EIEF

EINAUDI INSTITUTE FOR ECONOMICS AND FINANCE

EIEF Working Paper 20/03

March 2020

Sampling properties of the Bayesian posterior mean with an application to WALS estimation

by

Giuseppe De Luca

(University of Palermo)

Jan R. Magnus

(Vrije Universiteit Amsterdam)

Franco Peracchi

(Georgetown University and EIEF)

Sampling properties of the Bayesian posterior mean with an application to WALS estimation[*]

Giuseppe De Luca
University of Palermo, Palermo, Italy

Jan R. Magnus
Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

Franco Peracchi
Georgetown University, Washington, USA

March 4, 2020

Abstract

Many statistical and econometric learning methods rely on Bayesian ideas, often applied or reinterpreted in a frequentist setting. Two leading examples are shrinkage estimators and model averaging estimators, such as weighted-average least squares (WALS). In many instances, the accuracy of these learning methods in repeated samples is assessed using the variance of the posterior distribution of the parameters of interest given the data. This may be permissible when the sample size is large because, under the conditions of the Bernstein–von Mises theorem, the posterior variance agrees asymptotically with the frequentist variance. In finite samples, however, things are less clear. In this paper we explore this issue by first considering the frequentist properties (bias and variance) of the posterior mean in the important case of the normal location model, which consists of a single observation on a univariate Gaussian distribution with unknown mean and known variance. Based on these results, we derive new estimators of the frequentist bias and variance of the WALS estimator in finite samples. We then study the finite-sample performance of the proposed estimators by a Monte Carlo experiment with design derived from a real data application about the effect of abortion on crime rates.

Keywords: Normal location model; posterior moments and cumulants; higher-order delta method approximations; double-shrinkage estimators; WALS.

JEL classification: C11, C13, C15, C52, I21.

[*] Corresponding author: Giuseppe De Luca (giuseppe.deluca@unipa.it). We thank Domenico Giannone and Giorgio Primiceri for useful discussions.

1 Introduction

Many statistical and econometric learning methods rely on Bayesian ideas, often applied or reinterpreted in a frequentist setting. Examples include shrinkage estimators, such as ridge regression (Hoerl and Kennard 1970), smoothing splines (Reinsch 1967), and the least absolute shrinkage and selection operator (LASSO) introduced by Tibshirani (1996). They also include Bayesian model averaging estimators reinterpreted in a frequentist setting (see, e.g., Raftery et al. 1997, and Clyde 2000), as well as Bayesian-frequentist 'fusions' such as the Bayesian Averaging of Classical Estimates (BACE) of Sala-i-Martin et al. (2004), the weighted-average least squares (WALS) estimator of Magnus et al. (2010), and the Bayesian averaging of maximum likelihood (BAML) estimators of Moral-Benito (2012).

In many instances, the accuracy of these learning methods in repeated samples is assessed using the variance of the posterior distribution of the parameters of interest given the data. This may be permissible when the sample size is large because, under the conditions of the Bernstein–von Mises theorem (van der Vaart 1998), the posterior variance agrees asymptotically with the frequentist variance. In finite samples, however, things are much less clear.

To explore these issues we first consider the frequentist properties (bias and variance) of the posterior mean — the Bayesian point estimator under quadratic loss — in the stylized but important case represented by the normal location model, which consists of a single observation on a univariate Gaussian distribution with unknown mean and known variance. The results of our finite-sample analysis are perhaps somewhat counterintuitive, as they may seem in contradiction with the large sample implications of the Bernstein–von Mises theorem. We show that, for any positive and bounded prior, the posterior variance can be interpreted as a first-order delta method (DM) approximation to the standard deviation, not the variance, of the posterior mean under repeated sampling. This important result is not new, as it is immediate for the posterior mean under conjugate Gaussian priors, for which the DM approximation is exact, and follows easily from available results on the posterior cumulant-generating function (Pericchi et al. 1993) or from the general accuracy formula in Efron (2015). However, it has received little attention in the statistical and econometric literature.

We extend this result by deriving analytical DM approximations of any order to the bias and variance of the posterior mean in the normal location model. Such approximations depend crucially on higher-order posterior cumulants, so we also offer a recursive formula which facilitates the nontrivial task of computing these summaries of the posterior distribution. We describe how the DM approximations help to better understand the link between the frequentist and Bayesian approaches to inference, and how higher-order posterior cumulants contribute to improve the accuracy of our approximations. In addition to analytical comparisons, we evaluate numerically the importance of the higher-order refinement terms by focusing on specific prior densities (Gaussian, Laplace, Weibull, and Subbotin) in the class of (reflected) generalized gamma distributions. We show that posterior skewness and (excess) kurtosis lead to sizable adjustments in the second and third-order DM approximations to the bias and variance of the posterior mean. Moreover, as the order of the expansion increases, the approximated bias and variance profiles converge to those obtained via our Monte Carlo tabulations.

Since sampling moments of the posterior mean depend in general on the unknown location parameter, we discuss two plug-in methods for estimating the bias and variance of the posterior mean, respectively based on the (frequentist) ML estimator and the (Bayesian) posterior mean. In finite samples, choosing between the two methods raises a bias-precision trade-off: the plug-in ML estimators have better risk performance for sufficiently large values of the location parameter, while the plug-in Bayesian estimators have better risk performance for sufficiently small values of the location parameter. The DM approximation to the bias of the posterior mean suggests that the plug-in Bayesian estimators can be interpreted as double-shrinkage estimators because of the double evaluation of the posterior mean function in the leading term of the estimated bias.

The normal location model plays an important role in the WALS estimator introduced by Magnus et al. (2010) to deal with the problem of uncertainty about the regressors in a Gaussian linear model. WALS is a Bayesian combination of frequentist estimators and has been shown to enjoy important theoretical and computational advantages over other strictly Bayesian or strictly frequentist model-averaging estimators (Magnus and De Luca 2016). After implementing preliminary transformations of the regressors, the parameters of each model are estimated by constrained least squares under a frequentist perspective, while the weighting scheme is developed under a Bayesian perspective to obtain desirable theoretical properties, such as admissibility, bounded risk, robustness, near-optimality in terms of minimax regret, and a proper treatment of ignorance. One problem of this 'Bayesian-frequentist fusion' is the difficulty in evaluating the sampling properties of the resulting estimator in finite samples.

Our results for the normal location model can be directly applied to estimating the frequentist bias and variance of the WALS estimator. We show that the previous estimator of its sampling variance is upward biased, that is, the WALS estimator is more precise than originally thought. Our estimators of bias are also new and can be useful for several inferential purposes, such as constructing a bias-corrected WALS estimator or constructing WALS confidence intervals for the regression coefficients. Here we emphasize a particular usage of the estimated bias that is typically ignored in applied work, namely its role in assessing the precision of biased estimators.

We study these issues in an empirical application that looks at the effect of legalized abortion on crime rates (Donohue and Levitt 2001). In this example, we find that WALS estimates are qualitatively similar to the post-double selection estimates obtained by Belloni et al. (2014) and their more recent follow-up studies. Unlike these studies, however, we focus on the importance of evaluating biased estimators in terms of mean squared error (MSE) to show how the traditional approach of comparing only the standard errors may lead to misleading conclusions. Further, we assess the finite-sample performance of the new estimators of the bias and variance of the WALS estimator by a Monte Carlo experiment whose design is based on the above real data application about the effect of legalized abortion on crime rates.

The remainder of the paper is organized as follows. Section 2 derives recursive formulae for the posterior moments and the posterior cumulants in the normal location model. Section 3 shows how these results can then be used to assess the frequentist properties of the posterior mean. Section 4 evaluates these issues numerically based on specific priors in the class of (reflected) generalized gamma distributions. Section 5 introduces the WALS approach to model averaging. Section 6 applies the results of the previous three sections to investigate the frequentist properties (bias and variance) of the WALS estimator in finite samples. Section 7 presents our empirical application, while Section 8 presents our Monte Carlo experiment. Section 9 concludes.

2 Posterior moments for the normal location model

Consider drawing a single observation x from the Gaussian (univariate) distribution with unknown mean η and known variance which, without loss of generality, we set equal to one. The problem of estimating η from x is known as the *normal location problem.* The likelihood function for this model is equal to $\phi(x-\eta)$, where $\phi(\cdot)$ denotes the density of the standard Gaussian distribution. In the Bayesian analysis of the problem, uncertainty (or prior information) about η is represented by a proper prior density $\pi(\cdot)$ on $\mathbb{R}$ which is assumed to be positive and bounded. Defining the functions

$$A_h(x) = \int_{-\infty}^{\infty} (x-\eta)^h \phi(x-\eta)\, \pi(\eta)\, d\eta \qquad (h = 0, 1, \dots),$$

we can combine the likelihood and the prior to obtain the posterior density of η given x,

$$p(\eta|x) = \frac{\phi(x-\eta)\, \pi(\eta)}{A_0(x)}.$$

Our first result provides a recursive formula for the posterior moments of η given x.

Proposition 1 *Given an observation $x \sim \mathcal{N}(\eta, 1)$ and a prior density $\pi(\eta) \geq 0$ which is bounded for $\eta \in \mathbb{R}$, the hth (noncentral) posterior moment of η is given by*

$$\mu_h(x) = \mathbb{E}[\eta^h|x] = \int_{-\infty}^{\infty} \eta^h\, p(\eta|x)\, d\eta = g_h(x) - \sum_{j=1}^{h} (-1)^j \binom{h}{j} x^j \mu_{h-j}(x) \qquad (h = 1, 2, \dots),$$

where $\mu_0(x) = 1$ and $g_h(x) = (-1)^h\, \mathbb{E}[(x-\eta)^h|x] = (-1)^h A_h(x)/A_0(x)$ for $h = 0, 1, \dots$.

Proposition 1 generalizes earlier results by Pericchi and Smith (1992) about the posterior mean and variance to posterior moments of any order. In particular, the posterior mean and variance are given by

$$m(x) = \mathbb{E}[\eta|x] = x + g_1(x), \qquad v^2(x) = \mathbb{E}[\eta^2|x] - [m(x)]^2 = g_2(x) - g_1^2(x),$$

where we used the fact that $\mathbb{E}[\eta^2|x] = g_2(x) + 2xg_1(x) + x^2$. As noted by Kumar and Magnus (2013), the first derivatives of the functions $A_h(x)$ and $g_h(x)$ satisfy the recursions

$$A_h'(x) = \frac{dA_h(x)}{dx} = hA_{h-1}(x) - A_{h+1}(x)$$

and

$$g_h'(x) = (-1)^h \left[\frac{A_h'(x)}{A_0(x)} - \frac{A_0'(x)}{A_0(x)} \frac{A_h(x)}{A_0(x)}\right] = g_{h+1}(x) - g_1(x)g_h(x) - hg_{h-1}(x) \qquad (1)$$

with $A_0'(x) = -A_1(x)$ and $g_0'(x) = 0$ as starting values. Hence, in agreement with Pericchi and Smith (1992, Theorem 1) and Kumar and Magnus (2013, Lemma 3.4), we can also write

$$m(x) = x + \frac{d \log A_0(x)}{dx}, \qquad v^2(x) = m'(x) = 1 + \frac{d^2 \log A_0(x)}{dx^2}.$$

Because $v^2(x)$ is positive, this result implies that the posterior mean $m(x)$ is increasing in x.

