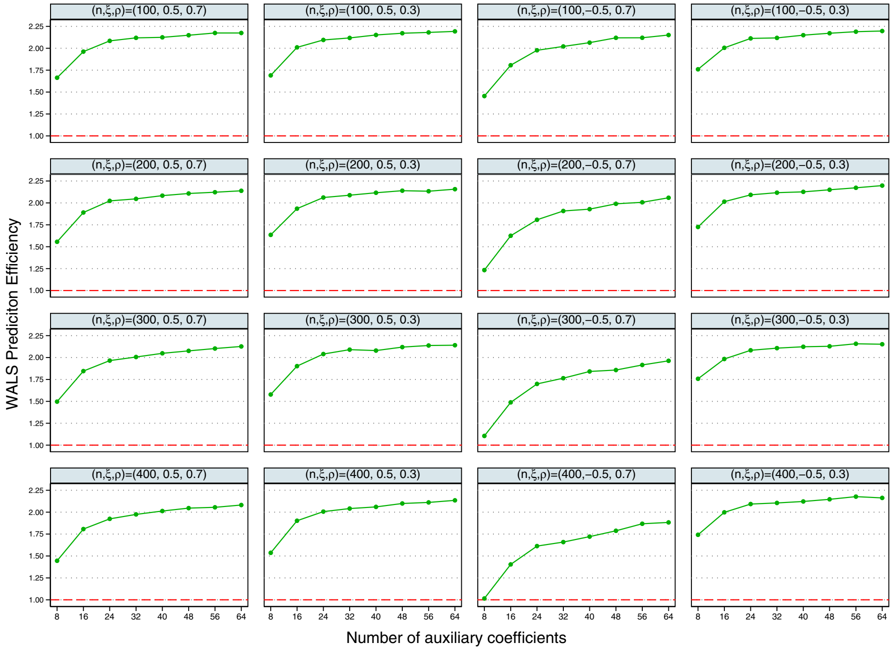

Fig. 9 Efficiency of the WALS predictor of $\mathbb{E}[y_f]$ relative to the LS-U predictor in the simulation designs with homoskedastic normal errors under alternative values of n, k_2, ξ, and ρ. *Notes*. Same as Fig. 4, but on the vertical axis we now plot the WALS prediction efficiency (i.e. the ratio between the mean squared prediction errors of LS-U and WALS)

where $\check{\boldsymbol{V}}$ is the Monte Carlo variance of $\check{\boldsymbol{\beta}}$ estimated from $\boldsymbol{B}^*$, the $R \times k$ matrix containing the replications of the bias-corrected WALS estimator in step (iv) of the algorithm described in Appendix B. This approach assumes normality of the bias-corrected WALS estimator, which is why it is called naive. The other, which we call the simulation-based approach, does not assume normality of the bias-corrected WALS estimator and builds the prediction interval directly from the quantiles of the empirical distribution of the elements of the vector $\boldsymbol{B}^*\boldsymbol{x}_f$. This prediction interval need not be symmetric around $\boldsymbol{x}'_f\check{\boldsymbol{\beta}}$.

Fig. 9 presents the relative efficiency of the WALS predictor of $\mathbb{E}[y_f] = \boldsymbol{x}'_f\boldsymbol{\beta}$ relative to the LS-U predictor in the 128 simulation designs with homoskedastic normal errors under alternative values of n, k_2, ξ, and ρ. In each design, $\boldsymbol{x}_f$ is drawn randomly from a multivariate normal distribution with mean zero and variance $\sigma_x^2\boldsymbol{\Sigma}_x(\rho)$ and then kept fixed for all replications of the same simulation design. Thus, $\boldsymbol{x}_f$ changes with k_2 and ρ. The figure has the same format as Fig. 4, except that efficiency is now measured by the ratio of the mean squared prediction errors (LS-U relative to WALS). WALS clearly dominates LS-U in all designs, and by an even larger margin than what we have seen for the focus parameter. As expected, the relative efficiency of WALS increases with the number of auxiliary parameters in the DGP. The typical profile of the relative efficiency of the WALS predictor is concave in k_2, revealing very large

