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
EIEF and University of Rome Tor Vergata, Italy

February 15, 2022

Abstract

Many statistical and econometric learning methods rely on Bayesian ideas. When applied in a frequentist setting, their precision is often assessed using the posterior variance. This is permissible asymptotically, but not necessarily in finite samples. We explore this issue focusing on weighted-average least squares (WALS), a Bayesian-frequentist 'fusion'. Exploiting the sampling properties of the posterior mean in the normal location model, we derive estimators of the finite-sample bias and variance of WALS. We study the performance of the proposed estimators in an empirical application and a closely related Monte Carlo experiment which analyze the impact of legalized abortion on crime.

Keywords: Normal location model; posterior moments and cumulants; double-shrinkage estimators; WALS.

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

* Corresponding author: Giuseppe De Luca (giuseppe.deluca@unipa.it).

1 Introduction

Many statistical and econometric learning methods rely on Bayesian ideas. Examples include shrinkage estimators, such as smoothing splines (Reinsch 1967), ridge regression (Hoerl and Kennard 1970), and the least absolute shrinkage and selection operator (LASSO) (Tibshirani 1996). They also include the frequentist use of Bayesian model averaging estimators (Raftery et al. 1997), as well as Bayesian-frequentist 'fusions' such as the Bayesian averaging of classical estimates 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 estimators of Moral-Benito (2012). In many instances, the precision of these methods in repeated samples is assessed using the variance of the posterior distribution of the parameters of interest given the data. This is 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.

The present paper explores this issue focusing on WALS. There are several reasons why we concentrate on WALS. First, WALS provides estimates of the parameters of a standard linear regression model, which represents the workhorse of applied econometrics. Second, WALS is a weighted average of frequentist estimators and has been shown to enjoy important theoretical and computational advantages over other model-averaging estimators (Magnus and De Luca 2016). Third, the WALS 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. Like other Bayesian-frequentist 'fusions', a key problem is how to evaluate the sampling properties in finite samples. The main concern with unbiased estimators is sampling variability, which explains the classical emphasis on estimating standard errors. With biased estimators, instead, one wishes to estimate both sampling variability and bias.

In deriving the frequentist properties of WALS, an important role is played by the normal location model, which consists of a single observation from a univariate normal distribution with unknown mean and known variance. This stylized model allows us to examine the exact sampling bias and variance of the posterior mean by Monte Carlo tabulations, focusing on the general class of (reflected) generalized gamma priors. To interpret the Monte Carlo results, we exploit Taylor series

approximations of the posterior mean in the normal location model which in turn depend on higher-order posterior cumulants. We describe how these 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 the analytical approximations.

Since the sampling moments of the posterior mean generally depend on the unknown location parameter, we discuss two alternative plug-in methods for estimating the bias and variance of the posterior mean, respectively based on the (frequentist) maximum likelihood (ML) estimator and the (Bayesian) posterior mean. In finite samples, choosing between the two methods entails a bias-precision trade-off: the plug-in ML estimators have better risk performance for large values of the location parameter, while the plug-in Bayesian estimators have better risk performance for small values of the location parameter. Analytic approximations to the bias of the posterior mean suggest 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.

We show how the results for the normal location model can be applied to estimating the frequentist bias and variance of WALS. In particular, we show that the previous estimator of its sampling variance is upward biased, that is, WALS is more precise than originally thought. The estimators of the bias, which are also new, can then be used to estimate the mean squared error (MSE) of WALS.

Our results for the normal location model have wider applicability and extend to other shrinkage methods that can be reduced to estimating the location parameters in the diagonal form of a Gaussian sequence model (Johnstone 2019, Chapter 2). Examples include the empirical Bayes thresholding estimators proposed by Johnstone and Silverman (2004) and the Bayesian model averaging estimators proposed by Lee and Oh (2013). These estimators exploit sparse mixture priors and can be regarded as Bayesian versions of the thresholding estimators used in the wavelet literature. In fact, the estimation strategy adopted in WALS is quite similar to that used in wavelet shrinkage estimation (Johnstone 2019, Chapter 7), as both approaches are characterized by an initial (semi)orthogonal transformation, a processing step which operates coordinatewise through some shrinkage rule in the Gaussian sequence model, followed by an inverse transformation step.

As an empirical application of the WALS approach, we analyze the data used by Donohue and Levitt (2001) in their influential paper on the impact of legalized abortion on crime. For this

particular example, we find that WALS leads to the same policy conclusions as the post-double-selection estimates of Belloni et al. (2014a, 2014b), namely that the evidence in favor of a causal effect of abortion on crime is not robust to the presence of nonlinear trends. We also assess the finite-sample performance of our new estimators of the bias and variance of WALS by a Monte Carlo experiment whose design is based on this empirical example.

Our paper contributes to the growing literature on post-shrinkage and post-averaging inference. Post-shrinkage inference is mostly conducted from a frequentist viewpoint. For smooth shrinkage estimators based on ℓ_2 penalization (such as ridge estimators or smoothing splines) inference about the model parameters is straightforward because sampling distributions are asymptotically normal under general conditions. This is not the case for shrinkage estimators based on ℓ_p penalization with $p < 2$, as the sampling distributions are unwieldy. Here the distinction between a low-dimensional vector of focus parameters and a high-dimensional vector of auxiliary or nuisance parameters plays an important role, especially in problems of causal inference. For example, the double-selection approach of Belloni et al. (2014a, 2014b) and Chernozhukov et al. (2018) successfully separates the problem of inference about the causal parameters of interest from the much more difficult problem of inference about a nuisance function involving a large number of auxiliary parameters, possibly larger than the sample size. This orthogonality or "immunization property" is attained through a careful choice of the moment conditions used to estimate the model parameters. WALS is different because inference about the focus parameters is directly linked to inference about the auxiliary parameters via the equivalence theorem (Magnus and Durbin 1999), that is, the MSE of estimators of the focus parameters depends on the MSE of estimators of the nuisance parameters.

Inference about averaging estimators is conducted from either a Bayesian or a frequentist viewpoint depending on whether model averaging is Bayesian (BMA) or frequentist (FMA). In BMA (Steel 2020), inference is based on the posterior distribution of the parameter of interest, which is a weighted average of posterior distributions under the various models weighted by posterior model probabilities. In FMA (Claeskens and Hjort 2008), inference is instead based on the distribution of the averaging estimator under repeated sampling. The interpretation is clearly different in the two cases, especially in finite samples. As emphasized by Hjort and Claeskens (2003), Claeskens and Hjort (2008), Hansen (2014), and Zhang and Liu (2019), among others, the asymptotic distribution of the averaging estimator is generally nonnormal due to the randomness of the data-dependent

weights. Furthermore, the averaging estimator may have a nonnegligible bias. Instead of focusing on large-sample properties, which are typically established under a local misspecification framework in which the estimation bias vanishes at the rate $n^{-1/2}$, our analysis is in the spirit of Liang et al. (2011), who study the finite-sample risk of a FMA estimator based on weights that minimize the trace of an unbiased estimator of its MSE. The main difference is that, because of the semi-orthogonal transformation and the Bayesian shrinkage step, the MSE of WALS depends directly on the MSE of the posterior mean in the normal location model.

The remainder of the paper is organized as follows. Section 2 summarizes the WALS approach to model averaging. Section 3 discusses the normal location problem from the frequentist and Bayesian viewpoints. Section 4 focuses on the estimation of the sampling bias and variance of the posterior mean as a shrinkage estimator of the mean in the normal location model. Section 5 applies the results of the previous three sections to investigate the bias and variance of WALS in finite samples. Section 6 presents our empirical application, while Section 7 presents our Monte Carlo experiment. Section 8 concludes. There are four appendices.

2 The WALS approach

In applied econometrics we typically consider not one model but a whole family of models, usually the 'unrestricted' model and many of its subsets. We then use the same data to select the model *and* to estimate the parameters of interest. The properties of our estimates are thus conditional on the selected model, but we typically act as if the estimates were unconditional. This is common procedure but it is not quite right. Model averaging attempts to view model selection and estimation as one joint procedure so that the resulting estimates are indeed unconditional.

In this paper, we focus on model averaging procedures for the linear regression model

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

where y $(n \times 1)$ is the vector of observations on the outcome of interest, X_1 $(n \times k_1)$ and X_2 $(n \times k_2)$ are matrices of nonrandom regressors, β_1 and β_2 are unknown parameter vectors, and ϵ is a vector of random disturbances. The k_1 columns of X_1 contain the 'focus regressors' which we want in the model on theoretical or other grounds, while the k_2 columns of X_2 contain the 'auxiliary regressors'

of which we are less certain. These auxiliary regressors could be controls that are added to avoid omitted variable bias or transformations and interactions of a set of original regressors, such as indicator variables, polynomials, B-splines, etc. 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.

The assumptions of normality and linearity in the parameters are simplifying restrictions. In WALS, normality is employed because of its exponential structure and the fact that uncorrelatedness implies independence, and linearity in the parameters is required to obtain explicit expressions for the semi-orthogonal transformation of the auxiliary regressors. Both normality and linearity are approximations: the normal distribution is a second-order approximation around the mode to any well-behaved density, while a linear-in-parameters specification can approximate arbitrarily well a conditional expectation function that depends nonlinearly on a set of controls.[1] Even though a full generalization to nonlinear models with nonnormal errors appears to be far from trivial, one step in this direction is the extension to generalized linear models provided by De Luca et al. (2018).

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. If $\widehat{\beta}_{1j}$ and $\widehat{\beta}_{2j}$ are the ordinary least-squares (OLS) estimators of β_1 and β_2 in model j, then standard model averaging estimators take the form

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

where the λ_j are nonnegative data-dependent model weights that add up to one. Even for moderate values of k_2, the computational burden of calculating the 2^{k_2} OLS estimates and the associated weights can be substantial. For example, if $k_2 = 20$, the model space contains 1,048,576 models.

