Weighted-average least squares (WALS): Confidence and prediction intervals*

Giuseppe De Luca
University of Palermo, Palermo, Italy (giuseppe.deluca@unipa.it)

Jan R. Magnus
Vrije Universiteit Amsterdam and Tinbergen Institute,
Amsterdam, The Netherlands (jan@janmagnus.nl)

Franco Peracchi
University of Rome Tor Vergata and EIEF, Rome, Italy (peracchi@uniroma2.it)

May 13, 2021

Abstract: We extend the results of De Luca *et al.* (2021) to inference for linear regression models based on weighted-average least squares (WALS), a frequentist model averaging approach with a Bayesian flavor. We concentrate on 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 representing the nuisance component of the model. In our Monte Carlo simulations we compare the performance of WALS with that of several competing estimators, including the unrestricted least-squares estimator (with all auxiliary regressors) and the restricted least-squares estimator (with no auxiliary regressors), two post-selection estimators based on alternative model selection criteria (the Akaike and Bayesian information criteria), various versions of frequentist model averaging estimators (Mallows and jackknife), and one version of a popular shrinkage estimator (the adaptive LASSO). We discuss confidence intervals for the focus parameter and prediction intervals for the outcome of interest, and conclude that the WALS approach leads to superior confidence and prediction intervals, but only if we apply a bias correction.

Keywords: Linear model; post-selection estimators; adaptive lasso; frequentist model averaging; WALS; confidence intervals; prediction intervals; Monte Carlo simulations.

JEL classification: C11, C12, C18, C21, C52.

*Corresponding author: Jan R. Magnus (jan@janmagnus.nl). We thank Paolo Li Donni, Chu-An Liu, and Xinyu Zhang for useful discussions.

1 Introduction

Data are generated by a potentially complex process, the so-called data-generating process (DGP), usually represented by a joint probability distribution over the sample space. The investigator does not know the DGP, so she uses models. These models should be 'close' to the DGP, the measure of closeness depending on what purpose the investigator has in mind. The best model for one purpose is not necessarily the best model for another purpose (Hjort and Claeskens 2003, Hansen 2005). The branch of statistical theory that attempts to find the best model for a given purpose from the available data is called 'model selection'. Like any other data-driven statistical decision, model selection is subject to sampling uncertainty which, if ignored, can lead to overestimating the accuracy of parameter estimates (Kabaila and Mainzer 2018).

In contrast, 'model averaging' is not concerned with finding the best model, but with the best estimator of those features of the DGP that are of interest to the investigator.[1] This estimator is based on a set of models, each model producing one estimator and one measure of model uncertainty. Model averaging combines all these estimators with weights that take into account the uncertainty about each model. No model is selected as 'the best' as estimation is based on the contribution of all models.

There is now a large literature on model averaging, both from a frequentist and a Bayesian perspective; see Steel (2020) for a thorough and extensive survey. Our approach, weighted-average least squares (WALS), is frequentist but with a Bayesian flavor. It applies to a linear regression model in which we are certain about including a set of core or 'focus regressors', but uncertain about the number and identity of a set of additional controls or 'auxiliary regressors'. For each model in the model space, the coefficients on the focus and auxiliary regressors (the focus and auxiliary parameters) are estimated by constrained least squares, hence by a frequentist procedure. However, after implementing a semi-orthogonal transformation of the auxiliary regressors, the WALS weighting scheme is developed following a Bayesian approach in order to obtain desirable theoretical properties. The final result is a model averaging estimator that assumes an intermediate position between Bayesian and frequentist model averaging.

Most model averaging estimators are biased in ways that are not properly captured by the local misspecification framework, which assumes that the specification error vanishes with the sample size n at the convergence rate of $\sqrt{n}$. Furthermore, the sampling distribution of most model averaging estimators is not well approximated by the normal distribution and there is increasing evidence

[1] Confusingly, the word 'model averaging' is a misnomer since we do not average over models but over estimators.

that, even after correcting for bias, inference based on model averaging can be misleading when relying on the normal approximation (Hjort and Claeskens 2003, Claeskens and Hjort 2008, Hansen 2014, Liu 2015, DiTraglia 2016, Zhang and Liu 2019). The purpose of the current paper is to find out whether this is also the case in WALS and, if so, what can be done about it.

The 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 for WALS confidence and prediction intervals. This approach yields re-centered intervals, using the bias-corrected posterior mean as a frequentist estimator of the normal location parameter.

We assess the finite-sample performance of this approach by an extensive Monte Carlo experiment, where in some cases we can cross-check our simulation results with exact results or analytic approximations. To facilitate comparisons with the simulation study by Zhang and Liu (2019), hereafter ZL, we stay close to their framework and consider a finite model space that is assumed to contain the DGP (M-closed environment) but has little additional structure. Unlike ZL, who restrict their attention to inference about a single *auxiliary* parameter, we first concentrate on 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. We compare the performance of WALS with that of several competing estimators, including the unrestricted least-squares estimator (with all auxiliary regressors) and the restricted least-squares estimator (with no auxiliary regressors), two post-selection estimators based on alternative model selection criteria (the Akaike and Bayesian information criteria), various versions of frequentist model averaging estimators (Mallows and jackknife), and one version of a popular shrinkage estimator (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, 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 by simulation.

The remainder of this paper is organized as follows. In Section 2 we introduce the framework and describe the estimators which we wish to consider. One of these estimators is WALS, which we describe in some detail in Section 3. In Section 4 we discuss how to construct confidence intervals for these estimators. The Monte Carlo experiment is described in Section 5. Sections 6–8 contain the simulation results, separately for point estimates (Section 6), confidence intervals (Section 7), and

prediction intervals (Section 8). Section 9 concludes. Appendix A contains the abbreviations used in the paper, and Appendix B describes the main algorithm for simulation-based WALS confidence intervals.

2 Framework and estimators

Our framework is the linear regression model

$$y = X\beta + \epsilon = X_1\beta_1 + X_2\beta_2 + \epsilon, \tag{1}$$

where y $(n \times 1)$ is the vector of observations on the outcome of interest, X_1 $(n \times k_1)$ and X_2 $(n \times k_2)$ are matrices of nonrandom regressors, β_1 and β_2 are unknown parameter vectors, and ϵ is a vector of random disturbances. The k_1 columns of X_1 contain the 'focus regressors' which we want in the model on theoretical or other grounds, while the k_2 columns of 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, such as indicator variables, polynomials, B-splines, etc. We assume that $k_1 \geqslant 1$, $k_2 \geqslant 0$, and that $X = (X_1, X_2)$ has full column-rank $k = k_1 + k_2 \leqslant n$. The disturbance vector ϵ has zero mean and a positive definite variance matrix, diagonal but not necessarily equal to $\sigma^2 I_n$. The DGP thus allows for nonnormality and heteroskedasticity.

When $k_2 = 0$ there is no model uncertainty and we simply estimate the model with only the focus regressors, but when $k_2 = 1$ we have two models to consider depending on whether we include or exclude the auxiliary regressor. In general, there are 2^{k_2} possible models that contain all focus regressors and a (possibly empty) subset of the auxiliary regressors. If $\widehat{\beta}_{1j}$ and $\widehat{\beta}_{2j}$ are the least-squares (LS) estimators of β_1 and β_2 in model j, then the model averaging estimators take the form

$$\widehat{\beta}_1 = \sum_{j=1}^{2^{k_2}} \lambda_j \widehat{\beta}_{1j}, \qquad \widehat{\beta}_2 = \sum_{j=1}^{2^{k_2}} \lambda_j \widehat{\beta}_{2j}, \tag{2}$$

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 weights can be substantial. For example, when $k_2 = 20$ the model space contains more than one million models.

One possibility is to reduce the number of models by *preordering*, as suggested by Hansen (2007). In this approach we order the auxiliary regressors *a priori* such that we only consider $k_2 + 1$ nested models where model p contains the focus regressors and the first p auxiliary regressors

$(p = 0, 1, \ldots, k_2)$. Except for a few special 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 remains open (see, however, Hansen 2014, p. 498) and if we use preliminary regressions to order the regressors then the statistical noise generated by these preliminary investigations should not be ignored.

Another possibility to reduce the computational burden is weighted-average least squares, as proposed in Magnus *et al.* (2010). We shall discuss the WALS method in more detail in Section 3. Before we discuss WALS, let us review the other estimators which will play a role in this paper. Most of these estimators also appear in ZL, and we purposely try to stay close to that paper for reasons of comparability.

Least squares (LS). The unrestricted LS estimator $\widehat{\beta}_u = (X'X)^{-1}X'y$ is denoted by LS-U and includes all k_2 auxiliary regressors. The restricted LS estimator $\widehat{\beta}_r = (\widehat{\beta}'_{1,r}, 0')'$, where $\widehat{\beta}_{1,r} = (X_1'X_1)^{-1}X_1'y$, is denoted by LS-R and includes no auxiliary regressors. These estimators require neither preordering nor model selection.

Information criterion (IC). As implemented in ZL, these estimators require preordering and the assumption that the errors in (1) are homoskedastic. After preordering, the pth model ($p = 0, 1, \ldots, k_2$) has k_1 focus regressors and p auxiliary regressors. Let

$$M_p^* = I_n - X_p^*(X_p^{*'}X_p^*)^{-1}X_p^{*'} \tag{3}$$

be the usual idempotent matrix in model p, where $X_p^* = (X_1, X_{2,p})$ denotes the matrix containing the first $k_1 + p$ regressors, and let $\widehat{\sigma}_p^2 = y'M_p^*y/n$ be the maximum likelihood (ML) estimator of the error variance in model p. The Akaike IC for model p is

$$\mathrm{AIC}_p = n\log(\widehat{\sigma}_p^2) + 2(k_1 + p) \tag{4}$$

and the Bayesian IC for model p is

$$\mathrm{BIC}_p = n\log(\widehat{\sigma}_p^2) + (\log n)(k_1 + p). \tag{5}$$

The IC-A model selection estimator is the LS estimator in the model with the lowest value of AIC_p, and the IC-B model selection estimator is the LS estimator in the model with the lowest value of BIC_p. There is no model averaging here, only model selection. Model averaging based on smoothed AIC and BIC weights was considered by Buckland *et al.* (1997) and Burham and Anderson (2002).

Adaptive LASSO (ALASSO). The adaptive LASSO estimator proposed by Zou (2006) does not rely on preordering. It solves the optimization problem

$$\min_{\beta}\left((y-X\beta)'(y-X\beta)+\psi_n\sum_{l=1}^{k}\frac{|\beta_l|}{\widehat{\beta}_{l,u}^2}\right), \tag{6}$$

where β_l is the lth component of β, $\widehat{\beta}_{l,u}$ is the lth component of the unrestricted LS estimator of β, and ψ_n is a tuning parameter selected by the generalized cross-validation method (Li 1987, Andrews 1991). The ALASSO estimator does not distinguish between focus and auxiliary regressors, but it could easily be modified by only penalizing the coefficients on the auxiliary regressors.

Mallows. The Mallows model averaging (MMA) estimator was introduced by Hansen (2007). It relies on preordering and assumes that the errors in (1) are homoskedastic with known variance σ^2 which we set equal to s_u^2, the unbiased LS estimator of σ^2 in the unrestricted model. Let $M^*(w) = \sum_{p=0}^{k_2} w_p M_p^*$, where M_p^* is defined in (3). Then the MMA weights are obtained by solving

$$\min_{w}\left(y'M^{*'}(w)M^*(w)y+2s_u^2\sum_{p=0}^{k_2}w_p(k_1+p)\right) \tag{7}$$

subject to $\sum_p w_p = 1$ and $0 \leqslant w_p \leqslant 1$ for all p. If we denote the optimal weights by $\widehat{w}_p$ then the MMA estimator takes the form

$$\widehat{\beta}_{\text{MMA}} = \sum_{p=0}^{k_2}\widehat{w}_p(X_p^{*'}X_p^*)^{-1}X_p^{*'}y. \tag{8}$$

The MMA estimator is asymptotically efficient (in the mean squared error sense) for nested models under homoskedasticity, but not under heteroskedasticity (Hansen 2007).

