TI 2022-022/III
Tinbergen Institute Discussion Paper

Asymptotic properties of the weighted-average least squares (WALS) estimator

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

1 University of Palermo

2 Vrije Universiteit Amsterdam, Tinbergen Institute

3 University of Rome Tor Vergata, EIEF

Asymptotic properties of the weighted-average least squares (WALS) estimator*

Giuseppe De Luca
University of Palermo, Italy

Jan R. Magnus
Vrije Universiteit Amsterdam and Tinbergen Institute,
The Netherlands

Franco Peracchi
University of Rome Tor Vergata and EIEF, Italy

March 2, 2022

*Corresponding author: Franco Peracchi (franco.peracchi@uniroma2.it). We are grateful to Aad van der Vaart for helpful comments.

Abstract: We investigate the asymptotic behavior of the WALS estimator, a model-averaging estimator with attractive finite-sample and computational properties. WALS is closely related to the normal location model, and hence much of the paper concerns the asymptotic behavior of the estimator of the unknown mean in the normal local model. Since we adopt a frequentist-Bayesian approach, this specializes to the asymptotic behavior of the posterior mean as a frequentist estimator of the normal location parameter. We emphasize two challenging issues. First, our definition of ignorance in the Bayesian step involves a prior on the t-ratio rather than on the parameter itself. Second, instead of assuming a local misspecification framework, we consider a standard asymptotic setup with fixed parameters. We show that, under suitable conditions on the prior, the WALS estimator is $\sqrt{n}$-consistent and its asymptotic distribution essentially coincides with that of the unrestricted least-squares estimator. Monte Carlo simulations confirm our theoretical results.

Keywords: Model averaging, normal location model, consistency, asymptotic normality, WALS

JEL classification: C11, C13, C51, C52

1 Introduction

Applied econometricians typically perform model selection and estimation sequentially or iteratively, not jointly. Given the data, the econometrician selects a model and then, within the chosen model, estimates the parameters of interest and carries out inference ignoring the data-driven model selection step. This is called 'pretesting', and it is not good (see, e.g., Pötscher 1991 and Leeb and Pötscher 2005).

There are various approaches to solving this problem. One is regularization via an ℓ_1 penalty, such as the LASSO (Tibshirani 1996), resulting in both shrinkage and selection. This approach tends to work well when there is a large number of potential regressors, possibly larger than the sample size, but the data-generation process (DGP) is 'sparse', i.e. most of its parameters are zero and only a few of them are large in magnitude. Another approach is regularization via an ℓ_2 penalty, such as ridge regression (Hoerl and Kennard 1970), resulting in shrinkage but no selection. This approach tends to work well when the DGP is 'dense', i.e. most of its parameters are nonzero but small in magnitude. The distinction between 'sparse' and 'dense' modeling is not always easy, partly because sparsity is not invariant to transformations of the regressors (Giannone et al. 2021). One may also consider hybrid cases, such as the 'sparse-plus-dense' representation analyzed by Chernozhukov et al. (2021).

Another approach is 'model averaging', which can be seen as the continuous version of pretesting, where the weights given to the various models are not zero or one, but rather continuous functions of appropriate diagnostics, such as t-ratios. In model averaging one does not select a single 'best' performing model out of the available set of models, but combines the information from all models to improve inference, e.g. predictions of the outcome of interest or estimation of the structural or 'focus' parameters in the model.

There are various model averaging methods, both Bayesian and frequentist. In Bayesian model averaging (see, e.g., Steel 2020), model weights are typically based on posterior model probabilities. In frequentist model averaging (see, e.g., Claeskens and Hjort 2008), the weighting scheme is typically

based on some optimality criterion. In either case, the key difficulty is how to handle the complexity of the model space. For example, if there is uncertainty about which of k regressors to include and there is no natural way of ordering them, then the model space contains 2^k submodels. Even with values of k well below the sample sizes now standard in economic applications, e.g. with $k = 30$ or 40, the model space becomes exceedingly large.

In the current paper we concentrate on the weighted-average least squares (WALS) estimator developed by Magnus et al. (2010), a frequentist model averaging method with an important Bayesian flavor. From its introduction in 2010, the WALS approach has been studied, extended, and applied in a number of papers; see inter alia Dardanoni et al. (2011), Amini and Parmeter (2012), Magnus and Wang (2014), Magnus and De Luca (2016), Magnus et al. (2016), De Luca et al. (2018, 2021a, 2021b), Duval et al. (2021), and Magkonis et al. (2021). WALS is attractive because it performs well in finite samples, offers a close-to-practice notion of prior ignorance (called 'neutrality'), and is not restricted to sequences of nested models. Equally important, it is numerically stable and fast to compute because it employs a preliminary transformation of the regressors which reduces the complexity of the model space from 2^k to k.

What is missing so far is a suitable asymptotic theory for WALS, and our purpose in the current paper is to provide such a theory. We emphasize two aspects of this asymptotic theory. First, we place our prior not on the parameter of interest (as is common in Bayesian analysis) but on its t-ratio. This is justified by the desire of using a proper notion of prior ignorance which closely resembles the usual frequentist model selection methods, say general-to-specific or specific-to-general, but it implies that the prior on the parameter of interest now depends on the sample size n and this complicates the analysis.

Second, instead of assuming a local misspecification framework (Hjort and Claeskens 2003, Zhang and Liang 2011, Hansen 2014, De Luca et al. 2018), we consider a standard asymptotic setup with fixed parameters. The local misspecification framework is convenient because it assumes that the auxiliary (nuisance) regression parameters shrink to zero with the sample size

n at the rate $n^{-1/2}$, giving rise to a well-defined trade-off between asymptotic bias and variance, but it is unrealistic (Ishwaran and Rao 2003, Raftery and Zheng 2003) because we expect to converge to the unrestricted model, not to the restricted model. Our asymptotic theory for WALS is developed in an $\mathcal{M}$-closed environment where the unknown DGP is included in the model space considered by the investigator. For simplicity, we also assume that the model space does not expand with n, although this additional assumption can easily be relaxed.

In line with the more recent literature on the asymptotic properties of other frequentist model averaging estimators (Zhang and Liu 2019, Zhu et al. 2019, Zhang et al. 2020), we show that, under suitable conditions on the prior for the t-ratio, the WALS estimator is $\sqrt{n}$-consistent and its asymptotic distribution essentially coincides with that of the unrestricted least-squares estimator.

We begin by summarizing the WALS approach to linear regression in Section 2. From this we see that an essential element in the theory is the 'normal location model', and that the asymptotic theory developed in this model carries over, more or less straightforwardly, to the WALS estimator. In Section 3 we distinguish between two Bayesian approaches to the normal location problem: one places a prior on the unknown location parameter and leads to Bayesian model averaging, the other places a prior on its 'theoretical' t-ratio and leads to WALS. In Section 4, we discuss two properties of the prior on the t-ratio: robustness and neutrality. The asymptotic behavior of our Bayesian estimator of the normal location parameter is studied in Sections 5 and 6. In Section 7 we return to WALS and show how the asymptotic theory of the normal location model carries over to the original regression model. Some Monte Carlo simulations are presented in Section 8, where we investigate the speed of convergence in relation to the choice of prior and the large-sample performance of the bias-correction strategy proposed in De Luca et al. (2021a). Section 9 concludes. All proofs are in the Appendix.

2 WALS and the normal location model

We assume a homoskedastic linear model, which we write as

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

where $X = (X_1 : X_2)$ is an $n \times k$ matrix of full column-rank $k = k_1 + k_2 < n$ and ϵ is an $n \times 1$ vector of independent and identically distributed disturbances which we assume to be distributed as $\mathrm{N}(0, \sigma^2 I_n)$, where I_n denotes the identity matrix of order n. The k_1 components of β_1 are the 'focus' parameters (the parameters of interest) and the k_2 components of β_2 are the 'auxiliary' (nuisance) parameters. The normality assumption plays a role in finite samples, but in this paper we are interested in asymptotics and the underlying normality is assumed for convenience only.

As in De Luca et al. (2018), we first implement the following one-to-one transformations of the matrix X_2 of auxiliary regressors and the vector β_2 of auxiliary parameters:

$$Z_2 = X_2\Delta_2\Psi^{-1/2}, \qquad \gamma_2 = \Psi^{1/2}\Delta_2^{-1}\beta_2, \tag{2}$$

where Δ_2 is a diagonal $k_2 \times k_2$ matrix such that the diagonal elements of the positive definite matrix $\Psi = \Delta_2 X_2' M_1 X_2 \Delta_2/n$ are all equal to one and $M_1 = I_n - X_1(X_1'X_1)^{-1}X_1'$.

Next, we rescale the matrix X_1 of focus regressors and the vector β_1 of focus parameters:

$$Z_1 = X_1\Delta_1, \qquad \gamma_1 = \Delta_1^{-1}\beta_1, \tag{3}$$

where Δ_1 is a diagonal $k_1 \times k_1$ matrix such that the diagonal elements of $Z_1'Z_1/n$ are all equal to one. Since $Z_1\gamma_1 = X_1\beta_1$ and $Z_2\gamma_2 = X_2\beta_2$, we can now write model (1) as

$$y = Z_1\gamma_1 + Z_2\gamma_2 + \epsilon, \tag{4}$$

where $Z_2'M_1Z_2/n = I_{k_2}$. The transformations in (2) ensure that the k_2 components of the least-squares (LS) estimator of γ_2 in model (4) are independent, while the rescaling in (3) serves only to increase the numerical accuracy of the inversion and eigenvalue routines.

The WALS estimator is obtained by averaging the LS estimators $\hat{\gamma}_{1j}$ and $\hat{\gamma}_{2j}$ of γ_1 and γ_2 over the J models in the model space:

$$\tilde{\gamma}_1 = \sum_{j=1}^{J} \lambda_j \hat{\gamma}_{1j}, \qquad \tilde{\gamma}_2 = \sum_{j=1}^{J} \lambda_j \hat{\gamma}_{2j}, \tag{5}$$

where the λ_j are nonnegative data-dependent model weights that add up to one.