Unlike other model averaging estimators, the WALS approach exploits a semi-orthogonal transformation of the auxiliary regressors that reduces the computational burden from order 2^{k_2} to order k_2, coupled with a rescaling of the focus regressors that improves the accuracy of inversion and eigenvalue routines. Specifically, we transform X_2 and β_2 by defining $Z_2 = X_2 \Delta_2 \Psi^{-1/2}$ and $\gamma_2 = \Psi^{1/2} \Delta_2^{-1} \beta_2$, where Δ_2 is a diagonal $k_2 \times k_2$ matrix such that all diagonal elements of the symmetric and positive definite matrix $\Psi = \Delta_2 X_2' M_1 X_2 \Delta_2$ are equal to one and $M_1 = I_n - X_1 (X_1' X_1)^{-1} X_1'$.

[1] It does however exclude models that are nonlinear in the parameters, such as logit, probit, and Poisson models.

We also rescale X_1 and β_1 by defining $Z_1 = X_1\Delta_1$ and $\gamma_1 = \Delta_1^{-1}\beta_1$, where Δ_1 is a diagonal $k_1 \times k_1$ matrix such that all diagonal elements of $Z_1'Z_1$ are equal to one. Since $Z_1\gamma_1 = X_1\beta_1$ and $Z_2\gamma_2 = X_2\beta_2$, we may then write model (1) equivalently as

$$y = Z_1\gamma_1 + Z_2\gamma_2 + \epsilon. \tag{3}$$

The fact that $Z_2'M_1Z_2 = I_{k_2}$ brings several important advantages. First, if $\widehat{\gamma}_{1j}$ and $\widehat{\gamma}_{2j}$ are the ordinary OLS estimators of γ_1 and γ_2 in model j, then the WALS estimators can be written as

$$\widehat{\gamma}_1 = \sum_{j=1}^{2^{k_2}} \lambda_j \widehat{\gamma}_{1j} = \widehat{\gamma}_{1r} - QW\widehat{\gamma}_{2u}, \qquad \widehat{\gamma}_2 = \sum_{j=1}^{2^{k_2}} \lambda_j \widehat{\gamma}_{2j} = W\widehat{\gamma}_{2u}, \tag{4}$$

where $\widehat{\gamma}_{1r} = (Z_1'Z_1)^{-1}Z_1'y$ is the estimator of γ_1 in the fully restricted model with $\gamma_2 = 0$, $\widehat{\gamma}_{2u} = Z_2'M_1y$ is the estimator of γ_2 in the fully unrestricted model, $Q = (Z_1'Z_1)^{-1}Z_1'Z_2$, $W = \sum_j \lambda_j W_j$, and $W_j = I_{k_2} - R_jR_j'$, where R_j is a $k_2 \times r_j$ selection matrix of rank $0 \leq r_j \leq k_2$ — that is, $R_j' = [I_{r_j} : 0]$ or a column-permutation thereof — representing the r_j exclusion restrictions implied by model $j = 1, \ldots, 2^{k_2}$.

Second, the dependence of $\widehat{\gamma}_1$ and $\widehat{\gamma}_2$ on the estimates from all the 2^{k_2} models in the model space is completely captured by the random diagonal matrix $W = \sum_j \lambda_j W_j$, whose k_2 diagonal elements w_h are partial sums of the λ_j 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 compute the restricted estimates of γ_1, the unrestricted estimates of γ_2, and determine the set of k_2 WALS weights w_h, not the considerably larger set of 2^{k_2} model weights λ_j. Unlike other approaches we are not ignoring any of the 2^{k_2} models in the original model space, which is feasible because each model contributes to the model averaging estimates only through the k_2 diagonal elements of the matrix W.

Third, Theorem 2 of Magnus and Durbin (1999) 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.

Fourth, the components of $\widehat{\gamma}_2 = W\widehat{\gamma}_{2u}$ are shrinkage estimators of the components of γ_2 as $0 \leq w_h \leq 1$, and the components of $\widehat{\gamma}_{2u} = Z_2'M_1y$ are independent as $\widehat{\gamma}_{2u} \sim \mathcal{N}(\gamma_2, \sigma^2 I_{k_2})$. Hence, if

we restrict each w_h to depend only on the hth component of $\widehat{\gamma}_{2u}$, then the shrinkage estimators in $\widehat{\gamma}_2$ will also be independent. Under this additional restriction (discussed in detail in Magnus and De Luca 2016), our k_2-dimensional problem reduces to k_2 (identical) one-dimensional problems, namely: given one observation $x \sim \mathcal{N}(\eta, \sigma^2)$, what is the estimator $m(x)$ of η with minimum MSE? This is the so-called *normal location problem*. Irrespective of whether the assumption of normal errors holds, the normal location problem well approximates the actual shrinkage estimation problem when the sample size is moderate to large. Since the risk properties of $m(x)$ are little affected by estimating the variance parameter (Danilov 2005, Magnus and De Luca 2016), we also assume that σ^2 is known.

The normal location problem is an important ingredient in the WALS procedure. We take a Bayesian approach to it, which allows a proper treatment of admissibility, bounded risk, robustness, near-optimality in terms of minimax regret, and ignorance about η. We thus add a Bayesian ingredient to a frequentist analysis in the spirit of Wright (2008, p. 330), who writes:

> 'One does not have to be a subjectivist Bayesian to believe in the usefulness of BMA or of Bayesian shrinkage techniques more generally. A frequentist econometrician can interpret these methods as pragmatic devices that may be useful for out-of-sample forecasting in the face of model and parameter uncertainty.'

The Bayesian ingredient requires two elements: (i) a neutral prior with bounded risk (such as the Laplace, Subbotin, or Weibull priors discussed in Appendix B), and (ii) the k_2-vector of t-ratios $x = \widehat{\gamma}_{2u}/s_u$, where $s_u^2 = y'M_1(I_n - Z_2Z_2')M_1y/(n-k)$ is the unrestricted estimator of σ^2 in model (3). 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 mean $m_h = m(x_h)$ as an estimator of η_h. The WALS estimators of γ_1 and γ_2 are then

$$\widehat{\gamma}_1 = \widehat{\gamma}_{1r} - Q\widehat{\gamma}_2, \qquad \widehat{\gamma}_2 = s_u m, \tag{5}$$

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

$$\widehat{\beta}_1 = \Delta_1\widehat{\gamma}_1, \qquad \widehat{\beta}_2 = \Delta_2\Psi^{-1/2}\widehat{\gamma}_2. \tag{6}$$

The mixture of Bayesian and frequentist approaches requires special attention when assessing the sampling properties of our model averaging estimator. Treating the posterior mean $m(x)$ as a frequentist estimator of η raises several questions. Is it legitimate to use the posterior variance $v^2(x)$ as a proxy to the sampling variance $\text{var}[m(x)]$? And, can we estimate the bias, variance, and MSE of $m(x)$? These and related issues are discussed in the next two sections.

3 The normal location problem

Thus motivated, we consider the problem of estimating η from a single observation x drawn from a univariate normal distribution with unknown mean η and known variance which, without loss of generality, we set equal to one. Hence the likelihood is $\phi(x-\eta)$, where $\phi(\cdot)$ denotes the density of the standard-normal distribution.

3.1 The frequentist viewpoint

A frequentist would simply ask how to estimate η based on one observation x from the $\mathcal{N}(\eta,1)$ distribution. At first glance the answer is obvious: $\widehat{\eta}_1 = x$. This estimator is unbiased and has variance one so that $\text{MSE}(\widehat{\eta}_1) = 1$. But one may also consider a second estimator $\widehat{\eta}_2 = 0$ with $\text{MSE}(\widehat{\eta}_2) = \eta^2$, which is lower than the first when $|\eta| < 1$ and higher when $|\eta| > 1$.

This simple insight leads to the idea of a shrinkage estimator, say $\widehat{\eta}_3 = w\widehat{\eta}_1 + (1-w)\widehat{\eta}_2 = wx$, where w depends on x, is monotonically nondecreasing for $x \geq 0$, and satisfies $0 \leq w(x) \leq 1$ and $w(-x) = w(x)$. Examples are the pretest estimator, the ridge estimator, and the LASSO, respectively given by

$$\widehat{\eta}_P = x\,1\{|x| > c\}, \qquad \widehat{\eta}_R = \frac{x}{1+c}, \qquad \widehat{\eta}_L = (x-c)\,1\{x > c\} + (x+c)\,1\{x \leq -c\}, \tag{7}$$

where $1\{A\}$ is the indicator function of the event A and $c > 0$ in all cases. The ridge estimator is obtained by minimizing $(x-\eta)^2 + c\eta^2$ with respect to η, and the LASSO by minimizing $(x-\eta)^2/2 + c|\eta|$. In this stylized setting the LASSO coincides with the Burr estimator obtained by minimizing the MSE over a wide class of weight functions based on the three-parameter Burr cumulative distribution function (Magnus 2002). These three shrinkage estimators are, however, not completely satisfactory: the pretest estimator is not continuous, the LASSO/Burr estimator is

continuous but not differentiable at $x = \pm c$, and the bias of the ridge estimator is unbounded. A more satisfactory approach requires a Bayesian viewpoint.

3.2 The Bayesian viewpoint

In a Bayesian context, uncertainty about η is represented by a proper prior density $\pi(\cdot)$ on $\mathbb{R}$ which is assumed to be positive and bounded. Combining likelihood and prior gives the posterior density of η given x,

$$p(\eta|x) = \frac{\phi(x-\eta)\,\pi(\eta)}{A_0(x)}, \qquad A_0(x) = \int_{-\infty}^{\infty} \phi(x-\eta)\,\pi(\eta)\,d\eta. \tag{8}$$

Pericchi and Smith (1992, Theorem 1) show that, for any positive and bounded prior density $\pi(\eta)$, the posterior mean and variance of η are given by

$$m(x) = \mathrm{E}[\eta|x] = x + \frac{d\log A_0(x)}{dx}, \qquad v^2(x) = \mathrm{E}[\eta^2|x] - [m(x)]^2 = 1 + \frac{d^2\log A_0(x)}{dx^2}, \tag{9}$$

where expectations are taken with respect to the conditional distribution of η given x. Thus, $m'(x) = v^2(x) > 0$, which shows that the posterior mean $m(x)$ is increasing in x. The expression for $m(x)$ in (9) is known as the Brown–Tweedie formula (Robbins 1956, Brown 1971). As shown by Johnstone (2019, p. 30), this representation is useful to deduce shrinkage properties of the posterior mean from assumptions on the tails of the prior.