Proposition 1 can be also viewed as a special case of the posterior moment-generating function derived by Pericchi et al. (1993, Proposition 2.1) under the more general setup where the likelihood comes from the exponential family. Our recursive formula is restricted to the Gaussian likelihood, which has the practical advantage of facilitating the computation of higher-order posterior moments. Similar considerations extend to the following proposition which provides a recursive formula for the derivatives of the posterior mean and therefore complements the results about the posterior cumulant-generating function derived by Pericchi et al. (1993, Proposition 2.2).

Proposition 2 *If $m(x) = x + g_1(x)$ is the posterior mean of η given x, then*

$$m^{(h)}(x) = \frac{d^h m(x)}{dx^h} = c_{h+1}(x) \qquad (h = 1, 2, \dots),$$

where $c_h(x)$ denotes the hth posterior cumulant of η given x. Moreover, $c_1(x) = g_1(x)$ and

$$c_{h+1}(x) = g_{h+1}(x) - \sum_{j=0}^{h-1} \binom{h}{j} c_{j+1}(x) g_{h-j}(x) \qquad (h = 1, 2, \dots). \tag{2}$$

Higher-order posterior cumulants play a role in the areas of Bayesian robustness and approximation. They are also important because empirical data are often characterized by skewed distributions with fat tails (Fernández and Steel 1998). In the next section we emphasize another important role of higher-order posterior cumulants which has received little attention in the Bayesian literature, namely the fact that they help assessing the frequentist properties of the posterior mean.

3 Sampling properties of the posterior mean

Under quadratic loss, the posterior mean $m(x) = x + g_1(x)$ is the Bayesian point estimator of η. Now suppose that, irrespective of its Bayesian provenance, $m(x)$ is used as an estimator of η. The idea of reinterpreting Bayesian learning methods in a frequentist setting is in fact quite common in statistics and econometrics. Other examples include shrinkage estimators, such as ridge regression, smoothing splines, and the LASSO; and model averaging estimators, such as WALS.

Like Efron (2015) we wish to study the accuracy of the posterior mean from a frequentist perspective. The sampling bias and variance of $m(x)$ are defined, respectively, as

$$\delta(\eta) = \mathbb{E}[m(x)|\eta] - \eta = \mathbb{E}[g_1(x)|\eta], \qquad \sigma^2(\eta) = \mathbb{E}\left[m^2(x)|\eta\right] - (\mathbb{E}[m(x)|\eta])^2,$$

and its mean squared error as $\text{MSE}(\eta) = \mathbb{E}\left[(m(x) - \eta)^2|\eta\right] = \sigma^2(\eta) + \delta^2(\eta)$. We wish to estimate $\delta(\eta)$ and $\sigma^2(\eta)$. In deciding on suitable estimators two problems occur. First, except for the Gaussian prior, the sampling moments of $m(x)$ do not admit closed-form expressions, and hence we have to either approximate or simulate these moments. Second, the moments depend in general on the unknown location parameter η, and hence we have to replace η by some estimator $\hat{\eta}$. We discuss these two issues in Sections 3.1 and 3.2, respectively.

3.1 Approximations to the sampling moments of $m(x)$

Let us start by assuming that the prior is Gaussian with zero mean and finite variance $\omega^2 > 0$. This is convenient because the posterior is then also Gaussian with mean $m(x) = wx$ and variance $v^2(x) = w$, where $w = \omega^2/(1 + \omega^2)$. If we now think of $m(x) = wx$ as a frequentist estimator of η, then the bias and variance of $m(x)$ are

$$\delta(\eta) = (w - 1)\eta = -\frac{\eta}{1 + \omega^2}, \qquad \sigma^2(\eta) = w^2 = \frac{\omega^4}{(1 + \omega^2)^2}, \tag{3}$$

respectively. Notice that the posterior variance $v^2(x) = w$ of η is equal to the standard deviation $\sigma(\eta) = w$ of $m(x)$. This may seem a peculiar feature of the conjugate Gaussian prior, but it isn't. In fact, the result holds approximatively for any positive and bounded prior density; see Section 3.1.1.

The Gaussian prior is convenient but often unsuitable, because the difference $x - m(x) = (1 - w)x$ does not vanish when $x \to \infty$, but rather increases linearly in x. In other words, a Gaussian prior is not discounted when confronted with an observation with which it drastically disagrees and, in this sense, is regarded as nonrobust for the normal location model (see, e.g., Kumar and Magnus 2013 and the large literature quoted therein). Equivalently from (3), the bias $\delta(\eta)$ of $m(x)$ is a linear, hence unbounded, function of η.

To ensure that $\delta(\eta)$ is a bounded function of η, $m(x)$ must be a nonlinear function of x, which raises the issue of how to approximate its sampling moments. Two general approaches to this problem are analytical delta method approximations, discussed in Section 3.1.1, and numerical Monte Carlo approximations, discussed in Section 3.1.2. In both cases, we ignore for the moment the fact that η in unknown and needs to be estimated; this issue is addressed in Section 3.2.

3.1.1 Delta method approximation

The following proposition presents our main result on the analytical delta method (DM) approach.

Proposition 3 *If the posterior mean $m(x)$ is used as estimator of η, then the delta method approximations of order $h+1$ ($h \geq 1$) to its bias and sampling variance are given recursively by*

$$\widehat{\delta}_{h+1}(\eta) = \widehat{\delta}_h(\eta) + q_{h+1}c_{h+2}(\eta), \qquad \widehat{\sigma}^2_{h+1}(\eta) = \widehat{\sigma}^2_h(\eta) + Q_{h+1}c_{h+2}(\eta),$$

where

$$Q_{h+1} = \left(\binom{2h+2}{h+1}q_{2h+2} - q^2_{h+1}\right)c_{h+2}(\eta) + 2\sum_{j=1}^{h}\left(\binom{h+1+j}{h+1}q_{h+1+j} - q_{h+1}q_j\right)c_{j+1}(\eta)$$

and

$$q_j = \begin{cases} \dfrac{1}{2^{j/2}(j/2)!} & \text{if } j \text{ even,} \\ 0 & \text{if } j \text{ odd.} \end{cases}$$

The starting values are

$$\widehat{\delta}_1(\eta) = m(\eta) - \eta, \qquad \widehat{\sigma}^2_1(\eta) = c^2_2(\eta),$$

and $m(\eta) = [m(x)]_{x=\eta}$ and $c_j(\eta) = [c_j(x)]_{x=\eta}$ denote, respectively, the posterior mean and the posterior cumulant of order j evaluated at $x = \eta$.

Proposition 3 shows that the bias and variance of the posterior mean depend crucially on the higher-order posterior cumulants of η. In particular, for $h = 2$, we have

$$\widehat{\delta}_2(\eta) = \widehat{\delta}_1(\eta) + \frac{1}{2}c_3(\eta), \qquad \widehat{\sigma}^2_2(\eta) = \widehat{\sigma}^2_1(\eta) + \frac{1}{2}c^2_3(\eta),$$

and, for $h = 3$,

$$\widehat{\delta}_3(\eta) = \widehat{\delta}_2(\eta), \qquad \widehat{\sigma}_3^2(\eta) = \widehat{\sigma}_2^2(\eta) + \frac{5}{12}c_4^2(\eta) + c_2(\eta)c_4(\eta).$$

Defining the posterior skewness and (excess) kurtosis as

$$\tau(x) = \frac{c_3(x)}{\left(v^2(x)\right)^{3/2}}, \qquad \kappa(x) = \frac{c_4(x)}{\left(v^2(x)\right)^2},$$

the second- and third-order approximations ($h = 2$ and $h = 3$) can be written equivalently as

$$\widehat{\delta}_2(\eta) = m(\eta) - \eta + \frac{1}{2}\tau(\eta)v^3(\eta), \qquad \widehat{\sigma}_2^2(\eta) = v^4(\eta)\left(1 + \frac{1}{2}\tau^2(\eta)v^2(\eta)\right)$$

and

$$\widehat{\delta}_3(\eta) = \widehat{\delta}_2(\eta), \qquad \widehat{\sigma}_3^2(\eta) = v^4(\eta)\left(1 + \frac{1}{2}\tau^2(\eta)v^2(\eta) + \kappa(\eta)v^2(\eta) + \frac{5}{12}\kappa^2(\eta)v^4(\eta)\right), \tag{4}$$

where $v(\eta) = [v(x)]_{x=\eta}$, $\tau(\eta) = [\tau(x)]_{x=\eta}$, and $\kappa(\eta) = [\kappa(x)]_{x=\eta}$.

We see that $\widehat{\sigma}_1^2(\eta) = v^4(\eta)$ coincides with the first-order approximation to $\sigma^2(\eta)$ obtained by the general accuracy formula of Efron (2015, Theorem 1). Thus, for any positive and bounded prior, the posterior variance represents an approximation to the standard deviation, not the variance, of $m(x)$. At first sight, this result may seem counter-intuitive and in contradiction with the large sample implications of the Bernstein–von Mises theorem. As shown in Appendix B for the Gaussian and Laplace priors, this apparent contradiction is due to the fact that, when $n > 1$, the posterior variance and the sampling variance of $m(x)$ are both of order n^{-1} and both depend on additional terms that converge to zero as $n \to \infty$. Thus, asymptotically, the two variances coincide.

The second-order expansion generalizes the first-order DM approximations $\widehat{\delta}_1(\eta)$ and $\widehat{\sigma}_1^2(\eta)$ by introducing some additional terms which depend on the third posterior cumulant, that is, on the posterior variance and the posterior skewness. The sign of the additional term in $\widehat{\delta}_2(\eta)$ depends on the sign of $\tau(\eta)$, while the additional term in $\widehat{\sigma}_2^2(\eta)$ is always nonnegative, so that $\widehat{\sigma}_2^2(\eta) \geq \widehat{\sigma}_1^2(\eta)$ for any $\eta \in \mathbb{R}$.

The third-order expansion does not further improve the DM approximation to $\delta(\eta)$, because $\widehat{\delta}_3(\eta) = \widehat{\delta}_2(\eta)$ due to the fact that $q_3 = 0$. The additional term in $\widehat{\sigma}_3^2(\eta)$ depends on the second and fourth posterior cumulants, that is on the posterior variance and the posterior (excess) kurtosis. We shall see in Section 4 that this term can be either positive or negative and may lead to a substantial improvement in the accuracy of DM approximations to $\sigma^2(\eta)$.

3.1.2 Monte Carlo approximation

The DM approximations provide closed-form relationships which help us to better understand the link between the frequentist and Bayesian approaches to inference. But what can we say about their accuracy? Which order of the expansion is sufficient in practice? And how sensitive are these issues to alternative choices of the prior density $\pi(\eta)$? We shall address these questions in the numerical analysis of Section 4.

Monte Carlo (MC) simulation offers an alternative approach for tabulating the unknown functional forms of $\delta(\eta)$ and $\sigma^2(\eta)$. The advantage over analytical DM expansions is that MC approximation errors can be made arbitrarily small using a sufficiently large number of independent draws from the $\mathcal{N}(\eta, 1)$ distribution. To implement the simulations we employ the following algorithm:

(i) For a given value of η, generate a vector $\widetilde{x} = (\widetilde{x}_1, \ldots, \widetilde{x}_J)$ of J independent draws from the $\mathcal{N}(\eta, 1)$ distribution.