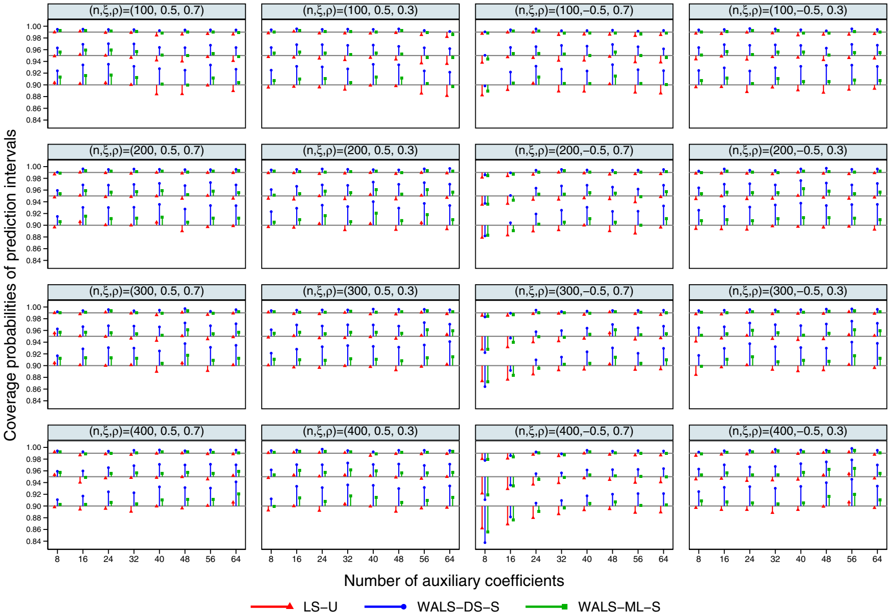

Fig. 10 Coverage probabilities of prediction interval of $\mathbb{E}[y_f]$ in the simulation designs with homoskedastic normal errors and alternative values of n, k_2, ξ, and ρ. *Notes*. Same as Fig. 4, but on the vertical axis we now plot the coverage probabilities of the LS-U (red line with triangle), WALS-DS-S (blue line with circle), and WALS-ML-S (green lines with square) prediction intervals for the 90%, 95% and 99% nominal probabilities.

gains when moving from a small number ($k_2 = 8$) to a moderate number ($k_2 = 24$) of auxiliary parameters.

Fig. 10 shows the actual coverage probabilities of the prediction intervals for LS-U and WALS for nominal probabilities of 90%, 95%, and 99%. For WALS we only present the simulation-based intervals, both DS and ML, because the naive prediction intervals are always very close. This figure is the analog of Fig. 7 and, perhaps not surprisingly, prediction interval coverage errors are only slightly larger than confidence interval coverage errors. There is only one design ($n = 400$, $\xi = -0.5$, $\rho = 0.7$) out of the 128 considered for which the coverage error is sizable (around 6%), and this coverage error is not much larger than for LS-U in the same design.

Fig. 11 plots the relative lengths of the 95% prediction intervals based on LS-U and WALS, hence the analog of Fig. 8. The disadvantage of using LS-U is now even more evident than before. LS-U prediction intervals are 2–3% larger than WALS-ML and 5–10% larger than WALS-DS. Furthermore, the relative length of the LS-U prediction intervals, viewed as a function of k_2, is concave for all designs, again revealing large gains when moving from $k_2 = 8$ to $k_2 = 24$.