The next result generalizes Pericchi and Smith's results by providing a recursive formula for high-order posterior moments of η.

Proposition 1 *Given one observation $x \sim \mathcal{N}(\eta, 1)$ and a positive and bounded prior density $\pi(\eta)$, the hth (noncentral) posterior moment of η is*

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

where $\mu_0(x) = 1$ and

$$g_h(x) = (-1)^h\,\mathrm{E}[(x-\eta)^h|x] = (-1)^h \frac{d^h \log A_0(x)}{dx^h} \quad (h = 1, 2, \dots).$$

Pericchi et al. (1993, Proposition 2.2) show that the hth derivative of $m(x)$ is equal to the

$(h+1)$st posterior cumulant of η. Higher-order posterior cumulants play a role in the areas of Bayesian robustness and approximation and in the empirical analysis of data characterized by skewed distributions with fat tails. Below, in Section 4.2, 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.

Under quadratic loss, the posterior mean $m(x)$ is the Bayesian point estimator of η. Our purpose is to consider $m(x)$ as a frequentist shrinkage estimator of η. The idea of reinterpreting Bayesian learning methods in a frequentist setting is quite common in statistics and econometrics. For example, both the ridge and the LASSO/Burr estimators introduced in Section 3.1 may be derived from a Bayesian viewpoint: the ridge estimator as the posterior mode when η has a normal prior with mean zero and variance $1/c$ (in which case the posterior mean, median and mode are all equal), and the LASSO/Burr estimator as the posterior mode when η has a Laplace prior (Park and Casella 2008). We can also obtain a number of thresholding estimators as posterior medians when η has a sparse mixture prior of the form $\pi(\eta) = (1-\omega)\delta_0(\eta) + \omega g(\eta)$, where $\omega \in (0,1)$, $\delta_0(\eta)$ is a degenerate density that assigns all its mass to the point $\eta = 0$ and $g(\eta)$ is a continuous density (Johnstone and Silverman 2004, Lee and Oh 2013). Both the posterior mode and the posterior median are useful if one is interested in model selection, where thresholding is important. But here we are interested in estimating the focus parameters as well as we can, not in selecting a best model; that is, we are interested in model averaging rather than model selection. Compared to the posterior mode and median, the posterior mean is typically smooth because it cannot have a thresholding zone (Johnstone 2019, p. 31) and is therefore more appropriate for model averaging.

4 Sampling properties of the posterior mean

We are interested in the sampling bias and variance of the posterior mean, defined as

$$\delta(\eta) = \mathrm{E}[m(x)|\eta] - \eta = \mathrm{E}[g_1(x)|\eta], \qquad \sigma^2(\eta) = \mathrm{E}\left[m(x)^2|\eta\right] - (\mathrm{E}[m(x)|\eta])^2, \tag{10}$$

where expectations are now taken with respect to the conditional distribution of x given η. The MSE of $m(x)$ is $\sigma^2(\eta) + \delta^2(\eta)$. In deciding on suitable estimators of $\delta(\eta)$ and $\sigma^2(\eta)$ two problems occur. First, except for normal priors, the sampling moments of $m(x)$ do not admit closed-form

expressions. Second, they depend on the unknown location parameter η, so we must replace η by an estimator. We discuss the first problem in Sections 4.1 and 4.2, and the second in Section 4.3. In practice, we recommend to obtain $\delta(\eta)$ and $\sigma^2(\eta)$ via the Monte Carlo approach of Section 4.1. The posterior cumulant approach in Section 4.2 is only provided to better understand the Monte Carlo results.

4.1 Monte Carlo tabulations

To ensure that $\delta(\eta)$ is a bounded function, $m(x)$ must be a nonlinear function of x, so its sampling moments must be either approximated analytically or obtained numerically. We discuss Monte Carlo methods in the current subsection and approximations in the next.

In order to obtain detailed tabulations of $\delta(\eta)$ and $\sigma^2(\eta)$ over a range of values of η 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 $\widetilde{x}$, then compute $\overline{m}_1 = \sum_{j=1}^{J} \widetilde{m}_j/J$ and $\overline{m}_2 = \sum_{j=1}^{J} \widetilde{m}_j^2/J$, and approximate the bias and variance of $m(x)$ by $\overline{\delta}_J(\eta) = \overline{m}_1 - \eta$ and $\overline{\sigma}_J^2(\eta) = \overline{m}_2 - \overline{m}_1^2$, 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, $\overline{\delta}_J(\eta)$ and $\overline{\sigma}_J^2(\eta)$.

Our algorithm is simple and transparent. It is not restricted to a particular set of priors and can easily be extended to other widely used Bayesian estimators of η, such as the posterior median or the posterior mode. Most importantly, the approximation errors can be made arbitrarily small by using a sufficiently large number of independent draws from the $\mathcal{N}(\eta, 1)$ distribution.

For the purpose of our application to WALS, we use $J = 1,000,000$ draws and allow η to range in the interval $[0, 30]$ with stepsize $\Delta\eta = 0.01$. We consider four neutral priors from the flexible three-parameter family of reflected generalized gamma distributions described in Appendix B: normal, Laplace, Subbotin, and Weibull. The normal prior is included as a benchmark to evaluate the accuracy of the calculations when the posterior mean is computed by numerical Gauss-Laguerre

quadrature methods with $1,000$ points and its bias and variance are tabulated by our Monte Carlo algorithm with $J = 1,000,000$ draws.[2]

FIGURES 1 and 2 HERE

The solid red lines in Figures 1 and 2 illustrate, respectively, the Monte Carlo tabulations of $\delta(\eta)$ and $\sigma^2(\eta)$ for the four priors under consideration. Since all priors are symmetric around zero, the bias $\delta(\eta)$ is always an odd function, that is $\delta(-\eta) = -\delta(\eta)$ with $\delta(0) = 0$; in contrast, the variance $\sigma^2(\eta)$ is an even function, that is $\sigma^2(-\eta) = \sigma^2(\eta)$, with a minimum at $\eta = 0$ (Magnus 2002, Theorems A.3 and A.4). For $\eta \geq 0$, under the Laplace prior, $\delta(\eta)$ is nonincreasing and convex, and converges to a minimum of -0.69 for large values of η. In contrast, $\sigma^2(\eta)$ is nondecreasing and convex-concave (first convex, then concave), and converges to a maximum of one. Under the Weibull and Subbotin priors, $\delta(\eta)$ and $\sigma^2(\eta)$ are nonmonotonic. Specifically, $\delta(\eta)$ reaches a minimum of -0.59 at $\eta = 3.17$ for the Weibull prior and a minimum of -0.61 at $\eta = 3.24$ for the Subbotin prior, while $\sigma^2(\eta)$ reaches a maximum of 1.05 at $\eta = 4.30$ for the Weibull prior and a maximum of 1.06 at $\eta = 4.54$ for the Subbotin prior. Calculations based on the normal prior suggest that the simulation error is of order 10^{-8} for both $\delta(\eta)$ and $\sigma^2(\eta)$.

4.2 Analytical approximations

The Monte Carlo results in Section 4.1 are exact (or as exact as we wish). To gain further insight we also provide analytical approximations to $\delta(\eta)$ and $\sigma^2(\eta)$ by means of a Taylor series expansion of $m(x)$ around η. Such an expansion resembles the 'delta method' which is typically presented in terms of $n^{-1/2}$ or n^{-1} for a function m of a consistent estimator x_n of η. Our case is a little different because $n = 1$, so the question of consistency doesn't arise. It turns out that a low-order expansion already provides fairly accurate approximations to the exact Monte Carlo results.

Proposition 2 *Given one observation $x \sim \mathcal{N}(\eta, 1)$ and a nonnegative bounded prior density $\pi(\eta)$, let $m(x)$ be the posterior mean of η given x. If $m(x)$ is used as estimator of η, the analytical*

[2] For the Laplace prior, the computing time of the algorithm is about one hour thanks to the closed-form expression for the posterior mean. For the other priors the computing time is about one week. The calculations were performed in Stata by using a workstation with one Intel(R) Core(TM) i7-4790 CPU/3.60 GHz processor and 32 GB of RAM. All routines and tabulations are available from the authors upon request.

approximations of order $h+1$ ($h \geq 1$) to its bias and sampling variance are given recursively by

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

where

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

and

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

The starting values are

$$\delta_1(\eta) = m(\eta) - \eta, \qquad \sigma^2_1(\eta) = c_2(\eta),$$

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 $x = \eta$.

We can always use Taylor's theorem (with remainder) to obtain a Taylor expansion for $m(x)$ up to a finite order, since $m(x)$ is infinitely many times differentiable (in fact, analytic) on $\mathbb{R}$, and this is what we do in Proposition 2. However, the fact that m is analytic on $\mathbb{R}$ does not imply that m is analytic on the complex plane $\mathbb{C}$, and the latter is required for the infinite Taylor series to converge. Whether or not m is analytic on $\mathbb{C}$ depends on the prior π. For the normal prior this is indeed the case and the radius of convergence is infinite. But for the three priors of interest to us (Laplace, Weibull, Subbotin) the radius is positive but not infinite. We discuss this technical issue in more detail in Appendix C.