Unlike some other model-averaging estimators, WALS is not restricted to a sequence of nested models. Hence, the model space consists of the $J = 2^{k_2}$ models that contain all focus regressors and a subset of the auxiliary regressors in (4). Using $Z_2' M_1 Z_2 / n = I_{k_2}$, we obtain

$$\hat{\gamma}_{1j} = \hat{\gamma}_{1r} - Q W_j \hat{\gamma}_{2u}, \qquad \hat{\gamma}_{2j} = W_j \hat{\gamma}_{2u} \qquad (j = 1, \ldots, J), \tag{6}$$

where $\hat{\gamma}_{1r} = (Z_1' Z_1)^{-1} Z_1' y$ is the LS estimator of γ_1 in the fully restricted model (with $\gamma_2 = 0$), $\hat{\gamma}_{2u} = Z_2' M_1 y / n$ is the LS estimator of γ_2 in the unrestricted model, $Q = (Z_1' Z_1)^{-1} Z_1' Z_2$, $W_j = I_{k_2} - R_j R_j'$ is a diagonal matrix whose elements are equal to zero or one, and R_j is a $k_2 \times r_j$ selection matrix of rank $0 \leqslant r_j \leqslant k_2$ representing the r_j exclusion restrictions implied by model j, that is, $R_j' = [I_{r_j} : 0]$ or a column-permutation thereof. The WALS estimator can then be written as

$$\tilde{\gamma}_1 = \hat{\gamma}_{1r} - Q W \hat{\gamma}_{2u}, \qquad \tilde{\gamma}_2 = W \hat{\gamma}_{2u}, \tag{7}$$

where

$$W = \sum_{j=1}^{J} \lambda_j W_j = \begin{pmatrix} w_1 & 0 & \ldots & 0 \\ 0 & w_2 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & w_{k_2} \end{pmatrix} \tag{8}$$

is a random diagonal matrix whose k_2 diagonal elements w_h (the 'WALS weights') are partial sums of the model weights, and

$$\hat{\gamma}_{2u} = \gamma_2 + \frac{Z_2' M_1 \epsilon}{n} \sim \mathrm{N}\left(\gamma_2, \frac{\sigma^2}{n} I_{k_2}\right), \tag{9}$$

using model (4) and the normality of ϵ. If we relax the normality of ϵ, then (9) still holds asymptotically under the standard conditions of the central limit theorem.

Since the WALS weights w_h are bounded between zero and one, the components of the WALS estimator $\tilde{\gamma}_2$ are shrinkage estimators of the components of γ_2. We want to find a set of WALS weights such that the WALS estimator has smallest mean squared error (MSE) matrix. The 'equivalence theorem' of Magnus and Durbin (1999) implies that the MSE matrix of the WALS estimator of γ_1 depends on the MSE matrix of the WALS estimator of γ_2. Thus, it is enough to focus on the WALS estimator $\tilde{\gamma}_2$ of γ_2.

We know from (7) that $\tilde{\gamma}_2 = W\hat{\gamma}_{2u}$ and that the components $(\hat{\gamma}_{2u})_h$ of $\hat{\gamma}_{2u}$ are distributed independently as

$$(\hat{\gamma}_{2u})_h \sim \mathrm{N}\left(\gamma_{2h}, \frac{\sigma^2}{n}\right), \tag{10}$$

where γ_{2h} is the hth component of γ_2. Equivalently, we can write

$$\frac{(\hat{\gamma}_{2u})_h}{\sigma/\sqrt{n}} \sim \mathrm{N}\left(\frac{\gamma_{2h}}{\sigma/\sqrt{n}}, 1\right), \tag{11}$$

where (10) emphasizes the parameter estimate, while (11) emphasizes the associated t-ratio (assuming that σ is known).

Under the additional restriction that the hth WALS weight w_h only depends on the hth component of $\hat{\gamma}_{2u}$, the individual components of the WALS estimator $\tilde{\gamma}_2$ are independent and so the k_2-dimensional problem of finding an estimator of γ_2 with smallest MSE matrix reduces to k_2 identical one-dimensional problems of the following type: given data $x_n \sim \mathrm{N}(\theta, \sigma^2/n)$ and a shrinkage estimator $m(x_n) = w(\sqrt{n}\, x_n)\, x_n$ of the scalar location parameter θ, find the shrinkage function $w(\cdot)$ such that $m(x_n)$ has minimum MSE. This stylized setting is known as the *normal location problem*. For theoretical considerations on admissibility, bounded risk, robustness, near-optimality in terms of minimax regret, and ignorance about θ, we adopt a Bayesian approach to the normal location problem.

Many of the previous WALS studies have concentrated on the choice of a suitable prior for this Bayesian step (Kumar and Magnus 2013, Magnus and

De Luca 2016) and the finite-sample properties of the posterior mean of θ given x_n as a frequentist estimator of θ (De Luca et al. 2021a). In the current paper, we study the large sample properties of the posterior mean of θ given x_n, viewed as a frequentist estimator of θ, and the development of a valid asymptotic theory for WALS. Unlike the asymptotic analysis carried out by De Luca et al. (2018) for WALS estimation of generalized linear models, we derive the asymptotic properties of the WALS estimator without assuming that the auxiliary parameters are in a shrinking neighborhood of zero (the local misspecification framework).

3 The normal location model: two Bayesian approaches

In the normal location model we aim to estimate a finite location parameter θ from one observation x_n, distributed as

$$x_n|\theta \sim \mathrm{N}\left(\theta, \frac{\sigma^2}{n}\right), \tag{12}$$

where σ is a finite and strictly positive scale parameter which we assume (at first) to be known. The obvious estimator of θ is x_n itself, which is unbiased and consistent, and is sometimes called the 'usual' estimator. Another estimator, sometimes called the 'silly' estimator, is 0 (zero). Since $\mathrm{MSE}(x_n) = \sigma^2/n$ and $\mathrm{MSE}(0) = \theta^2$ we prefer the silly estimator (in the MSE sense) if and only if $|\theta| < \sigma/\sqrt{n}$.

Defining

$$x_n^* = \frac{x_n}{\sigma/\sqrt{n}}, \qquad \theta_n^* = \frac{\theta}{\sigma/\sqrt{n}}, \tag{13}$$

we can write (12) equivalently as

$$x_n^*|\theta_n^* \sim \mathrm{N}(\theta_n^*, 1). \tag{14}$$

Equations (12) and (14) are the stylized versions of Equations (10) and (11) from the previous section.

The transformed random variable x_n^* is the t-ratio (when σ is known) for testing the hypothesis $\theta = 0$, while θ_n^* may be called the 'theoretical t-ratio'. There is no difference, at least no essential difference, between the approach via (12) and the approach via (14). If, however, we add a prior, then it *does* make a difference whether we place the prior on the parameter θ or on the theoretical t-ratio θ_n^*.

The standard Bayesian approach places a prior on θ and this leads to Bayesian model averaging (see e.g. Steel 2020). The prior does not depend on the sample size and the posterior mean of θ given x_n is a consistent estimator of θ (van der Vaart 1998, Chapter 10) because, as the sample size increases, the data information will become increasingly important and will dominate the prior information which remains constant.

Since, as we have seen, $\text{MSE}(0) < \text{MSE}(x_n)$ if and only if $|\theta_n^*| < 1$ and, more generally, since model selection and model averaging typically depend on diagnostics (such as t-ratios) rather than on parameter estimates, it makes sense to place a prior on θ_n^* rather than on θ, and this is indeed what we shall do. This approach, which plays a key role in the Bayesian shrinkage step of the WALS procedure, though intuitive, is not standard.[1]

Our approach does not guarantee that the posterior mean of θ is consistent for θ. Consider, for example, a normal prior $\theta_n^* \sim \text{N}(0, \tau^{*2})$, where τ^* is a finite scale parameter which does not depend on n. Combining the prior with the likelihood in (14), gives the posterior distribution

$$\theta_n^* | x_n^* \sim \text{N}(\lambda^* x_n^*, \lambda^*), \qquad \lambda^* = \frac{\tau^{*2}}{\tau^{*2} + 1}. \tag{15}$$

The posterior mean $\mathbb{E}(\theta_n^* | x_n^*) = \lambda^* x_n^*$ implies the posterior mean $m_n = \mathbb{E}(\theta | x_n) = \lambda^* x_n$. Viewing m_n as a (frequentist) estimator of θ, its sampling bias and variance are

$$\mathbb{E}(m_n) - \theta = (\lambda^* - 1)\theta, \qquad \mathbb{V}(m_n) = \lambda^{*2} \sigma^2 / n, \tag{16}$$

so that m_n is in general not consistent for θ because, although the variance vanishes as $n \to \infty$, the bias doesn't. Consistency occurs only if $\theta = 0$ or if

[1] A similar idea in a different context was advocated by Hjort (1986).

we are willing to assume 'local misspecification':

$$\theta = \delta/\sqrt{n} \qquad (\delta \neq 0). \tag{17}$$

Neither assumption is satisfactory because they imply, either exactly or asymptotically, that the data are generated by the fully restricted model rather than by the unrestricted model. In other words, in the local misspecification framework the model space shrinks rather than expands when the sample size increases.

Given that m_n is in general not consistent for θ, what do we need to ensure the asymptotic validity of the WALS procedure and how can we justify the idea of putting a prior on the t-ratio rather than on the parameter itself? To answer these questions, we begin with the likelihood (14) and a prior density π^* for the theoretical t-ratio θ_n^*. Together they give the posterior density of $\theta_n^*|x_n^*$ and, in particular, the posterior mean

$$m_n^* = m(x_n^*) = \mathbb{E}(\theta_n^*|x_n^*). \tag{18}$$

Our interest is in the asymptotic properties of m_n^* as a frequentist estimator of θ_n^* and of $m_n = (\sigma/\sqrt{n})m_n^*$ as a frequentist estimator of θ. The sampling properties (bias and variance) of m_n^* in finite samples have recently been studied by De Luca et al. (2021a).