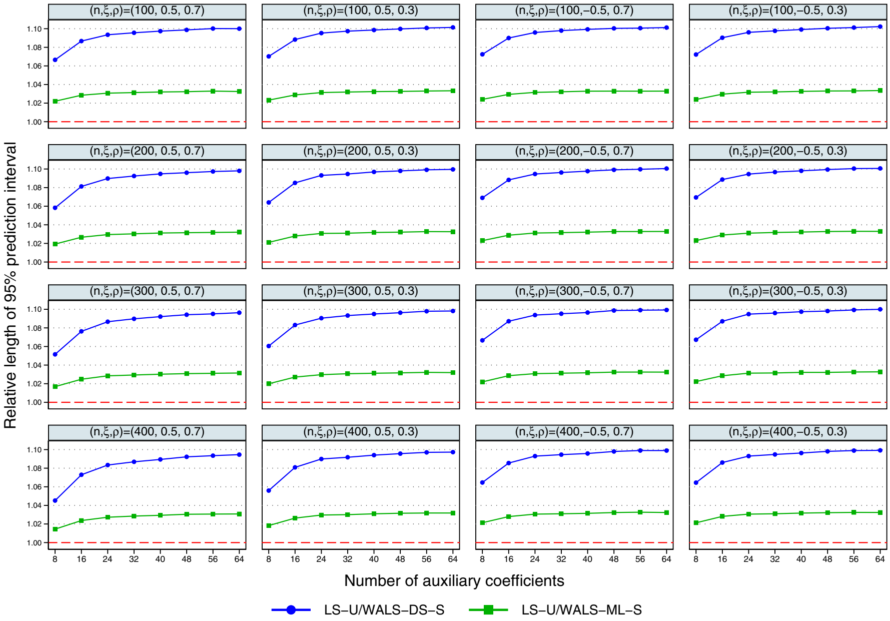

Fig. 11 Relative lengths of the 95% prediction interval of $\mathbb{E}[y_f]$ in the simulation designs with homoskedastic normal errors and alternative values of n, k_2, ξ, and ρ. *Notes*. Same as Fig. 4, but on the vertical axis we now plot the relative lengths of the 95% WALS-DS-S and WALS-ML-S prediction intervals (i.e. LS-U divided by WALS-DS-S and LS-U divided by WALS-ML-S)

8 Conclusions

In this paper we extend the theory of WALS estimation to inference by proposing a simulation-based method for confidence and prediction intervals. To highlight the properties of WALS and put them in perspective we also consider its main competitors. We discuss both confidence intervals for a focus parameter and prediction intervals for the outcome of interest by an extensive set of Monte Carlo experiments that allow for increasing complexity of the model space and include heteroskedastic, skewed, and thick-tailed error distributions.

In the homoskedastic case the dominating estimator is WALS, whose bias is more than offset by a smaller variance, especially when the sample size is small, thus capturing the essence of model averaging. In the heteroskedastic case, the performance of all estimators deteriorates but their relative position in terms of MSE changes little. With a large sample size, the preferred estimator is the unrestricted estimator LS-U, closely followed by ALASSO and WALS.

Regarding coverage probabilities, we find that LS-U and WALS perform well, while all other methods are much less accurate. Comparing the length of confidence intervals, WALS performs slightly better than the LS-U estimator, though differences are small. Finally, regarding prediction intervals, WALS clearly dominates LS-U. The relative efficiency of WALS increases with the number k_2 of auxiliary parameters and its typical profile is concave in k_2. Coverage errors of prediction intervals are only slightly larger than of confidence intervals, and when we compare the relative lengths of 95% prediction intervals based on LS-U and WALS the dominance of WALS is even stronger.

Post-selection/averaging inference is a challenging issue, which is likely to play a prominent role in future developments and applications of model selection/averaging techniques. In addition to estimating the coefficients of interest accurately, many economic problems require us to evaluate the precision of the estimated relationships and their statistical significance. Our new methods for WALS confidence and prediction intervals provide an easy, accurate, and computational convenient solution for these difficult tasks. For the latest set of Stata, R, and Python routines covering WALS inference of (univariate) generalized linear models (linear, logistic and Poisson regressions) we refer the reader to De Luca et al. (2022).

Acknowledgements Giuseppe De Luca acknowledges financial support from the MIUR PRIN (Grant Number PRJ-0324).

Funding Open access funding provided by Universitá degli Studi di Palermo.

Code availability All routines are available from the authors upon request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Appendix A: The Nine Estimators

In this appendix, we formalize the nine estimators which were introduced in Sect. 2 and listed in Table 1.

Least squares (LS). The LS-U estimator of $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$ is $\widehat{\boldsymbol{\beta}}_u = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ and the LS-R estimator is $\widehat{\boldsymbol{\beta}}_r = (\widehat{\boldsymbol{\beta}}_{1,r}', \boldsymbol{0}')'$, where $\widehat{\boldsymbol{\beta}}_{1,r} = (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{y}$.

Post-selection estimators based on information criteria. Suppose that, after preordering, the pth model ($p = 0, 1, \ldots, k_2$) has k_1 focus regressors and p auxiliary regressors. Assume that the underlying error is homoskedastic and let $\widehat{\sigma}_p^2 = \boldsymbol{y}'\boldsymbol{M}_p^*\boldsymbol{y}/n$ be the maximum likelihood (ML) estimator of its variance, where

$$\boldsymbol{M}_p^* = \boldsymbol{I}_n - \boldsymbol{X}_p^*(\boldsymbol{X}_p^{*'}\boldsymbol{X}_p^*)^{-1}\boldsymbol{X}_p^{*'} \tag{5}$$

is the usual idempotent matrix in model p, $\boldsymbol{I}_n$ is the identity matrix of order n, and $\boldsymbol{X}_p^* = (\boldsymbol{X}_1, \boldsymbol{X}_{2,p})$ is the matrix containing the first $k_1 + p$ regressors. The AIC for model p is

$$\mathrm{AIC}_p = n \log(\widehat{\sigma}_p^2) + 2(k_1 + p)$$

and the BIC is

$$\mathrm{BIC}_p = n \log(\widehat{\sigma}_p^2) + (\log n)(k_1 + p).$$

The IC-A and IC-B estimators are the LS estimators in the models with the lowest value of AIC_p and BIC_p respectively.