Propositions 1 and 2 generalize the Brown–Tweedie formula to higher-order derivatives and establish a connection with higher-order cumulants. More specifically, Proposition 2 shows that the bias and variance of the posterior mean depend crucially on the posterior cumulants of η. For $h = 2$, we have

$$\delta_2(\eta) = \delta_1(\eta) + \frac{1}{2}c_3(\eta), \qquad \sigma^2_2(\eta) = \sigma^2_1(\eta) + \frac{1}{2}c^2_3(\eta). \tag{11}$$

And, for $h = 3$,

$$\delta_3(\eta) = \delta_2(\eta), \qquad \sigma_3^2(\eta) = \sigma_2^2(\eta) + \frac{5}{12}c_4^2(\eta) + c_2(\eta)c_4(\eta). \tag{12}$$

Defining the posterior skewness and (excess) kurtosis as $\tau(x) = c_3(x)/v^3(x)$ and $\kappa(x) = c_4(x)/v^4(x)$, respectively, the second- and third-order analytical approximations ($h = 2$ and $h = 3$) can equivalently be written as

$$\delta_2(\eta) = m(\eta) - \eta + \frac{1}{2}\tau(\eta)\, v^3(\eta), \qquad \sigma_2^2(\eta) = v^4(\eta)\left[1 + \frac{1}{2}\tau^2(\eta)v^2(\eta)\right] \tag{13}$$

and

$$\delta_3(\eta) = \delta_2(\eta), \qquad \sigma_3^2(\eta) = v^4(\eta)\left[1 + \frac{1}{2}\tau^2(\eta)v^2(\eta) + \kappa(\eta)\, v^2(\eta) + \frac{5}{12}\kappa^2(\eta)v^4(\eta)\right], \tag{14}$$

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

We see that $\sigma_1^2(\eta) = v^4(\eta)$ coincides with the first-order approximation to $\sigma^2(\eta)$ obtained by the general accuracy formula in Efron (2015, Theorem 1). Thus, for any positive and bounded prior density, the posterior variance $v^2(\eta)$ represents an approximation to the sampling standard deviation, not the sampling 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 D, this apparent contradiction is due to the fact that, when $n > 1$, the posterior variance of η 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 analytical approximations $\delta_1(\eta)$ and $\sigma_1^2(\eta)$ by introducing additional terms which depend on the posterior variance and the posterior skewness. The sign of the additional term in $\delta_2(\eta)$ depends on the sign of $\tau(\eta)$, while the additional term in $\sigma_2^2(\eta)$ is always nonnegative, so $\sigma_2^2(\eta) \geq \sigma_1^2(\eta)$ for all η.

The third-order expansion does not further improve the approximation to $\delta(\eta)$, because $\delta_3(\eta) = \delta_2(\eta)$ due to the fact that $q_3 = 0$. The additional term in $\sigma_3^2(\eta)$ depends on the posterior variance and the posterior (excess) kurtosis. This term can be either positive or negative and may lead to a substantial improvement in the accuracy of the analytical approximations to $\sigma^2(\eta)$.

The analytical approximations in this section provide closed-form relationships which help understand better the link between the frequentist and Bayesian approaches to inference. But what

can we say about their accuracy? And how sensitive are the results to alternative choices of the prior density? In Figures 1 and 2, we assess the accuracy of different approximations to $\delta(\eta)$ and $\sigma^2(\eta)$ by comparing the analytical approximations of orders $h = 1, 2, 3$ with the Monte Carlo (MC) tabulations based on $J = 1,000,000$ draws. In each panel, TS_1, TS_2, and TS_3 represent the first-, second-, and third-order truncated series approximations to $\delta(\eta)$ and $\sigma^2(\eta)$, respectively. For the conjugate normal prior, our analytical approximations are exact because the posterior mean is linear. For the other priors, the first- and second-order analytical approximations are still poor, but the third-order approximation is already quite close to the truth.

4.3 Estimating the sampling moments of $m(x)$

The sampling bias $\delta(\eta)$ and variance $\sigma^2(\eta)$ of $m(x)$ depend on η which is unknown. Thus, according to the plug-in principle, we replace η by an estimator, say $\widehat{\eta}$, which is a function of x. We consider two estimators of η: $\widehat{\eta} = x$, which is a natural choice (see, e.g., Efron 2015) because x is the unbiased ML estimator of η; and $\widehat{\eta} = m(x)$, which is the Bayesian point estimator of η. The first estimator leads to the plug-in ML estimators $\delta(x)$ and $\sigma^2(x)$, while the second estimator leads to the plug-in double-shrinkage estimators (so named because of the double use of the posterior mean function in the leading term of the estimated bias) $\delta(m(x))$ and $\sigma^2(m(x))$. We have experimented with other plug-in estimators, but these are less natural and perform less well.

The choice between these alternative estimators of $\delta(\eta)$ and $\sigma^2(\eta)$ is similar to the choice between x and $m(x)$ as estimators of η, 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 has good risk properties around $|\eta| = 1$, which is the value of central interest. The central value $|\eta| = 1$ is important because we know that the restricted least-squares estimator dominates the unrestricted least-squares estimator (in the mean squared error sense) if and only if $\eta^2 < 1$ (Magnus and Durbin 1999, Theorem 1; Magnus 2002, pp. 226 and 229). This is also the reason why we choose our priors such that

$$\Pr[\eta < -1] = \Pr[-1 < \eta < 0] = \Pr[0 < \eta < 1] = \Pr[\eta > 1] = 1/4.$$

However, if $\widehat{\eta}$ is a good estimator of η, then $f(\widehat{\eta})$ is not necessarily a good estimator of $f(\eta)$, unless f is linear. 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 two plug-in estimators of $\delta(\eta)$ and $\sigma^2(\eta)$ by a simple Monte Carlo experiment.

The design of the experiment is as follows. Because of symmetry we need only consider $\eta \geq 0$. For any η in the interval $[0, 10]$ with stepsize $\Delta\eta = 0.01$ we generate a vector $x = (x_1, \ldots, x_R)$ of $R = 100,000$ independent draws from the $\mathcal{N}(\eta, 1)$ distribution. For each element x_r of x ($r = 1, \ldots, R$) we then compute the ML estimates $\delta(x_r)$ and $\sigma^2(x_r)$ and the double-shrinkage estimates $\delta(m(x_r))$ and $\sigma^2(m(x_r))$. Then we approximate the bias and root mean squared error (RMSE) profiles of the ML and double-shrinkage estimators using their Monte Carlo replications.

FIGURE 3 HERE

Figure 3 presents the Monte Carlo results for the ML and double-shrinkage estimators of the bias $\delta(\eta)$ of $m(x)$ under the Laplace and Weibull priors. Together with the profiles of the bias (upper panels) and RMSE (lower panels) of the two plug-in estimators we plot for each prior the Monte Carlo profiles of $\delta(\eta)$ already presented in the upper-right and bottom-left panels of Figure 1. Recall that the bias is an odd function. 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 ML estimator $\delta(x)$ of $\delta(\eta)$ will be upward biased due to Jensen's inequality. The estimator is unbiased at $\eta = 0$, where $\delta(\eta) = 0$ and $\delta(x)$ takes positive and negative values with equal probabilities, and for large values of η (say, $\eta > 6$), where $\delta(x)$ is roughly constant. Since $m(x)$ is biased towards zero and $\delta(\eta)$ is nonincreasing, the double-shrinkage estimator $\delta(m(x))$ presents an additional source of positive bias due to the shrinkage estimation of η. So, the ML estimator performs better than the double-shrinkage estimator in terms of bias. In terms of MSE, the double-shrinkage estimator performs better than ML for small and medium values of η (roughly, $\eta < 2$), while the opposite is true for larger values of η. Similar considerations apply in the case of the Weibull prior.

FIGURE 4 HERE

Figure 4 presents the Monte Carlo results for the ML and double-shrinkage estimators of the variance $\sigma^2(\eta)$ of $m(x)$ under the Laplace and Weibull priors, together with the Monte Carlo profiles of $\sigma^2(\eta)$ already presented in the upper-right and bottom-left panels of Figure 2. Unlike the bias,

the variance is an even function. Under the Laplace prior, $\sigma^2(\eta)$ is nondecreasing and convex-concave for $\eta > 0$. In the case of ML 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. For the double-shrinkage estimator 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 and negative for larger values of η. Regarding the MSE we see, as with the estimation of $\delta(\eta)$, that the double-shrinkage estimator performs better than ML for small and medium values of η (roughly, $\eta < 2$), while the opposite is true for larger values of η. Similar considerations apply to the Weibull prior.

Both plug-in estimators perform well. Since we think of $|\eta| = 1$ as the key value of η (see the discussion earlier in this subsection) and the double-shrinkage estimator performs better than ML at $\eta = 1$, we have a slight preference for the double-shrinkage estimator compared to ML.

5 Sampling properties of the WALS estimator

Our summary of the WALS methodology in Section 2 led to the estimators of the γ's and β's in (5) and (6). We did not, however, discuss their sampling variances. Earlier papers on the development of WALS estimated these variances using the diagonal $k_2 \times k_2$ matrix $V = \text{diag}(v_1^2, \ldots, v_{k_2}^2)$ 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 $\widehat{\gamma}_2$ were computed as

$$\widehat{\text{var}}[\widehat{\gamma}_1] = s_u^2 (Z_1' Z_1)^{-1} + Q\, \widehat{\text{var}}[\widehat{\gamma}_2]\, Q', \qquad \widehat{\text{var}}[\widehat{\gamma}_2] = s_u^2 V, \tag{15}$$

and the estimated covariance as $\widehat{\text{cov}}[\widehat{\gamma}_1, \widehat{\gamma}_2] = -Q\, \widehat{\text{var}}[\widehat{\gamma}_2]$, where $Q = (Z_1' Z_1)^{-1} Z_1' Z_2$. As a consequence, the variances of $\widehat{\beta}_1$ and $\widehat{\beta}_2$ were estimated by

$$\widehat{\text{var}}[\widehat{\beta}_1] = \Delta_1\, \widehat{\text{var}}[\widehat{\gamma}_1]\, \Delta_1, \quad \widehat{\text{var}}[\widehat{\beta}_2] = \Delta_2 \Psi^{-1/2}\, \widehat{\text{var}}[\widehat{\gamma}_2]\, \Psi^{-1/2} \Delta_2 \tag{16}$$

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

But this is not quite right. As discussed in Section 4, 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 approximation to the standard deviation of m_h, so the diagonal elements of the matrix V should not be v_h^2 but v_h^4. But this is only a first-order approximation. Section 4 provides the theory for the appropriate estimation of the diagonal elements of V. To estimate the sampling variance of WALS we should use (15) and (16), where the diagonal matrix V is redefined so that its hth diagonal element equals the double-shrinkage estimator $\sigma^2(m(x_h))$ (or the ML estimator $\sigma^2(x_h)$) of the sampling variance of m_h.