(ii) Given a prior $\pi(\eta)$, compute the values of the posterior mean $\widetilde{m}_j = m(\widetilde{x}_j)$ for each element $\widetilde{x}_j$ of $\widetilde{x}$ $(j = 1, \ldots, J)$, then compute $M_1 = \sum_{j=1}^J \widetilde{m}_j/J$ and $M_2 = \sum_{j=1}^J \widetilde{m}_j^2/J$, and approximate the bias and variance of $m(x)$ by $\widetilde{\delta}_J(\eta) = M_1 - \eta$ and $\widetilde{\sigma}_J^2(\eta) = M_2 - M_1^2$, respectively.

(iii) Repeat the first two steps for selected values of η in a known interval $[\eta_1, \eta_2]$ with a given stepsize $\Delta\eta$ and store the values of η, $\widetilde{\delta}_J(\eta)$ and $\widetilde{\sigma}_J^2(\eta)$.

When the prior density is symmetric around zero, the posterior density is symmetric around its mean and the same tabulations with an opposite sign of $\widetilde{\delta}_J(\eta)$ are valid for positive and negative values of η. In that case we restrict the algorithm to positive values of η by setting $\eta_1 = 0$.

3.2 Plug-in estimators of the sampling moments of $m(x)$

So far we have discussed how to approximate the functional forms of $\delta(\eta)$ and $\sigma^2(\eta)$, either by analytical DM expansions or by numerical MC tabulations. These approximations depend however on η, which is unknown. So we have to replace η by an estimator, say $\widehat{\eta}$, which is function of x.

For the third-order DM approximations (4) the plug-in estimators of $\delta(\eta)$ and $\sigma^2(\eta)$ are

$$\widehat{\delta}_{3,\widehat{\eta}}(x) = m(\widehat{\eta}) - \widehat{\eta} + \frac{1}{2}\tau(\widehat{\eta})v^3(\widehat{\eta}), \quad \widehat{\sigma}^2_{3,\widehat{\eta}}(x) = v^4(\widehat{\eta})\left(1 + \frac{1}{2}\tau^2(\widehat{\eta})v^2(\widehat{\eta}) + \kappa(\widehat{\eta})v^2(\widehat{\eta}) + \frac{5}{12}\kappa^2(\widehat{\eta})v^4(\widehat{\eta})\right).$$

We shall consider two estimators of η. First, $\widehat{\eta} = x$, which is the most common choice (see, e.g., Efron 2015), because x is the unbiased maximum likelihood (ML) estimator of η. This leads to the 'delta method maximum likelihood' (DMML) estimators $\widehat{\delta}_{3,x}(x)$ and $\widehat{\sigma}^2_{3,x}(x)$ of $\delta(\eta)$ and $\sigma^2(\eta)$. Second, $\widehat{\eta} = m(x)$, which is the Bayesian point estimator of η and leads to the 'delta method double-shrinkage' (DMDS) estimators (because of the double evaluation of the posterior mean function in the leading term of the estimated bias) $\widehat{\delta}_{3,m}(x)$ and $\widehat{\sigma}^2_{3,m}(x)$ of $\delta(\eta)$ and $\sigma^2(\eta)$.

The choice between the DMML and DMDS estimators of $\delta(\eta)$ and $\sigma^2(\eta)$ is similar to the choice between x and $m(x)$ as an estimator of η (see, e.g., Magnus and De Luca 2016) and is motivated by finite-sample considerations about their bias-precision trade-off. The ML estimator x has zero bias and unit variance for all values of η. Under quadratic loss, its risk has good properties when $|\eta|$ is large, but not when η is close to zero. The posterior mean $m(x)$ is biased, but it has good risk properties around $|\eta| = 1$, which is the value of central interest.

Similarly, using the MC tabulations from Section 3.1.2, we can define the 'Monte Carlo maximum likelihood' (MCML) and the 'Monte Carlo double-shrinkage' (MCDS) estimators $\widetilde{\delta}_{J,\widetilde{\eta}}(x)$ and $\widetilde{\sigma}^2_{J,\widetilde{\eta}}(x)$ of $\delta(\eta)$ and $\sigma^2(\eta)$, depending on whether η is estimated by $\widetilde{\eta}(x) = x$ or $\widetilde{\eta}(x) = m(x)$.

4 Numerical results for generalized gamma priors

In Sections 2 and 3 we only assumed that the prior density $\pi(\eta)$ is positive and bounded. To gain further insight on the problem of estimating $\delta(\eta)$ and $\sigma^2(\eta)$, we shall restrict our attention to a flexible and mathematically tractable three-parameter class of priors belonging to the (reflected) generalized gamma distributions:

$$\pi(\eta; a, b, c) = \frac{cb^d}{2\Gamma(d)}|\eta|^{-a}\exp\left(-b|\eta|^c\right) \qquad (\eta \in \mathbb{R}),$$

where $0 \leq a < 1$, $b > 0$, $c > 0$, and $d = (1-a)/c$. In addition to the one-parameter family of Gaussian distributions ($a = 0$, $c = 2$) with mean zero and variance $\omega^2 = (2b)^{-1}$, this class includes as special cases the one-parameter family of Laplace distributions ($a = 0$, $c = 1$) and the two-parameter families of the Subbotin ($a = 0$, also known as the exponential power distribution) and the (reflected) Weibull ($a = 1 - c$) distributions.

The Laplace prior, like the Gaussian prior, admits closed-form expressions for the posterior mean and variance of η given x (Pericchi and Smith 1992):

$$m(x) = x - bh(x), \qquad v^2(x) = 1 + b^2 \left(1 - (h(x))^2\right) - \frac{b\,(1 + h(x))\,\phi(x - b)}{\Phi(x - b)},$$

where $\Phi(\cdot)$ denotes the distribution function of the standard Gaussian distribution, $\psi(x) = [\Phi(-x - b)]/[\Phi(x - b)]$, and $h(x) = [1 - e^{2bx}\psi(x)]/[1 + e^{2bx}\psi(x)]$ is a monotonically increasing bounded function with $h(-x) = -h(x)$, $h(0) = 0$, and $h(\infty) = 1$. Closed-form expressions for arbitrary moments and quantiles of the posterior distribution of η given x in the normal location model with Laplace priors have recently been derived by De Luca et al. (2020). Unlike Gaussian priors, Laplace priors lead to an estimator of η which is admissible and has bounded risk.

The Laplace prior, however, is not robust because $x - m(x) = bh(x) \to b > 0$ as $x \to \infty$, a property that it shares with the Gaussian prior. In contrast, Weibull and Subbotin priors are robust because $x - m(x) \to 0$ as $x \to \infty$ (Kumar and Magnus 2013), but the resulting posterior moments can only be determined numerically, for example through Gauss-Laguerre quadrature methods.

The choice of the free prior parameters is based on two criteria. For all priors, we first fix the parameter b to ensure a proper treatment of ignorance about η. Our notion of ignorance relies upon the concept of neutrality which requires the prior median of η to be zero and the prior median of $|\eta|$ to be one. Magnus and De Luca (2016) show that these conditions hold with $b = .2275$ for the Gaussian prior and $b = \log 2$ for the Laplace and reflected Weibull priors. For the Subbotin prior we don't obtain an explicit value, but neutrality restricts $b = b(c)$ to be a nonlinear function of c.

For the reflected Weibull and Subbotin priors we fix the parameter c on the basis of the minimax regret criterion. Let $m(x; c)$ be the class of posterior means associated with different values of c. Under squared error loss, the regret criterion for this class of estimators is defined as

$$\text{regret}(\eta; c) = \text{risk}(\eta; c) - \frac{\eta^2}{1 + \eta^2} = \int_{-\infty}^{\infty} (m(x; c) - \eta)^2\, \phi(x - \eta) dx - \frac{\eta^2}{1 + \eta^2},$$

where $\eta^2/(1 + \eta^2)$ is the lower bound of the risk of $m(x; c)$. By minimizing the maximum regret criterion, Magnus and De Luca (2016) find that the optimal neutral prior has $c = 0.7995$ ($b = 0.9377$) for the Subbotin distribution and $c = 0.8876$ for the Weibull distribution.

4.1 Accuracy of the DM and MC approximations

We now assess the accuracy of different approximations to $\delta(\eta)$ and $\sigma^2(\eta)$ by comparing the DM approximations of orders $h = 1, 2, 3$ with the MC tabulations based on $J = 100$ and $J = 1,000,000$ draws. For these comparisons we use the Gaussian, Laplace, Subbotin, and Weibull priors described in the previous section. In our implementation of the MC algorithm, we restrict η to the interval $[0, 30]$ with stepsize $\Delta\eta = 0.01$. For the Laplace prior, the computing time of the algorithm with $1,000,000$ draws is about one hour thanks to the closed-form expressions for the posterior moments; for the Subbotin and reflected Weibull priors, the computing time is about one week.

Figures 1 and 2 illustrate the DM and MC approximations to $\delta(\eta)$ and $\sigma^2(\eta)$ for the four priors under consideration. The four panels illustrate the bias $\delta(\eta)$ and the variance $\sigma^2(\eta)$, respectively, of the posterior mean $m(x)$ as an estimator of η in the $x \sim \mathcal{N}(\eta, 1)$ model under alternative choices of the prior density $\pi(\eta)$: Gaussian, Laplace, Weibull, and Subbotin. In each panel, DM1–DM3 represent the first-, second-, and third-order DM approximations to $\delta(\eta)$ and $\sigma^2(\eta)$, respectively; and MC1 and MC2 represent the MC tabulations based on $J = 100$ and $J = 1,000,000$ pseudo-random draws, respectively.

MC tabulations based on $J = 100$ draws are still imprecise, but with $J = 1,000,000$ draws the MC approximation error is of the order 10^{-8} for both $\delta(\eta)$ and $\sigma^2(\eta)$, and we may for all practical purposes take the MC approximation based on $1,000,000$ draws as exact, so that $\widetilde{\delta}_J(\eta) = \delta(\eta)$ and $\widetilde{\sigma}^2_J(\eta) = \sigma^2(\eta)$ for any $\eta \in \mathbb{R}$. For the conjugate Gaussian prior, all DM approximations are exact because the posterior distribution of η given x is also Gaussian. For the other priors, the first- and second-order DM approximations are still poor, but the third-order approximation is already quite close to the truth, and if we increase the order of the expansion further, then the approximated bias and variance profiles converge to the MC profiles based on $J = 1,000,000$ draws.

4.2 Monte Carlo evaluation of plug-in estimators

Suppose we have an estimator $\widehat{\eta}$ of an unknown parameter η and we consider this to be a 'good' estimator, then does it follow that $f(\widehat{\eta})$ is also a good estimator of $f(\eta)$? In general, there is no guarantee. For example, if $\widehat{\eta}$ is an unbiased estimator of η, then $\widehat{\eta}^2$ is not an unbiased estimator of η^2; in fact $\mathbb{E}[\widehat{\eta}^2] \geq \eta^2$ by Jensen's inequality. In our case we have two estimators of η, namely x and $m(x)$, and two functions of interest, namely $\delta(\eta)$ and $\sigma^2(\eta)$. We evaluate the finite-sample performance of the plug-in estimators of $\delta(\eta)$ and $\sigma^2(\eta)$ by a simple Monte Carlo experiment. For any η in the interval $[0, 10]$ with stepsize $\Delta\eta = 0.01$ we generate a vector $x = (x_1, \dots, x_R)$ of $R = 100,000$ independent draws from the $\mathcal{N}(\eta, 1)$ distribution. For any element x_r of x $(r = 1, \dots, R)$, we then compute the MCML estimates $\widetilde{\delta}_{J,x_r} = \widetilde{\delta}_J(x_r)$ and $\widetilde{\sigma}^2_{J,x_r} = \widetilde{\sigma}^2_J(x_r)$ and the MCDS estimates $\widetilde{\delta}_{J,m_r} = \widetilde{\delta}_J(m_r)$ and $\widetilde{\sigma}^2_{J,m_r} = \widetilde{\sigma}^2(m_r)$. Then we approximate the bias and root MSE (RMSE) profiles of the MCML and MCDS estimators using their Monte Carlo replications.