Adaptive LASSO (ALASSO). This estimator solves the optimization problem

$$\min_{\boldsymbol{\beta}} \left((\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \sum_{l=1}^{k} \psi_{l,n} |\beta_l| \right),$$

where β_l is the lth component of $\boldsymbol{\beta}$ and the weight $\psi_{l,n} = \psi_n / \widehat{\beta}_{l,u}^2$ depends on a tuning penalty parameter ψ_n selected by generalized cross-validation and the lth component $\widehat{\beta}_{l,u}$ of $\widehat{\boldsymbol{\beta}}_u$.

Mallows model averaging (MMA). To reduce the computational burden this estimator is typically based on preordering. Letting s_u^2 be the unbiased LS estimator of σ^2 in the unrestricted model and $\boldsymbol{M}^*(\boldsymbol{w}) = \sum_{p=0}^{k_2} w_p \boldsymbol{M}_p^*$, where $\boldsymbol{M}_p^*$ is defined in (5), the MMA weights are obtained by solving

$$\min_{\boldsymbol{w}} \left(\boldsymbol{y}' \boldsymbol{M}^{*'}(\boldsymbol{w}) \boldsymbol{M}^*(\boldsymbol{w}) \boldsymbol{y} + 2 s_u^2 \sum_{p=0}^{k_2} w_p (k_1 + p) \right)$$

subject to $\sum_p w_p = 1$ and $0 \leq w_p \leq 1$ for all p. Denoting the optimal weights by $\widehat{\boldsymbol{w}} = (\widehat{w}_0, \ldots, \widehat{w}_{k_2})$, the MMA estimator takes the form

$$\widehat{\boldsymbol{\beta}}_{\mathrm{MMA}} = \sum_{p=0}^{k_2} \widehat{w}_p (\boldsymbol{X}_p^{*'} \boldsymbol{X}_p^*)^{-1} \boldsymbol{X}_p^{*'} \boldsymbol{y}. \tag{6}$$

Jackknife model averaging (JMA). Let $\boldsymbol{D}_p$ be the diagonal matrix containing the diagonal elements of $\boldsymbol{M}_p^*$ on its diagonal and zeros elsewhere and define $\boldsymbol{M}^\dagger(\boldsymbol{w}) = \sum_{p=0}^{k_2} w_p \boldsymbol{D}_p^{-1} \boldsymbol{M}_p^*$.[5] Then the JMA weights are obtained by solving

$$\min_{\boldsymbol{w}} \left(\boldsymbol{y}' \boldsymbol{M}^{\dagger'}(\boldsymbol{w}) \boldsymbol{M}^\dagger(\boldsymbol{w}) \boldsymbol{y} \right) \tag{7}$$

[5] To ensure nonsingularity of $\boldsymbol{D}_p$ we must add the requirement that the ith unit vector in $\mathbb{R}^n$ (the vector whose ith component is 1 and all other components are 0) does not lie in the column space of $\boldsymbol{X}$ for any i.

subject to $\sum_p w_p = 1$ and $0 \leq w_p \leq 1$ for all p. The modified JMA (JMA-M) estimator is defined by weights that solve

$$\min_{\boldsymbol{w}} \left(\boldsymbol{y}' \boldsymbol{M}^{\dagger'}(\boldsymbol{w}) \boldsymbol{M}^{\dagger}(\boldsymbol{w}) \boldsymbol{y} + \psi_n \sum_{p=0}^{k_2} w_p (k_1 + p) \right) \tag{8}$$

subject to the same constraints as in (7), where the tuning parameter ψ_n is set equal to $\log n$. The JMA and JMA-M estimators take the same form as (6) where $\widehat{w}_p$ is given by the solution of (7) and (8), respectively.