In a similar fashion we now use the plug-in estimators of the bias of the posterior mean to estimate the bias and the MSE of WALS. 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 either the double-shrinkage estimate $\delta(m(x_h))$ or the ML estimate $\delta(x_h)$. As shown in Section 4.3, these estimators are generally biased but their RMSEs are bounded and their biases are relatively small. For example, under the Laplace prior, we have $|\,\mathrm{E}[\widehat{\delta}_h] - \delta_h| \leq 0.05$ for the ML estimator and $|\,\mathrm{E}[\widehat{\delta}_h] - \delta_h| \leq 0.15$ for the double-shrinkage estimator (see Figure 3). In both cases, the maximum bias is reached at $|\eta_h| = 1.84$ where $|\delta_h| = 0.51$, and hence $|\,\mathrm{E}[\widehat{\delta}_h]/\delta_h - 1| = 10\%$ for the ML estimator and $|\,\mathrm{E}[\widehat{\delta}_h]/\delta_h - 1| = 28\%$ for the double-shrinkage estimator. This suggests that we can think of $\widehat{\delta}_h$ as a nearly unbiased estimator of δ_h, especially for the ML estimator. After estimating the bias of m by $\widehat{\delta} = (\widehat{\delta}_1, \ldots, \widehat{\delta}_{k_2})$, we estimate the bias $d_2 = \mathrm{E}[\widehat{\gamma}_2] - \gamma_2$ of $\widehat{\gamma}_2$ by $\widehat{d}_2 = s_u\widehat{\delta}$ and the bias $b_2 = \mathrm{E}[\widehat{\beta}_2] - \beta_2$ of $\widehat{\beta}_2$ by $\widehat{b}_2 = \Delta_2\Psi^{-1/2}\widehat{d}_2$. Provided that the unknown data-generating process is nested in the unrestricted model (3), we can also estimate the bias of $\widehat{\gamma}_1$,

$$d_1 = \mathrm{E}[\widehat{\gamma}_1] - \gamma_1 = \mathrm{E}[\widehat{\gamma}_{1r}] - \gamma_1 - Q\,\mathrm{E}[\widehat{\gamma}_2] = Q\gamma_2 - Q(\gamma_2 + d_2) = -Qd_2, \tag{17}$$

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

6 Empirical application: The impact of legalized abortion on crime

As an empirical application of the WALS approach, we analyze the same data as used by Donohue and Levitt (2001) and Belloni et al. (2014b) to assess whether the legalization of abortion in the

USA in the early 1970s caused a reduction of crime during the 1990s.[3] The key regressor of interest is a measure of the abortion rate relevant for various types of crime, determined by the ages of criminals when they tend to commit these crimes. The data consist of a balanced panel of 50 US states (D.C. excluded) over the period 1985–1997, with a total of 650 state-year observations.

There are several reasons why we choose this empirical application. First, it allows us to compare the WALS approach with two model selection approaches, namely the original approach of Donohue and Levitt (2001) based on an *ad hoc* sensitivity analysis and the approach of Belloni et al. (2014b) based on an automatic (i.e., data-driven) procedure. Second, our empirical application deals with a setting that is becoming increasingly relevant in applied economics, namely one in which the number of auxiliary regressors is large relative to the sample size. Third, it illustrates how introducing estimates of the MSE criterion may substantially alter the results of an analysis relative to the traditional approach of comparing only the estimated standard errors. Fourth, it provides the basis for the simulations in Section 7. Last but not least, it addresses a relevant and controversial issue with important policy implications.

Based on their model selection strategy, Donohue and Levitt (2001) conclude 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. They provide two arguments for a causal impact of abortion on crime: (i) a cohort-size effect, as higher abortion rates in a given cohort reduce the number of young people who are most at risk for committing crimes in the future; and (ii) a selection-effect, as legalized abortion gives women more control over child planning, thus increasing the probability of a favorable environment to raise their children. Belloni et al. (2014b) argue that the main concern with these conclusions is that state-level abortion rates in the early 1970s were not randomly assigned, and failing to adequately control for factors that are associated with state-level abortion and crime rates may lead to omitted variable bias.

In addition to state and time fixed effects, the baseline specification in Donohue and Levitt (2001) only included eight time-varying state-specific controls (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 an indicator for the existence of a concealed weapons law). To reduce concerns about serial correlation and omitted

[3] Although the analysis in Belloni et al. (2014b) is very similar to the analysis in Belloni et al. (2014a), we focus on the former because of the availability of the replication files at doi.org/10.3886/E113924V1.

variable bias, we follow the strategy proposed by Belloni et al. (2014b). Thus, we eliminate the state fixed effects by estimating the model in first differences. We also consider a much broader set of control variables to account for nonlinear trends that may depend on time-varying state-level characteristics. Our focus regressors include the first-difference of the abortion rate and a full set of time fixed effects, 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 time-varying controls considered by Donohue and Levitt, 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).[4]

TABLE 1 HERE

The three panels of Table 1 show the estimated impact of legalized abortion (the estimated coefficient on the first differences of the abortion rates) for violent, property, and murder crimes respectively. The table compares the WALS estimates based on the Laplace (WALS-L) and Weibull (WALS-W) priors with four other estimates: the least-squares estimates from the unrestricted model that includes all focus and auxiliary regressors (LS-U), the least-squares estimates from the fully restricted model that includes only the focus regressors (LS-R), the least-squares estimates from 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 Donohue and Levitt in their original paper (LS-I), and the least-squares estimates from the intermediate model that includes the focus regressors and the subset of auxiliary regressors selected by the double-selection procedure (LS-D). The LS-U, LS-I and LS-D estimates coincide with those reported in Table 1 of Belloni et al. (2014b).[5]

In addition to the estimated coefficients, we present the estimated bias, standard error, and RMSE of the various estimators based on the assumption that the unknown data-generating process is nested in the unrestricted model. This assumption implies that the LS-U estimator is unbiased, so

[4] Because of a coding error in their Stata program, Belloni et al. (2014b) exclude interactions between squared initial differences of the eight time-varying controls and the quadratic trend terms.

[5] Unlike Belloni et al. (2014b), who also include controls that are collinear (as they are automatically removed by their user-written Stata routine `lassoShooting`), we first eliminate collinear controls using the `regress` and `_check_omit` commands in Stata. In the models for property and murder crimes, this leads to small differences in the set of controls chosen by the double-selection procedure, which in turn produces small differences in the LS-D estimate of the effect of abortion on murder crime.

we can estimate the bias of all other estimators by the difference with respect to the LS-U estimator. For example, we estimate the bias $b_{1r} = \mathrm{E}[\widehat{\beta}_{1r}] - \beta_1 = (X_1'X_1)^{-1}X_1'X_2\beta_2$ of the LS-R estimator by $\widehat{b}_{1r} = \widehat{\beta}_{1r} - \widehat{\beta}_{1u} = (X_1'X_1)^{-1}X_1'X_2\widehat{\beta}_{2u}$. For the standard errors of the four least-squares estimators, we report both the classical standard errors and the standard errors clustered at the state level. For the two WALS estimators we compare the estimated bias based on the difference with respect to the LS-U estimate and the previously used posterior standard errors (i.e. the estimated standard errors based on the square root of the diagonal elements of the matrix V in (15) and (16)) with the double-shrinkage and ML estimates of the bias and the standard errors discussed in Section 5. For the two WALS estimators we do not report the standard errors clustered at the state level as this would require extending our theoretical results to dependent observations. To our knowledge, the problem of computing clustered standard errors for model averaging estimators is still unexplored. Similarly, very little is known about the clustered standard errors of the LS-D estimator because the double-selection procedure of Belloni et al. (2014b) does not account for serial correlation in the data and the reported standard errors only reflect the effects of clustering in the selected model. For simplicity, we focus our discussion on the comparison of classical standard errors and RMSEs.

The small differences between the LS-R and LS-I estimates are similar to those found by Donohue and Levitt (2001) in their sensitivity analysis. Although unbiased, the LS-U estimate has a large standard error and is the worst in terms of estimated RMSE. The double-selection procedure drastically reduces the number of controls from the original 294 to only a few (between 7 and 9 depending on the type of crime considered). The LS-D estimates have much lower standard errors than the LS-U estimates, but they are about twice the size of the LS-R and LS-I estimates. The classical standard errors of the least-squares estimators are small, but do not account for the noise generated by model selection. The two WALS estimates are less biased but also less precise than the LS-R, LS-I, and LS-D estimates because their standard errors are unconditional. In agreement with our theoretical results we also observe that the posterior standard errors are always larger than the double-shrinkage and ML estimates of the standard errors. In terms of estimated RMSE, the preferred estimators are LS-I/LS-R in the model for property crimes and the two WALS estimators in the model for murder crimes. In the model for violent crimes, the estimated RMSE of these four estimators is about the same.

The two WALS estimates lead to the same policy conclusion as the LS-D estimate of Belloni

et al. (2014b), namely that the empirical evidence for a causal effect of abortion on crime is not robust to the presence of nonlinear trends. In the present case, given the data and the specification employed by Belloni et al. (2014b), it turns out that only a small number of auxiliary variables is important, the rest of them being essentially noise. This explains why the restricted estimator LS-R and the intermediate estimator LS-I perform well, in fact better than the LS-D estimator, in terms of both estimated bias and standard error. Comparisons between the estimated RMSEs of LS-R/LS-I and WALS are less clear cut, possibly because of the extremely rich model space considered. Thus, unlike Belloni et al. (2014b), we emphasize that the above conclusions are subject to considerable sampling uncertainty.

7 Monte Carlo simulations

In the previous section we considered a setting in which the data-generating process is unknown. We now turn to Monte Carlo simulations, in which this process is known.

We employ the same first-difference model as used in the empirical application in Section 6, with the same sample size (600 state-year observations) and the same design. For each type of crime, the model coefficients are set equal to our unrestricted least-squares estimates and the simulated crime rates by state and year are obtained by adding to the estimated value of the linear predictor a pseudo-random draw from a normal distribution with mean zero and variance equal to the classical least-squares estimate of the error variance.