The left and right panels in Figures 3 and 4 illustrate the bias (left) and RMSE (right) profiles of the MCML and MCDS estimators of the bias $\delta(\eta)$ (Figure 3) and the variance $\sigma^2(\eta)$ (Figure 4) of $m(x)$ under the Laplace (upper panels) and Weibull (lower panels) priors.

In Figure 3 we estimate the bias $\delta(\eta)$ of $m(x)$. Note that $\delta(\eta)$ is an odd function, that is, $\delta(-\eta) = -\delta(\eta)$. Under the Laplace prior, $\delta(\eta)$ is nonincreasing and convex for $\eta \geq 0$. This implies that, even though x is unbiased for η, the MCML estimator $\widetilde{\delta}_{J,x}(x)$ of $\delta(\eta)$ will be upward biased due to Jensen's inequality. The estimator is unbiased at $\eta = 0$, where $\delta(\eta) = 0$ and $\widetilde{\delta}_{J,x}(x)$ takes positive and negative values with equal probabilities, and for large values of η (say, $\eta > 6$), where $\widetilde{\delta}_{J,x}(x)$ is roughly constant. Since $m(x)$ is biased towards zero and $\delta(\eta)$ is nonincreasing, the MCDS estimator $\widetilde{\delta}_{J,m}(x)$ presents an additional source of positive bias due to the shrinkage estimation of η. So, from the point of bias we prefer MCML over MCDS. However, from the point of MSE it is less clear. For small and medium values of η (roughly, $\eta < 2$), we prefer MCDS over MCML. Similar considerations apply to the reflected Weibull prior. Since the key value of η is one and our prior implies that $\mathbb{P}(\eta < -1) = \mathbb{P}(-1 < \eta < 0) = \mathbb{P}(0 < \eta < 1) = \mathbb{P}(\eta > 1) = 1/4$, we have a slight preference for the MCDS estimator.

In Figure 4 we estimate the variance $\sigma^2(\eta)$ of $m(x)$. In this case $\sigma^2(\eta)$ is an even function, that is, $\sigma^2(-\eta) = \sigma^2(\eta)$. Under the Laplace prior, $\sigma^2(\eta)$ is nondecreasing and concave for $\eta > 0$. In the case of MCML there is only one source of bias (nonlinearity) due to Jensen's inequality. The bias is positive for small values of η and negative for larger values of η. But in the case of MCDS there are two sources of bias (nonlinearity and shrinkage). Shrinkage implies a negative bias, while nonlinearity implies a positive bias for small values of η and a negative bias for larger values of η.

The net result is positive for small values of η (reaching a maximum at $\eta = 0$) and negative for larger values of η. Regarding the MSE we see, as with the estimation of $\delta(\eta)$, that for small and medium values of η (roughly, $\eta < 2$), we prefer MCDS over MCML. Similar considerations apply to the reflected Weibull prior. We conclude that we have a slight preference of MCDS over MCML.

5 The WALS approach to model averaging

The normal location model plays a crucial role in the WALS approach, which we summarize briefly below; for a fuller description see Magnus and De Luca (2016). The basic framework in WALS is the linear regression model

$$y = X_1\beta_1 + X_2\beta_2 + \epsilon, \tag{5}$$

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. We assume that $k_1 \geq 1$, $k_2 \geq 1$, $X = (X_1, X_2)$ has full column-rank $k = k_1 + k_2 \leq n$, and that the disturbances are independent and identically distributed as $\mathcal{N}(0, \sigma^2 I_n)$, where I_n denotes the identity matrix of order n.

Because of the uncertainty on which auxiliary regressors to include, there are 2^{k_2} possible models that contain all focus regressors and a subset of the auxiliary regressors. We represent the jth model as (5) with the added restriction $R_j^\top \beta_2 = 0$, where R_j denotes a $k_2 \times r_j$ matrix of rank $0 \leq r_j \leq k_2$ such that $R_j^\top = [I_{r_j} : 0]$ or a column-permutation thereof. If $\widehat{\beta}_{1j}$ and $\widehat{\beta}_{2j}$ are the LS estimators of β_1 and β_2 in model j, the model averaging estimators are of the form

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

where the λ_j are data-dependent model weights satisfying the restrictions $0 \leq \lambda_j \leq 1$, $\sum_j \lambda_j = 1$ and $\lambda_j = \lambda_j(M_1 y)$ with $M_1 = I_n - X_1(X_1^\top X_1)^{-1} X_1^\top$.

Unlike other model averaging estimators, the WALS approach exploits a preliminary rescaling of the focus regressors and a semiorthogonal transformation of the auxiliary regressors to reduce the computational burden from order 2^{k_2} to order k_2. Specifically, we rescale X_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 $Z_1^\top Z_1$ are equal to one. We also transform X_2 by defining $Z_2 = X_2\Delta_2\Xi^{-1/2}$ and $\gamma_2 = \Xi^{1/2}\Delta_2^{-1}\beta_2$, where Δ_2 is a diagonal $k_2 \times k_2$ matrix such that all diagonal elements of the symmetric and positive definite matrix $\Xi = \Delta_2 X_2^\top M_1 X_2 \Delta_2$ are equal to one. Since $Z_1\gamma_1 = X_1\beta_1$ and $Z_2\gamma_2 = X_2\beta_2$, the model (5) after these transformations may equivalently be written as

$$y = Z_1\gamma_1 + Z_2\gamma_2 + \epsilon. \tag{6}$$

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

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

where $\widehat{\gamma}_{1r} = (Z_1^\top Z_1)^{-1}Z_1^\top y$, $\widehat{\gamma}_{2u} = Z_2^\top M_1 y$, $Q = (Z_1^\top Z_1)^{-1}Z_1^\top Z_2$, $W = \sum_j \lambda_j W_j$, and $W_j = I_{k_2} - R_j R_j^\top$. Further, the WALS estimators of β_1 and β_2 can be directly obtained by the relationships $\beta_1 = \Delta_1\gamma_1$ and $\beta_2 = \Delta_2\Xi^{-1/2}\gamma_2$.

Second, the equivalence theorem (Magnus and Durbin 1999, Theorem 2) implies that the 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.

Third, the dependence of $\widehat{\gamma}_1$ and $\widehat{\gamma}_2$ on model j is completely captured by the random diagonal $k_2 \times k_2$ matrix $W = \sum_j \lambda_j W_j$, whose diagonal elements w_h are partial sums of the λ_j because 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 $\widehat{\gamma}_1$ and $\widehat{\gamma}_2$ is of order k_2 as we only need to determine the set of k_2 WALS weights w_h, not the considerably larger set of 2^{k_2} model weights λ_j.

Fourth, the components of $\widehat{\gamma}_2 = W\widehat{\gamma}_{2u}$ are shrinkage estimators of the components of γ_2, as $0 \le w_h \le 1$, and the components of $\widehat{\gamma}_{2u} = Z_2^\top M_1 y$ are independent, as $\widehat{\gamma}_{2u} \sim \mathcal{N}(\gamma_2, \sigma^2 I_{k_2})$. Hence, if we strengthen the condition $\lambda_j = \lambda_j(M_1 y)$ and assume that each w_h depends only on the hth component of $\widehat{\gamma}_{2u}$, then the shrinkage estimators in $\widehat{\gamma}_2$ will also be independent. Under this additional assumption, 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? Since the estimation of the variance parameter has little impact on the risk properties of $m(x)$ (Danilov 2005), we also assume that σ^2 is known. The baseline problem of the WALS weighting scheme is then equivalent to the normal location problem studied in Sections 2 and 3.

The WALS weighting scheme is based on a Bayesian approach because of theoretical considerations related to admissibility, bounded risk, robustness, near-optimality in terms of minimax regret, and a proper treatment of ignorance about η. This Bayesian step requires two key ingredients. First, a neutral prior with bounded risk, such as the Laplace, Subbotin, or Weibull priors discussed in Section 4. Second, the k_2-vector of t-ratios $x = \widehat{\gamma}_{2u}/s_u$, where $s_u^2 = y^\top M_1(I_n - Z_2Z_2^\top)M_1y/(n-k)$ is the classical estimator of σ^2 in model (6).

For each of the k_2 components x_h of x, we assume that $x_h \sim \mathcal{N}(\eta_h, 1)$, so the Bayesian approach to the normal location problem yields the posterior means $m_h = m(x_h)$ as estimators of η_h for $h = 1, \dots, k_2$. The WALS estimators of γ_1 and γ_2 are then given by

$$\widehat{\gamma}_1 = \widehat{\gamma}_{1r} - Q\widehat{\gamma}_2, \qquad \widehat{\gamma}_2 = s_u m, \tag{7}$$

and the WALS estimators of β_1 and β_2 by

$$\widehat{\beta}_1 = \Delta_1\widehat{\gamma}_1, \qquad \widehat{\beta}_2 = \Delta_2\Xi^{-1/2}\widehat{\gamma}_2. \tag{8}$$

6 Sampling properties of the WALS estimator

Our summary of the WALS methodology led to the estimators of the γ's and β's in (7) and (8). But how about their sampling variances? Earlier papers on the development of WALS have estimated these variances using the diagonal $k_2 \times k_2$ matrix $V = \operatorname{diag}(v_1^2, \dots, v_{k_2}^2)$ with diagonal elements equal to the posterior variances $v_h^2 = v^2(x_h)$. More precisely, by exploiting the fact that $\widehat{\gamma}_{1r}$ and $\widehat{\gamma}_{2u}$ are independent, the estimated variances of $\widehat{\gamma}_1$ and $\widehat{\gamma}_2$ have been computed as

$$\widehat{\mathbb{V}}[\widehat{\gamma}_1] = s_u^2(Z_1^\top Z_1)^{-1} + Q\,\widehat{\mathbb{V}}[\widehat{\gamma}_2]\,Q^\top, \qquad \widehat{\mathbb{V}}[\widehat{\gamma}_2] = s_u^2 V, \tag{9}$$

and the estimated covariance as $\widehat{\mathbb{C}}[\widehat{\gamma}_1, \widehat{\gamma}_2] = -Q\,\widehat{\mathbb{V}}[\widehat{\gamma}_2]$, where $Q = (Z_1^\top Z_1)^{-1} Z_1^\top Z_2$. As a consequence, the variances of $\widehat{\beta}_1$ and $\widehat{\beta}_2$ have been estimated by

$$\widehat{\mathbb{V}}[\widehat{\beta}_1] = \Delta_1\,\widehat{\mathbb{V}}[\widehat{\gamma}_1]\,\Delta_1, \quad \widehat{\mathbb{V}}[\widehat{\beta}_2] = \Delta_2 \Xi^{-1/2}\,\widehat{\mathbb{V}}[\widehat{\gamma}_2]\,\Xi^{-1/2}\Delta_2 \tag{10}$$

and the covariance by $\widehat{\mathbb{C}}[\widehat{\beta}_1, \widehat{\beta}_2] = \Delta_1\,\widehat{\mathbb{C}}[\widehat{\gamma}_1, \widehat{\gamma}_2]\,\Xi^{-1/2}\Delta_2$.