Weighted-average least squares (WALS). We first transform $\boldsymbol{X}_2$ and $\boldsymbol{\beta}_2$ by defining $\boldsymbol{Z}_2 = \boldsymbol{X}_2 \boldsymbol{\Delta}_2 \boldsymbol{\Psi}^{-1/2}$ and $\boldsymbol{\gamma}_2 = \boldsymbol{\Psi}^{1/2} \boldsymbol{\Delta}_2^{-1} \boldsymbol{\beta}_2$, where $\boldsymbol{\Delta}_2$ is a diagonal $k_2 \times k_2$ matrix such that all diagonal elements of $\boldsymbol{\Psi} = \boldsymbol{\Delta}_2 \boldsymbol{X}_2' \boldsymbol{M}_1 \boldsymbol{X}_2 \boldsymbol{\Delta}_2$ are equal to one and $\boldsymbol{M}_1 = \boldsymbol{I}_n - \boldsymbol{X}_1 (\boldsymbol{X}_1' \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1'$. We also rescale $\boldsymbol{X}_1$ and $\boldsymbol{\beta}_1$ by defining $\boldsymbol{Z}_1 = \boldsymbol{X}_1 \boldsymbol{\Delta}_1$ and $\boldsymbol{\gamma}_1 = \boldsymbol{\Delta}_1^{-1} \boldsymbol{\beta}_1$, where $\boldsymbol{\Delta}_1$ is a diagonal $k_1 \times k_1$ matrix such that all diagonal elements of $\boldsymbol{Z}_1' \boldsymbol{Z}_1$ are equal to one. The equivalence between models (1) and (3) follows from the fact that $\boldsymbol{Z}_1 \boldsymbol{\gamma}_1 = \boldsymbol{X}_1 \boldsymbol{\beta}_1$ and $\boldsymbol{Z}_2 \boldsymbol{\gamma}_2 = \boldsymbol{X}_2 \boldsymbol{\beta}_2$. In addition, the transformations ensure that $\boldsymbol{Z}_2' \boldsymbol{M}_1 \boldsymbol{Z}_2 = \boldsymbol{I}_{k_2}$.

Averaging the LS estimators $\widehat{\boldsymbol{\gamma}}_{1j}$ and $\widehat{\boldsymbol{\gamma}}_{2j}$ over the $J = 2^{k_2}$ models in the model space gives the WALS estimators

$$\widehat{\boldsymbol{\gamma}}_1 = \sum_{j=1}^{J} \lambda_j \widehat{\boldsymbol{\gamma}}_{1j} = \widehat{\boldsymbol{\gamma}}_{1,r} - \boldsymbol{Q} \boldsymbol{W} \widehat{\boldsymbol{\gamma}}_{2,u}, \qquad \widehat{\boldsymbol{\gamma}}_2 = \sum_{j=1}^{J} \lambda_j \widehat{\boldsymbol{\gamma}}_{2j} = \boldsymbol{W} \widehat{\boldsymbol{\gamma}}_{2,u}, \tag{9}$$

where the $\lambda_j = \lambda_j(\widehat{\boldsymbol{\gamma}}_{2,u})$ are nonnegative data-dependent model weights that add up to one, $\widehat{\boldsymbol{\gamma}}_{1,r} = (\boldsymbol{Z}_1' \boldsymbol{Z}_1)^{-1} \boldsymbol{Z}_1' \boldsymbol{y}$ is the LS-R estimator of $\boldsymbol{\gamma}_1$, $\widehat{\boldsymbol{\gamma}}_{2,u} = \boldsymbol{Z}_2' \boldsymbol{M}_1 \boldsymbol{y}$ is the LS-U estimator of $\boldsymbol{\gamma}_2$, $\boldsymbol{Q} = (\boldsymbol{Z}_1' \boldsymbol{Z}_1)^{-1} \boldsymbol{Z}_1' \boldsymbol{Z}_2$, $\boldsymbol{W} = \sum_j \lambda_j (\boldsymbol{I}_{k_2} - \boldsymbol{S}_j \boldsymbol{S}_j')$, $\boldsymbol{S}_j$ is a $k_2 \times r_j$ selection matrix of rank $0 \leq r_j \leq k_2$ (that is, $\boldsymbol{S}_j' = [\boldsymbol{I}_{r_j} : \boldsymbol{0}]$ or a column-permutation thereof), and r_j is the number of exclusion restrictions implied by model j.