The objective of our Monte Carlo simulations is twofold: evaluating the finite-sample performance of various least-squares and WALS estimators of the parameter of interest, and evaluating the finite-sample performance of the various estimators of their sampling moments. The parameter of interest is the coefficient on the first-difference of the abortion rate which, under the assumed model, measures the causal effect of legalized abortion on crime. This is equal to 0.071 for violent crimes, -0.161 for property crimes, and -1.327 for murder crimes. We compare six estimators of the coefficient of interest: the four least-squares estimators (LS-U, LS-R, LS-I, and LS-D) and the two WALS estimators (WALS-L and WALS-W). The true bias, standard error, and RMSE of the six estimators are approximated by drawing 5,000 Monte Carlo samples. We also assess the finite-sample performance of various estimators of the true sampling moments of the six estimators.

Specifically, for the LS-R, LS-I, and LS-D estimators, we estimate the true bias by the difference with respect to the LS-U estimate and the true standard error by the classical least-squares estimate. For the two WALS estimators, we compare the estimated bias based on the difference with respect to the LS-U estimate and the posterior standard errors with our new double-shrinkage and ML estimates. Since each of these estimates has its own bias, standard error and RMSE, we also report their relative bias, relative standard error, and relative RMSE by taking ratios with respect to the true bias, standard error and RMSE.

TABLE 2 HERE

Table 2 presents the true bias, standard error, and RMSE of the six estimators of the coefficient of interest in the models for each type of crime. 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 standard error. In line with the results of the previous section, we also find that the LS-D estimator is more biased and less precise than the LS-R and LS-I estimators, and that the two WALS estimators have a lower bias but a higher standard error than the LS-R, LS-I and LS-D estimators. Based on the RMSE criterion, the preferred estimators are LS-I and LS-R in the models for violent and property crimes, and the two WALS estimators in the model for murder crimes. These results are not surprising because the assumed data-generating process contains many irrelevant regressors. The emphasis of our Monte Carlo experiment is not on comparing the finite-sample performance of point estimators of the coefficient of interest under different choices of the data-generating process, but rather on comparing the finite-sample performance of various estimators of their sampling moments. In particular, we wish to find out how our new double-shrinkage and ML estimators of the bias and standard error of WALS behave.

TABLE 3 HERE

Table 3 presents the relative bias, standard error, and RMSE of bias estimators for the LS-R, LS-I, LS-D, WALS-L and WALS-W estimators of the coefficient of interest. Although essentially unbiased, the bias estimators based on the difference with respect to the LS-U estimates are rather imprecise due to the uncertainty in estimating the auxiliary coefficients in the unrestricted model. In particular, our double-shrinkage and ML estimators are always more biased than those based on the difference with respect to the LS-U estimator, but they also lead to substantial improvements

in terms of relative standard error and RMSE. As predicted by our results in Section 4.3, the ML estimator of the bias is generally less biased than the corresponding double-shrinkage estimator which, however, is always preferred in terms of relative RMSE.

TABLE 4 HERE

Table 4 presents the relative bias, standard error, and RMSE of the estimated standard errors of our six estimators. The Monte Carlo results confirm that our new double-shrinkage and ML estimators of the standard errors of WALS reduce the substantial upward bias of the previously used estimator (labeled as PSE). The relative performance of the RMSE of these new estimators is comparable to that of the classical standard errors for the correctly specified LS-U estimator. On the other hand, the estimated standard errors for the models chosen by the double-selection procedure of Belloni et al. (2014b) perform poorly in all simulation designs.

8 Conclusions

In this paper we analyzed the finite-sample properties of WALS, a partly Bayesian partly frequentist model averaging estimator which accounts for model uncertainty in a normal linear setup. In particular, we obtained estimators of the bias and variance of WALS by exploiting the corresponding estimators of the sampling bias and variance of the posterior mean in the normal location model.

For priors belonging to the class of (reflected) generalized gamma distributions, such as the Laplace, Weibull and Subbotin priors, the sampling moments of the posterior mean do not admit closed-form expressions. To address this issue we used both Monte Carlo tabulations and analytical approximations based on Taylor series expansions. The Monte Carlo tabulations are very accurate and could easily be extended to other types of priors, other aspects of the sampling distributions of the posterior mean, and other shrinkage estimators of the normal location parameter. Analytical approximations help to better understand the tabulated functional forms of the bias and variance by exploiting the link between derivatives of the posterior mean and higher-order posterior cumulants.

Since the sampling moments of the posterior mean depend on the unknown location parameter, we 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. Simulations show that the former has a relative advantage in terms of bias and good risk performance for large values of

the normal location parameter, while the latter leads to better risk performance for small values of the normal location parameter.

We illustrated the importance of the new estimators of the bias and variance of WALS in a real-data application which analyzes the impact of legalized abortion on crime rates. Results from a related Monte Carlo experiment shows that our estimators of the sampling moments of WALS perform well in finite samples.

Further work is required to investigate the implications of our findings for the WALS approach to inference (i.e., the construction of confidence intervals and testing strategies). Although bias corrections are important for constructing confidence intervals with correct coverage probabilities, we do not recommend a naive approach based on debiased estimates and critical values from the normal distribution, as this would take the classical normal approximation at face value and would ignore the additional uncertainty due to the estimated bias. Preliminary results in this direction suggest that the issue can be addressed by a Monte Carlo approach which exploits our new plug-in estimators of the bias of the posterior mean in the normal location model.

Acknowledgements

We thank Domenico Giannone, Henk Pijls, and Giorgio Primiceri for useful discussions, and an Associate Editor and three anonymous referees for their positive and constructive comments. Giuseppe De Luca acknowledges financial support from MIUR PRIN PRJ-0324.

References

Belloni, A., Chernozhukov, V., Hansen, C., 2014a. Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies* 81, 608–650.

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

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

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J., 2018. Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal* 21, C1–C68.

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

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., Peracchi, F., 2018. Weighted-average least squares estimation of generalized linear models. *Journal of Econometrics* 204, 1–17.

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

Donohue, J. J., 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.

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

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

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

Johnstone, I. M., 2019. *Gaussian estimation: Sequence and wavelet models.* Available at: statweb.stanford.edu/ imj/GE_09_16_19.pdf.

Johnstone, I. M., Silverman, B. W., 2004. Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Annals of Statistics* 32, 1594–1649.

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

Lee, J., Oh, H.-S., 2013. Bayesian regression based on principal components for high-dimensional data. *Journal of Multivariate Analysis* 117, 175–192.

Liang, H., Zou, G., Wan, A. T. K., Zhang, X., 2011. Optimal weight choice for frequentist model average estimators. *Journal of the American Statistical Association* 106, 1053–1066.

Magnus, J. R., 2002. Estimation of the mean of a univariate normal distribution with known variance. *Econometrics Journal* 5, 225–236.

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

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

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

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

Park, T., Casella, G., 2008. The Bayesian Lasso. *Journal of the American Statistical Association* 103, 681–686.

Pericchi, L. R., Sansó, B., 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., 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., Hoeting, J. A., 1997. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* 92, 179–191.

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

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.

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

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 University Press, New York.

Wright, J. H., 2008. Bayesian model averaging and exchange rate forecasts. *Journal of Econometrics* 146, 329–341.

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

Table 1: Impact of legalized abortion on crime

Type of crime	Estimator	Estimated impact	Method for bias/SE	Estimated sampling moments				
				Bias	SE	SE_c	RMSE	$RMSE_c$
Violent	LS-U	0.071	DU	0.000	0.318	0.284	0.318	0.284
	LS-R	−0.157	DU	−0.228	0.046	0.033	0.232	0.230
	LS-I	−0.157	DU	−0.227	0.047	0.034	0.232	0.230
	LS-D	−0.171	DU	−0.242	0.113	0.117	0.267	0.269
	WALS-L	−0.007	DU	−0.078				
			PSE		0.265			
			DS	−0.046	0.224		0.229	
			ML	−0.067	0.234		0.244	
	WALS-W	−0.012	DU	−0.083				
			PSE		0.263			
			DS	−0.046	0.223		0.227	
			ML	−0.067	0.237		0.246	
Property	LS-U	−0.161	DU	0.000	0.135	0.106	0.135	0.106
	LS-R	−0.100	DU	0.061	0.024	0.022	0.066	0.065
	LS-I	−0.106	DU	0.055	0.024	0.021	0.060	0.059
	LS-D	−0.061	DU	0.100	0.042	0.058	0.108	0.115
	WALS-L	−0.134	DU	0.027				
			PSE		0.114			
			DS	0.013	0.097		0.098	
			ML	0.022	0.102		0.104	
	WALS-W	−0.130	DU	0.031				
			PSE		0.114			
			DS	0.012	0.097		0.098	
			ML	0.024	0.103		0.106	
Murder	LS-U	−1.327	DU	0.000	1.485	0.932	1.485	0.932
	LS-R	−0.215	DU	1.112	0.184	0.052	1.127	1.113
	LS-I	−0.218	DU	1.109	0.185	0.068	1.124	1.111
	LS-D	−0.192	DU	1.135	0.416	0.176	1.209	1.149
	WALS-L	−0.849	DU	0.478				
			PSE		1.179			
			DS	0.220	1.016		1.040	
			ML	0.386	1.035		1.104	
	WALS-W	−0.783	DU	0.543				
			PSE		1.146			
			DS	0.213	0.997		1.019	
			ML	0.411	1.024		1.103	

Notes. LS-U and LS-R are the least-squares estimators in the unrestricted and fully restricted models, LS-I is the least-squares estimator in the intermediate model with the eight controls used by Donohue and Levitt (2001), LS-D is the least-squares estimator in the intermediate model with the controls selected by the double-selection procedure of Belloni et al. (2014), and WALS-L and WALS-W are the WALS estimators based on the Laplace and Weibull priors. All models are estimated in first-differences as explained in Section 6. Methods for estimating the bias/SE: difference with respect to LS-U (DU), posterior standard error (PSE), double-shrinkage (DS), maximum likelihood (ML). SE_c and $RMSE_c$ denote, respectively, the standard error and the RMSE clustered at the state level.