Jackknife. The jackknife model averaging (JMA) estimator (Hansen and Racine 2012) also relies on preordering but it allows for heteroskedasticity.[2] Let D_p be the diagonal matrix containing the diagonal elements of M_p^* on its diagonal and zeros elsewhere and define $M^\dagger(w) = \sum_{p=0}^{k_2} w_p D_p^{-1} M_p^*$.[3] Then the JMA weights are obtained by solving

$$\min_{w}\ y'M^{\dagger'}(w)M^\dagger(w)y \tag{9}$$

[2]In our framework, MMA and JMA are set to be nested, as in ZL. See however Wan *et al.* (2010), Hansen and Racine (2012), and Zhang (2021) for asymptotic efficiency results in a non-nested framework.

[3]The diagonal elements of D_p are all nonnegative, but not necessarily strictly positive. To ensure nonsingularity of D_p we must add the requirement that the ith unit vector in $\mathbb{R}^n$ (the vector whose ith entry is 1 and 0 elsewhere) does not lie in the column space of X for any i.

with respect to w subject to $\sum_p w_p = 1$ and $0 \leqslant w_p \leqslant 1$ for all p. The JMA estimator is asymptotically efficient for nested models under heteroskedasticity. Under homoskedasticity, the MMA and JMA estimators have the same (nonstandard) limiting distribution (ZL, p. 824). The JMA-M (modified JMA) estimator introduced by ZL is defined by weights that solve

$$\min_w \left(y' M^{\dagger'}(w) M^{\dagger}(w) y + \psi_n \sum_{p=0}^{k_2} w_p (k_1 + p) \right) \tag{10}$$

subject to the same constraints as in (9), where the tuning parameter ψ_n is set equal to $\log n$, as in ZL. The JMA and JMA-M estimators take the same form as (8) where $\widehat{w}_p$ is given by the solution of (9) and (10), respectively.

3 The WALS approach

The WALS estimator was introduced in Magnus *et al.* (2010) and reviewed in Magnus and De Luca (2016). Unlike other model averaging estimators, the WALS approach exploits a semi-orthogonal transformation of the auxiliary regressors that reduces the computational burden from order 2^{k_2} to order k_2, coupled with a rescaling of the focus regressors that improves the accuracy of inversion and eigenvalue routines. Specifically, we transform X_2 and β_2 by defining $Z_2 = X_2 \Delta_2 \Psi^{-1/2}$ and $\gamma_2 = \Psi^{1/2} \Delta_2^{-1} \beta_2$, where Δ_2 is a diagonal $k_2 \times k_2$ matrix such that all diagonal elements of $\Psi = \Delta_2 X_2' M_1 X_2 \Delta_2$ are equal to one and $M_1 = I_n - X_1 (X_1' X_1)^{-1} X_1'$. We also rescale X_1 and β_1 by defining $Z_1 = X_1 \Delta_1$ and $\gamma_1 = \Delta_1^{-1} \beta_1$, where Δ_1 is a diagonal $k_1 \times k_1$ matrix such that all diagonal elements of $Z_1' Z_1$ are equal to one. Since $Z_1 \gamma_1 = X_1 \beta_1$ and $Z_2 \gamma_2 = X_2 \beta_2$, we may then write model (1) equivalently as

$$y = Z_1 \gamma_1 + Z_2 \gamma_2 + \epsilon. \tag{11}$$

The fact that $Z_2' M_1 Z_2 = I_{k_2}$ brings several important advantages. First, if $\widehat{\gamma}_{1j}$ and $\widehat{\gamma}_{2j}$ are the LS estimators of γ_1 and γ_2 in model j, then the WALS estimators can be written as

$$\widehat{\gamma}_1 = \sum_{j=1}^{2^{k_2}} \lambda_j \widehat{\gamma}_{1j} = \widehat{\gamma}_{1,r} - Q W \widehat{\gamma}_{2,u}, \qquad \widehat{\gamma}_2 = \sum_{j=1}^{2^{k_2}} \lambda_j \widehat{\gamma}_{2j} = W \widehat{\gamma}_{2,u}, \tag{12}$$

where $\widehat{\gamma}_{1,r} = (Z_1' Z_1)^{-1} Z_1' y$ is the estimator of γ_1 in the restricted model (with $\gamma_2 = 0$), $\widehat{\gamma}_{2,u} = Z_2' M_1 y$ is the estimator of γ_2 in the unrestricted model, $Q = (Z_1' Z_1)^{-1} Z_1' Z_2$, $W = \sum_j \lambda_j W_j$, and $W_j = I_{k_2} - S_j S_j'$, where S_j is a $k_2 \times r_j$ selection matrix of rank $0 \leqslant r_j \leqslant k_2$ — that is, $S_j' = [I_{r_j} : 0]$ or a column-permutation thereof — representing the r_j exclusion restrictions implied by model $j = 1, \ldots, 2^{k_2}$.

Second, the dependence of $\widehat{\gamma}_1$ and $\widehat{\gamma}_2$ on the estimators from all 2^{k_2} models in the model space is completely captured by the random diagonal matrix $W = \sum_j \lambda_j W_j$, whose k_2 diagonal elements w_h are partial sums of the λ_j since the W_j are nonrandom diagonal matrices with $k_2 - r_j$ ones and r_j zeros on the diagonal. It follows that the computational burden of calculating $\widehat{\gamma}_1$ and $\widehat{\gamma}_2$ is of order k_2, as we only need to compute the restricted estimate $\widehat{\gamma}_{1,r}$ and the unrestricted estimate $\widehat{\gamma}_{2,u}$, and determine the set of k_2 WALS weights w_h; not the considerably larger set of 2^{k_2} model weights λ_j. Unlike some other approaches we are not ignoring any of the 2^{k_2} models in the original model space because each model contributes to the model averaging estimates through the k_2 diagonal elements of the matrix W.

Third, Theorem 2 of Magnus and Durbin (1999) implies that the mean squared error (MSE) of $\widehat{\gamma}_1$ depends on the MSE of $\widehat{\gamma}_2$. Thus, if we can choose the λ_j optimally such that $\widehat{\gamma}_2$ is a 'good' estimator of γ_2 (in the MSE sense), then *the same* weights will also provide a 'good' estimator of γ_1.

Fourth, the components of $\widehat{\gamma}_2 = W\widehat{\gamma}_{2,u}$ are shrinkage estimators of the components of γ_2, as $0 \leqslant w_h \leqslant 1$. Under the additional assumption that the errors in (11) are homoskedastic and normal, $\widehat{\gamma}_{2,u} \sim \mathcal{N}(\gamma_2, \sigma^2 I_{k_2})$. Hence, if we restrict each w_h to depend only on the hth component of $\widehat{\gamma}_{2,u}$, then the shrinkage estimators in $\widehat{\gamma}_2$ will also be independent. Under this additional restriction (discussed in detail in Magnus and De Luca 2016), our k_2-dimensional problem reduces to k_2 (identical) one-dimensional problems, namely: given one observation $x \sim \mathcal{N}(\eta, \sigma^2)$, what is the estimator $m(x)$ of η with minimum MSE? This is the so-called *normal location problem*. Since the risk properties of $m(x)$ are little affected by estimating the variance parameter (Danilov 2005, Magnus and De Luca 2016), we also assume that σ^2 is known.

The normal location problem is an important ingredient in the WALS procedure. We take a Bayesian approach to it, which allows a proper treatment of admissibility, bounded risk, robustness, near-optimality in terms of minimax regret, and ignorance about η. The Bayesian approach requires two elements: a distribution for the k_2-vector of t-ratios $x = \widehat{\gamma}_{2,u}/s_u$ where, as before, $s_u^2 = y'M_1(I_n - Z_2Z_2')M_1y/(n-k)$ is the LS estimator of the error variance in (11), and a neutral prior with bounded risk for the normal location parameter η. The concept of neutrality relates to the notion of ignorance about η (Kumar and Magnus 2013, Magnus and De Luca 2016). Specifically, it requires the prior median of η to be zero and the prior median of $|\eta|$ to be one.

Assuming, as for the MMA estimator, that the errors in (11) are homoskedastic with known variance equal to s_u^2, the k_2 components x_h of x are independently distributed as $\mathcal{N}(\eta_h, 1)$. The

Bayesian approach to this normal location problem then yields the posterior mean $m_h = m(x_h)$ as an estimator of η_h. Hence, the WALS estimators of γ_1 and γ_2 are

$$\widehat{\gamma}_1 = \widehat{\gamma}_{1,r} - Q\widehat{\gamma}_2, \qquad \widehat{\gamma}_2 = s_u m, \tag{13}$$

with $m = (m_1, \ldots, m_{k_2})'$, and the WALS estimators of β_1 and β_2 are

$$\widehat{\beta}_1 = \Delta_1 \widehat{\gamma}_1, \qquad \widehat{\beta}_2 = \Delta_2 \Psi^{-1/2} \widehat{\gamma}_2. \tag{14}$$

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 that $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are in general biased estimators of β_1 and β_2. Second, for any nonnegative bounded prior density, the variance $v_h^2(x_h)$ of the posterior distribution of η_h represents a first-order approximation to the sampling *standard deviation* of the posterior mean m_h (De Luca *et al.* 2021, Proposition 2). This somewhat counterintuitive result shows that care should be taken when assessing the sampling precision of the posterior mean m_h as a frequentist estimator of η_h.

De Luca *et al.* (2021) recently investigated these issues and proposed new plug-in estimators of the sampling moments of m_h. Specifically, they first used Monte Carlo methods to accurately tabulate the functional forms of the bias $\delta_h(\eta_h)$ and variance $\sigma_h^2(\eta_h)$ of the posterior mean m_h under three types of neutral priors with bounded risk belonging to the class of (reflected) generalized gamma distributions: Laplace, Weibull, and Subbotin. For each prior, they then compared two alternative plug-in methods for estimating the unknown location parameter η_h: the maximum likelihood estimator x_h and the Bayesian posterior mean m_h. The first estimator leads to the plug-in ML estimators $\delta_h(x_h)$ and $\sigma_h^2(x_h)$, the second to the plug-in double shrinkage estimators (so named because of the double use of the posterior mean function in the leading term of the analytical approximations to the estimated bias) $\delta_h(m_h)$ and $\sigma_h^2(m_h)$. Based on these plug-in estimators, De Luca *et al.* (2021) also derived new estimators for the sampling bias and variance of the WALS estimators (14). These findings have implications for WALS inference, i.e. the construction of confidence and prediction intervals, and these implications are the subject of the current paper.

4 Confidence intervals

We shall consider sixteen confidence intervals: ten from ZL and six based on WALS. In this section we write our parameter of interest as β_l, indicating the lth component of β, which could be either

a focus or an auxiliary parameter. As in ZL, we wish to construct $(1-\alpha)$-level confidence intervals for β_l and these take the form

$$\mathrm{CI}_n(\beta_l) = \left[\check{\beta}_l - \underline{c}_l, \check{\beta}_l + \overline{c}_l\right], \tag{15}$$

where $\check{\beta}_l$ is an estimate of β_l and the quantities $\underline{c}_l$ and $\overline{c}_l$ are chosen to attain the desired coverage level. If $\underline{c}_l = \overline{c}_l$ then the interval is called symmetric.