As in Kumar and Magnus (2013), we impose the following conditions on the prior density π^*:

(C1) π^* is symmetric around zero;
(C2) π^* is positive and non-increasing on $(0, \infty)$; and
(C3) π^* is differentiable, except possibly at 0.

These are mild regularity conditions on the shape of the prior, allowing a non-differentiable peak at zero. Kumar and Magnus (2013) show that, under these three conditions, the posterior mean function m satisfies the following properties:

(P1) m is odd: $m(-x) = -m(x)$ and $m(0) = 0$;
(P2) m is strictly increasing: $m(x_1) < m(x_2)$ if $x_1 < x_2$;

(P3) m is a shrinkage rule: $0 < m(x) < x$ for $x > 0$; and
(P4) m is unbounded: $m(x) \to \infty$ as $x \to \infty$.

4 Robustness and neutrality

An important requirement for posterior inference is that, when the data information is sufficiently strong, the prior should have bounded influence on $m(x)$ (Sansó and Pericchi 1992). Although $m(x)$ is bounded from above by x, conditions (C1)–(C3) are not sufficient to characterize this additional property. To see this, let's introduce the discrepancy function

$$g(x) = x - m(x). \tag{19}$$

Under the normal prior $\theta_n^* \sim \mathrm{N}(0, \tau^{*2})$, we obtain the posterior (15), so that $m(x) = \lambda^* x$ and hence $g(x) = (1 - \lambda^*)x$, which is not bounded. Let

$$\omega^*(\theta) = -\frac{d \log \pi^*(\theta)}{d\theta} = -\frac{\pi^{*\prime}(\theta)}{\pi^*(\theta)}. \tag{20}$$

Then, under the normal prior, $\omega^*(\theta) = \theta/\tau^{*2}$, which does not converge to a finite constant as $\theta \to \infty$.

Thus motivated, let's impose two further conditions on the prior π^*.

(C4) $\omega^*(\theta) \to \omega_0^*$ as $\theta \to \infty$, where $\omega_0^* \geqslant 0$ is some finite constant; and
(C5) π^* satisfies $\int_0^1 \pi^*(\theta)\, d\theta = \int_1^\infty \pi^*(\theta)\, d\theta$.

Condition (C4) is relevant for the prior to have bounded influence on the posterior mean, while Condition (C5) is relevant for a notion of prior ignorance, which we call 'neutrality'.

Kumar and Magnus (2013, Theorem 4.1) showed that $g(x) \to 0$ if and only if $\omega_0^* = 0$. The next result is in the same spirit and shows in addition that the two functions have the same speed of convergence.

Proposition **1** *Under conditions (C1)–(C4),*

$$\lim_{x \to \infty} \frac{g(x)}{\omega^*(x)} = 1.$$

This proposition implies that, under (C1)–(C4), g is bounded and, in fact, that $g(x)$ converges to ω_0^* as $x \to \infty$, so that π^* has bounded influence on the posterior mean. It also implies a stronger property called (Bayesian) robustness, which requires the discrepancy between x and $m(x)$ to vanish as $x \to \infty$, so that prior information is essentially ignored when x is sufficiently large (Lindley 1968, Dawid 1973, Choy and Smith 1997). This follows from the fact that, when $\omega_0^* = 0$, $g(x)$ converges to 0 as $x \to \infty$.

Let us briefly discuss the concept of 'neutrality', which is important in WALS although it only plays a minor role in the asymptotic theory. In a Bayesian context one typically has to formalize the concept of prior ignorance. A flat (improper) prior is often used, as it can be computationally convenient. But a flat prior does not capture the idea of prior ignorance. In our context, ignorance means that we are ignorant whether or not θ_n^* is smaller than one in absolute value, that is, whether or not the restricted LS estimator has a lower MSE than the unrestricted LS estimator. Thus we say that π^* is 'neutral' when it is symmetric around zero and

$$\Pr(|\theta_n^*| < 1) = \frac{1}{2}. \tag{21}$$

Conditions (C1) and (C5) imply neutrality.

We can write (21) equivalently in terms of the original θ parameter:

$$\Pr\left(|\theta| < \frac{\sigma}{\sqrt{n}}\right) = \frac{1}{2}, \tag{22}$$

from which we see that the prior distribution of θ is asymptotically of the mixed discrete-continuous type with $\Pr(\theta = 0) = 1/2$ and $\Pr(\theta > 0) = \Pr(\theta < 0) = 1/4$. This is similar to the 'spike and slab' prior originally proposed by Mitchell and Beauchamp (1988), which is becoming increasingly popular in the application of Bayesian regularization methods (see, e.g., Abadie and Kasy 2019, Giannone et al. 2021).

A Bayesian interpretation of the local misspecification framework (17), where $\theta = \delta/\sqrt{n}$ for some $\delta \neq 0$, would correspond to a prior on θ of the form

$$\Pr\left(|\theta| < \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha_n, \tag{23}$$

where $\alpha_n \to 0$ as $n \to \infty$. In this case, the prior probability that the theoretical t-ratio is less than one in absolute value approaches one as $n \to \infty$, which is the opposite of what we want. After all, in an $\mathcal{M}$-closed environment with a fixed model space, more data lead to higher t-ratios, so that including auxiliary variables becomes more profitable. In the end we wish to converge to the unrestricted model, not to the fully restricted model as under local misspecification.

As in previous WALS studies we assume that our prior on θ_n^* belongs to the class of reflected generalized gamma distributions with density

$$\pi^*(\theta; a, b, c) = \frac{cb^d}{2\Gamma(d)}|\theta|^{-a}\exp\left(-b|\theta|^c\right) \qquad (-\infty < \theta < \infty), \tag{24}$$

where $0 \leqslant a < 1$, $b > 0$, $c > 0$, $d = (1-a)/c$, and $\Gamma(d)$ is the gamma function. This class of priors includes as special cases the one-parameter family of normal distributions ($a = 0$, $c = 2$, $d = 1/2$) with mean zero and variance $1/(2b)$, the one-parameter family of Laplace distributions ($a = 0$, $c = 1$, $d = 1$), and the two-parameter families of reflected Weibull ($a = 1-c$, $d = 1$) and Subbotin distributions ($a = 0$, $d = 1/c$). All these priors satisfy regularity conditions (C1)–(C3).

As shown in Kumar and Magnus (2013) and Magnus and De Luca (2016), a reflected generalized gamma prior is robust if and only if $0 < c < 1$, and is neutral if and only if

$$\Gamma(b, d) = \frac{1}{\Gamma(d)}\int_0^b t^{d-1}e^{-t}\,dt = 1/2, \tag{25}$$

where $\Gamma(b, d)$ is the (lower) incomplete gamma function. The Weibull and Subbotin priors are robust, but the Laplace prior is not (although it has bounded influence) since $\omega_0^* = b \neq 0$, and the normal prior has unbounded influence since $\omega_0^* \to \infty$.

Neutrality leads to $b \approx 0.2275$ for the normal prior and to $b = \log 2$ for the Laplace and Weibull priors. For the Subbotin prior, neutrality restricts b to be a nonlinear function of c given by $\Gamma(b, 1/c) = 1/2$.

For the Weibull and Subbotin priors we also fix the free prior parameter c by the minimax regret criterion for m_n^*, where regret is defined as the

difference between the MSE of m_n^* and the minimum MSE in estimating θ_n^*. Based on this criterion, Magnus and De Luca (2016) find that the 'optimal' neutral and robust priors have $c \approx 0.8876$ for the Weibull distribution and $c \approx 0.7995$ and $b \approx 0.9377$ for the Subbotin distribution.

5 Asymptotic behavior of $g(x_n^*)$ and $m(x_n^*)$

For $x \neq 0$ we define the function $w(x) = m(x)/x$, so that we can write

$$m(x) = w(x)\, x, \tag{26}$$

where conditions (C1)–(C4) imply that $w(x)$ is a symmetric shrinkage function, that is, it satisfies $w(-x) = w(x)$ and $0 < w(x) < 1$. When the prior is robust (i.e. when $\omega_0^* = 0$), then $w(x) \to 1$ as $x \to \infty$.

We have

$$m_n^* = m(x_n^*) = w(x_n^*)\, x_n^*, \qquad g_n^* = g(x_n^*) = x_n^* - m_n^*, \tag{27}$$

so that the posterior mean m_n^* may be regarded as a weighted sum of the t-ratio x_n^* and zero, the center of the prior distribution of θ_n^*. Our first interest is in characterizing the behavior of g_n^* as $n \to \infty$.

PROPOSITION **2** *Under conditions (C1)–(C4) we have*

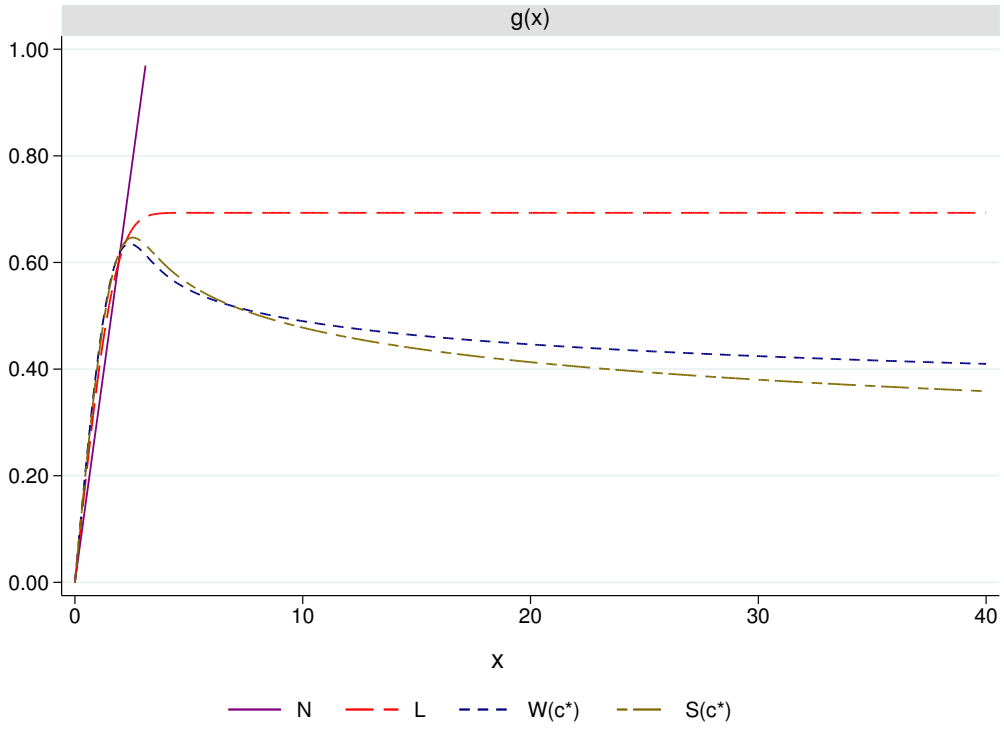
$$\begin{cases} g_n^* \xrightarrow{p} \omega_0^* & \text{if } \theta > 0, \\ g_n^* = (1 - w(z))\, z & \text{if } \theta = 0, \\ g_n^* \xrightarrow{p} -\omega_0^* & \text{if } \theta < 0, \end{cases}$$

where $z \sim \mathrm{N}(0, 1)$.