This, however, is not quite right. As discussed in Section 3, thinking of the posterior variance v_h^2 as the estimated variance of m_h is not correct in a frequentist world unless the sample size is very large, which it isn't because this part of the theory is based on a single observation. In the extreme case of a single observation, v_h^2 represents the first-order DMML estimator of the standard deviation of m_h, so the diagonal elements of V should not be v_h^2 but v_h^4.

The accuracy of the estimators can be further improved by using higher-order DM approximations, which in the limit lead to the MC tabulations studied in Section 3.1.2. Thus, to estimate the sampling variance of WALS, we now use (9) and (10), where the diagonal matrix V is redefined so that its hth diagonal element equals the MCDS estimator $\widetilde{\sigma}^2_{J,m}(x_h)$ (or the MCML estimator $\widetilde{\sigma}^2_{J,x}(x_h)$) of the sampling variance of m_h.

In a similar fashion we now use the plug-in estimators of the biases of the posterior means to estimate the bias (and hence the MSE) of the WALS estimators. For each of the k_2 components m_h of m, we compute first an estimate $\widehat{\delta}_h$ of the bias $\delta_h = \delta(\eta_h)$ of m_h using the MCDS estimate $\widetilde{\delta}_{J,m}(x_h)$ (or the MCML estimate $\widetilde{\delta}_{J,x}(x_h)$). As shown in Section 4.2, these estimators are generally biased but their RMSEs are bounded and their biases are relatively small. For example, under the Laplace prior, we have that $|\,\mathbb{E}[\widehat{\delta}_h] - \delta_h| \leq 0.0528$ for the MCML estimator and $|\,\mathbb{E}[\widehat{\delta}_h] - \delta_h| \leq 0.1457$ for the MCDS estimator (see Figure 3). In both cases, the maximum bias is reached at $|\eta_h| = 1.84$ where $|\delta_h| = 0.5124$, and hence $|\,\mathbb{E}[\widehat{\delta}_h]/\delta_h - 1| = 10.30\%$ for the MCML estimator and $|\,\mathbb{E}[\widehat{\delta}_h]/\delta_h - 1| = 28.43\%$ for the MCDS estimator. This suggests that we can think of $\widehat{\delta}_h$ as a nearly unbiased estimator of δ_h, especially for the MCML estimator. After estimating the bias of m by $\widehat{\delta} = (\widehat{\delta}_1, \dots, \widehat{\delta}_{k_2})$, we estimate the bias $d_2 = \mathbb{E}[\widehat{\gamma}_2] - \gamma_2$ of $\widehat{\gamma}_2$ by $\widehat{d}_2 = s_u \widehat{\delta}$ and the bias $b_2 = \mathbb{E}[\widehat{\beta}_2] - \beta_2$ of $\widehat{\beta}_2$ by $\widehat{b}_2 = \Delta_2 \Xi^{-1/2} \widehat{d}_2$. Provided that the unknown data-generation process (DGP) is nested in the unrestricted model (6), we can also estimate the bias of $\widehat{\gamma}_1$,

$$d_1 = \mathbb{E}[\widehat{\gamma}_1] - \gamma_1 = \mathbb{E}[\widehat{\gamma}_{1r}] - \gamma_1 - Q\,\mathbb{E}[\widehat{\gamma}_2] = Q\gamma_2 - Q(\gamma_2 + d_2) = -Qd_2,$$

by $\widehat{d}_1 = -Q\widehat{d}_2$ and the bias $b_1 = \mathbb{E}[\widehat{\beta}_1] - \beta_1$ of $\widehat{\beta}_1$ by $\widehat{b}_1 = \Delta_1\widehat{d}_1$.

7 Empirical application: Legalization of abortion and crime reduction

In an influential paper, Donohue and Levitt (2001), henceforth DL, used a panel data set of U.S. states from 1985 to 1997 to show that the legalization of abortion in the early 1970s played an important role in explaining the reduction of violent, property, and murder crimes during the 1990s. The evidence in favor of this causal relationship has been questioned in a number of follow-up studies (see, e.g., Foote and Goetz 2008 and Belloni et al. 2014). A major concern is that state-level abortion rates in the early 1970s were not randomly assigned. Thus, failing to control for factors that are associated with state-level abortion and crime rates may lead to omitted variable bias in the estimated effect of interest. In this section, we contribute to this debate by studying the sampling properties of various least squares (LS) and WALS estimators in the context of the flexible specifications proposed by Belloni et al. (2014), henceforth BCH.

The regressor of interest is a measure of the abortion rate relevant for each type of crime, determined by the ages of criminals when they tend to commit crimes. The baseline specification used by DL includes state and time effects as additional controls, plus eight time-varying and state-specific confounding factors (log of lagged prisoners per capita, log of lagged police per capita, per capita income, per capita beer consumption, unemployment rate, poverty rate, generosity of the AFDC welfare program at time $t - 15$, and a dummy for the existence of a concealed weapons law). To reduce serial correlation, BCH eliminate the state effects by analyzing models in first differences. They also introduce a rich set of control variables to account for a nonlinear trend that may depend on time-varying state-level characteristics. In this specification the focus regressors include the first-difference of the abortion rate and a full set of time dummies, while the auxiliary regressors include a total of 294 controls (initial levels and initial differences of the abortion rates, first differences, lagged levels, initial levels, initial differences and within-state averages of the eight controls considered by DL, squares of the aforementioned variables, all interactions of these variables with a quadratic trend, and all interactions among the first-differences of the eight time-varying controls).[1] After deleting Washington D.C. and taking first differences, the analysis is based on a balanced panel of 50 states over a 12-year period. For additional information on data definitions and transformations we refer the reader to the original papers of DL and BCH.

Table 1 shows the estimated coefficients on the first differences of the abortion rates in the models for violent, property, and murder crimes. For each type of crime, we compare the WALS estimates based on the Laplace (WALS-L) and Weibull (WALS-W) priors with the four LS estimates from the unrestricted model that includes all focus and auxiliary regressors (LS-U), the fully restricted model that includes only the focus regressors (LS-R), the intermediate model that includes the focus regressors and the subset of auxiliary regressors corresponding to the first differences of the eight time-varying controls used by DL (LS-I), and the intermediate model that includes the focus regressors and the subset of auxiliary regressors selected by the BCH's double-selection procedure (LS-DS). The LS-U, LS-I and LS-DS estimates coincide with those reported in BCH (Table 1).[2]

In addition to the estimated coefficients, we present the estimated bias, standard error (SE) and RMSE of the various LS and WALS estimators based on the assumption that the unknown DGP is nested in the unrestricted model. The assumption is crucial for most sensitivity analyses where the investigator assesses (formally or informally) whether the estimated coefficients of interest are robust to deviations from a baseline model. This assumption implies that the LS-U estimator is unbiased, so we can estimate the bias of the other LS estimators unbiasedly by the observed differences in the estimated coefficients with respect to the LS-U estimates. For example, we estimate the bias $b_{1r} = \mathbb{E}[\widehat{\beta}_{1r}] - \beta_1 = (X_1^\top X_1)^{-1} X_1^\top X_2 \beta_2$ of the LS-R estimator $\widehat{\beta}_{1r}$ by $\widehat{b}_{1r} = (X_1^\top X_1)^{-1} X_1^\top X_2 \widehat{\beta}_{2u} = \widehat{\beta}_{1r} - \widehat{\beta}_{1u}$. As for the SE (and hence RMSE) of the LS estimators, we report both the classical SE and the SE clustered at the state-level (SE$_c$ and RMSE$_c$). For the WALS estimators, we compute the MCDS and MCML estimates of the bias and the (classical) SE discussed in Section 6, but not the SE clustered at the state-level which would require extending our theoretical results to dependent data. To our knowledge, the problem of computing clustered SE for model averaging estimators is still unexplored. Similarly, very little is known about the SE$_c$ of the LS-DS estimator because

[1] The full set of auxiliary variables used by BCH includes 294 noncollinear variables, not 284 variables as incorrectly reported in their papers. In practice, because of a coding error, their Stata program also excludes interactions between squared initial differences of the eight time-varying controls and the quadratic trend terms. For comparability reasons, we use exactly the same controls of BCH.

[2] Unlike BCH, we adopt a common procedure to exclude the collinear controls in the various estimation routines. In the models for property and murder crimes, this leads to small differences in the controls selected by the BCH's double-selection procedure. In turn, we also find small differences in the LS-DS estimate of abortion on murder crime.

the double selection procedure of BCH does not account for serial correlation of the data and the reported SE$_c$ reflects only the effects of clustering in the selected model. An alternative approach could be to compute the LS and WALS estimates after some preliminary data transformation (e.g. Prais-Winsten or Cochrane-Orcutt) which attempts to remove serial correlation from the outcome and the regressors. The underlying WALS theory has been developed in Magnus et al. (2011). However, this alternative approach would assume that the preliminary model needed to estimate the serial correlation coefficients is correctly specified. For simplicity, we shall focus our discussion on the comparisons of the classical SE and RMSE.

In line with previous studies, we find that the small differences between the LS-R and LS-I estimates are basis for the robustness of the results provided by the DL sensitivity analysis. Although unbiased, the LS-U estimator has a large SE. Actually, if we take formally into account the bias-precision trade-off in the choice of the control variables, as suggested by BCH, then this is the worst estimator in terms of RMSE. The BCH double selection procedure drastically reduces the uncertainty due to the choice of the 294 auxiliary variables by selecting a few controls (between 7 and 9) that are important to predict either the outcome or the treatment variable of interest in each model. The SEs of the LS-DS estimator are much lower with respect to the LS-U estimator, but are about twice those of the LS-R and LS-I estimators. Based on these findings, BCH conclude that the empirical evidence in favor of the causal effect of abortion on crime is not robust to the presence of nonlinear trends. However, as it is clear from our results on the estimated bias and RMSE, this conclusion neglects one important point: according to the assumed model space the LS-DS is never preferred to LS-R and LR-I estimators, neither in terms of bias nor in terms of SE. Thus, why should we question the robustness of the DL's findings based on a 'worse' estimator of the coefficient of interest? Probably, trying to control for 294 additional controls in a sample of 600 observations is a very ambitious task for both the LS-U and the LS-DS estimators. Similar considerations extend to the WALS-L and WALS-W estimators, which lead to the same policy implications of the LS-DS estimator. Estimated sampling moments suggest that the WALS estimators are less biased, but also less precise, than the LS-R, LS-I, and LS-DS estimators. In terms of RMSE, the preferred estimators are LS-I/LS-R in the model for property crimes and WALS-W/WALS-L in the model for murder crimes. In the model for violent crimes, these four estimators have similar estimates of the RMSE. It is therefore difficult to establish which is the preferred estimator, which adds ambiguity to the results because different estimators lead to different policy implications.