Least squares (LS). The confidence interval for β_l in the unrestricted model is obtained by setting $\check{\beta}_l$ equal to the lth component of the LS-U estimator $\widehat{\beta}_u$ and letting $\underline{c}_l = \overline{c}_l = z_{1-\alpha/2}\, s(\check{\beta}_l)$, where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$th quantile of the standard normal distribution and $s(\check{\beta}_l)$ denotes the (classical or heteroskedasticity-robust) standard error of $\check{\beta}_l$. The confidence interval for β_l in the restricted model is obtained in a similar way except that $\check{\beta}_l$ is now the lth component of $\widehat{\beta}_r$.

Information criterion (IC). For the IC-A and IC-B estimators, let $\widehat{p}$ be the number of auxiliary regressors selected by the Akaike or Bayesian IC, $\widehat{\beta}_l(\widehat{p})$ the LS estimator of β_l in the selected model, and $s(\widehat{\beta}_l(\widehat{p}))$ the associated standard error. Then, $\check{\beta}_l = \widehat{\beta}_l(\widehat{p})$ and $\underline{c}_l = \overline{c}_l = z_{1-\alpha/2}\, s(\check{\beta}_l)$. ZL call these confidence intervals 'naive' because they ignore model selection noise.

Adaptive LASSO (ALASSO). Here $\check{\beta}_l$ is the ALASSO estimator of β_l 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 rely on the asymptotic validity of the bootstrap for the ALASSO estimator, established by Chatterjee and Lahiri (2011) and Camponovo (2015) for alternative versions of the bootstrap.

Mallows. There are two variants based on the Mallows estimator $\check{\beta}_l$ of β_l. In MMA-B we have $\underline{c}_l = \overline{c}_l = n^{-1/2} q_l^*(\alpha)$, 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. In MMA-S we have $\underline{c}_l = n^{-1/2} q_l(1-\alpha/2)$ 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 ZL (Theorem 2). The first interval is symmetric, the second is not.

Jackknife. There are two variants based on the jackknife estimator and one based on the modified jackknife estimator. Let $\check{\beta}_l$ be the jackknife estimator of β_l. In JMA-B we have, similar to MMA, $\underline{c}_l = \overline{c}_l = n^{-1/2} q_l^*(\alpha)$, 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. In JMA-S we have $\underline{c}_l = n^{-1/2} q_l(1-\alpha/2)$ and $\overline{c}_l = -n^{-1/2} q_l(\alpha/2)$, where $q_l(\alpha)$ is now based on ZL (Theorem 4).

For the modified estimator we let $\breve{\beta}_l$ be the JMA-M estimator of β_l 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 (using ZL's definition), that is, the model obtained from the ordered sequence of models by deleting all redundant regressors (where the parameter is zero) which appear *at the end* of the sequence. (Hence, there may still be redundant regressors in the just-fitted model, but the last included regressor is *not* redundant.)[4] Symmetry of these confidence intervals is justified by the asymptotic normality of the JMA-M estimator (ZL, Theorem 5).

WALS. We shall compare three types of WALS confidence intervals for β_l: uncentered-and-naive (UN), centered-and-naive (CN), and simulation-based (S).

In the uncentered-and-naive confidence interval we set $\breve{\beta}_l$ equal to the WALS estimator $\widehat{\beta}_l$ and let $\underline{c}_l = \overline{c}_l = z_{1-\alpha/2}\, s(\widehat{\beta}_l)$, where $s(\widehat{\beta}_l)$ is either the plug-in maximum likelihood (ML) estimator or the plug-in double-shrinkage (DS) estimator of its standard error. Uncentered-and-naive confidence 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 of β_l.

In the centered-and-naive confidence interval we apply a bias correction and set $\breve{\beta}_l$ equal to the bias-corrected WALS estimator

$$\breve{\beta}_l = \widehat{\beta}_l - b(\widehat{\beta}_l), \tag{16}$$

where $b(\widehat{\beta}_l)$ is either the plug-in ML estimator or the plug-in double-shrinkage estimator of the bias of $\widehat{\beta}_l$. As in the uncentered-and-naive approach we set $\underline{c}_l = \overline{c}_l = z_{1-\alpha/2}\, s$, but now, in contrast to the uncentered-and-naive approach, $s = s(\breve{\beta}_l)$ depends on the bias-corrected WALS estimator and is computed by the simulation-based algorithm discussed in Appendix B. As for the uncentered-and-naive confidence interval, the centered-and-naive confidence interval takes the classical normal approximation at face value (hence naive), but it re-centers to correct for estimation bias and it accounts for randomness in the estimated bias.

The simulation-based approach also yields a re-centered confidence interval by using the bias-corrected posterior mean as a frequentist estimator of the normal location parameter, and it accounts for the randomness of the bias-corrected posterior mean by exploiting a large set of pseudo-random Monte Carlo replications. But unlike the centered-and-naive approach it produces

[4]The just-fitted model is unknown in practice, so s_l^* is not a feasible estimator. In applications, one could treat the model that receives the largest weight as the just-identified model because the weights of under-fitted and over-fitted models should be small. In the simulations we follow ZL 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.

a quantile-based confidence interval that does not require critical values from the normal distribution and is not necessarily symmetric. We expect the simulation-based confidence interval to be superior to the other two. The algorithm underlying the centered-and-naive and simulation-based methods is provided in Appendix B.

5 Monte Carlo setup

The simulation setup closely follows ZL with some exceptions which we shall list and explain later in this section. We have $k_1 = 2$ focus regressors, x_{11} and x_{12} (of which the first is the constant term), and k_2 auxiliary regressors, $x_{21}, \ldots, x_{2k_2}$. Our parameter of interest is the coefficient β_{12} on the second focus regressor x_{12}, which may be interpreted as the causal effect of x_{12} on y.

The $k_2 + 1$ regressors $x_{12}, x_{21}, \ldots, x_{2k_2}$ are drawn from a multivariate normal distribution with mean zero and variance $\sigma_x^2 \, \Sigma_x(\rho)$, where

$$\Sigma_x(\rho) = \begin{pmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \vdots & \vdots & & \vdots \\ \rho & \rho & \ldots & 1 \end{pmatrix} \qquad (-\frac{1}{k_2} < \rho < 1). \tag{17}$$

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 t-distribution with five degrees of freedom or a skewed t^*-distribution (also with $d = 5$ degrees of freedom), defined as

$$f(t^*; \mu, \sigma, d, \lambda) = \frac{\Gamma(\frac{d+1}{2})}{\pi^{1/2}(\theta\sigma)\Gamma(\frac{d}{2})} \left(1 + \frac{(t^* - \mu)^2}{(1 + \lambda \, \mathrm{sgn}(t^* - \mu))^2 (\theta\sigma)^2}\right)^{-\frac{d+1}{2}}, \tag{18}$$

where $d > 2$, $|\lambda| < 1$, $\mathrm{sgn}(\cdot)$ is the sign function, and

$$\theta = \frac{[\pi^{1/2}(d-2)^{1/2}\Gamma(\frac{d-2}{2})] \, (d-2)^{1/2}}{\sqrt{(1+3\lambda^2)[\pi^{1/2}(d-2)^{1/2}\Gamma(\frac{d-2}{2})]^2 - [4\lambda\Gamma(\frac{d-1}{2})]^2}} \tag{19}$$

is a normalizing constant to ensure that $\mathbb{V}(t^*) = \sigma^2$ (Hansen 1994, Theodossiou 1998). The parameter λ controls the skewness of the distribution (defined for $d > 3$). For $\lambda = 0$ we obtain the standard t-distribution with d degrees of freedom by setting $\mu = 0$ and $\sigma = d^{1/2}/\theta = (d/(d-2))^{1/2}$. When $\lambda > 0$ the distribution is skewed to the right; and when $\lambda < 0$ it is skewed to the left. In addition to the standard normal distribution and the $t(5)$-distribution we shall consider two versions of the skewed t^*-distribution with five degrees of freedom, one with $\lambda = 0.5$ (moderate positive skewness) and one with $\lambda = 0.8$ (large positive skewness).

Table 1: 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	$t^*(0,1,5,0.8)$	$2.5\,\tau_i$

We consider four homoskedastic and four heteroskedastic error distributions, as given in Table 1. In the homoskedastic cases we take $\sigma_i = 2.5$ when the distribution of u_i has variance one. For the $t(5)$-distribution the variance is $5/3$ and hence we need a correction factor $\sqrt{15/4} = 2.5/\sqrt{5/3}$. In the heteroskedastic cases we define

$$\tau_i = \frac{1 + 2|x_{12}^{(i)}| + 4|x_{21}^{(i)}|}{1 + 6\sigma_x\sqrt{2/\pi}}, \tag{20}$$

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

Table 2: Four configurations of the $k_2 = 8$ auxiliary parameters

Conf.	β_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 set $k_2 = 8$ so that we have $2^{k_2} = 256$ possible models that include the two focus regressors and a subset of the eight auxiliary regressors. We fix $\beta_1 = (1, 1)'$ and consider four configurations of the eight auxiliary coefficients, as shown in Table 2.

Our setup is intentionally similar to the setup in ZL with three important exceptions:

- Our parameter of interest is one of the focus parameters, not one of the auxiliary parameters as in ZL, because it is focus parameters that we are primarily interested in.

- ZL ignore the possibility of skewness in the error distribution. In fact, of the eight cases in Table 1 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 ZL's case 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 2, 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 1 with the four configurations of the auxiliary parameters in Table 2, 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 ZL), we compute the bias, variance, and MSE of the nine estimators discussed in Sections 2 and 3: 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.[5] Our simulation data were generated in MATLAB to exploit the availability of the `sgtrnd` routine for computing pseudo-random draws from the skewed t^*-distribution. Since WALS has been shown to be quite robust to different choices of the prior (De Luca *et al.* 2018, 2021), we restrict our attention to WALS based on the Laplace prior because the Laplace prior allows closed-form expressions for the posterior moments (De Luca *et al.* 2020) and this has computational advantages.

6 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. This section discusses point estimates. Confidence intervals and prediction intervals are discussed in Sections 7 and 8, respectively.

[5]The MATLAB routines were kindly provided by X. Zhang and C.-A. Liu. All Stata routines are available from the authors upon request.

FIGURES 1–2 HERE

In Figures 1 and 2 we present the first two sampling moments of the nine estimators for $n = 100$. The sixteen plots in Figure 1 represent the homoskedastic designs, the sixteen plots in Figure 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 unrestricted estimator LS-U (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 1 and 2.

The similarity of the sixteen plots in Figure 1 is remarkable. The estimators LS-U, LS-R, ALASSO, and WALS are not affected by preordering, hence their moments and MSEs are the same across configurations. But this is not the case for the other five estimators: IC-A, IC-B, MMA, JMA, and JMA-M. For these other estimators 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 restricted estimator LS-R 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 unrestricted estimator in configurations (a) and (c), but considerably worse in configurations (b) and (d). As predicted by the asymptotic theory discussed in Section 2, 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 ZL, we don't find that JMA-M 'dominates other estimators in most cases'; in fact, JMA is 7–14% more efficient relative to JMA-M (as measured by the ratio of their MSEs) in the sixteen designs of Figure 1.

The dominating estimator is WALS, whose excess bias relative to LS-U (which is unbiased) 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 to violations of the normality assumption, probably due to the fact that $n = 100$ is already large enough to justify asymptotic approximations to the normal location problem based on the central limit theorem.

Now consider the case of heteroskedastic errors, still for $n = 100$, as plotted in Figure 2. All

models are now misspecified, also the model based on the normal distribution. In our setup this leads to a deterioration of the MSE by about 30% (averaged over all estimators and all designs with and without heteroskedasticity), but the ordering of estimators remains unaltered. Curiously and contrary to what is predicted by the asymptotic theory, MMA is 2% more efficient than JMA under heteroskedasticity. WALS is still the preferred estimator in terms of MSE.

FIGURE 3 HERE

When the number of observations increases, then things change. Since we work in an M-closed environment, the number of models is fixed and does not increase with the number of observations. The unrestricted model produces unbiased estimators whose variance (and hence MSE) decreases at the rate n^{-1}. So eventually LS-U will dominate unless we let the number of models increase as well (as we shall do later).