The asymptotic behavior of g_n^* thus depends on whether or not $\theta = 0$ and whether or not the prior π^* is robust. In particular, for robust priors, $g_n^* = o_p(1)$ if $\theta \neq 0$ and $g_n^* = O_p(1)$ if $\theta = 0$.

Figure 1 illustrates the behavior of the discrepancy function g under the neutral normal, Laplace, and 'optimal' (in the minimax regret sense) Weibull and Subbotin priors. For small values of x (say, $x \leqslant 5$), the differences

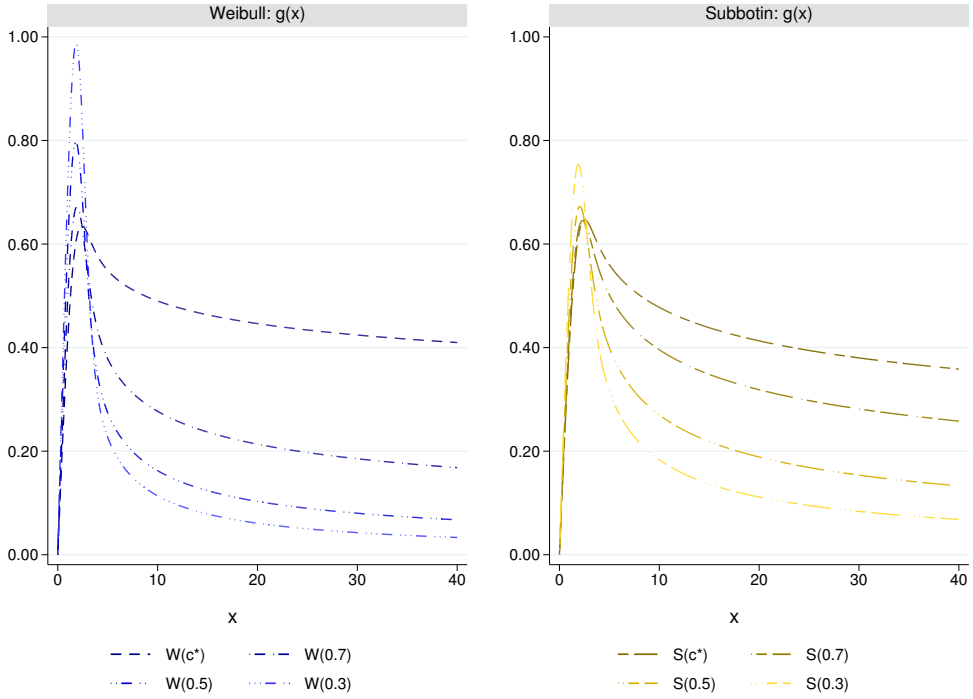
Figure 1: The discrepancy function $g(x)$ under normal (N), Laplace (L), and 'optimal' Weibull ($W(c^*)$) and Subbotin ($S(c^*)$) priors



between the discrepancy functions of the four priors are small, but they become larger when x increases. More specifically, g diverges to infinity for the normal prior, it converges to a constant $b = \log 2$ for the Laplace prior, and it converges to zero for the robust Weibull and Subbotin priors. This is all in accordance to Proposition 2. The proposition does not, however, tell us how fast $g(x)$ converges to zero under the Weibull and Subbotin priors. The answer is: rather slowly.

We investigate the slow convergence further in Figure 2, which plots the discrepancy functions under the Weibull (left panel) and Subbotin (right panel) priors for alternative values of the prior parameter c. In all cases, the prior parameter b is chosen to satisfy the neutrality condition (25). The figure shows that, if we choose a smaller value for c, then the discrepancy

Figure 2: The discrepancy function $g(x)$ under Weibull and Subbotin priors for alternative values of the prior parameter c



functions of both priors become larger at small values of x but converge to zero more rapidly as x increases. Thus, values of c below the minimax regret choice c^* imply a higher speed at which $g(x) \to 0$ as $x \to \infty$, but at the cost of a larger finite-sample bias.

The asymptotic behavior of m_n^* as an estimator of θ_n^* now follows easily.

PROPOSITION **3** *Under conditions (C1)–(C4) we have*

$$\begin{cases} m_n^* - \theta_n^* \xrightarrow{d} \mathrm{N}(-\omega_0^*, 1) & \text{if } \theta > 0, \\ m_n^* = w(z)\, z & \text{if } \theta = 0, \\ m_n^* - \theta_n^* \xrightarrow{d} \mathrm{N}(\omega_0^*, 1) & \text{if } \theta < 0, \end{cases}$$

where $z \sim \mathrm{N}(0, 1)$.

If $\theta \neq 0$, then π^* is not correctly centered and estimating θ_n^* by m_n^* suffers from a finite-sample attenuation bias (De Luca et al. 2021a). However, as the t-ratio x_n^* increases with the sample size, conditions (C1)–(C4) ensure that $\text{plim}\,|g_n^*| = \omega_0^*$. Hence, the bias of m_n^* is asymptotically bounded and converges to zero when π^* is robust. This also implies that m_n^* and x_n^* are asymptotically equivalent except for a possible shift ω_0^* when π^* is not robust.

If $\theta = 0$, then π^* is correctly centered and $m_n^* = w(z)z$ is unbiased and more efficient than $x_n^* = z$, no matter what the sample size is and whether the prior is robust or not. The distribution of m_n^* at $\theta = 0$ thus does not depend on n and is not standard-normal. It is symmetric around zero and hence all odd moments are zero but, since $\mathbb{E}(m_n^*)^{2h} = \mathbb{E}(w(z)^{2h}z^{2h}) < \mathbb{E}(z^{2h})$, the even moments are all smaller than the corresponding moments of the standard-normal distribution. In particular, setting $h = 1$, we see that at $\theta = 0$ the posterior mean m_n^* is unbiased and more efficient than the usual estimator x_n^*.

Propositions 2 and 3 can easily be extended to the case when σ is unknown and estimated consistently by s_n. We only need to replace x_n^* by

$$x_n^{**} = \frac{x_n^*}{s_n/\sigma} = \frac{x_n}{s_n/\sqrt{n}}. \tag{28}$$

Specifically, if we redefine the posterior mean and the discrepancy function in (27) as

$$m_n^{**} = m(x_n^{**}) = w(x_n^{**})x_n^{**}, \qquad g_n^{**} = g(x_n^{**}) = x_n^{**} - m_n^{**}, \tag{29}$$

then Propositions 2 and 3 remain valid for g_n^{**} and m_n^{**}, respectively. The only difference is that, for $\theta = 0$, we should replace z by $z_n \xrightarrow{d} \text{N}(0,1)$.

6 A Bayesian shrinkage estimator of θ

These preliminary results enable us to address the estimation of θ in (12). Given a consistent estimator s_n of σ, our Bayesian shrinkage estimator of θ is

$$\hat{\theta}_n = \frac{s_n}{\sqrt{n}} m_n^{**} = \frac{s_n}{\sqrt{n}} w(x_n^{**})\, x_n^{**} = w(x_n^{**})\, x_n, \tag{30}$$

which deviates from the usual estimator x_n by the shrinkage factor $w(x_n^{**})$ evaluated at x_n^{**}.

Since

$$\sqrt{n}(x_n - \hat{\theta}_n)/s_n = \sqrt{n}(x_n - w(x_n^{**})x_n)/s_n = x_n^{**} - m_n^{**} = g_n^{**}, \tag{31}$$

we see from the extension of Proposition 2 that, under conditions (C1)–(C4),

$$\begin{cases} \sqrt{n}(x_n - \hat{\theta}_n)/s_n \xrightarrow{p} \omega_0^* & \text{if } \theta > 0, \\ \sqrt{n}(x_n - \hat{\theta}_n)/s_n = (1 - w(z_n))\, z_n & \text{if } \theta = 0, \\ \sqrt{n}(x_n - \hat{\theta}_n)/s_n \xrightarrow{p} -\omega_0^* & \text{if } \theta < 0, \end{cases} \tag{32}$$

where $\operatorname{plim} z_n = z \sim \mathrm{N}(0,1)$. Thus we obtain

PROPOSITION **4** *Under conditions (C1)–(C4), the shrinkage estimator $\hat{\theta}_n$ is consistent for all θ and asymptotically normal for all $\theta \neq 0$. In particular,*

$$\sqrt{n}(\hat{\theta}_n - \theta)/s_n \xrightarrow{d} \begin{cases} \mathrm{N}(-\omega_0^*, 1) & \text{if } \theta > 0, \\ w(z)\, z & \text{if } \theta = 0, \\ \mathrm{N}(\omega_0^*, 1) & \text{if } \theta < 0, \end{cases}$$

where $z \sim \mathrm{N}(0,1)$.