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

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

Notes. LS-U and LS-R are the least-squares estimators in the unrestricted and fully restricted models, LS-I is the least-squares estimator in the intermediate model with the eight controls used by Donohue and Levitt (2001), LS-D is the least-squares estimator in the intermediate model with the controls selected by the double-selection procedure of Belloni et al. (2014), and WALS-L and WALS-W are the WALS estimators based on the Laplace and Weibull priors.

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

		Violent crimes			Property crimes			Murder crimes		
Estimator	Method for bias	Relat. Bias	Relat. SE	Relat. RMSE	Relat. Bias	Relat. SE	Relat. RMSE	Relat. Bias	Relat. SE	Relat. RMSE
LS-R	DU	0.005	1.389	1.389	0.007	2.188	2.188	0.000	1.309	1.309
LS-I	DU	0.005	1.392	1.392	0.008	2.421	2.422	0.000	1.311	1.311
LS-D	DU	0.004	1.223	1.223	0.005	1.441	1.441	0.000	1.244	1.244
WALS-L	DU	0.016	1.386	1.386	0.020	1.697	1.697	0.001	1.210	1.210
	DS	0.323	0.947	1.000	−0.398	1.186	1.251	−0.401	0.768	0.867
	ML	0.127	1.218	1.225	−0.132	1.498	1.504	−0.145	1.046	1.056
WALS-W	DU	0.016	1.418	1.418	0.018	1.628	1.629	0.001	1.195	1.195
	DS	0.340	0.927	0.987	−0.430	1.087	1.169	−0.435	0.719	0.841
	ML	0.158	1.188	1.199	−0.173	1.369	1.380	−0.186	0.983	1.000

Notes. LS-U and LS-R are the least-squares estimators in the unrestricted and fully restricted models; LS-I is the least-squares estimator in the intermediate model with the eight controls used by Donohue and Levitt (2001); LS-D is the least-squares estimator in the intermediate model with the controls selected by the double-selection procedure of Belloni et al. (2014); WALS-L and WALS-W are the WALS estimators based on the Laplace and Weibull priors. Methods for estimating the bias: difference with respect to LS-U (DU), double-shrinkage (DS), maximum likelihood (ML).

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

		Violent crimes			Property crimes			Murder crimes		
Estimator	Method for SE	Relat. Bias	Relat. SE	Relat. RMSE	Relat. Bias	Relat. SE	Relat. RMSE	Relat. Bias	Relat. SE	Relat. RMSE
LS-U	LS	−0.001	0.041	0.041	−0.012	0.041	0.042	0.010	0.042	0.043
LS-R	LS	0.285	0.035	0.287	0.303	0.035	0.305	0.155	0.033	0.158
LS-I	LS	0.282	0.035	0.284	0.295	0.035	0.297	0.146	0.033	0.150
LS-D	LS	0.294	0.039	0.297	0.170	0.061	0.181	0.183	0.034	0.186
WALS-L	PSE	0.158	0.047	0.165	0.127	0.044	0.134	0.203	0.051	0.210
	DS	−0.012	0.040	0.042	−0.034	0.037	0.050	0.025	0.042	0.049
	ML	0.039	0.043	0.058	0.016	0.040	0.043	0.069	0.047	0.083
WALS-W	PSE	0.149	0.048	0.157	0.113	0.044	0.121	0.206	0.055	0.213
	DS	−0.017	0.042	0.046	−0.042	0.038	0.057	0.029	0.044	0.053
	ML	0.048	0.046	0.067	0.021	0.041	0.046	0.089	0.052	0.104

Notes. LS-U and LS-R are the least-squares estimators in the unrestricted and fully restricted models; LS-I is the least-squares estimator in the intermediate model with the eight controls used by Donohue and Levitt (2001); LS-D is the least-squares estimator in the intermediate model with the controls selected by the double-selection procedure of Belloni et al. (2014); WALS-L and WALS-W are the WALS estimators based on the Laplace and Weibull priors. Methods for estimating the SE: least-squares (LS), posterior variance as estimated sampling variance of the posterior standard error (PSE), double-shrinkage (DS), maximum likelihood (ML).

Figure 1: Monte Carlo (MC) and truncated series (TS) approximations to the bias $\delta(\eta)$ of the posterior mean $m(x)$ under normal, Laplace, Weibull, and Subbotin priors.

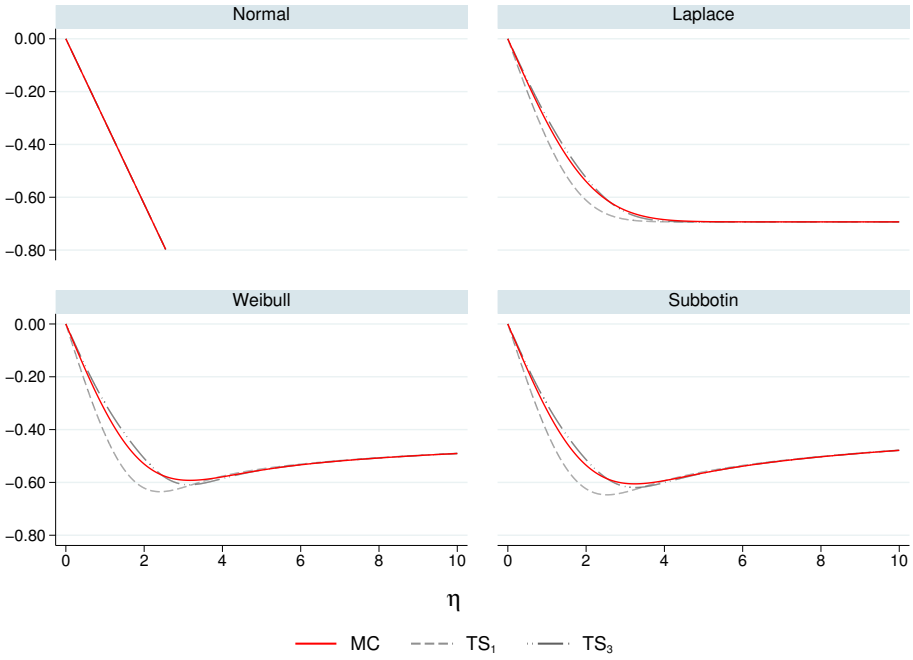


Figure 2: Monte Carlo (MC) and truncated series (TS) approximations to the variance $\sigma^2(\eta)$ of the posterior mean $m(x)$ under normal, Laplace, Weibull, and Subbotin priors.

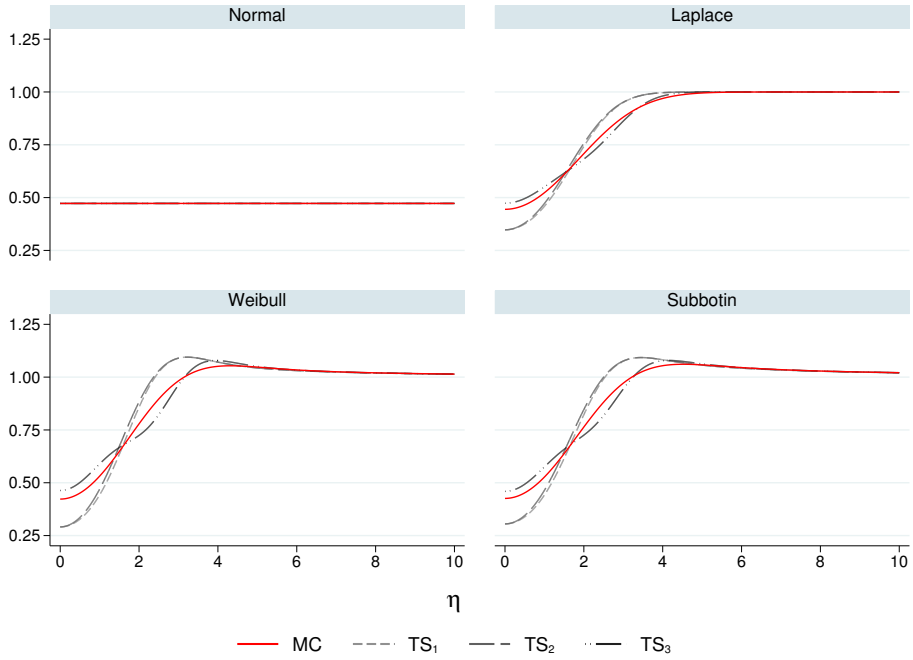


Figure 3: Bias and RMSE of the Maximum Likelihood (ML) and Double-Shrinkage (DS) estimators of the bias $\delta(\eta)$ of the posterior mean $m(x)$ under Laplace and Weibull priors.