In Figure 3 we only present designs 1 and 5 (both based on the normal distribution) 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 over the four distributions, 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 unrestricted estimator LS-U is unbiased, while the restricted estimator LS-R has a bias which does not vanish asymptotically but rather converges to a limit; as a result, the bias in LS-R is almost constant between $n = 100$ and $n = 400$. 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 unrestricted estimator LS-U, 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) the ordering is important and 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 $16, 24, 32, \ldots, 64$ by setting $\beta_2 = (\xi, \xi^2, \ldots, \xi^{k_2/2}, 0, 0, \ldots, 0)'$. 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, we allow in addition to $\sigma_x^2 = \rho = 0.7$ (high correlation) also $\sigma_x^2 = \rho = 0.3$ (low correlation). In total, our second Monte Carlo experiment includes 128 simulation designs for the different combinations of n, k_2, ξ, and ρ. For each design, we consider again 5,000 Monte Carlo replications.

In the extended design we restrict ourselves to distribution 1 (homoskedastic normal errors) and to only two estimators: the unrestricted LS estimator LS-U and WALS.

FIGURE 4 HERE

In Figure 4 we consider the efficiency of the WALS estimator relative to the LS-U estimator, given by the ratio $\text{MSE}(\widehat{\beta}_{12,u})/\text{MSE}(\widehat{\beta}_{12,\text{WALS}})$. The smaller (better) is $\text{MSE}(\widehat{\beta}_{12,\text{WALS}})$, the higher is the efficiency of WALS relative to LS-U. 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. It turns out that the parameter values used in ZL, which we have followed to allow comparisons, are not the most favorable for WALS. The four plots in column 1 are the same as in the main simulation study with $k_2 = 8$, $\xi = 0.5$, and $\rho = 0.7$, except that we have now added the intermediate values $n = 200$ and $n = 300$. We see that LS-U dominates when n is larger than about 250. But when ξ is negative (column 2) or when the correlation is small (column 3) or both (column 4), then WALS dominates LS-U for (much) larger values of n, certainly larger than 400. As expected, we also see that an increase in the number of auxiliary variables increases the efficiency of WALS relative to LS-U.

7 Monte Carlo results: confidence intervals

In the previous section we presented and discussed the finite-sample performance of the nine point estimators defined in Section 2. Our main concern in this paper, however, is not with estimation but with inference, and hence we now turn to confidence intervals. We compare sixteen methods as discussed in Section 4, ten from ZL and six based on WALS.[6] Our parameter of interest is still β_{12}, the coefficient of the second focus regressor, and we consider confidence intervals for β_{12} of

[6]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, and those for WALS (DS-S, ML-S, DS-CN, and ML-CN) on 5,000 Monte Carlo replications. Despite the larger number of replications, the simulation-based algorithm for WALS is considerably faster than the other algorithms.

the form (15) with nominal coverage probability of (at least) $1-\alpha$. For given α (10%, 5%, 1%), we can calculate $\breve{\beta}_{12}$, $\underline{c}_{12}(\alpha)$, and $\overline{c}_{12}(\alpha)$ for each method and each replication of the 32 simulation designs. By averaging over the 5,000 Monte Carlo replications of each simulation design, we then obtain the coverage probability (the relative frequency that the interval contains the true value of β_{12}) and the length of the interval. Our first concern is how close to $1-\alpha$ this coverage probability is, our second concern is the average length of the confidence interval.

FIGURES 5–6 HERE

Figures 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 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 classical or heteroskedasticity-robust unrestricted least squares (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, in agreement with ZL's findings for the confidence intervals for an auxiliary coefficient. The bootstrapped confidence intervals for MMA and JMA are more accurate than the simulation-based algorithms proposed by ZL, but the underlying undercoverage errors are still sizeable (with $n = 100$, the undercoverage errors of MMA-B and JMA-B are -0.03 for $\alpha = 10\%$ and -0.02 for $\alpha = 5\%$) and they increase with the sample size (with $n = 400$, the undercoverage errors become -0.07 for $\alpha = 10\%$ and -0.05 for $\alpha = 5\%$). The confidence intervals of JMA-M also have nonnegligible undercoverage errors which tend to increase with the sample size. ALASSO performs well for $n = 400$, but the undercoverage errors of its 90% and 95% bootstrapped confidence intervals for $n = 100$ are rather large (-0.19 for $\alpha = 10\%$ and -0.08 for $\alpha = 5\%$). Apparently, the asymptotic validity of the bootstrap for ALASSO requires a large value of n, especially for confidence intervals with a small significance level. The uncentered-and-naive

confidence intervals for WALS do not perform well because they use critical values from the normal distribution and ignore the estimation bias. Ignoring the estimation bias is much more important than naively using critical values from the normal distribution, as is shown by first comparing uncentered-and-naive with centered-and-naive (large difference) and then centered-and-naive with simulation-based (small difference). Obviously to use the correct critical values is better, but the improvement is very small.

Table 3: Skewness and excess kurtosis of the bias-corrected WALS estimator of β_{12}

	Double shrinkage		Maximum likelihood	
	$n = 100$	$n = 400$	$n = 100$	$n = 400$
Skewness	−0.019	−0.026	−0.007	−0.010
Excess kurtosis	0.019	0.009	0.005	0.001

While we have established that taking the bias into account matters for constructing the correct confidence intervals using WALS, one may also wonder about the behavior of higher moments of the bias-corrected WALS estimator. In Table 3 we present the skewness and excess kurtosis of the estimator of the focus parameter β_{12}, where we average over the eight designs (four distributions, homo- and heteroskedastic) since the variation between the eight designs is negligible. These higher-order moments are again estimated by the simulation-based algorithm discussed in Appendix B.

The estimator is left-skewed and exhibits positive excess kurtosis, but the deviations from zero (the normal case) are very small. In comparison, the $\chi^2(8)$ distribution, which already looks quite 'normal', has skewness 1.0 and excess kurtosis 1.5, and the $\chi^2(32)$ distribution has skewness 0.5 and excess kurtosis 3/8. For an auxiliary parameter the deviation from normality is slightly higher, but still small. When n increases then the skewness increases somewhat while the excess kurtosis decreases. The ML-based estimator is slightly closer to normality than the DS-based estimator. Based on these findings we conclude that the bias-corrected WALS estimator is well approximated by a normal distribution, thus confirming the comparison between centered-and-naive and simulation-based earlier in this section.

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 under the extended design.

In the extended design defined in the previous section 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 double-shrinkage and maximum likelihood estimators of the bias of the posterior mean in the normal location model.[7]

FIGURES 7–8 HERE

The coverage probabilities of the three methods (LS-U, WALS-DS-S, WALS-ML-S) are compared in Figure 7 for the 90%, 95% and 99% confidence levels. The coverage errors of the three methods are in general small. For 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 double-shrinkage estimator of the bias of the posterior mean has always a larger bias than the plug-in maximum likelihood estimator (De Luca *et al.* 2021, Figure 3). Our results also suggest that the absolute value of the coverage errors for WALS-DS-S increases with α: it reaches a maximum of 0.006 when $\alpha = 1\%$, 0.018 when $\alpha = 5\%$, and 0.028 when $\alpha = 10\%$.

In Table 3 we investigated the skewness and excess kurtosis of the bias-corrected WALS estimator of the focus parameter β_{12}. In the extended design we also find that both skewness and excess kurtosis deviate very little from zero and that the ML-based estimator is slightly closer to normality than the DS-based estimator. For $\xi = -0.5$ the skewness is positive rather than negative. Hence the previous conclusion that the bias-corrected WALS estimator is well approximated by a normal distribution still holds in the extended design.

Next we compare the lengths of the confidence intervals in Figure 8, where we present the relative lengths (LS-U divided by WALS-DS-S and LS-U divided by WALS-ML-S) for the 95% level only, since the results for the 90% and 99% levels are indistinguishable from the 95% level. For all cases we have LS-U > WALS-ML-S > WALS-DS-S, so that the simulation-based WALS confidence intervals are always smaller than the classical LS-U confidence intervals, even for the designs where 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

[7] 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.

and about 5.4% for WALS-DS-S. This result agrees with the fact that, although more biased, the plug-in double-shrinkage estimator of the bias of the posterior mean has better MSE performance than the plug-in maximum likelihood estimator, at least when the unknown location parameter has a small or moderate value.

The relative gains of WALS on LS-U in terms of confidence interval length are much smaller than the relative gains in terms of MSE obtained from the point estimators, which agrees with the findings of Kabaila and Leeb (2006), Wang and Zhou (2013), and Ankargren and Jin (2018) 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, bias correction comes at the price of increased sampling variability, which is reflected in the length of the confidence interval.

8 Monte Carlo results: prediction intervals

Finally we consider the problem of predicting a single observation y_f given our framework $y = X\beta+\epsilon$ in (1) and given a covariate vector $x_f = (x_{1f}', x_{2f}')'$, that is,

$$y_f = x_f'\beta + \epsilon_f = x_{1f}'\beta_1 + x_{2f}'\beta_2 + \epsilon_f, \tag{21}$$

where ϵ and ϵ_f are jointly normally distributed, independent of each other, with zero means and variances $\mathbb{V}(\epsilon) = \sigma^2 I_n$ and $\mathbb{V}(\epsilon_f) = \sigma^2$. If $\widehat{\beta}_1$ and $\widehat{\beta}_2$ denote the WALS estimators of β_1 and β_2, then the WALS point predictor of y_f is defined as

$$\widehat{y}_f = x_{1f}'\widehat{\beta}_1 + x_{2f}'\widehat{\beta}_2, \tag{22}$$

and its prediction error is

$$\widehat{y}_f - y_f = x_{1f}'(\widehat{\beta}_1 - \beta_1) + x_{2f}'(\widehat{\beta}_2 - \beta_2) - \epsilon_f. \tag{23}$$

Note particularly the assumption of independence of ϵ and ϵ_f. Without this assumption the analysis is rather more complicated; see Magnus *et al.* (2016). Because of (13) and (14) the WALS point predictor of y_f may be viewed as a weighted average of the point predictors from all 2^{k_2} models in the model space.

We are interested in constructing a prediction interval for $\mathbb{E}(y_f) = x_{1f}'\beta_1 + x_{2f}'\beta_2$. The main difference between the confidence intervals introduced in Section 4 and the prediction intervals discussed here is that in the former case we are dealing with the sampling uncertainty on one focus

parameter only, while in the latter case we need to deal with the sampling uncertainty on all focus and auxiliary parameters.

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

$$x_f'\breve{\beta} - z_{1-\alpha/2}\sqrt{x_f' V_{\breve{\beta}}\, x_f} < \mathbb{E}(y_f) < x_f'\breve{\beta} + z_{1-\alpha/2}\sqrt{x_f' V_{\breve{\beta}}\, x_f}, \tag{24}$$

where $V_{\breve{\beta}}$ is the Monte Carlo variance of $\breve{\beta}$ estimated from 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 approach, 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 $B^* x_f$. This prediction interval need not be symmetric around $x_f'\breve{\beta}$.

FIGURES 9–11 HERE

Figure 9 presents the relative efficiency of the WALS point predictor of $\mathbb{E}(y_f) = x_f'\beta$ 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 x_f is drawn randomly from a multivariate normal distribution with mean zero and variance $\sigma_x^2 \Sigma_x(\rho)$ and then kept fixed for all replications of the same simulation design. Thus, x_f changes with k_2 and ρ.

The figure has the same format as Figure 4, except that efficiency is now measured by the ratio of the mean squared prediction error of WALS relative to LS-U. WALS clearly dominates LS-U in all designs, and by an even larger margin than what we have seen for the focus coefficient. As expected, the relative efficiency of WALS increases with the number k_2 of auxiliary coefficients in the DGP. The typical profile of the relative efficiency of the WALS predictor is concave in k_2, revealing very large gains when moving from a small number ($k_2 = 8$) to a moderate number ($k_2 = 24$) of auxiliary coefficients.