Under certain conditions, the first and second moments of $\sqrt{n}(\hat{\theta}_n - \theta)$ will converge to finite limits, and indeed we have

PROPOSITION **5** *Under conditions (C1)–(C4),*

$$\sqrt{n}\,\mathbb{E}(\hat{\theta}_n - \theta) \to \begin{cases} -\sigma\omega_0^* & \text{if } \theta > 0, \\ 0 & \text{if } \theta = 0, \\ \sigma\omega_0^* & \text{if } \theta < 0, \end{cases}$$

and

$$n\,\mathbb{V}(\hat{\theta}_n) \to \begin{cases} \sigma^2 & \text{if } \theta \neq 0, \\ \sigma^2\,\mathbb{E}(w(z)z)^2 & \text{if } \theta = 0. \end{cases}$$

Our Bayesian shrinkage estimator is thus asymptotically unbiased only when the prior is robust, and it is always more efficient than the usual estimator x_n^{**} at $\theta = 0$.

7 Implications for WALS

We now return to WALS, using the asymptotic results obtained for the normal location model. Our model is $y = X_1\beta_1 + X_2\beta_2 + \epsilon$, as in (1), which we transform to $y = Z_1\gamma_1 + Z_2\gamma_2 + \epsilon$, as in (4). The key to this transformation is that $Z_2'M_1Z_2/n = I_{k_2}$, so that the k_2 components of the LS estimator of γ_2 are independent.

Our purpose is to obtain the asymptotic distribution of the WALS estimator of $\beta = (\beta_1', \beta_2')'$ through the asymptotic distribution of the WALS estimator of $\gamma = (\gamma_1', \gamma_2')'$. To keep track of the sample size, we add an index n to all relevant data-dependent parameters and random variables.

Let $\hat{\gamma}_{2u,n}$ be the LS estimator of γ_2 in the unrestricted model, and let $\hat{\gamma}_{1r,n}$ be the LS estimator of γ_1 in the fully restricted model. Denoting the hth component of γ_2 by γ_{2h}, and the hth component of $\hat{\gamma}_{2u,n}$ by $(\hat{\gamma}_{2u,n})_h$, we define

$$\theta^*_{h,n} = \frac{\gamma_{2h}}{\sigma/\sqrt{n}}, \qquad x^{**}_{h,n} = \frac{(\hat{\gamma}_{2u,n})_h}{s_n/\sqrt{n}}, \tag{33}$$

as in (13) and (28), where s_n^2 is the LS estimator of σ^2 in the unrestricted model. Given a neutral prior on $\theta^*_{h,n}$, such as the Laplace, Weibull or Subbotin priors discussed in Section 4, the Bayesian approach to the normal location problem yields a consistent estimator

$$\hat{\theta}_{h,n} = \frac{s_n}{\sqrt{n}}\, m^{**}_{h,n} \tag{34}$$

of γ_{2h}, as in (30).

The WALS estimators of γ_1 and γ_2 can be written as

$$\tilde{\gamma}_{1,n} = \hat{\gamma}_{1r,n} - Q_n\tilde{\gamma}_{2,n}, \qquad \tilde{\gamma}_{2,n} = (s_n/\sqrt{n})m^{**}_n, \tag{35}$$

where $m^{**}_n = (m^{**}_{1,n}, \ldots, m^{**}_{k_2,n})'$, and the WALS estimators of β_1 and β_2 as

$$\tilde{\beta}_{1,n} = \Delta_{1,n}\tilde{\gamma}_{1,n}, \qquad \tilde{\beta}_{2,n} = \Delta_{2,n}\Psi_n^{-1/2}\tilde{\gamma}_{2,n}. \tag{36}$$

The probability limits of these estimators follow from those of

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} = \frac{1}{n}\begin{pmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{pmatrix}, \tag{37}$$

the continuity of eigenprojections and symmetric matrix functions (De Luca et al. 2018, Appendix B), and the asymptotic results obtained for the normal location model. Specifically, letting

$$\operatorname{plim} \Sigma = \bar{\Sigma} = \begin{pmatrix} \bar{\Sigma}_{11} & \bar{\Sigma}_{12} \\ \bar{\Sigma}_{21} & \bar{\Sigma}_{22} \end{pmatrix}, \tag{38}$$

we find that

$$\operatorname{plim} \Delta_{1,n} = \bar{\Delta}_1, \qquad \operatorname{plim} \frac{Z_1'Z_1}{n} = \bar{\Delta}_1 \bar{\Sigma}_{11} \bar{\Delta}_1 = \bar{\Sigma}^*_{11}, \tag{39}$$

where $\bar{\Delta}_1$ is a nonrandom diagonal matrix whose diagonal elements are equal to the inverse of the square root of the corresponding diagonal element of $\bar{\Sigma}_{11}$, and $\bar{\Sigma}^*_{11}$ is a nonrandom matrix with diagonal elements equal to one. Similarly, by the continuity of the inverse, we find that

$$\operatorname{plim} \Delta_{2,n} = \bar{\Delta}_2, \qquad \operatorname{plim} \Psi_n = \bar{\Delta}_2 (\bar{\Sigma}^{22})^{-1} \bar{\Delta}_2 = \bar{\Psi}, \tag{40}$$

where $\bar{\Delta}_2$ is a nonrandom diagonal matrix whose diagonal elements are equal to the square root of the corresponding diagonal element of $\bar{\Sigma}^{22} = (\bar{\Sigma}_{22} - \bar{\Sigma}_{21}\bar{\Sigma}_{11}^{-1}\bar{\Sigma}_{12})^{-1}$ (the bottom-right block of the matrix $\bar{\Sigma}^{-1}$), and $\bar{\Psi}$ is a nonrandom matrix with diagonal elements equal to one. The continuity of $\Psi_n^{-1/2}$ also implies that

$$\operatorname{plim} \frac{Z_1'Z_2}{n} = \bar{\Delta}_1 \bar{\Sigma}_{12} \bar{\Delta}_2 \bar{\Psi}^{-1/2} = \bar{\Sigma}^*_{12} \tag{41}$$

and

$$\operatorname{plim} \frac{Z_2'Z_2}{n} = \bar{\Psi}^{-1/2} \bar{\Delta}_2 \bar{\Sigma}_{22} \bar{\Delta}_2 \bar{\Psi}^{-1/2} = \bar{\Sigma}^*_{22}, \tag{42}$$

so that $\operatorname{plim} Q_n = \bar{\Sigma}^{*-1}_{11} \bar{\Sigma}^*_{12} = \bar{Q}^*$ and $\bar{\Sigma}^*_{22} - \bar{\Sigma}^*_{21} \bar{Q}^* = I_{k_2}$.

Now that the relationship between WALS and the normal location model has been made precise, we can invoke Proposition 4 to obtain the asymptotic distribution of the WALS estimator $\tilde{\gamma}_{2,n}$ of γ_2.

PROPOSITION **6** *The WALS estimator $\tilde{\gamma}_{2,n}$ is a consistent estimator of γ_2 and $\sqrt{n}(\tilde{\gamma}_{2,n} - \gamma_2)/s_n \xrightarrow{d} \mathcal{Z}_2$, where $\mathcal{Z}_2$ is a random $k_2 \times 1$ vector of inde-*

pendent components with hth component

$$\mathcal{Z}_{2h} = \begin{cases} z - \omega_0^* & \text{if } \gamma_{2h} > 0, \\ w(z)\, z & \text{if } \gamma_{2h} = 0, \\ z + \omega_0^* & \text{if } \gamma_{2h} < 0, \end{cases}$$

where $z \sim \mathrm{N}(0,1)$. *In particular, when the prior is robust,*

$$\mathcal{Z}_{2h} = \begin{cases} z & \text{if } \gamma_{2h} \neq 0, \\ w(z)\, z & \text{if } \gamma_{2h} = 0. \end{cases}$$

Recall that the WALS estimator $\tilde{\gamma}_{2,n}$ depends only on the unrestricted LS estimator $\hat{\gamma}_{2u,n}$ and is therefore independent of the fully restricted LS estimator $\hat{\gamma}_{1r,n}$. Since the latter may be written as

$$\hat{\gamma}_{1r,n} = \left(\frac{Z_1'Z_1}{n}\right)^{-1}\left(\frac{Z_1'y}{n}\right) = \gamma_1 + Q_n\gamma_2 + \left(\frac{Z_1'Z_1}{n}\right)^{-1}\left(\frac{Z_1'\epsilon}{n}\right), \tag{43}$$

we have, using (35) and (43),

$$\tilde{\gamma}_{1,n} - \gamma_1 = \hat{\gamma}_{1r,n} - \gamma_1 - Q_n\tilde{\gamma}_{2,n} = (Z_1'Z_1/n)^{-1}(Z_1'\epsilon/n) - Q_n(\tilde{\gamma}_{2,n} - \gamma_2), \tag{44}$$

and hence, using (44) and Proposition 6,

$$\sqrt{n}\begin{pmatrix} \tilde{\gamma}_{1,n} - \gamma_1 \\ \tilde{\gamma}_{2,n} - \gamma_2 \end{pmatrix} \xrightarrow{d} \sigma \begin{pmatrix} \bar{\Sigma}_{11}^{*-1/2} & -\bar{Q}^* \\ 0 & I_{k_2} \end{pmatrix} \begin{pmatrix} \mathcal{Z}_1 \\ \mathcal{Z}_2 \end{pmatrix}, \tag{45}$$

where $\mathcal{Z}_1 \sim \mathrm{N}(0, I_{k_1})$ and $\mathcal{Z}_2$ is defined in Proposition 6. The asymptotic distribution of the WALS estimator of β_1 and β_2 then follows.