8 Monte Carlo simulations

In this previous section we estimated parameters of interest in a real-life application. In such an application we don't know the truth. We now turn to MC simulations, where we *do* know the truth. This truth (the DGP) is based on the empirical application in the previous section. Specifically, for each type of crime, we set the parameters of the DGP equal to the unrestricted LS estimates for the model in first differences and then simulate the variation in the crime rates of interest by adding to the estimated linear predictor pseudo-random draws from the Gaussian distribution with mean zero and variance equal to the classical LS estimate s_u^2 of σ^2. We focus on estimating the coefficient on the first-difference of the abortion rate, which under the assumed DGP is equal to 0.071 for violent crimes, -0.161 for property crimes, and -1.327 for murder crimes.

For each model, we compare six estimators of the causal effect of interest: the four LS estimators (LS-U, LS-R, LS-I, and LS-DS) and the two WALS estimators (WALS-L and WALS-W). The true bias, SE and RMSE of each estimator are approximated using 5,000 Monte Carlo replications by using the LS estimates of the unrestricted model as true DGP. For each of these estimators we have one or more methods for estimating the underlying bias and SE: the LS estimators of the biases and SEs of the four LS estimators and the MCDS and MCML estimators of the biases and SEs of the two WALS estimators. In our Monte Carlo experiment we also study the bias, SE and RMSE of the LS, MCDS and MCML estimators of the biases and SEs of the six estimators of the causal effects of interest. Specifically, since each estimator has its own bias and SE, we report the relative bias, SE and RMSE of these three estimators of the sampling moments by taking ratios with respect to the true biases and the true SEs.

Table 2 presents the (true) bias, SE and RMSE of the six estimators of the causal effect for the three models on each type of crime. As expected, the bias of the LS-U estimator is always close to zero, but this estimator is never preferred in terms of RMSE due to its large SE. In line with the sampling moments estimated from the empirical application, we find that the LS-DS estimator is more biased and less precise than the LS-R and LS-I estimators, and that the two WALS estimators have lower bias and higher SE than the LS-R, LS-I and LS-DS estimators. According to the RMSE criterion, the preferred estimators are LS-I/LS-R in the models for violent and property crimes and WALS-L/WALS-W in the model for murder crimes.

In Table 3 we concentrate on estimating the bias. We present the relative bias, SE and RMSE of the LS, MCDS and MCML estimators of the biases of the LS-R, LS-I, LS-DS, WALS-L and WALS-W estimators of the causal effects of interest. Although unbiased, the LS estimators of the biases of the LS-R, LS-I and LS-DS estimators are rather imprecise as they depend directly on the LS estimators of the auxiliary coefficients under the unrestricted model. As predicted from our theoretical results, the MCML estimator of the bias of each WALS estimator is generally less biased than the corresponding MCDS estimator. The latter, however, is always preferred to the other estimators in terms of relative RMSE.

In Table 4 we consider the standard error, and we present the relative bias, SE and RMSE of the LS, MCDS and MCML estimators of the SEs of the six estimators of the causal effects of interest. Here, for the WALS-L and WALS-W estimators, we also report the finite-sample performance of the previously used estimator of the SEs (labeled as PV) which was computed from (9) and (10) using the posterior variances v_h^2 as diagonal elements of V. Our Monte Carlo results confirm that the new MCDS and MCML estimators of the SEs of the WALS estimators reduce the substantial upward bias of the previously used PV estimator. The relative RMSE performances of the new MCDS and MCML estimators of the SEs of the WALS estimators are comparable to those of the LS estimator of the SEs of the correctly specified LS-U estimator.

9 Conclusions

In this paper we have analyzed the finite-sample sampling properties (bias and variance) of the posterior mean in the normal location model using both analytical delta method approximations and numerical Monte Carlo tabulations. Our analytical results have shown how higher-order posterior cumulants contribute to improving the accuracy of delta method approximations to the bias and to the variance of the posterior mean. We have also provided recursive formulae to facilitate the nontrivial task of computing higher-order posterior moments and posterior cumulants, which are in turn the key ingredients needed to derive delta method approximations of any order.

Our numerical results reveal that high-order refinement terms have sizable effects. Moreover, as the order the expansion increases, the approximated bias and variance profiles converge to those resulting from accurate Monte Carlo tabulations. Since sampling moments of the posterior mean depend on the unknown location parameter, we have compared two plug-in strategies for estimating the frequentist bias and variance of the posterior mean: one based on the ML estimator and another on the posterior mean. Our simulations show that the former has a relative advantage in terms of bias and good risk performance for large values of the normal location parameters, while the latter leads to better risk performance for small values of the normal location parameter. The performance of these estimators is relatively unaffected by the prior under consideration and by the nonlinear profiles of the bias and variance of the underlying posterior mean.

Our theoretical and numerical results for the normal location model have direct implications for the sampling properties of the WALS estimator, a partly-Bayesian and partly-frequentist model averaging estimator which accounts for the problem of uncertainty about the regressors in a Gaussian linear model. We have derived estimators of the bias and variance of WALS that are based on considerations about the finite-sample sampling properties of the posterior mean in the normal location model. We illustrate the importance of these developments in a real data application that looks at the effect of legalized abortion on crime rates. Results from a related Monte Carlo experiment also reveal that the new estimators of the bias and variance of WALS have good finite-sample performance. Further work is required to investigate the implications of our findings for the WALS approach to inference (e.g., confidence intervals and testing strategies). Preliminary results in this direction appear to be promising.

References

Belloni A., Chernozhukov V., and Hansen C. (2014). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28: 29–50.

Clyde M. A. (2000). Model uncertainty and health effect studies for particular matter. *Environmetrics*, 11: 745763.

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. (2020). Posterior moments and quantiles for the normal location model with Laplace prior. *Communications in Statistics—Theory and Methods*, forthcoming.

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

Efron B. (2015). Frequentist accuracy of Bayesian estimates. *Journal of Royal Statistical Society: Series B*, 77: 617–646.

Fernández C., and Steel M. F. J. (1998). On Bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association*, 93: 359–371.

Foote C. L., and Goetz C. F. (2008). The impact of legalized abortion on crime: Comment. *Quarterly Journal of Economics*, 123: 407–23.

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

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

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., Wan A. T. K., and Zhang, X. (2011). Weighted average least squares estimation with nonspherical disturbances and an application to the Hong Kong housing market. *Computational Statistics & Data Analysis*, 55: 1331–1341.

Moral-Benito E. (2012). Determinants of economic growth: A Bayesian panel data approach. *Review of Economics and Statistics*, 94: 566–579.

Pericchi L. R., Sansó B., and Smith A. F. M. (1993). Posterior cumulant relationships in Bayesian inference involving the exponential family. *Journal of the American Statistical Association*, 88: 1419–1426.

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.

Raftery A. E., Madigan D., and Hoeting J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Society*, 92: 179–191.

Reinsch C. H. (1967). Smoothing by spline functions. *Numerische Mathematik*, 10: 177–183.

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

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 University Press, New York.

Table 1: Effect of abortion on crime

			Estimated sampling moments					
Type of crime	Estimator	Effect	Method	Bias	SE	SE$_c$	RMSE	RMSE$_c$
Violent	LS-U	0.071	LS	0.000	0.318	0.284	0.318	0.284
	LS-R	−0.157	LS	−0.228	0.046	0.033	0.232	0.230
	LS-I	−0.157	LS	−0.227	0.047	0.034	0.232	0.230
	LS-DS	−0.171	LS	−0.242	0.113	0.117	0.267	0.269
	WALS-L	−0.007	MCDS	−0.046	0.224		0.229	
			MCML	−0.067	0.234		0.244	
	WALS-W	−0.012	MCDS	−0.046	0.223		0.227	
			MCML	−0.067	0.237		0.246	
Property	LS-U	−0.161	LS	0.000	0.135	0.106	0.135	0.106
	LS-R	−0.100	LS	0.061	0.024	0.022	0.066	0.065
	LS-I	−0.106	LS	0.055	0.024	0.021	0.060	0.059
	LS-DS	−0.061	LS	0.100	0.042	0.058	0.108	0.115
	WALS-L	−0.134	MCDS	0.013	0.097		0.098	
			MCML	0.022	0.102		0.104	
	WALS-W	−0.130	MCDS	0.012	0.097		0.098	
			MCML	0.024	0.103		0.106	
Murder	LS-U	−1.327	LS	0.000	1.485	0.932	1.485	0.932
	LS-R	−0.215	LS	1.112	0.184	0.052	1.127	1.113
	LS-I	−0.218	LS	1.109	0.185	0.068	1.124	1.111
	LS-DS	−0.192	LS	1.135	0.416	0.176	1.209	1.149
	WALS-L	−0.849	MCDS	0.220	1.016		1.040	
			MCML	0.386	1.035		1.104	
	WALS-W	−0.783	MCDS	0.213	0.997		1.019	
			MCML	0.411	1.024		1.103	

Notes. LS-U and LS-R are the LS estimators of the effect of interest in the unrestricted and fully restricted models, respectively; LS-I is the LS estimator in the intermediate model with the eight time-varying controls used by DL; LS-DS is the LS estimator in the intermediate model with the subset of controls selected by BCH's double selection procedure; WALS-L and WALS-W are the WALS estimators based on the Laplace and Weibull priors. Estimators of the sampling moments: LS (least squares), MCDS (Monte Carlo double shrinkage), MCML (Monte Carlo maximum likelihood). All models are estimated in first-differences as explained in Section 7.

Table 2: Monte Carlo results for the estimators of the effect of abortion on crime

Type of crime	Effect	Estimator	Bias	SE	RMSE
Violent	0.071	LS-U	−0.001	0.319	0.319
		LS-R	−0.228	0.043	0.232
		LS-I	−0.227	0.043	0.231
		LS-DS	−0.248	0.105	0.269
		WALS-L	−0.066	0.235	0.244
		WALS-W	−0.067	0.237	0.246
Property	−0.161	LS-U	−0.000	0.136	0.136
		LS-R	0.062	0.022	0.065
		LS-I	0.056	0.022	0.060
		LS-DS	0.092	0.043	0.102
		WALS-L	0.022	0.103	0.106
		WALS-W	0.024	0.105	0.108
Murder	−1.327	LS-U	−0.000	1.471	1.471
		LS-R	1.113	0.200	1.131
		LS-I	1.110	0.204	1.129
		LS-DS	1.130	0.442	1.213
		WALS-L	0.385	1.027	1.097
		WALS-W	0.410	1.017	1.097

Notes. See Notes to Table 1.

Table 3: Monte Carlo results for the estimators of the biases of the estimated effects of abortion on crime

			Estimators of the bias			
Type of crime	Estimator	Bias	Method	R.Bias	R.SE	R.RMSE
Violent	LS-R	−0.228	LS	0.005	1.389	1.389
	LS-I	−0.227	LS	0.005	1.392	1.392
	LS-DS	−0.248	LS	0.004	1.223	1.223
	WALS-L	−0.066	MCDS	0.323	0.947	1.000
			MCML	0.127	1.218	1.225
	WALS-W	−0.067	MCDS	0.340	0.927	0.987
			MCML	0.158	1.188	1.199
Property	LS-R	0.062	LS	0.007	2.188	2.188
	LS-I	0.056	LS	0.008	2.421	2.422
	LS-DS	0.092	LS	0.005	1.441	1.441
	WALS-L	0.022	MCDS	−0.398	1.186	1.251
			MCML	−0.132	1.498	1.504
	WALS-W	0.024	MCDS	−0.430	1.087	1.169
			MCML	−0.173	1.369	1.380
Murder	LS-R	1.113	LS	0.000	1.309	1.309
	LS-I	1.110	LS	0.000	1.311	1.311
	LS-DS	1.130	LS	0.000	1.244	1.244
	WALS-L	0.385	MCDS	−0.401	0.768	0.867
			MCML	−0.145	1.046	1.056
	WALS-W	0.410	MCDS	−0.435	0.719	0.841
			MCML	−0.186	0.983	1.000

Notes. See Notes to Table 1. Estimators of the bias: LS (least squares), MCDS (Monte Carlo double shrinkage), MCML (Monte Carlo maximum likelihood).