Figure 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 simulated-based intervals, both DS and ML, because the naive and simulation-based prediction intervals are always very close. This figure is the analog of Figure 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.

Figure 11 plots the relative lengths of the 95% prediction intervals based on LS-U and WALS, hence the analog of Figure 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 the number of auxiliary coefficients, is concave for all designs, again revealing large gains when moving from $k_2 = 8$ to $k_2 = 24$.

9 Conclusions

In this paper we have attempted to extend the theory of WALS estimation to inference. A key ingredient in this extension is the use of bias correction in WALS, as introduced in De Luca *et al.* (2021). To highlight the properties of WALS and put them in perspective we also discussed and analyzed its main competitors.

One problem with Monte Carlo experiments is that a critical reader may suspect that the authors have selected precisely those experiments that make their favorite tool shine. This suspicion may or may not be justified but it is almost impossible to check for the reader. With this possible critique in mind we have chosen for an existing Monte Carlo setup, namely the one employed by ZL, with only a few well-reasoned changes.

The WALS procedure has many advantages: computational simplicity even with a large number of auxiliary variables; no need to preorder the variables as in Hansen's procedures; no need to make data-dependent choices of tuning parameters as in adaptive LASSO; a natural (Bayesian) treatment of ignorance; and excellent finite-sample performance of the estimator.

We discussed both confidence intervals for the focus parameter and prediction intervals for the outcome of interest, and compared the performance of WALS with that of alternative estimators, including the unrestricted and restricted least-squares estimators, post-selection estimators based on different model selection criteria, various frequentist model averaging estimators, and the adaptive LASSO. Our results rely on 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.

We find, in the homoskedastic case and with a relatively small sample size ($n = 100$), that the dominating estimator is WALS, whose excess bias relative to the unrestricted LS estimator

(which is unbiased) is more than offset by a smaller variance, thus capturing the essence of model averaging. In the heteroskedastic case, in which all models are misspecified, the performance of all estimators deteriorates but their relative position in terms of MSE changes little. Since we work in an M-closed environment, the model space remains fixed as the sample size increases. The preferred estimator is now the unrestricted estimator LS-U, followed by ALASSO as second-best, and WALS as third-best. These three estimators are not affected by the order of the auxiliary variables (preordering).

Regarding coverage probabilities, there are five methods that produce accurate coverage probabilities, namely classical or heteroskedasticity-robust unrestricted least squares and the four centered versions of WALS. All other methods are much less accurate. In particular, the naive confidence intervals for post-selection estimators have large undercoverage errors because they ignore model selection noise. Although ALASSO performs well for $n = 400$, the undercoverage errors of its bootstrapped confidence intervals for $n = 100$ are rather large. Comparing the length of confidence intervals, WALS performs slightly better than the unrestricted LS estimator, though differences are small. As in De Luca *et al.* (2021) we have a slight preference for the plug-in DS estimator over the plug-in ML estimator of the bias of the posterior mean. Given the fact that ML leads to relatively smaller coverage errors and DS to relatively shorter intervals, and since coverage errors tend to increase with α, a specialist user may wish to use the ML estimator for 90% intervals and the DS estimator for 95% and 99% intervals.

Finally, regarding prediction intervals, WALS clearly dominates LS-U. The relative efficiency of WALS increases with the number of auxiliary coefficients and its typical profile is concave in the number of auxiliary variables. 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.

Appendix A: Abbreviations

The following abbreviations are used in this paper:

AIC: Akaike information criterion
ALASSO: adaptive LASSO
BIC: Bayes information criterion
CI: confidence interval
CN: centered-and-naive (confidence interval in WALS)
DGP: data-generation process

DS double shrinkage
IC: information criterion
IC-A: Akaike IC (model selection estimator)
IC-B: Bayes IC (model selection estimator)
JMA: jackknife model averaging
JMA-B: bootstrap-based JMA
JMA-M: modified JMA
JMA-S: simulation-based JMA
LASSO: least absolute shrinkage and selection operator
LS: least squares
LS-R: restricted LS
LS-U: unrestricted LS
ML: maximum likelihood
MMA: Mallows model averaging
MMA-B: bootstrap-based MMA
MMA-S: simulation-based MMA
MSE: mean squared error
S: simulation-based (confidence interval in WALS)
UN: uncentered-and-naive (confidence interval in WALS)
WALS: weighted-average least squares
ZL: Zhang and Liu (2019)

Appendix B: Algorithm for the simulation-based WALS confidence intervals

Let $x = \widehat{\gamma}_{2,u}/s_u = (x_1, \ldots, x_{k_2})'$ be the k_2-vector of t-ratios from the unrestricted model and $\widehat{\eta} = (\widehat{\eta}_1, \ldots, \widehat{\eta}_{k_2})'$ an estimator of the k_2-vector of parameters $\eta = (\eta_1, \ldots, \eta_{k_2})'$ in the multivariate normal location model $x \sim \mathcal{N}_{k_2}(\eta, I_{k_2})$. The simulation-based WALS confidence intervals for $\beta = (\beta_1', \beta_2')'$ are obtained by the following algorithm:

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

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

$$\breve{m}_{rh}^* = m_{rh}^* - \delta_{rh}^* \qquad (r = 1, \ldots, R;\ h = 1, \ldots, k_2), \tag{25}$$

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

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

$$\widehat{\beta}_{1,r} = \Delta_1 (Z_1' Z_1)^{-1} Z_1' y, \qquad V_{1r} = s_u^2 \Delta_1 (Z_1' Z_1)^{-1} \Delta_1 \tag{26}$$

are the LS-R estimate of β_1 in the fully restricted model and its estimated variance matrix, respectively.

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

$$\breve{B}^*_1 = B^*_{1r} - s_u \breve{M}^* Z_2' Z_1 (Z_1' Z_1)^{-1} \Delta_1, \qquad \breve{B}^*_2 = s_u \breve{M}^* \Psi^{-1/2} \Delta_2. \tag{27}$$

(v) Compute the $(1-\alpha)$-level confidence interval for the generic component β_l of β $(l = 1, \ldots, k)$ 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 $\breve{b}^*_l$ of $\breve{B}^*$.

Remarks on the algorithm. To achieve good performance in terms of coverage probabilities, the initial estimator $\widehat{\eta}$ in the first step of the algorithm must be (approximately) unbiased for η. This leaves us three possible choices: (i) the ML estimator x, (ii) the double-shrinkage bias-corrected posterior mean $\breve{m}(x) = m(x) - \delta(m(x))$, and (iii) the maximum likelihood bias-corrected posterior mean $\breve{m}(x) = m(x) - \delta(x)$. In our experience, the differences between these three estimators are small. In the simulations, we used (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.

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

An important difference with the simulation-based MMA and JMA confidence intervals proposed by ZL is that they simulate from the limiting distribution of the model averaging estimator, while in simulation-based WALS 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), and these properties allow us to study the sampling distribution of the bias-corrected WALS estimator by simulations.

The $R \times k$ matrix $\breve{B}^*$ 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 used $\breve{B}^*$ to compute the standard error of $\breve{\beta}_l$ required in the centered-and-naive WALS confidence intervals, and the complete variance matrix of $\breve{\beta}$ required to implement the naive approach to prediction intervals. We also used $\breve{B}^*$ to obtain the skewness and kurtosis which we need to investigate deviations from normality.

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 coefficients 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

Andrews, D. W. K. (1991). Asymptotic optimality of generalized C_L, cross-validation, and generalized cross-validation in regression with heteroskedastic errors. *Journal of Econometrics*, 47, 359–377.

Ankargren, S., and Jin, S. (2018). On the least squares model averaging interval estimator. *Communications in Statistics—Theory and Methods*, 47: 118–132.

Buckland, S., Burnham, K., and Augustin, N. (1997). Model selection: An integral part of inference. *Biometrics*, 53: 603–618.

Burham, K. P., and Anderson, D. R. (2002). *Model Selection and Inference: A Practical Information-Theoretic Approach* (2nd ed.). Spinger, New York, USA.

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

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

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

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

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

De Luca, G., Magnus, J. R., and Peracchi, F. (2020). Posterior moments and quantiles for the normal location model with Laplace prior. *Communications in Statistics—Theory and Methods*, doi: 10.1080/03610926.2019.1710756.

De Luca, G., Magnus, J. R., and Peracchi, F. (2021). Sampling properties of the Bayesian posterior mean with an application to WALS estimation. *Journal of Econometrics*, forthcoming.

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

Hansen, B. E. (1994). Autoregressive conditional density estimation. *International Economic Review*, 35: 705–730.

Hansen, B. E. (2005). Challenges for econometric model selection. *Econometric Theory*, 21: 60–68.

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., and Racine, J. (2012). Jackknife model averaging. *Journal of Econometrics*, 167: 38–46.

Hjort, N. L., and Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association*, 78: 879–899.

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

Kumar, K., and Magnus, J. R. (2013). A characterization of Bayesian robustness for a normal location parameter. *Sankhya (Series B)*, 75: 216–237.

Li, K.-C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set. *The Annals of Statistics*, 15: 958–975.

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

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

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

Magnus, J. R., Wang, W., and Zhang, X. (2016). Weighted-average least squares prediction. *Econometric Reviews*, 35: 1040–1074.

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

Theodossiou, P. (1998). Financial data and the skewed generalized t distribution. *Management Science*, 44: 1650–1661.

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

Wang, H., and 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., and 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.

Figure 1: Squared bias and variance of the estimators of the focus coefficient β_{12} in the simulation designs with $k_2 = 8$, $n = 100$, and homoskedastic errors

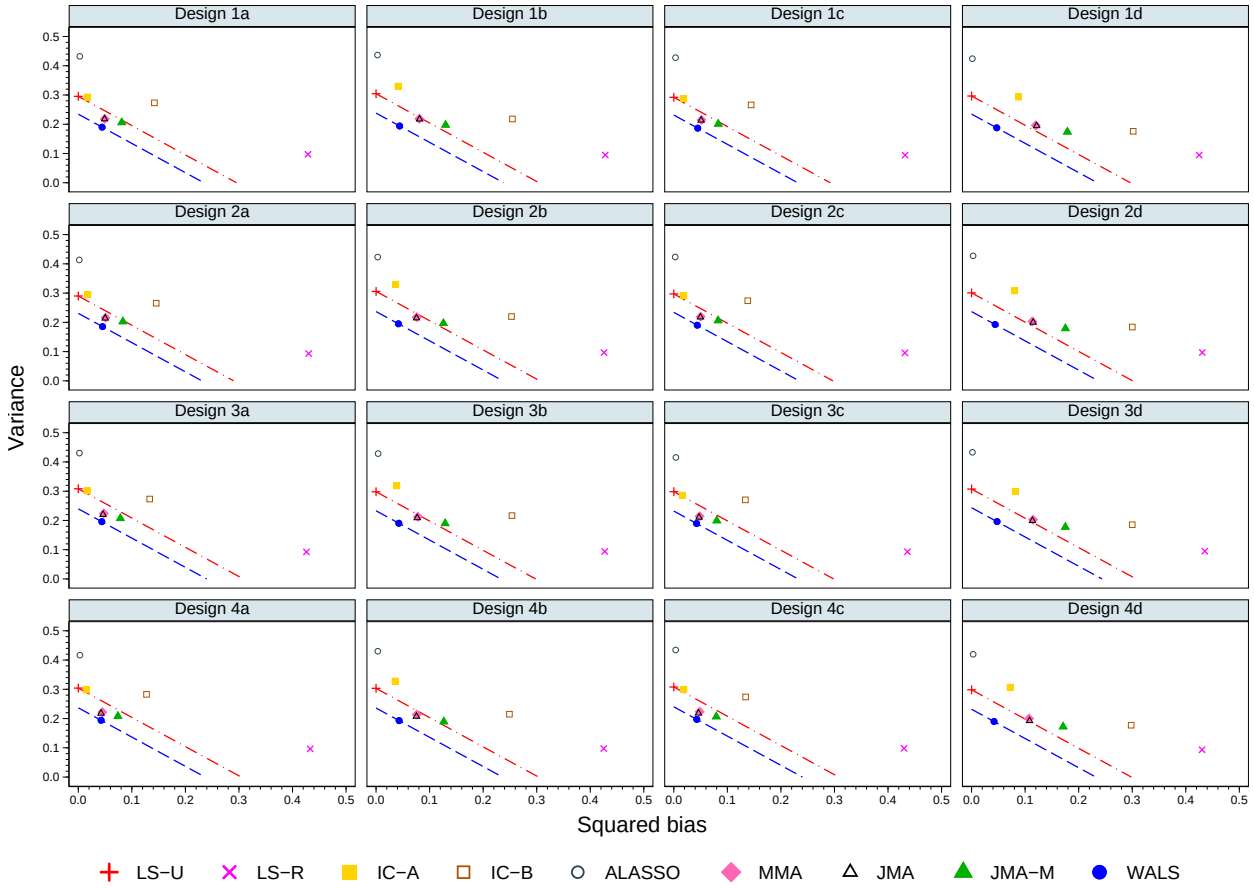

Figure 2: Squared bias and variance of the estimators of the focus coefficient β_{12} in the simulation designs with $k_2 = 8$, $n = 100$, and heteroskedastic errors