PROPOSITION **7** *The WALS estimators* $\tilde{\beta}_{1,n}$ *and* $\tilde{\beta}_{2,n}$ *are consistent for* β_1 *and* β_2, *and*

$$\sqrt{n}\begin{pmatrix} \tilde{\beta}_{1,n} - \beta_1 \\ \tilde{\beta}_{2,n} - \beta_2 \end{pmatrix} \xrightarrow{d} \sigma \begin{pmatrix} \bar{\Sigma}_{11}^{-1/2} & -\bar{\Sigma}_{11}^{-1}\bar{\Sigma}_{12}(\bar{\Sigma}^{22})^{1/2} \\ 0 & (\bar{\Sigma}^{22})^{1/2} \end{pmatrix} \begin{pmatrix} \mathcal{Z}_1 \\ \mathcal{Z}_2 \end{pmatrix}$$

where $\mathcal{Z}_1 \sim \mathrm{N}(0, I_{k_1})$, $\mathcal{Z}_2$ *is defined in Proposition 6, and* $\mathcal{Z}_1$ *and* $\mathcal{Z}_2$ *are independent.*

If $\omega_0^* = 0$ (robust prior) and all components of γ_2 are nonzero, then the asymptotic distribution of WALS equals that of the unrestricted LS estimator. Under restrictions (C1)–(C4) on the prior, this agrees with the standard Bayesian approach to model uncertainty. When $\beta_2 = 0$, corresponding to the limiting case implied by the local misspecification framework, the WALS estimator based on a robust prior is asymptotically more efficient than the unrestricted LS estimator. On the other hand, if ω_0^* is a nonzero finite constant, as in the case of the Laplace prior, the WALS estimator is asymptotically biased and less efficient, in the MSE sense, than the unrestricted LS estimator. This is one reason why robust priors are preferred over non-robust priors.

Our results on the asymptotic properties of the WALS estimator are in line with those of other frequentist model-averaging estimators, such as the Mallows model-averaging estimator (Hansen 2007) and the jackknife model-averaging estimator (Hansen and Racine 2012). Under a standard asymptotic setup with fixed parameters, Zhang and Liu (2019) have recently shown that these two estimators asymptotically assign zero weight to all under-fitted models. Similar results hold for other frequentist model-averaging estimators based on smoothed information criteria (Wang et al. 2019). These results imply that, when the DGP has a 'sparse' structure, the asymptotic distribution of these estimators is nonstandard because of the random positive weights assigned to just-fitted and over-fitted models. However, when the DGP has a 'dense' structure, such estimators are asymptotically equivalent to the unrestricted LS estimator. This is exactly what happens with the WALS estimator based on a robust prior, as the underlying model (4) is likely to be dense due to the transformations in (2).

Proposition 7 also provides useful insights on the issue of inference after WALS estimation. De Luca et al. (2021b) have recently proposed 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. This approach does not require asymptotic approximations and its intervals are not necessarily symmetric. The extensive set of Monte Carlo experiments in De Luca et al. (2021b) suggests that one can also construct valid intervals by a simpler 'centered-and-naive'

approach which corrects for the estimation bias and uses critical values from the normal distribution. The Monte Carlo evidence in the next section shows that this simpler approach is justified by the asymptotic approximations.

8 Monte Carlo evidence

So far we have developed the asymptotic theory, first for the normal location model, then for WALS. Several questions remain which we can answer, with appropriate caution, by Monte Carlo experimentation. How fast is the convergence of the WALS estimator (especially the WALS estimator of the focus parameters) to its asymptotic distribution? Is the convergence monotonic in n? Can we increase the speed at which the finite-sample bias converges to zero by using a robust prior with c smaller than the minimax regret solution c^*? And, what is the large-sample performance of the bias-correction strategy proposed by De Luca et al. (2021a, 2021b)?

We address these questions by considering a homoskedastic linear regression model with $k_1 = 2$ focus regressors: the constant term $x_{1,1}$ and $x_{1,2}$; and $k_2 = 8$ auxiliary regressors: $x_{2,1}, \ldots, x_{2,8}$.[2] For $-1/(k_2-1) < \rho < 1$ we define the $k_2 \times k_2$ equicorrelation matrix

$$\Omega_{k_2}(\rho) = \begin{pmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \vdots & \vdots & & \vdots \\ \rho & \rho & \ldots & 1 \end{pmatrix} = \nu_1 J + \nu_2 (I_{k_2} - J), \tag{46}$$

where

$$\nu_1 = 1 + (k_2 - 1)\rho, \qquad \nu_2 = 1 - \rho, \qquad J = \imath\imath'/k_2, \tag{47}$$

and $\imath$ denotes the $k_2 \times 1$ vector of ones. The nine regressors $x_{1,2}, x_{2,1}, \ldots, x_{2,8}$ are drawn from a multivariate normal distribution with mean zero and variance $\sigma_x^2\, \Omega_{k_2+1}(\rho)$. We fix $\beta_1 = (1,1)'$, $\beta_2 = (\xi, \xi^2, \xi^3, \xi^4, 0, 0, 0, 0)'$, $\sigma_x^2 = 0.7$,

[2] In addition to simulation designs with $k_2 = 8$, we also considered designs with $k_2 = 16$ and $k_2 = 32$. The results obtained are very similar and are available from the authors upon request.

$\rho = 0.7$, and we set $\xi = 0.5$. Then we consider eight simulation designs corresponding to four choices of the sample size: $n = 100$, $n = 400$, $n = 1,600$, and $n = 6,400$; and two distributions of the regression errors: standard-normal and skewed-t, where the latter has zero mean, unit variance, five degrees of freedom, and skewness parameter 0.8. Our parameter of interest is the focus coefficient $\beta_{1,2} = 1$ associated with $x_{1,2}$.

Based on 10,000 Monte Carlo replications for each design, we compare the simulated density of $\sqrt{n}(\tilde{\beta}_{1,2} - \beta_{1,2})$ with the asymptotic normal density given in Proposition 7. In all designs we have

$$\bar{\Sigma} = \begin{pmatrix} \bar{\Sigma}_{11} & \bar{\Sigma}_{12} \\ \bar{\Sigma}_{21} & \bar{\Sigma}_{22} \end{pmatrix} \tag{48}$$

with $\bar{\Sigma}_{11} = \text{diag}(1, \sigma_x^2)$, $\bar{\Sigma}_{21} = \rho\sigma_x^2(0 : \imath)$, and $\bar{\Sigma}_{22} = \sigma_x^2\Omega_{k_2}(\rho)$. This implies that the diagonal blocks of $\bar{\Sigma}^{-1}$ are given by

$$\begin{aligned} \bar{\Sigma}^{22} &= (\bar{\Sigma}_{22} - \bar{\Sigma}_{21}\bar{\Sigma}_{11}^{-1}\bar{\Sigma}_{12})^{-1} = (1/\sigma_x^2)(\Omega_{k_2}(\rho) - \rho^2\imath\imath')^{-1} \\ &= (1/\sigma_x^2)(\nu_1 J + \nu_2(I_{k_2} - J) - \rho^2\imath\imath')^{-1} \\ &= (1/\sigma_x^2)((\nu_1 - k_2\rho^2)J + \nu_2(I_{k_2} - J))^{-1} \\ &= (1/\sigma_x^2)((\nu_1 - k_2\rho^2)^{-1}J + \nu_2^{-1}(I_{k_2} - J)) \end{aligned} \tag{49}$$

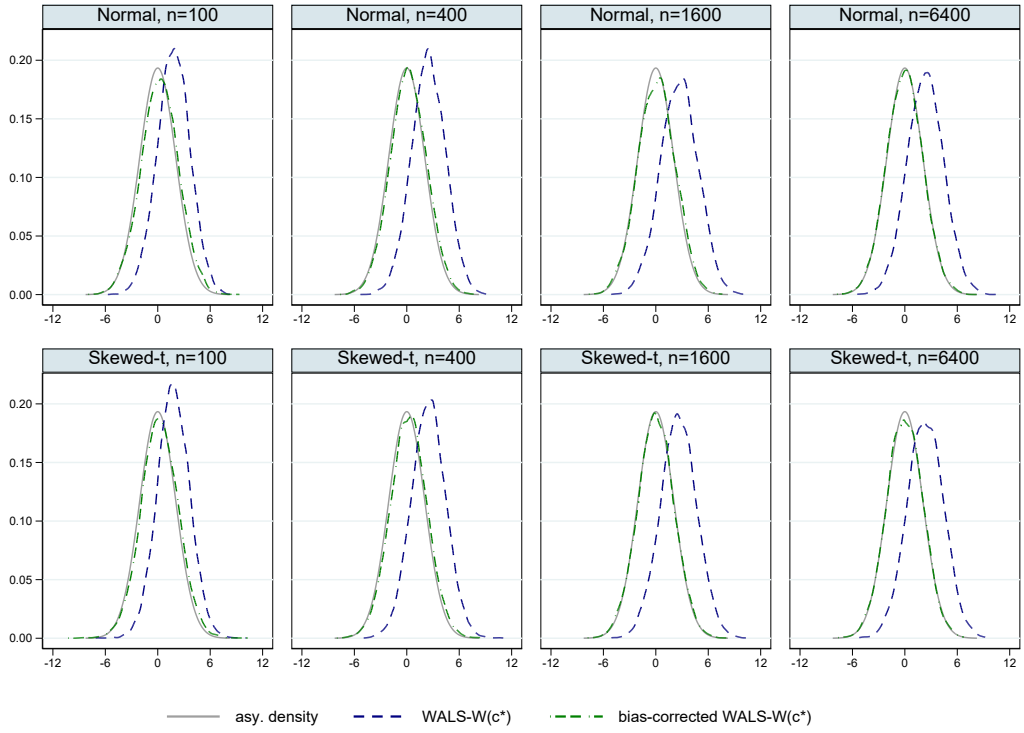
and

$$\begin{aligned} \bar{\Sigma}^{11} &= \bar{\Sigma}_{11}^{-1} + \bar{\Sigma}_{11}^{-1}\bar{\Sigma}_{12}\bar{\Sigma}^{22}\bar{\Sigma}_{21}\bar{\Sigma}_{11}^{-1} \\ &= \frac{1}{\sigma_x^2}\begin{pmatrix} \sigma_x^2 & 0' \\ 0 & 1 + \rho^2\sigma_x^2\imath'\bar{\Sigma}^{22}\imath \end{pmatrix} = \begin{pmatrix} 1 & 0' \\ 0 & \nu_1/[\sigma_x^2(\nu_1 - k_2\rho^2)] \end{pmatrix}. \end{aligned} \tag{50}$$

Since the regression errors have unit variance (that is, $\sigma^2 = 1$), the asymptotic variance of all WALS estimators of $\beta_{1,2}$ is equal to the second diagonal element of $\bar{\Sigma}^{11}$, so that $\bar{\sigma}_{1,2}^2 \approx 4.2569$.