Table 4: Monte Carlo results for the estimators of the standard errors of the estimated effects of abortion on crime

			Estimators of the SE			
Type of crime	Estimator	SE	Method	R.Bias	R.SE	R.RMSE
Violent	LS-U	0.319	LS	−0.001	0.041	0.041
	LS-R	0.043	LS	0.285	0.035	0.287
	LS-I	0.043	LS	0.282	0.035	0.284
	LS-DS	0.105	LS	0.294	0.039	0.297
	WALS-L	0.235	MCDS	−0.012	0.040	0.042
			MCML	0.039	0.043	0.058
			PV	0.158	0.047	0.165
	WALS-W	0.237	MCDS	−0.017	0.042	0.046
			MCML	0.048	0.046	0.067
			PV	0.149	0.048	0.157
Property	LS-U	0.136	LS	−0.012	0.041	0.042
	LS-R	0.022	LS	0.303	0.035	0.305
	LS-I	0.022	LS	0.295	0.035	0.297
	LS-DS	0.043	LS	0.170	0.061	0.181
	WALS-L	0.103	MCDS	−0.034	0.037	0.050
			MCML	0.016	0.040	0.043
			PV	0.127	0.044	0.134
	WALS-W	0.105	MCDS	−0.042	0.038	0.057
			MCML	0.021	0.041	0.046
			PV	0.113	0.044	0.121
Murder	LS-U	1.471	LS	0.010	0.042	0.043
	LS-R	0.200	LS	0.155	0.033	0.158
	LS-I	0.204	LS	0.146	0.033	0.150
	LS-DS	0.442	LS	0.183	0.034	0.186
	WALS-L	1.027	MCDS	0.025	0.042	0.049
			MCML	0.069	0.047	0.083
			PV	0.203	0.051	0.210
	WALS-W	1.017	MCDS	0.029	0.044	0.053
			MCML	0.089	0.052	0.104
			PV	0.206	0.055	0.213

Notes. See Notes to Table 1. Estimators of the SE: LS (least squares), MCDS (Monte Carlo double shrinkage), MCML (Monte Carlo maximum likelihood), PV (Posterior variance).

Figure 1: DM and MC approximations to the bias $\delta(\eta)$ of the posterior mean $m(x)$ under Gaussian, Laplace, reflected Weibull, and Subbotin priors.

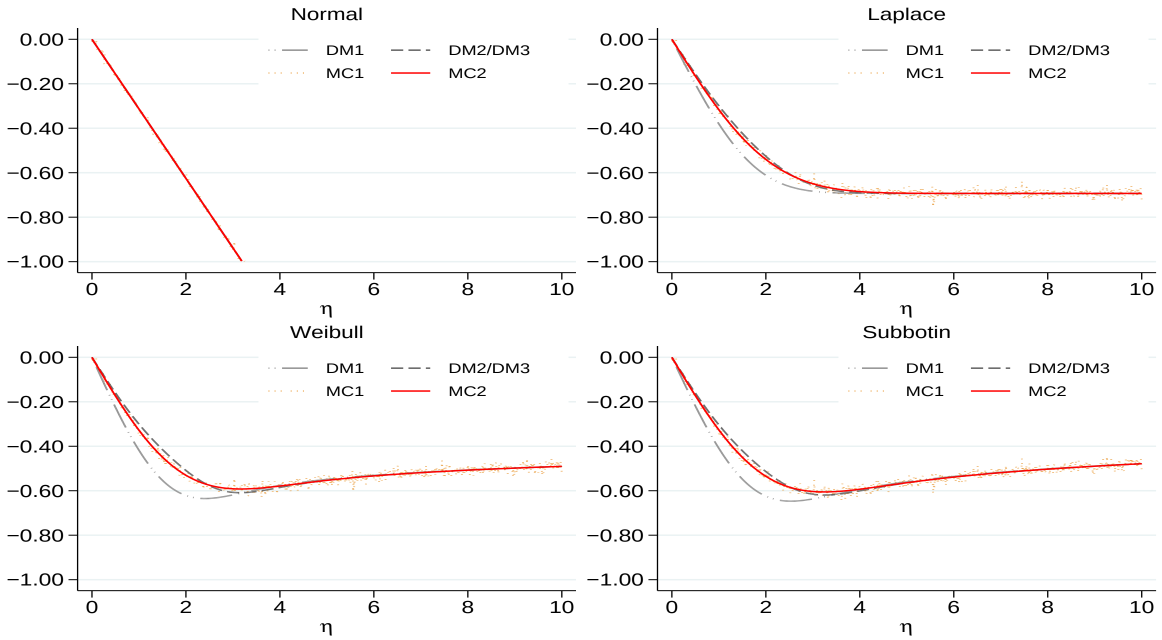


Figure 2: DM and MC approximations to the variance $\sigma^2(\eta)$ of the posterior mean $m(x)$ under Gaussian, Laplace, reflected Weibull, and Subbotin priors.

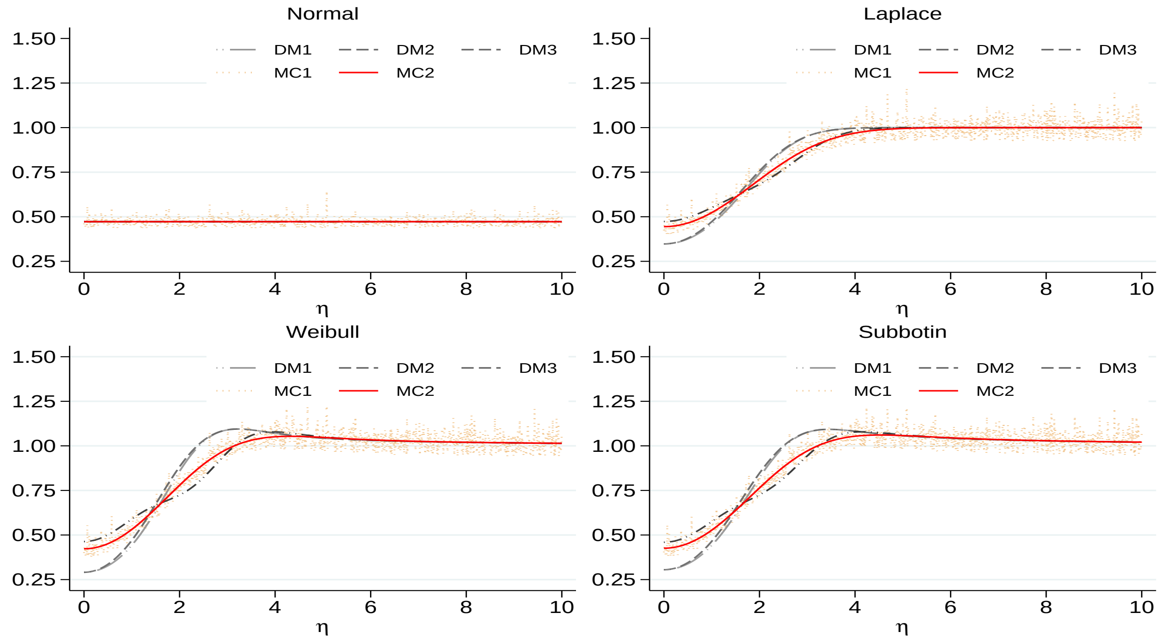


Figure 3: Bias and RMSE of the MCML and MCDS estimators of the bias $\delta(\eta)$ of the posterior mean $m(x)$ under Laplace and reflected Weibull priors.

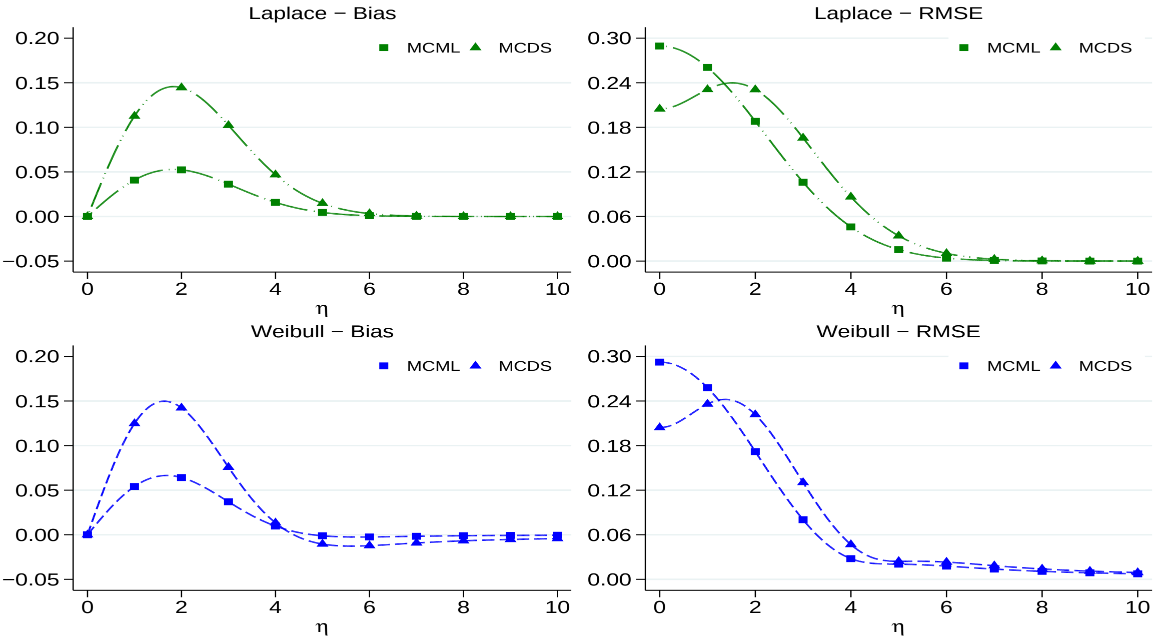
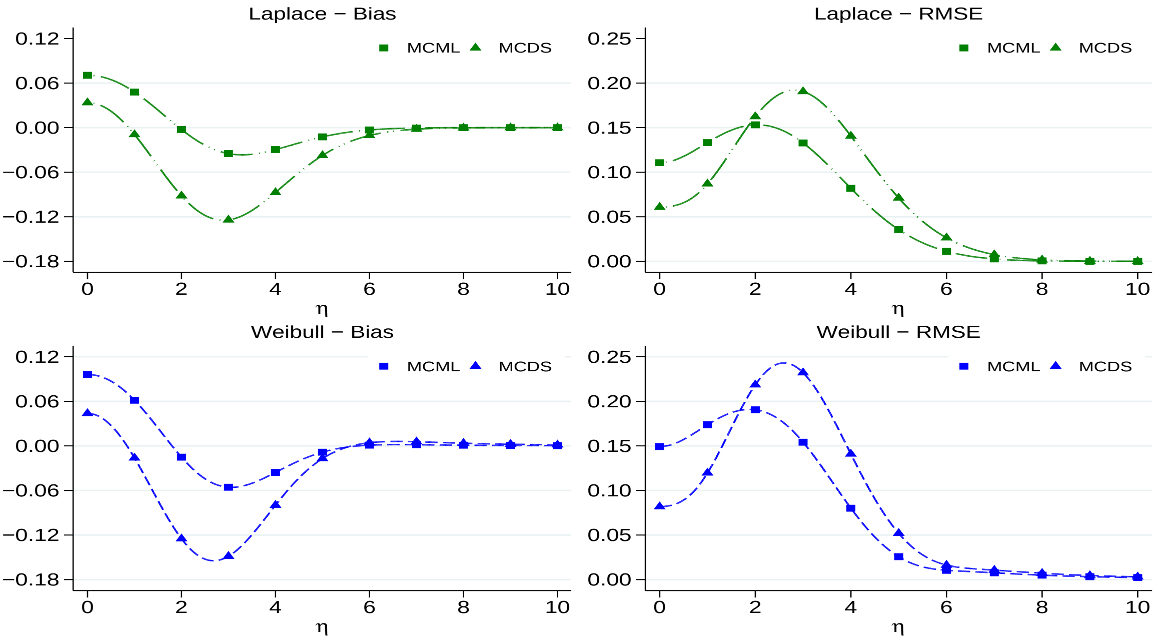


Figure 4: Bias and RMSE of the MCML and MCDS estimators of the sampling variance $\sigma^2(\eta)$ of the posterior mean $m(x)$ under Laplace and reflected Weibull priors.