The results in (9) show that the dependence of $\widehat{\boldsymbol{\gamma}}_1$ and $\widehat{\boldsymbol{\gamma}}_2$ on the estimators from all the available models is completely captured by the random diagonal matrix $\boldsymbol{W}$, whose k_2 diagonal elements w_h are partial sums of the λ_j. Since $0 \leq w_h \leq 1$, this implies that the components of $\widehat{\boldsymbol{\gamma}}_2$ in (9) are shrinkage estimators of the components of $\boldsymbol{\gamma}_2$. The assumption that $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$ implies that $\widehat{\boldsymbol{\gamma}}_{2,u} \sim \mathcal{N}(\boldsymbol{\gamma}_2, \sigma^2 \boldsymbol{I}_{k_2})$. Hence, if we restrict each w_h to depend only on the hth component of $\widehat{\boldsymbol{\gamma}}_{2,u}$, the components of $\widehat{\boldsymbol{\gamma}}_2$ will also be independent.

If we again treat the error variance σ^2 as known and equal to its unbiased LS estimator s_u^2, it follows that the k_2-vector of t-ratios $\boldsymbol{x} = \widehat{\boldsymbol{\gamma}}_{2,u}/s_u$ is distributed as $\mathcal{N}(\boldsymbol{\eta}, \boldsymbol{I}_{k_2})$, where $\boldsymbol{\eta} = \boldsymbol{\gamma}_2/\sigma_u$ is the k_2-vector of 'theoretical' t-ratios. The individual components x_h of $\boldsymbol{x}$ are therefore independently distributed as $\mathcal{N}(\eta_h, 1)$. As for the prior information on the vector $\boldsymbol{\eta}$, we assume that its components are i.i.d. with a density that is symmetric around zero, positive and nonincreasing on $(0, \infty)$, differentiable (except possibly at 0), and satisfies the 'neutrality condition' $\mathbb{P}[|\eta_h| < 1] = 1/2$.

Given the normal location model for x_h and the chosen prior for η_h, the Bayesian approach yields the posterior mean $m_h = m(x_h)$ as the estimator of η_h. The WALS estimators of $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ then take the form

$$\widehat{\boldsymbol{\gamma}}_1 = \widehat{\boldsymbol{\gamma}}_{1,r} - \boldsymbol{Q}\widehat{\boldsymbol{\gamma}}_2, \qquad \widehat{\boldsymbol{\gamma}}_2 = s_u \boldsymbol{m},$$

with $\boldsymbol{m} = (m_1, \ldots, m_{k_2})'$, and the WALS estimators of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are

$$\widehat{\boldsymbol{\beta}}_1 = \boldsymbol{\Delta}_1\widehat{\boldsymbol{\gamma}}_1, \qquad \widehat{\boldsymbol{\beta}}_2 = \boldsymbol{\Delta}_2\boldsymbol{\Psi}^{-1/2}\widehat{\boldsymbol{\gamma}}_2. \tag{10}$$

Appendix B: Algorithm for Simulation-Based WALS Confidence Intervals

Let $\boldsymbol{x} = \widehat{\boldsymbol{\gamma}}_{2,u}/s_u = (x_1, \ldots, x_{k_2})'$ be the k_2-vector of t-ratios from the unrestricted model and $\widehat{\boldsymbol{\eta}} = (\widehat{\eta}_1, \ldots, \widehat{\eta}_{k_2})'$ an estimator of the k_2-vector of parameters $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_{k_2})'$ in the multivariate normal location model $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\eta}, \boldsymbol{I}_{k_2})$.

The simulation-based WALS confidence intervals for $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$ are obtained by the following five-step algorithm:

(i) Compute $\widehat{\boldsymbol{\eta}}$ and use its generic element $\widehat{\eta}_h$ to generate the R-vectors $\boldsymbol{x}_h^* = (x_{1h}^*, \ldots, x_{Rh}^*)'$ of independent pseudo-random draws from the $\mathcal{N}(\widehat{\eta}_h, 1)$ distribution.

(ii) Compute the $R \times k_2$ matrix $\check{\boldsymbol{M}}^*$ of pseudo-random draws for the bias-corrected posterior means with generic element

$$\check{m}_{rh}^* = m_{rh}^* - \delta_{rh}^* \qquad (r = 1, \ldots, R;\ h = 1, \ldots, k_2),$$

where $m_{rh}^* = m(x_{rh}^*)$ is the posterior mean evaluated at x_{rh}^* and δ_{rh}^* is either the plug-in ML estimator $\delta(x_{rh}^*)$ or the plug-in DS estimator $\delta(m_{rh}^*)$ of the bias of m_{rh}^*.