Figure 3 shows the asymptotic and the simulated density of the WALS estimator based on a Weibull prior with tuning parameter c equal to the minimax regret value $c^* \approx 0.8876$. The upper panels refer to simulation designs with normal errors and the bottom panels to simulation designs with skewed-t errors. Moving from left to right, the sample size n quadruples each

Figure 3: Asymptotic and simulated densities of WALS and bias-corrected WALS: Weibull prior with $c \approx 0.8876$ (minimax regret)



time. Each panel plots three densities: the asymptotic density of $\tilde{\beta}_{1,2}$ (that is, a normal density with mean zero and variance $\bar{\sigma}^2_{1,2}$), a kernel density of the Monte Carlo replications of $\sqrt{n}(\tilde{\beta}_{1,2} - \beta_{1,2})$, and a third density to be discussed shortly.

Since the Weibull prior is robust, the asymptotic distribution is centered at zero. Comparing the asymptotic and the simulated distribution of WALS-W(c^*), we see that the latter has a normal shape (even for small values of n) but its location is off-centre. Formal tests of normality of the simulated distribution based on its skewness and excess kurtosis (D'Agostino et al. 1990) reject normality when the errors are skewed-t and $n = 100$ or $n = 400$, but don't reject normality when the errors are normal or n is large. This confirms convergence to normality and tells us something about the speed of

convergence.

The impression that the simulated density converges to the wrong point is of course wrong, because this would contradict our theory. But the convergence is very slow, because of the fact that the bias converges to zero very slowly. This, in turn, is a consequence of the slow convergence of $g(x)$ discussed in Section 5. This problem can be resolved by introducing a bias-corrected WALS estimator

$$\tilde{\beta}^*_{1,2} = \tilde{\beta}_{1,2} - \tilde{b}_{1,2}, \tag{51}$$

where $\tilde{b}_{1,2}$ is the maximum likelihood plug-in estimator of the bias of $\tilde{\beta}_{1,2}$ proposed in De Luca et al. (2021a). The Monte Carlo simulations in De Luca et al. (2021b) suggest that bias-correction plays a key role in constructing valid confidence and prediction intervals for WALS, and it plays a key role again in speeding up the convergence.

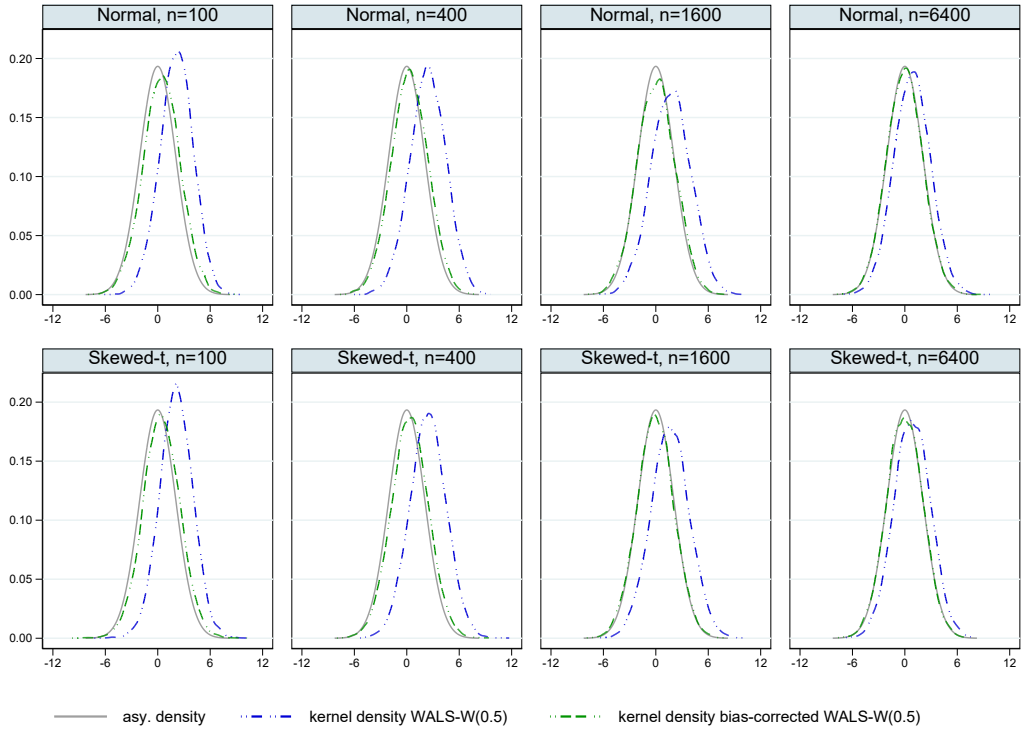
The third curve plotted in Figure 3 is the density of the Monte Carlo replications of $\sqrt{n}(\tilde{\beta}^*_{1,2} - \beta_{1,2})$. This density is now always correctly centered at zero. In small samples, bias-correction leads to an increase of the sampling variance but, as n increases, the bias-corrected estimator converges to a normal distribution with mean zero and variance $\bar{\sigma}^2_{12}$.

Figure 4 illustrates the results of an analogous Monte Carlo experiment based on the Weibull prior with $c = 0.5$ instead of $c = c^* \approx 0.8876$. As predicted by our asymptotic theory, a lower value of c increases the speed at which the finite-sample bias converges to zero, but the superiority of the bias-corrected WALS estimator remains.

9 Conclusions

The purpose of this paper was to obtain the asymptotic properties of the WALS estimator, a frequentist-Bayesian model-averaging estimator with attractive finite-sample properties. The theory of WALS is strongly connected to the theory of the normal location model, and therefore much of the current paper concerns this model, which is also of interest outside WALS. At first,

Figure 4: Asymptotic and simulated densities of WALS and bias-corrected WALS: Weibull prior with $c = 0.5$



it may seem strange that there is an asymptotic theory at all, because in the normal location model we ask how to estimate θ when we have *one* observation from a normal distribution with mean θ and known variance. This known variance, however, depends on n and this is how n becomes relevant.

In a Bayesian context we can place a prior on the parameter θ, but we can also place a prior on the t-ratio associated with θ. The former case is more common, but the latter case leads to a more transparent notion of prior ignorance. If prior knowledge is available, a Bayesian wishes to take this into account. If there is no prior knowledge, we can still be a Bayesian but we must define what we mean by ignorance. In WALS, ignorance about a parameter θ is defined by its 'theoretical' t-ratio θ_n^*. From the MSE point of view, it is better to include a regressor if the absolute value of its t-ratio

is greater than one and exclude it otherwise. So, ignorance is defined by a prior that places equal probability on $|\theta_n^*| > 1$ and $|\theta_n^*| < 1$. This is called neutrality.

In our Bayesian approach, based on a neutral prior for the t-ratio, the asymptotic theory is not straightforward because the underlying prior on θ depends on n. We showed that the WALS estimator is consistent if the chosen prior has bounded influence on the posterior mean, and that it has the same asymptotic distribution as the unrestricted LS estimator if the prior is also robust (and the DGP does not coincide with the fully restricted model). In particular, the stronger condition of robustness prevents biases in the asymptotic distribution of the WALS estimator. When the DGP coincides with the fully restricted model, the WALS estimator is always asymptotically more efficient than the unrestricted LS estimator.

The paper also allows us to compare the asymptotic theory of WALS, based on a fixed parameter setup, with the local misspecification framework. The latter is much used because it puts variance and squared bias on the same asymptotic scale, but it seems far removed from reality. In a model-averaging framework, local misspecification implies that the DGP *shrinks* towards the restricted model, while common sense suggests that it should *expand* to the unrestricted model. This is precisely what happens to the WALS estimator based on a neutral and robust prior: in an $\mathcal{M}$-closed environment where the model space does not expand with n, it converges to the unrestricted estimator.

Our asymptotic theory for WALS assumes that the number k_2 of auxiliary coefficients is fixed and imposes no restriction other than $k = k_1 + k_2 < n$. It is easy, however, to think of cases where k_2 increases with n and our theory can easily be extended to these cases provided the restrictions $k = k_1 + k_2 < n$ and $k_2/n \to 0$ as $n \to \infty$ are satisfied. The reason is that, after the preliminary transformations in (2) and (3), k_2 affects the asymptotic results for the normal location model only through the error variance σ^2 (which is likely to decrease with k_2, though not necessarily monotonically). Requiring $k_2/n \to 0$ as $n \to \infty$ thus ensures that σ^2 can be estimated consistently by its unbiased estimator s_n^2 in the unrestricted model.

In addition to confirming our theoretical results, Monte Carlo simulations based on the neutral and 'optimal' (in the minimax regret sense) Weibull prior suggest that the finite-sample bias of the WALS estimator converges to zero slowly. More rapid convergence to the asymptotic distribution can be achieved by using a prior parameter c smaller than the minimax regret solution c^* or by applying the bias-correction strategy recently proposed in De Luca et al. (2021a, 2021b).