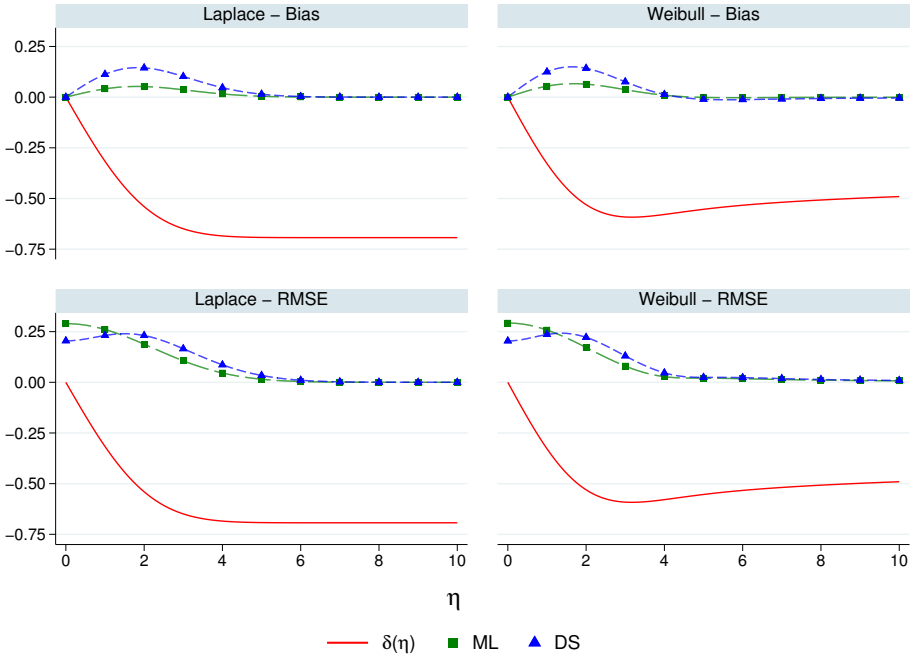
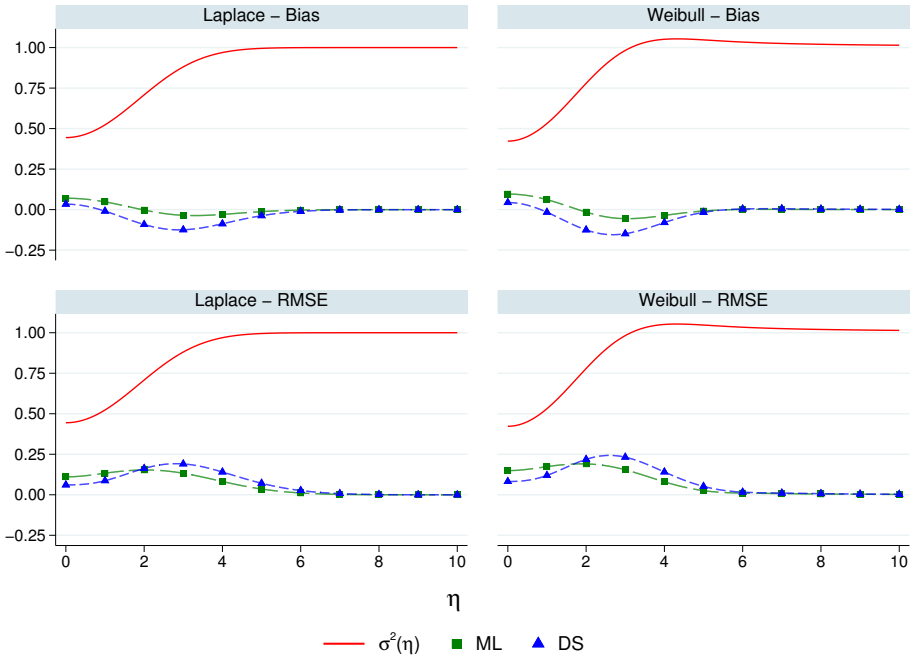


Figure 4: Bias and RMSE of the Maximum Likelihood (ML) and Double-Shrinkage (DS) estimators of the sampling variance $\sigma^2(\eta)$ of the posterior mean $m(x)$ under Laplace and Weibull priors.



Appendix A: Proofs

Proof of Proposition 1: The stated assumptions on the prior guarantee that the function $A_0(x)$ exist and admit 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.

Proof of Proposition 2: 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 2.2 in Pericchi et al. (1993) implies that $a_j(\eta)$ $(j \geq 1)$ is equal to the posterior cumulant of order $j+1$ evaluated at η, that is $a_j(\eta) = c_{j+1}(\eta)$. We then obtain the approximations

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

and

$$\begin{aligned}
\sigma_h^2(\eta) = \mathrm{var}\left[m_h(x)|\eta\right] &= \sum_{j=1}^{h}\sum_{k=1}^{h} \frac{a_j(\eta) a_k(\eta) \,\mathrm{cov}(z^j, z^k)}{j!k!} \\
&= \sum_{j=1}^{h} \left[\binom{2j}{j} q_{2j} - q_j^2\right] c_{j+1}(\eta)^2 + 2 \sum_{k<j} \left[\binom{j+k}{j} q_{j+k} - q_j q_k\right] c_{j+1}(\eta) c_{k+1}(\eta),
\end{aligned}$$

where q_j denotes the jth moment of the standard-normal distribution. The results follow.

Appendix B: Generalized gamma priors

In our application to WALS, we focus on priors belonging to a rich and mathematically tractable three-parameter class of priors, namely the (reflected) generalized gamma distributions with density

$$\pi(\eta; a, b, c) = \frac{cb^d}{2\Gamma(d)}|\eta|^{-a}\exp\left(-b|\eta|^c\right) \qquad (\eta \in \mathbb{R}), \tag{B.1}$$

where $0 \leq a < 1$, $b > 0$, $c > 0$, $d = (1-a)/c$, and $\Gamma(d)$ is the gamma function. In addition to the one-parameter family of normal 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 normal prior with zero mean and finite variance $\omega^2 > 0$ is convenient because the posterior is then also normal 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 sampling bias and variance of $m(x)$ are

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

respectively. Notice that the posterior variance of η is not equal to the sampling variance of $m(x)$, but rather to its standard deviation. This is not a peculiar feature of the normal prior but, as shown in Section 4.2, holds approximatively for any positive and bounded prior density.

The normal 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 normal 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 (B.2), the bias $\delta(\eta)$ of $m(x)$ is a linear, hence unbounded, function of η.

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

$$m(x) = x - bh(x), \qquad v^2(x) = 1 + b^2\left[1 - h(x)^2\right] - \frac{b\left(1+h(x)\right)\phi(x-b)}{\Phi(x-b)}, \tag{B.3}$$

where $\Phi(\cdot)$ denotes the distribution function of the standard-normal 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 normal 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 normal prior.

In contrast, the 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.

Our choice of the free prior parameters in (B.1) 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 $b = 0.23$ for the normal 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

$$\text{regret}(\eta; c) = \text{risk}(\eta; c) - \frac{\eta^2}{1+\eta^2} = \int_{-\infty}^{\infty} (m(x; c) - \eta)^2 \, \phi(x - \eta) dx - \frac{\eta^2}{1+\eta^2}, \tag{B.4}$$

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 $c = 0.80$ ($b = 0.94$) for the Subbotin distribution and $c = 0.89$ for the Weibull distribution.

Appendix C: Convergence of the Taylor series of $m(x)$

Given the posterior density $p(\eta|x)$ and the integration constant $A_0(x)$ defined in (8), we obtain the posterior mean $m(x)$ defined in (9). We are interested in the Taylor series of $m(x)$ at some point ξ.

In particular, we want to know whether the Taylor series is convergent for all x and, if not, what the radius of convergence is.

Since $m(x) = x + A_0'(x)/A_0(x)$, let us first investigate $A_0(x)$, which we rewrite as

$$A_0(x) = e^{-x^2/2} \int_{-\infty}^{\infty} e^{x\eta} \phi(\eta)\pi(\eta)\, d\eta.$$

This shows that $A_0(x) > 0$ and hence that $m(x)$ is analytic on $\mathbb{R}$. To prove the convergence of the Taylor series of $m(x)$ we need a stronger result, namely that m is analytic on $\mathbb{C}$.

Replacing x (real) by $z = x + iy$ (complex), we obtain

$$A_0(z) = e^{-z^2/2} \int_{-\infty}^{\infty} e^{z\eta} \phi(\eta)\pi(\eta)\, d\eta = e^{-z^2/2} \int_{-\infty}^{\infty} e^{x\eta} \left(\cos(y\eta) + i \sin(y\eta)\right) \phi(\eta)\pi(\eta)\, d\eta,$$

from which it follows that $A_0(z)$ is analytic on $\mathbb{C}$. Hence, $m(z)$ is analytic on $\mathbb{C}$ outside the zeros of $A_0(z)$, and these zeros are given as the solutions of the equations

$$F_1(x,y) = \int_{-\infty}^{\infty} e^{x\eta} \cos(y\eta) \phi(\eta)\pi(\eta)\, d\eta = 0, \qquad F_2(x,y) = \int_{-\infty}^{\infty} e^{x\eta} \sin(y\eta) \phi(\eta)\pi(\eta)\, d\eta = 0,$$

or equivalently as the solutions of

$$G_1(x,y) = \int_{0}^{\infty} \cosh(x\eta) \cos(y\eta) \phi(\eta)\pi(\eta)\, d\eta = 0 \tag{C.1}$$

and

$$G_2(x,y) = \int_{0}^{\infty} \sinh(x\eta) \sin(y\eta) \phi(\eta)\pi(\eta)\, d\eta = 0. \tag{C.2}$$

Let $z_j = x_j + iy_j$ $(j = 1, \ldots, J)$ be the points where $G_1(x_j, y_j) = G_2(x_j, y_j) = 0$, that is, where $A_0(z_j) = 0$. For given (real) ξ the radius of convergence is then given by

$$R(\xi) = \min_j |z_j - \xi| = \min_j \sqrt{(x_j - \xi)^2 + y_j^2}. \tag{C.3}$$

The Taylor series of $m(x)$ around $x = \xi$ thus converges for $|x - \xi| < R(\xi)$ and diverges for $|x - \xi| > R(\xi)$.

The radius $R(\xi)$ depends on the prior π. If the prior is normal or uniform then the radius is infinite and the Taylor series converges for all x. But the Taylor series based on the Laplace,

Weibull, and Subbotin priors have finite (and similar) radii of convergence. For the Laplace prior the minimum radius of convergence equals 2.97, which is reached at $\xi = \pm 2.25$; for Weibull the minimum is 2.59, reached at $\xi = \pm 2.22$; and for Subbotin the minimum is 2.70, reached at $\xi = \pm 2.29$.

Appendix D: An apparent contradiction

The results in Section 4.2 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 $v_n^2(x) = w_n/n$, where $w_n = \omega^2/(\omega^2 + 1/n)$. The frequentist variance of $m_n(x)$ is $\sigma_n^2 = \text{var}[m_n(x)] = w_n^2/n$, 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 normal prior. When we consider another prior, say the Laplace prior defined by $\pi(\eta) = b\, e^{-b|\eta|}/2$ with $b > 0$, then the same reasoning applies using the formulae in De Luca et al. (2020).