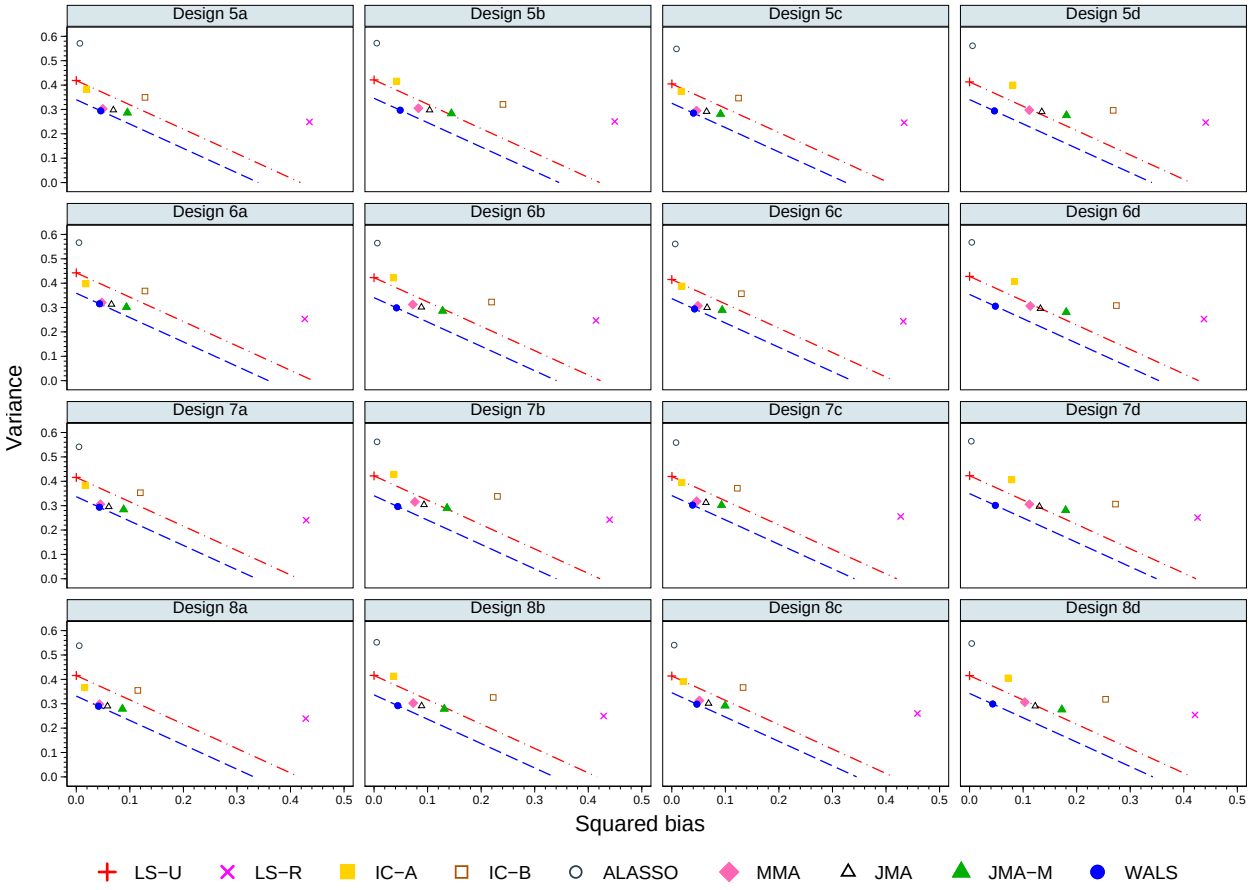

Figure 3: Squared bias and variance of the estimators of the focus coefficient β_{12} in the simulation designs with $k_2 = 8$, $n = 400$, and normal errors

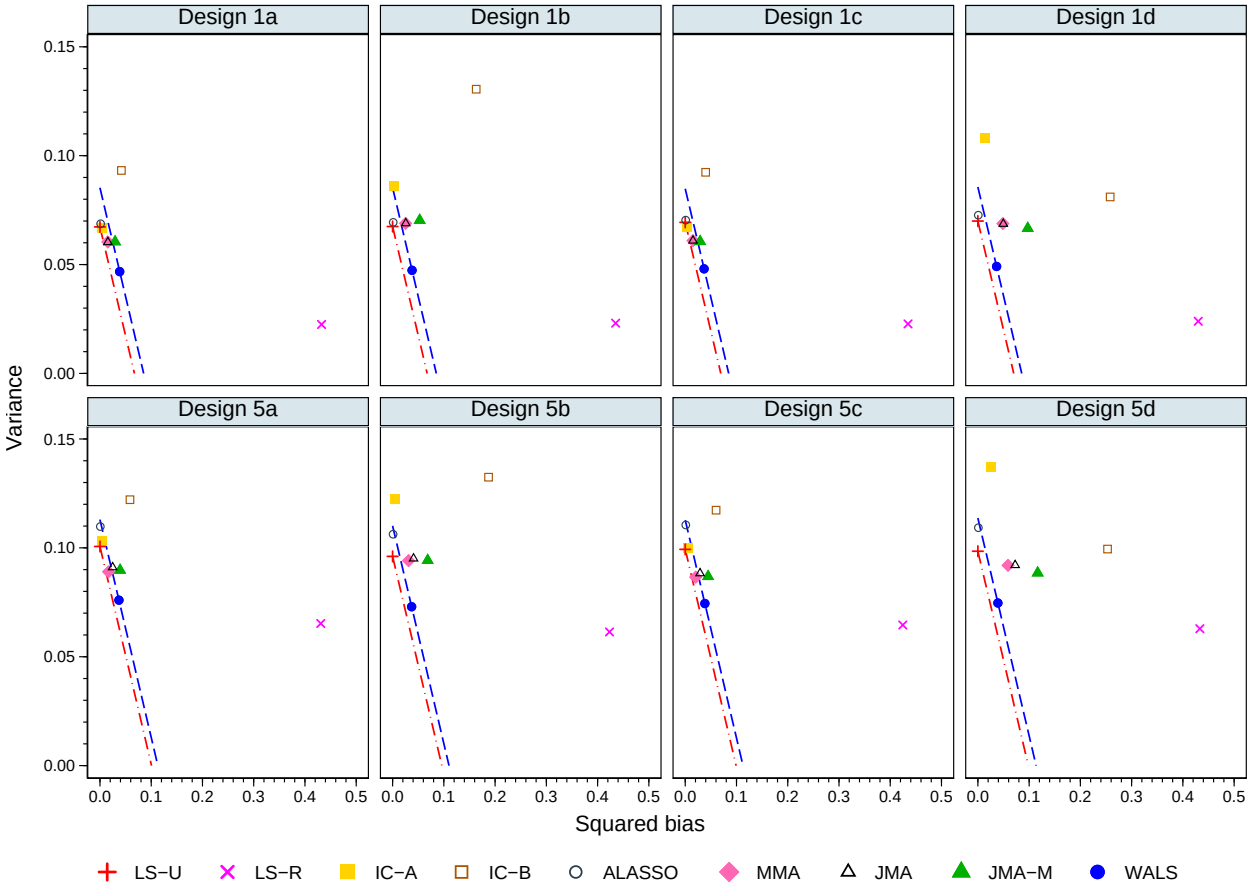

Figure 4: Efficiency of WALS relative to LS-U of the estimator of β_{12} in the simulation designs with homoskedastic normal errors under alternative values of n, k_2, ξ, and ρ

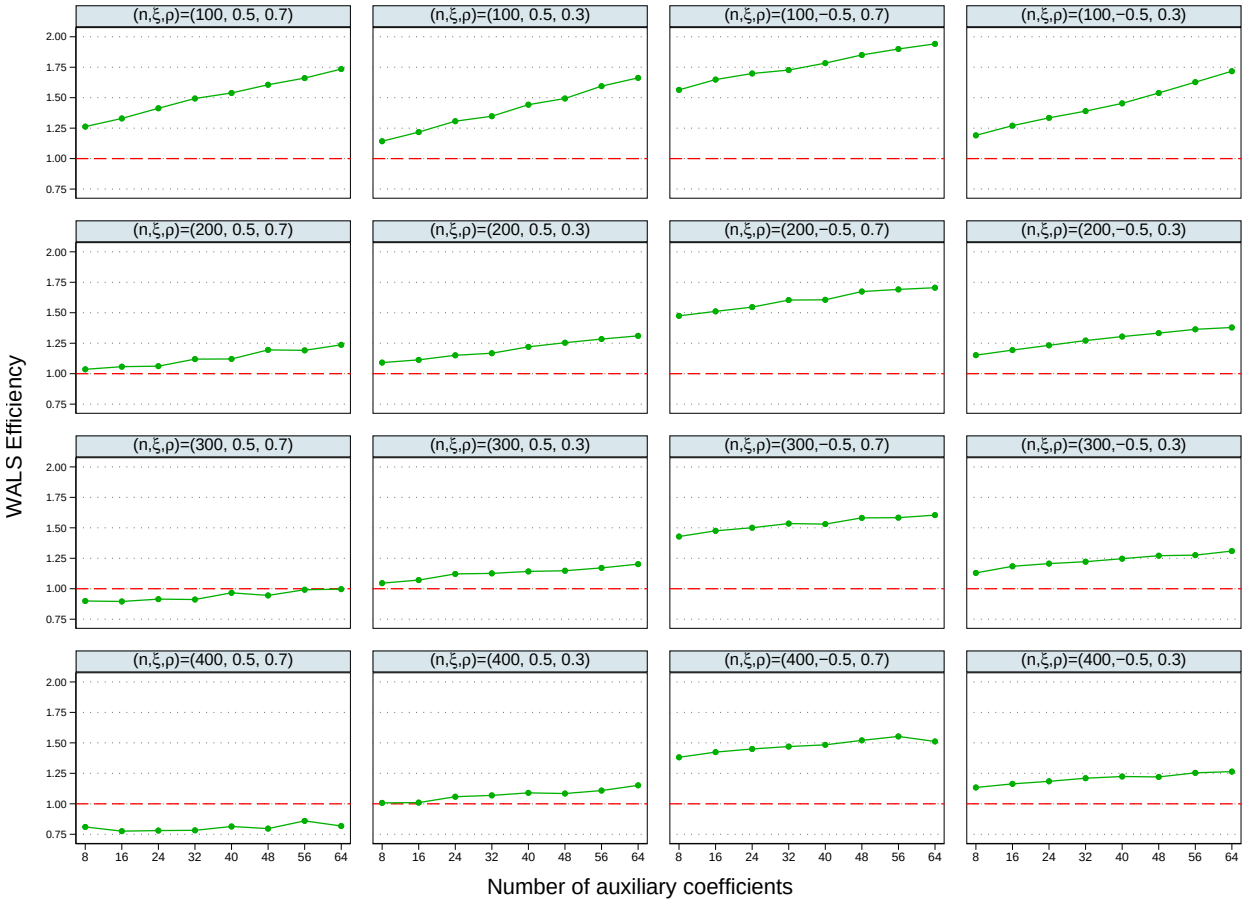

Figure 5: Coverage probability and length of the confidence intervals for the focus coefficient β_{12} in the simulation designs with $k_2 = 8$ and $n = 100$