A Proofs

Proposition 1. The stated assumptions on the prior guarantee that the function $A_h(x)$ exists and admits derivatives of any order (Pericchi and Smith 1992, Appendix A). We have

$$(x-\eta)^h = \sum_{j=0}^{h} \binom{h}{j} x^j (-\eta)^{h-j} = (-1)^h \eta^h + \sum_{j=1}^{h} (-1)^{h-j} \binom{h}{j} x^j \eta^{h-j}$$

from the binomial theorem, so that

$$\eta^h = (-1)^h (x-\eta)^h - \sum_{j=1}^{h} (-1)^j \binom{h}{j} x^j \eta^{h-j}.$$

Taking expectations, conditional on x, the result follows.

Proposition 2. The fact that $c_{h+1}(x) = m^{(h)}(x)$, $h = 1, 2, \ldots$, follows from Pericchi et al. (1993, Proposition 2.2). To prove the recursion (2), we first prove it for $h = 1$ and $h = 2$:

$$\begin{aligned}
c_2 &= m'(x) = 1 + g_1' = 1 + g_2 - g_1^2 - 1 = g_2 - g_1^2 = g_2 - c_1 g_1, \\
c_3 &= m''(x) = g_2' - 2g_1 g_1' = (g_3 - g_1 g_2 - 2g_1) - 2g_1(g_2 - g_1^2 - 1) \\
&= g_3 - g_1 g_2 - 2(g_2 - g_1^2) g_1 = g_3 - c_1 g_2 - 2c_2 g_1,
\end{aligned}$$

with $g_h = g_h(x)$ and $c_h = c_h(x)$. Then we prove that if (2) holds at h and $h+1$, then it also holds at $h+2$. Since $c_1' = c_2 - 1$ and $c_j' = c_{j+1}$ for $j \geq 2$, the recursion (1) implies that

$$\begin{aligned}
c_{h+2} = c_{h+1}' &= g_{h+1}' - c_1' g_h - \sum_{j=1}^{h-1} \binom{h}{j} c_{j+1}' g_{h-j} - \sum_{j=0}^{h-1} \binom{h}{j} c_{j+1} g_{h-j}' \\
&= g_{h+1}' - (c_2 - 1) g_h - \sum_{j=2}^{h} \binom{h}{j-1} c_{j+1} g_{h-j+1} - \sum_{j=0}^{h-1} \binom{h}{j} c_{j+1} g_{h-j}' \\
&= g_{h+2} - c_1 g_{h+1} - (h+1) g_h - (c_2 - 1) g_h - \sum_{j=2}^{h} \binom{h}{j-1} c_{j+1} g_{h-j+1} \\
&\quad - \sum_{j=0}^{h-1} \binom{h}{j} c_{j+1} \left(g_{h-j+1} - c_1 g_{h-j} - (h-j) g_{h-j-1} \right) \\
&= g_{h+2} - \sum_{j=0}^{h} \binom{h+1}{j} c_{j+1} g_{h-j+1} - \Delta_{h+2},
\end{aligned}$$

where

$$\begin{aligned}
\Delta_{h+2} &= -\sum_{j=0}^{h}\binom{h+1}{j}c_{j+1}g_{h-j+1} + c_1 g_{h+1} + hg_h + c_2 g_h \\
&\quad + \sum_{j=2}^{h}\binom{h+1}{j}c_{j+1}g_{h-j+1} - \sum_{j=2}^{h}\binom{h}{j}c_{j+1}g_{h-j+1} \\
&\quad + \sum_{j=0}^{h-1}\binom{h}{j}c_{j+1}\left(g_{h-j+1} - c_1 g_{h-j} - (h-j)g_{h-j-1}\right) \\
&= hg_h - c_1 c_{h+1} + c_1 g_{h+1} - c_1\sum_{j=0}^{h-1}\binom{h}{j}c_{j+1}g_{h-j} - \sum_{j=0}^{h-1}\binom{h}{j}(h-j)c_{j+1}g_{h-j-1} \\
&= hg_h - \sum_{j=0}^{h-1}\binom{h}{j}(h-j)c_{j+1}g_{h-j-1} = hg_h - h\sum_{j=0}^{h-1}\binom{h-1}{j}c_{j+1}g_{h-j-1} \\
&= hg_h - h\left(\sum_{j=0}^{h-2}\binom{h-1}{j}c_{j+1}g_{h-j-1} + c_h g_0\right) = hg_h - h(g_h - c_h + c_h) = 0
\end{aligned}$$

and we have used the fact that

$$\binom{h}{j-1} = \binom{h+1}{j} - \binom{h}{j}, \qquad (h-j)\binom{h}{j} = h\binom{h-1}{j},$$

and the induction assumption that the formula holds at h and $h+1$.

Proposition 3. Let $z = x - \eta \sim \mathcal{N}(0,1)$ and consider a Taylor series expansion of $m(x)$ around η of order $h \geq 1$:

$$m_h(x) = m(\eta) + \sum_{j=1}^{h}\frac{a_j(\eta)z^j}{j!},$$

where the $a_j(\eta) = \left[d^j m(x)/dx^j\right]_{x=\eta} = m^{(j)}(\eta)$ are nonrandom constants which depend on η but not on x. Proposition 2 implies that $a_j(\eta)$ $(j \geq 1)$ is equal to the posterior cumulant of order $j+1$ evaluated at η, that is $a_j(\eta) = c_{j+1}(\eta)$. Thus, using the fact that

$$q_j = \frac{\mathbb{E}[z^j]}{j!} = \begin{cases} \dfrac{1}{2^{j/2}(j/2)!} & \text{if } j \text{ even,} \\ 0 & \text{if } j \text{ odd,} \end{cases}$$

we obtain the following delta method approximations:

$$\widehat{\delta}_h(\eta) = \mathbb{E}\left[m_h(x)|\eta\right] - \eta = m(\eta) - \eta + \sum_{j=1}^{h} c_{j+1}(\eta) q_j$$

and

$$\begin{aligned}\widehat{\sigma}_h^2(\eta) = \mathbb{V}\left[m_h(x)|\eta\right] &= \sum_{j=1}^{h}\sum_{k=1}^{h} \frac{a_j(\eta)a_k(\eta)\,\mathbb{C}(z^j, z^k)}{j!k!} \\ &= \sum_{j=1}^{h}\left(\binom{2j}{j} q_{2j} - q_j^2\right) c_{j+1}^2(\eta) + 2\sum_{k<j}\left(\binom{j+k}{j} q_{j+k} - q_j q_k\right) c_{j+1}(\eta) c_{k+1}(\eta).\end{aligned}$$

The results follow.

B An apparent contradiction

The results in Section 3.1.1 highlight a puzzling contradiction. We have the posterior mean $m(x)$ and the posterior variance $v^2(x)$. If we interpret $m(x)$ as an estimator of η, then this estimator has a (frequentist) variance $\sigma^2(\eta)$. We have seen that the *variance* $v^2(x)$ represents a first-order approximation to the frequentist *standard deviation* $\sigma(\eta)$. But we also know, from the Bernstein–von Mises theorem, that $v^2(x)$ and $\sigma^2(\eta)$ converge to each other. How can these two facts be reconciled?

To understand this apparent contradiction, consider a sample $x = (x_1, \ldots, x_n)$, rather than a single observation, from the $\mathcal{N}(\eta, 1)$ distribution. The simplest case is when the prior on η is $\mathcal{N}(0, \omega^2)$. In that case, the posterior mean and variance are given by $m_n(x) = w_n \bar{x}_n$ and $nv_n^2 = w_n$, where $w_n = \omega^2/(\omega^2 + 1/n)$. The frequentist variance of $m_n(x)$ is $\sigma_n^2 = \mathbb{V}[m_n(x)] = w_n^2/n$, and hence we have $v_1^2 = \sigma_1$ for $n = 1$. But when $n > 1$, both variances are of order $1/n$ and we have $w_n \to 1$ as $n \to \infty$ so that

$$n(\sigma_n^2 - v_n^2) = w_n^2 - w_n = w_n(w_n - 1) \to 0$$

as $n \to \infty$. This explains the apparent contradiction, at least in the case of a Gaussian prior.

Now consider another prior, the Laplace prior defined by $\pi(\eta) = b\, e^{-b|\eta|}/2$ with $c > 0$. As shown by De Luca et al. (2020), the posterior mean and variance of η are now

$$m_n(x) = \bar{x}_n - \frac{bh_n}{n}, \qquad nv_n^2(x) = 1 + \frac{b^2(1-h_n^2)}{n} - \frac{b(1+h_n)r(p_n)}{n^{1/2}},$$

where

$$\psi_n = \frac{1-\Phi(q_n)}{\Phi(p_n)}, \qquad h_n = \frac{1-e^{2b\bar{x}_n\psi_n}}{1+e^{2b\bar{x}_n\psi_n}}, \qquad r(p_n) = \frac{\phi(p_n)}{\Phi(p_n)},$$

and

$$p_n = n^{1/2}(\bar{x}_n - b/n), \qquad q_n = n^{1/2}(\bar{x}_n + b/n).$$

Given the posterior mean $m_n(x)$ we have

$$n\sigma_n^2(\eta) = n\,\mathbb{V}[m_n(x)] = 1 + \frac{b^2}{n}\mathbb{V}[h_n] - 2b\,\mathbb{C}[\bar{x}_n, h_n].$$

Both the posterior and the sampling variance are of order $1/n$ with

$$n(\sigma_n^2(\eta) - v_n^2(x)) \to 0,$$

since h_n is bounded with finite variance, $h_n^2 \to 1$, $r(p_n) \to 0$ as $n \to \infty$, and

$$\mathbb{C}^2[\bar{x}_n, h_n] \leq \mathbb{V}[\bar{x}_n]\,\mathbb{V}[h_n] = \frac{\mathbb{V}[h_n]}{n} \to 0.$$