Appendix: Proofs

Proof of Proposition 1: By the Brown–Tweedie formula (Robbins 1956, Brown 1971, Pericchi and Smith 1992), we know that

$$m(x) = x + \frac{d \log A_0(x)}{dx},$$

where

$$A_0(x) = \int_{-\infty}^{\infty} \phi(x-\theta)\pi^*(\theta)\, d\theta = \int_{-\infty}^{\infty} \phi(u)\pi^*(u+x)\, du.$$

Defining

$$A_1(x) = \int_{-\infty}^{\infty} (x-\theta)\phi(x-\theta)\pi^*(\theta)\, d\theta = -\int_{-\infty}^{\infty} u\phi(u)\pi^*(u+x)\, du = -A_0'(x),$$

we can then rewrite the discrepancy function as

$$g(x) = x - m(x) = \frac{A_1(x)}{A_0(x)}.$$

By the definition of $\omega^*(\theta) = -(d/d\theta)\log \pi^*(\theta) = -\pi^{*\prime}(\theta)/\pi^*(\theta)$ in (C4), it also follows that

$$\frac{A_0'(x)}{\pi^{*\prime}(x)} = \frac{A_1(x)}{\omega^*(x)\pi^*(x)}.$$

Since $A_0(x)$ and $\pi^*(x)$ both converge to zero as $x \to \infty$, l'Hôpital's rule gives

$$\lim_{x\to\infty} \frac{A_0(x)}{\pi^*(x)} = \lim_{x\to\infty} \frac{A_0'(x)}{\pi^{*\prime}(x)} = \lim_{x\to\infty} \frac{A_1(x)}{\omega^*(x)\pi^*(x)},$$

and hence

$$\lim_{x\to\infty} \frac{g(x)}{\omega^*(x)} = \lim_{x\to\infty} \frac{A_1(x)}{A_0(x)\omega^*(x)} = 1.$$

Proof of Proposition 2: Let $x_n^* = \theta_n^* + z$ with $z \sim \mathrm{N}(0,1)$. For $\theta > 0$ we have

$$\Pr(x_n^* > M) = \Pr(z > M - \theta_n^*) \to 1 \text{ for every } M > 0,$$

which we write as $\operatorname{plim} x_n^* = \infty$. Then, by the (generalized) continuous mapping theorem,

$$\operatorname{plim}(x_n^* - m_n^*) = \operatorname{plim} g(x_n^*) = g(\operatorname{plim} x_n^*) = g(\infty) = \omega_0^*.$$

This is not completely trivial because the usual continuous mapping theorem, which says that $\operatorname{plim} x_n^* = x$ implies $\operatorname{plim} g(x_n^*) = g(x)$, does not apply here since $x = \infty$. Hence a generalization is required; see van der Vaart (1998, Theorem 18.11 and Example 18.4) and Billingsley (1999, Theorem 2.7).

Similarly, for $\theta < 0$,

$$\Pr(x_n^* < -M) = \Pr(z < -M - \theta_n^*) \to 1 \text{ for every } M > 0,$$

which we write as $\operatorname{plim} x_n^* = -\infty$. Then also $\operatorname{plim}(x_n^* - m_n^*) = -\omega_0^*$ using property (P1).

For $\theta = 0$, we have $x_n^* = z$ so that $m_n^* = w(z)\, z$ and $g_n^* = (1 - w(z))\, z$.

Proof of Proposition 3: We write

$$m_n^* - \theta_n^* = (x_n^* - \theta_n^*) - g(x_n^*).$$

Since $x_n^* - \theta_n^* \sim \mathrm{N}(0,1)$, the result follows from Proposition 2.

Proof of Proposition 4: When $\theta \neq 0$, the asymptotic normality follows from Proposition 3 and the fact that s_n is a consistent estimator of σ, while consistency follows from the asymptotic normality (van der Vaart, 1998, Chapter 2, Problem 18). When $\theta = 0$, we have $\hat{\theta}_n - \theta = O_p(n^{-1/2})$ and hence $\hat{\theta}_n - \theta = o_p(1)$.

Proof of Proposition 5: The first and second moments of $\sqrt{n}(\hat{\theta}_n - \theta)$ will converge to a finite limit if and only if the sequence of random variables g_n^{**} is asymptotically uniformly integrable (Van der Vaart, 1998, Section 2.5). A sufficient condition for this is that

$$\limsup_{n \to \infty} \mathbb{E}\, |g(x_n^{**})|^{2+\delta} < \infty$$

for some $\delta > 0$, and this is guaranteed by the fact that g is a bounded function, which is a consequence of Proposition 1. The asymptotic mean and variance of $\sqrt{n}(\hat{\theta}_n - \theta)$ then follow.

Proof of Proposition 6: This follows directly from Proposition 4.

Proof of Proposition 7: Using (45) we have

$$\begin{aligned}\sqrt{n}\begin{pmatrix}\tilde{\beta}_{1,n}-\beta_1\\ \tilde{\beta}_{2,n}-\beta_2\end{pmatrix} &= \begin{pmatrix}\Delta_{1,n} & 0\\ 0 & \Delta_{2,n}\Psi_n^{-1/2}\end{pmatrix}\sqrt{n}\begin{pmatrix}\tilde{\gamma}_{1,n}-\gamma_1\\ \tilde{\gamma}_{2,n}-\gamma_2\end{pmatrix}\\ &\xrightarrow{d} \sigma\begin{pmatrix}\bar{\Delta}_1\bar{\Sigma}_{11}^{*-1/2} & -\bar{\Delta}_1\bar{Q}^*\\ 0 & \bar{\Delta}_2\bar{\Psi}^{-1/2}\end{pmatrix}\begin{pmatrix}\mathcal{Z}_1\\ \mathcal{Z}_2\end{pmatrix}\\ &= \sigma\begin{pmatrix}\bar{\Sigma}_{11}^{-1/2} & -\bar{\Sigma}_{11}^{-1}\bar{\Sigma}_{12}(\bar{\Sigma}^{22})^{1/2}\\ 0 & (\bar{\Sigma}^{22})^{1/2}\end{pmatrix}\begin{pmatrix}\mathcal{Z}_1\\ \mathcal{Z}_2\end{pmatrix}.\end{aligned}$$

References

Abadie, A., and Kasy, M. (2019). Choosing among regularized estimators in empirical economics: The risk of machine learning. *Review of Economics and Statistics*, 101, 1–20.

Amini, S. M., and Parmeter, C. F. (2012). Comparison of model averaging techniques: Assessing growth determinants. *Journal of Applied Econometrics*, 27, 870–876.

Billingsley, P. (1999). *Convergence of Probability Measures*, Second Edition. Wiley, New York, USA.

Brown, L. D. (1971). Admissible estimators, recurrent diffusions and insoluble boundary value problems. *Annals of Mathematical Statistics*, 42, 855–903.

Choy, S. T. B., and Smith, A. F. M. (1997). On robust analysis of a normal location parameter. *Journal of the Royal Statistical Society, Series B*, 59, 463–474.

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

D'Agostino, R. B., Belanger, A. J., and D'Agostino, R. B., Jr. (1990). A suggestion for using powerful and informative tests of normality. *The American Statistician*, 44(4), 316–321.

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

Dawid, A. P. (1973). Posterior expectations for large observations. *Biometrika*, 60, 664–667.

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. (2021a). Sampling properties of the Bayesian posterior mean with an application to WALS estimation. *Journal of Econometrics*, in press.

De Luca, G., Magnus, J. R., and Peracchi, F. (2021b). Weighted-average least squares (WALS): Confidence and prediction intervals. *EIEF Discussion Paper* 21/08.

Duval, R., Furceri, D., and 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.

Giannone, D., Lenza, M., and Primiceri, G. E. (2021). Economic predictions with big data: The illusion of sparsity. *Econometrica*, 89, 2409–2437.

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. (1986). Bayes estimators and asymptotic efficiency in parametric counting process models. *Scandinavian Journal of Statistics*, 13, 63–85.

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

Hoerl, A. E., and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12, 55–67.

Ishwaran, H., and Rao, J. S. (2003). Discussion to "Frequentist model average estimators" and "The focussed information criterion" by Hjort, N. L. and Claeskens, G. *Journal of the American Statistical Association*, 98, 922–925.

Kumar, K., and Magnus, J. R. (2013). A characterization of Bayesian robustness for a normal location parameter. *Sankhyā: The Indian Journal of Statistics*, 75, 216–237.

Leeb H., and Pötscher B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, 21, 21–59.

Lindley, D. V. (1968). The choice of variables in multiple regression (with discussion). *Journal of the Royal Statistical Society, Series B*, 30, 31–66.

Magkonis, G., Zekente, K.-M., and Logothetis, V. (2021). Does the left spend more? An econometric survey of partisan politics. *Oxford Bulletin of Economics and Statistics*, 83, 1077–1099.

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., and Wang, W. (2014). Concept-based Bayesian model averaging and growth empirics. *Oxford Bulletin of Economics and Statistics*, 76, 874–897.

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

Mitchell, T. J., and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83, 1023–1032.

Pericchi, L. R., and Smith, A. F. M. (1992). Exact and approximate posterior moments for a normal location parameter. *Journal of the Royal Statistical Society (Series B)*, 54, 793–804.

Pötscher B. M. (1991). Effects of model selection on inference. *Econometric Theory*, 7, 163–185.

Raftery, A. E., and Zheng, Y. (2003). Discussion: Performance of Bayesian model averaging. *Journal of the American Statistical Association*, 98, 931–938.

Robbins, H. (1956). An empirical Bayes approach to statistics. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, vol. I, pp. 157–163. University of California Press, Berkeley and Los Angeles, CA.

Sansó, B., and Pericchi, L. R. (1992). Near ignorance classes of log-concave priors for the location model. *Test*, 1, 39–46.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58, 267–288.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge, UK.

Wang, M., Zhang, X., Wan, A. T. K., and Zou, G. (2019). On the asymptotic distribution of model averaging based on information criterion. arXiv:1910.12208.

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

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

Zhang, X., Zou, G., Liang, H., and Carroll, R. J. (2020). Parsimonious model averaging with a diverging number of parameters. *Journal of the American Statistical Association*, 115, 972–984.

Zhu, R., Wan, A. T. K., Zhang, X. and Zou, G. (2019). A Mallows-type model averaging estimator for the varying-coefficient partially linear model. *Journal of the American Statistical Association*, 114, 882–892.