(iii) Generate the $R \times k_1$ matrix $\boldsymbol{B}_{1r}^*$ of independent pseudo-random draws from the distribution $\mathcal{N}(\widehat{\boldsymbol{\beta}}_{1r}, \boldsymbol{V}_{1r})$, where

$$\widehat{\boldsymbol{\beta}}_{1,r} = \boldsymbol{\Delta}_1(\boldsymbol{Z}_1'\boldsymbol{Z}_1)^{-1}\boldsymbol{Z}_1'\boldsymbol{y}, \qquad \boldsymbol{V}_{1r} = s_u^2\boldsymbol{\Delta}_1(\boldsymbol{Z}_1'\boldsymbol{Z}_1)^{-1}\boldsymbol{\Delta}_1$$

are the LS-R estimate of $\boldsymbol{\beta}_1$ in the fully restricted model and its estimated variance matrix, respectively.

(iv) Compute the $R \times k$ matrix $\check{\boldsymbol{B}}^* = (\check{\boldsymbol{B}}_1^*, \check{\boldsymbol{B}}_2^*)$ of pseudo-random draws for the bias-corrected WALS estimator $\check{\boldsymbol{\beta}} = (\check{\boldsymbol{\beta}}_1', \check{\boldsymbol{\beta}}_2')'$ of $\boldsymbol{\beta}$, where

$$\check{\boldsymbol{B}}_1^* = \boldsymbol{B}_{1r}^* - s_u\check{\boldsymbol{M}}^*\boldsymbol{Z}_2'\boldsymbol{Z}_1(\boldsymbol{Z}_1'\boldsymbol{Z}_1)^{-1}\boldsymbol{\Delta}_1, \qquad \check{\boldsymbol{B}}_2^* = s_u\check{\boldsymbol{M}}^*\boldsymbol{\Psi}^{-1/2}\boldsymbol{\Delta}_2.$$

(v) Compute the $(1-\alpha)$-level confidence interval for the generic component β_l of $\boldsymbol{\beta}$ $(l = 1, \ldots, k)$ as $[q_l^*(\alpha/2), q_l^*(1-\alpha/2)]$, where $q_l^*(\alpha/2)$ and $q_l^*(1-\alpha/2)$ are, respectively, the $\alpha/2$ and $(1-\alpha/2)$ empirical percentiles of the R replications corresponding to the lth column $\check{\boldsymbol{b}}_l^*$ of $\check{\boldsymbol{B}}^*$.

Remark 1 To achieve good performance in terms of coverage probabilities, the initial estimator $\widehat{\boldsymbol{\eta}}$ in the first step of the algorithm must be (approximately) unbiased for $\boldsymbol{\eta}$. This leaves us three possible choices: (i) the ML estimator $\boldsymbol{x}$, (ii) the DS bias-corrected posterior mean $\boldsymbol{m}(\boldsymbol{x})-\boldsymbol{\delta}(\boldsymbol{m}(\boldsymbol{x}))$, and (iii) the ML bias-corrected posterior mean $\boldsymbol{m}(\boldsymbol{x})-\boldsymbol{\delta}(\boldsymbol{x})$. In our experience, the differences between these three estimators are small. In the simulations, we use (ii) for the WALS-DS-S confidence intervals and (iii) for the WALS-ML-S confidence intervals. The main difference between these two methods is the choice of the plug-in estimator for the bias of the posterior mean in the second stage of the algorithm, namely $\delta(m_{rh}^*)$ for WALS-DS-S or $\delta(x_{rh}^*)$ for WALS-ML-S.

Remark 2 Like in other parametric bootstrap approaches, our simulation-based method ignores uncertainty caused by randomness of the regressors. Thus, as typically assumed in the WALS theory for point estimation, we treat the regressors as fixed.

Remark 3 An important difference with the MMA-S and JMA-S confidence intervals proposed by Zhang and Liu (2019) is that they simulate from the limiting distribution of the model averaging estimator, while in the simulation-based WALS algorithm we don't. The WALS confidence intervals are based on the finite-sample properties of the plug-in estimators of the frequentist bias of the posterior mean in the normal location model (De Luca et al. 2021).

Remark 4 The $R \times k$ matrix $\check{\boldsymbol{B}}^*$ of Monte Carlo replications obtained from step (iv) of the algorithm can be used to estimate any aspect of the sampling distribution of the bias-corrected WALS estimator. For example, we use it to compute the standard error of $\check{\beta}_l$ in the CN method for confidence intervals, and the complete variance matrix of $\check{\boldsymbol{\beta}}$ in the naive approach to prediction intervals.