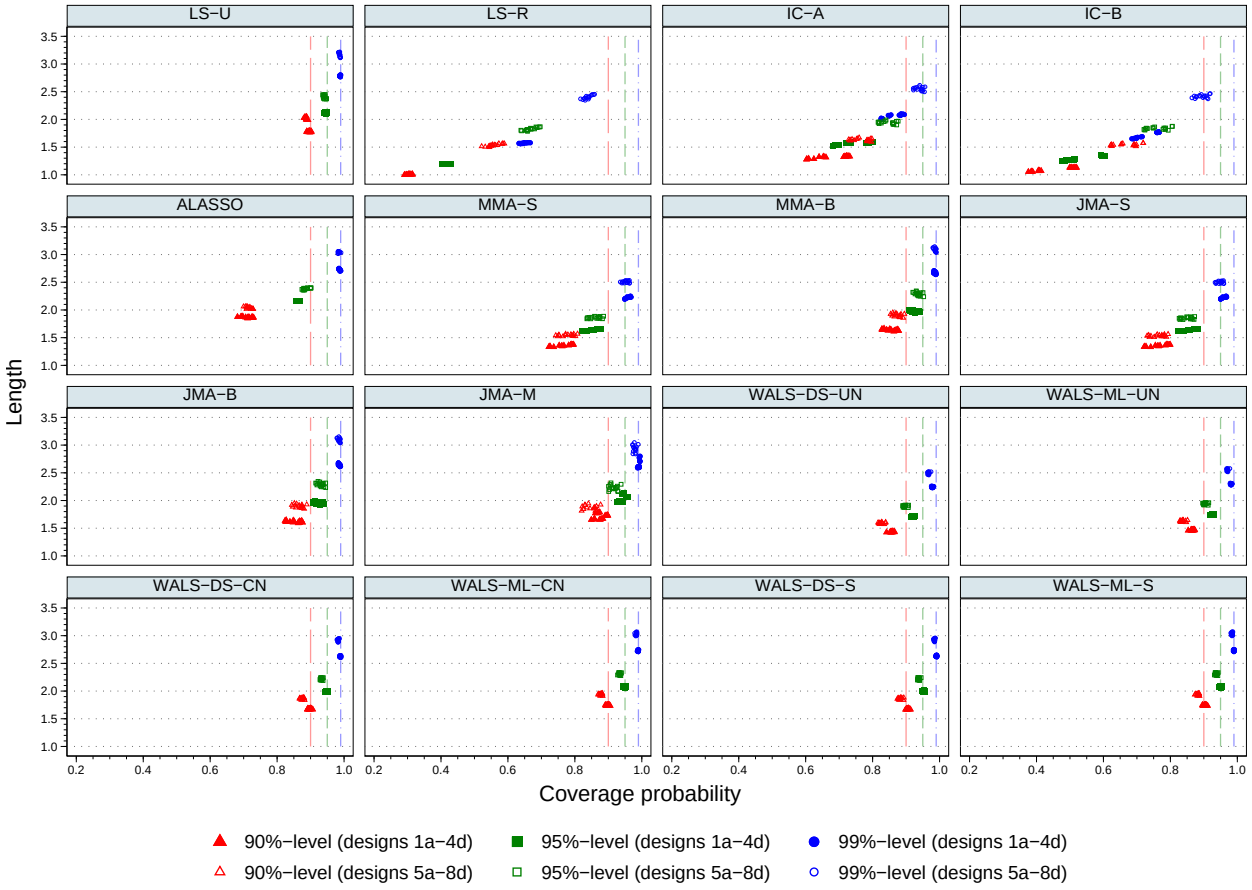

Figure 6: Coverage probability and length of the confidence intervals for the focus coefficient β_{12} in the simulation designs with $k_2 = 8$ and $n = 400$

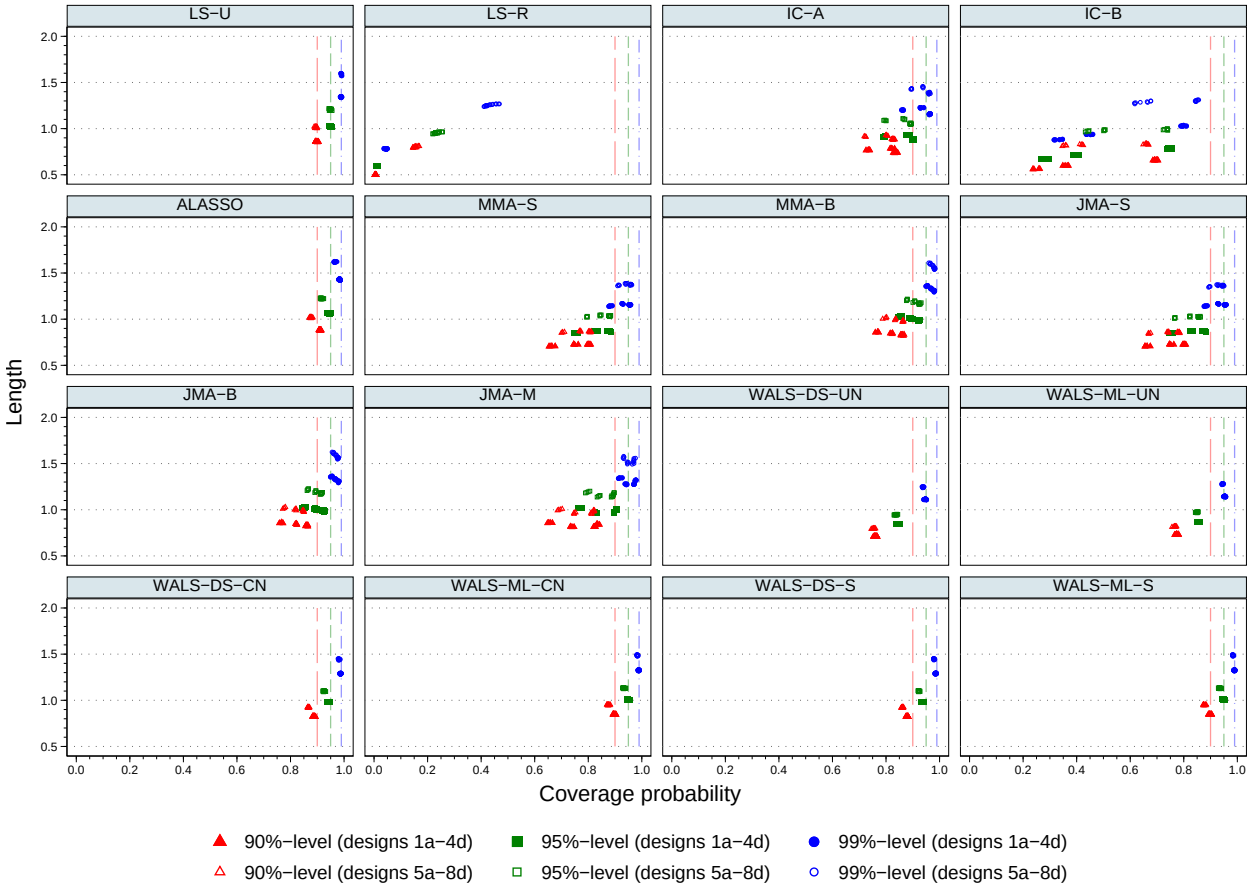

Figure 7: Coverage probabilities of confidence interval of β_{12} in the simulation designs with homoskedastic normal errors and alternative values of n, k_2, ξ, and ρ

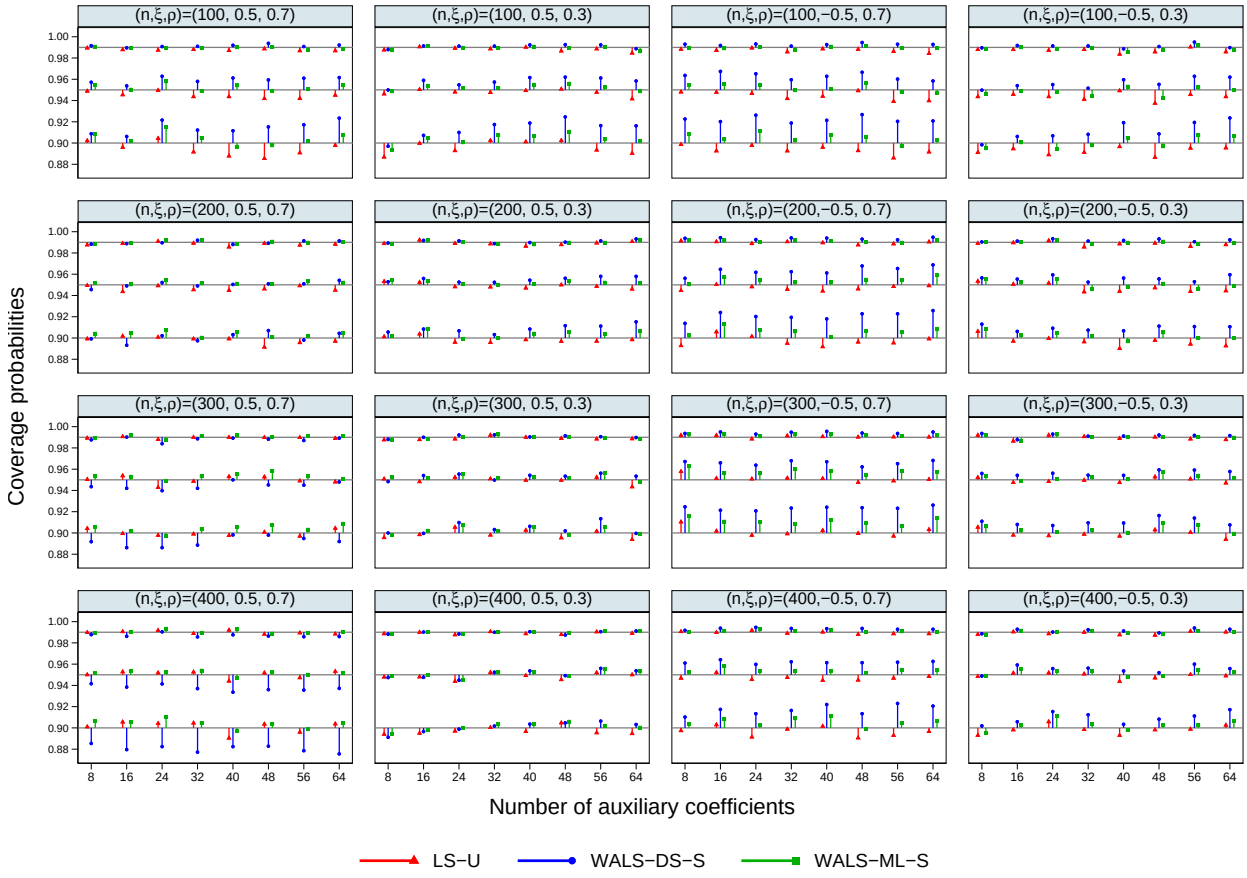

Figure 8: Relative lengths of the 95% confidence interval of β_{12} in the simulation designs with homoskedastic normal errors and alternative values of n, k_2, ξ, and ρ

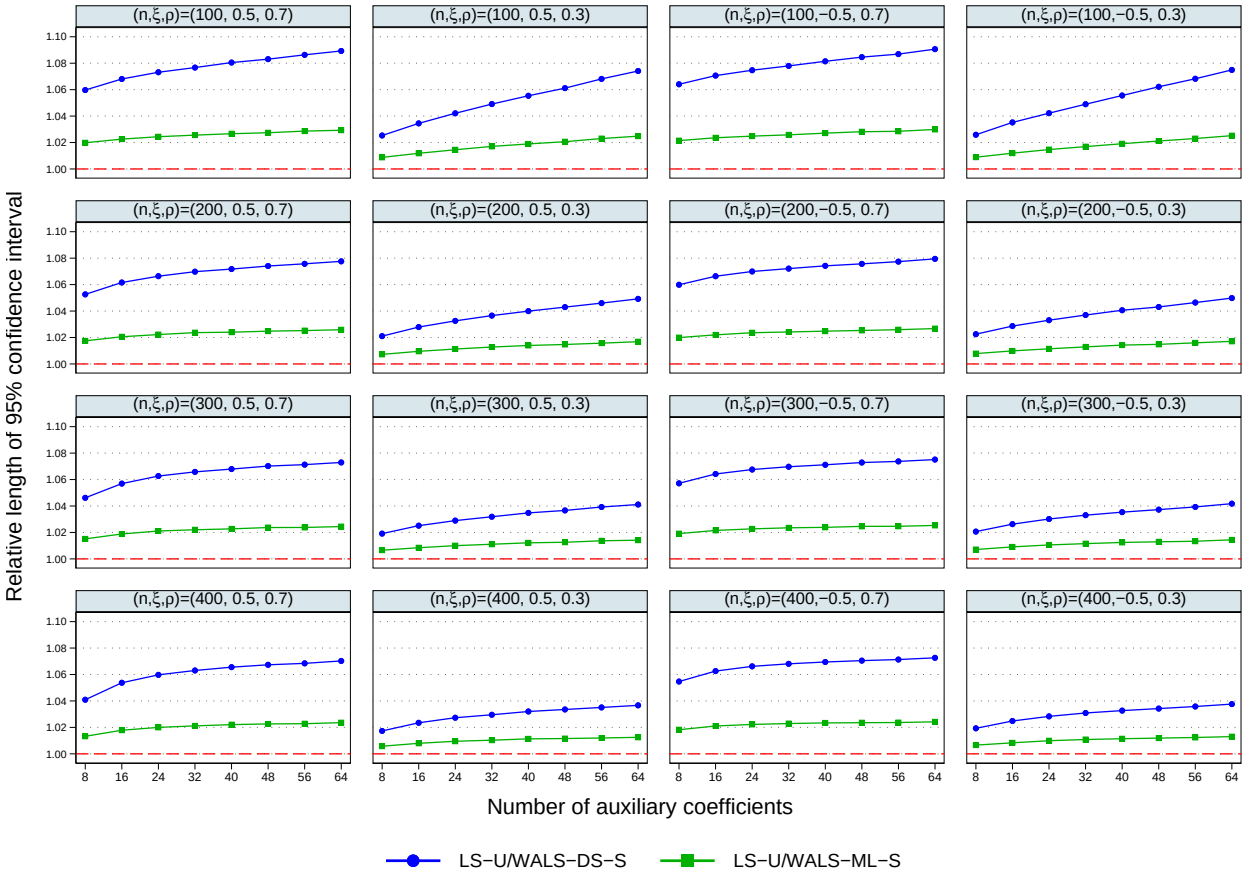

Figure 9: Efficiency of the WALS predictor of $\mathbb{E}(y_f)$ relative to the LS-U predictor in the simulation designs with homoskedastic normal errors under alternative values of n, k_2, ξ, and ρ

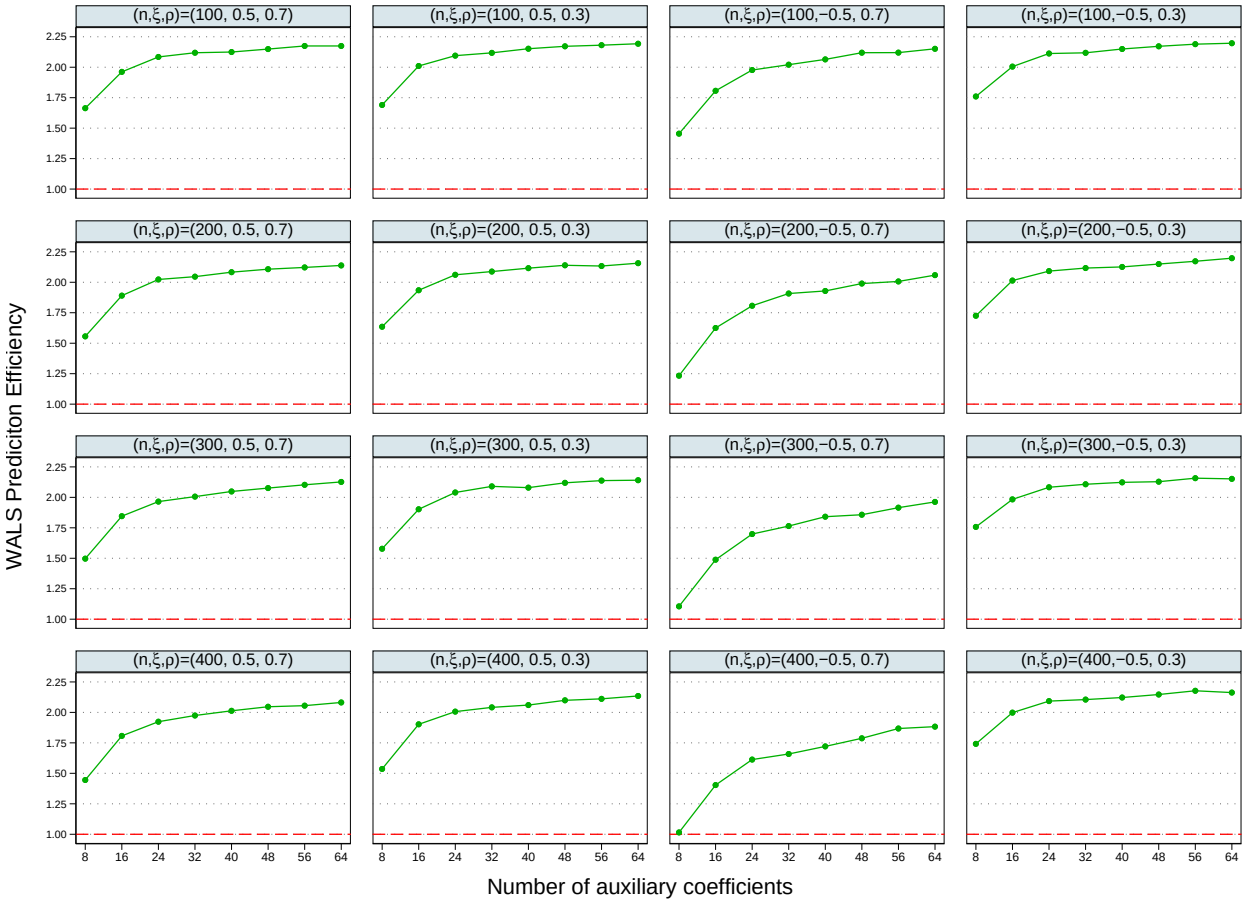

Figure 10: Coverage probabilities of prediction interval of $\mathbb{E}(y_f)$ in the simulation designs with homoskedastic normal errors and alternative values of n, k_2, ξ, and ρ

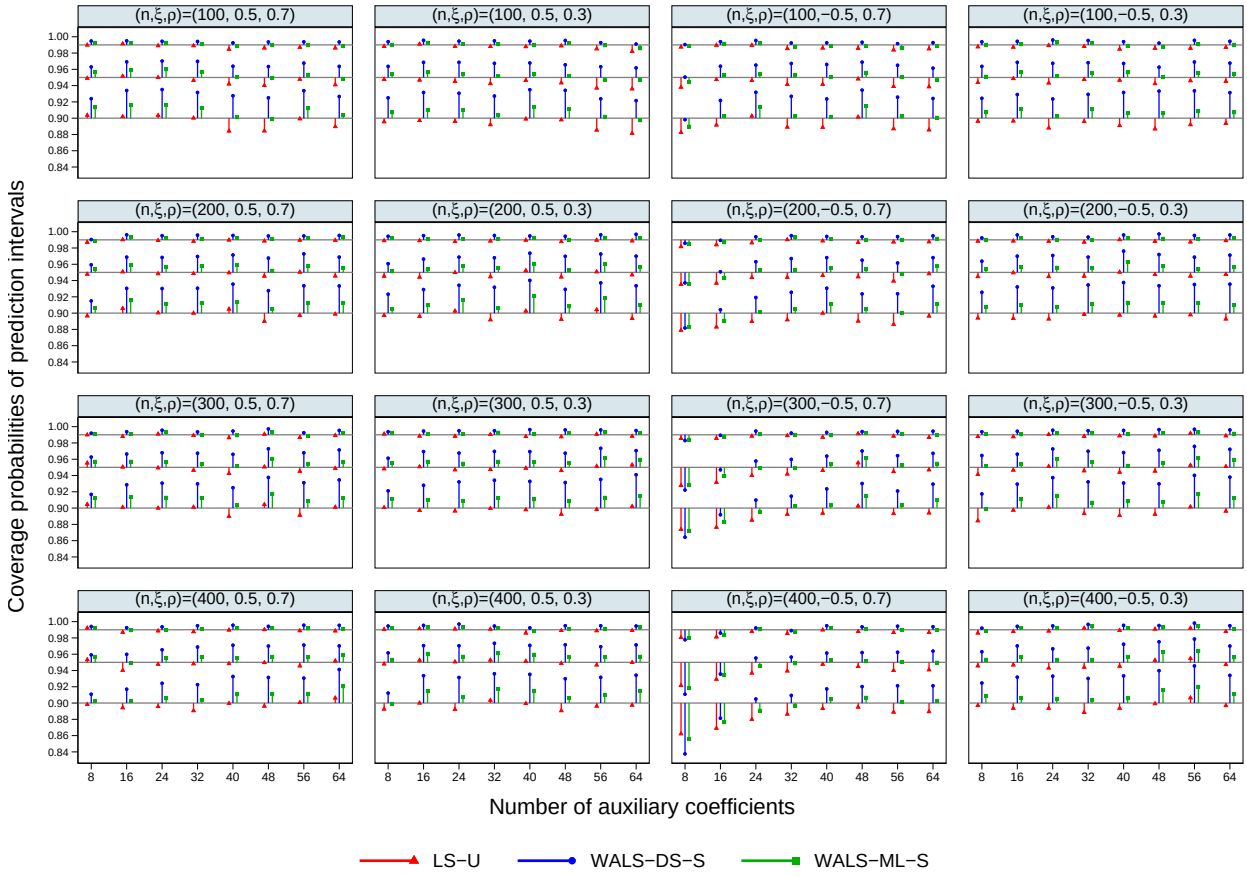

Figure 11: Relative lengths of the 95% prediction interval of $\mathbb{E}(y_f)$ in the simulation designs with homoskedastic normal errors and alternative values of n, k_2, ξ, and ρ