Remark 5 Our algorithm is very fast, especially with the Laplace prior. For example, in applications with $n = 400$ observations and $k_2 = 40$ auxiliary regressors, we can compute point estimates, their estimated moments, and confidence intervals for all parameters based on 100,000 Monte Carlo replications in about 3.5 seconds by using a workstation with one Intel(R) Core(TM) i7-4790 CPU/3.60 GHz processor and 32 GB of RAM.

References

Camponovo, L. (2015). On the validity of the pairs bootstrap for lasso estimators. *Biometrika, 102,* 981–987.

Chatterjee, A., & Lahiri, S. N. (2011). Bootstrapping Lasso estimators. *Journal of the American Statistical Association, 106,* 608–625.

Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging*. Cambridge University Press.

Danilov, D. (2005). Estimation of the mean of a univariate normal distribution when the variance is not known. *Econometrics Journal, 8,* 277–291.

Dardanoni, V., Modica, S., & Peracchi, F. (2011). Regression with imputed covariates: A generalized missing-indicator approach. *Journal of Econometrics, 162,* 362–368.

De Luca, G., Magnus, J. R., & Peracchi, F. (2018). Weighted-average least squares estimation of generalized linear models. *Journal of Econometrics, 204,* 1–17.

De Luca, G., Magnus, J. R., & Peracchi, F. (2021). Sampling properties of the Bayesian posterior mean with an application to WALS estimation. *Journal of Econometrics* (to appear). https://doi.org/10.1016/j.jeconom.2021.04.008.

De Luca, G., Magnus, J. R., & Yue, Y. (2022). *Computational aspects of weighted-average least squares (WALS) inference*. Mimeo.

DiTraglia, F. (2016). Using invalid instruments on purpose: Focused moment selection and averaging for GMM. *Journal of Econometrics, 195,* 187–208.

Donohue, J. J., & Levitt, S. D. (2001). The impact of legalized abortion on crime. *Quarterly Journal of Economics, 116,* 379–420.

Duval, R., Furceri, D., & Miethe, J. (2021). Robust political economy correlates of major product and labor market reforms in advanced economies: Evidence from BAMLE for logit models. *Journal of Applied Econometrics, 36,* 98–124.

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

Hansen, B. E. (2014). Model averaging, asymptotic risk, and regressor groups. *Quantitative Economics, 5,* 495–530.

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

Kabaila, P., & Leeb, H. (2006). On the large-sample minimal coverage probability of confidence intervals after model selection. *Journal of the American Statistical Association, 101,* 619–629.

Kabaila, P., & Mainzer, R. (2018). Two sources of poor coverage of confidence intervals after model selection. *Statistics and Probability Letters, 140,* 185–190.

Knight, K., & Fu, W. (2000). Asymptotics for Lasso-type estimators. *Annals of Statistics, 28,* 1356–1378.

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

Magnus, J. R., & De Luca, G. (2016). Weighted-average least squares (WALS): A survey. *Journal of Economic Surveys, 30,* 117–148.

Magnus, J. R., & Durbin, J. (1999). Estimation of regression coefficients of interest when other regression coefficients are of no interest. *Econometrica, 67,* 639–643.

Magnus, J. R., Powell, O., & Prüfer, P. (2010). A comparison of two averaging techniques with an application to growth empirics. *Journal of Econometrics, 154,* 139–153.

Sala-i-Martin, X., Doppelhofer, G., & Miller, R. I. (2004). Determinants of long-term growth: A Bayesian averaging of classical estimates (BACE) approach. *American Economic Review, 94,* 813–835.

Sousa, J. M., & Sousa, R. M. (2019). Asset returns under model uncertainty: Evidence from the Euro area, the US and the UK. *Computational Economics, 54,* 139–176.

Steel, M. F. J. (2020). Model averaging and its use in economics. *Journal of Economic Literature, 58,* 644–719.

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

Wang, H., & Zhou, S. Z. F. (2013). Interval estimation by frequentist model averaging. *Communications in Statistics-Theory and Methods, 42,* 4342–4356.

Zhang, X. (2021). A new study on asymptotic optimality of least squares model averaging. *Econometric Theory, 37,* 388–407.

Zhang, X., & Liu, C.-A. (2019). Inference after model averaging in linear regression models. *Econometric Theory, 35,* 816–841.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association, 101,* 1418–1429.