# Bayesian estimation of the normal location model: A non-standard approach

Giuseppe De Luca
University of Palermo, Italy

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

Franco Peracchi
EIEF and University of Rome Tor Vergata, Italy

February 19, 2024

*Abstract:* We consider the estimation of the location parameter $\theta$ in the normal location model, and study the sampling properties of estimators derived from a Bayesian approach that places a prior on a scaled version of $\theta$ interpreted as the 'population $t$-ratio'. Our non-standard approach is motivated by the fact that in model selection and model averaging, it is the $t$-ratio rather than the parameter estimate that plays an important role. We show that the finite-sample distribution of these estimators is not centered at $\theta$ and is generally non-normal. We also show that the speed at which the estimation bias vanishes as the sample size increases critically depends on the choice of prior. In the asymptotic theory, we prove uniform $\sqrt{n}$-consistency of our estimators and obtain their asymptotic distribution under a general moving-parameter setup that includes both the fixed-parameter and the local-parameter settings as special cases. Our results have direct implications for the WALS estimator of Magnus et al. (2010). They may also be helpful for other model selection or model averaging procedures.

*Corresponding author:* Franco Peracchi
`franco.peracchi@uniroma2.it`
EIEF and University of Rome Tor Vergata, Italy

# 1    Motivation

We reconsider possibly the simplest of all estimation problems, namely the estimation of a parameter $\theta$ based on a single observation $x$, normally distributed with unknown mean $\theta$ and unit variance, that is,

$$x \sim \mathrm{N}(\theta, 1). \tag{1}$$

Model (1) is known as the (univariate) *normal location model* and the problem of estimating the location parameter $\theta$ in this model is known as the *normal location problem*. The normal location problem has been studied from various angles; see, for example, Pericchi and Smith 1992; Magnus 1999, 2002; Johnstone and Silverman 2004; Danilov 2005; Kumar and Magnus 2013; DasGupta and Johnstone 2014; Johnstone 2019; and De Luca et al. 2022.

We offer two motivations for studying this seemingly trivial problem. First, the obvious solution of estimating $\theta$ by the maximum likelihood estimator $\hat{\theta} = x$ may not be a good choice. Second, despite its simplicity, this problem is closely related to the more challenging issue of determining the effects of model selection in linear regression models.

Let us consider both motivations in more detail. The maximum likelihood estimator $\hat{\theta}$, sometimes called the 'usual' estimator, is unbiased, has unit variance and, under squared error loss, its risk is equal to its mean squared error $\mathrm{MSE}(\hat{\theta}) = 1$. It is the unique minimax estimator of $\theta$ (van der Vaart 1998, Proposition 8.6) and it is admissible (Berger 1985, p. 548). One may therefore wonder why we wish to consider alternative estimators. To answer this question, let us compare $\hat{\theta}$ to another estimator of $\theta$, namely the 'silly' estimator $\breve{\theta} = 0$. Since $\mathrm{MSE}(\breve{\theta}) = \theta^2$, we find that the silly estimator is 'better' than the usual estimator in the mean squared error sense if and only if $|\theta| < 1$. This simple comparison shows that, although no estimator dominates the usual estimator for all values of $\theta$, it is easy to find other estimators with smaller risk in a non-negligible region of the parameter space.

Ideally, we would like to find an admissible estimator which behaves like $\breve{\theta}$ for 'sufficiently small' values of $\theta$, and like $\hat{\theta}$ for 'sufficiently large' values of $\theta$. This intuitive idea leads to the 'pretest' estimator

$$\bar{\theta} = \begin{cases} x & \text{if } |x| > c, \\ 0 & \text{if } |x| \leqslant c, \end{cases} \tag{2}$$

3

for some $c \geqslant 0$ (possibly $c = 1$), which we can also write as

$$\bar{\theta} = \lambda(x)x \qquad \lambda(x) = \begin{cases} 1 & \text{if } |x| > c, \\ 0 & \text{if } |x| \leqslant c. \end{cases} \tag{3}$$

The function $\lambda(x)$ may be interpreted either as a shrinkage function that shrinks the usual estimator of $\theta$ towards zero, or as a weight function employed to combine the usual and the silly estimators of $\theta$ with weights $\lambda(x)$ and $1 - \lambda(x)$ respectively. Unfortunately, $\bar{\theta}$ is inadmissible (it has a discontinuity at $x = \pm c$), and it has other undesirable properties as well; see Judge and Bock 1983; Pötscher 1991; Kabaila 1995; Magnus 1999; Leeb and Pötscher 2003, 2005, 2006a; Kabaila and Leeb 2006; and Efron 2014.

A continuous version of the pretest estimator is obtained when we allow $\lambda(x)$ to increase smoothly with $|x|$. The question then arises how to choose the function $\lambda(x)$. A simple example is

$$\lambda(x) = \frac{x^2}{x^2 + 1}, \tag{4}$$

but there are many alternative specifications. Magnus (2002) shows that this smooth generalization of the pretest estimator already produces an estimator of $\theta$ with better sampling properties. However, the estimator based on (4) is still inadmissible (Strawderman and Cohen 1971, p. 278).

Although admissibility is a frequentist concept, it is intimately connected to Bayesian ideas. In fact, subject only to mild conditions, any (generalized) Bayes estimator is admissible (Berger 1985, Section 4.8). Thus, instead of specifying the function $\lambda(x)$, we may take a Bayesian approach by specifying a prior distribution $\pi(\theta)$ for $\theta$. Under squared error loss, the posterior mean $\tilde{\theta} = \mathbb{E}(\theta|x)$ is our estimator of $\theta$, and we are interested in its *frequentist* properties. Thus, our approach is to use the prior as a vehicle to obtain a frequentist estimator that enjoys good sampling properties. For previous studies on the frequentist properties of Bayesian estimators see, for example, Joanes and Peers (1974), Carlin and Louis (2000, Chapter 4), Johnstone and Silverman (2004), Efron (2012, 2015), and De Luca et al. (2022).

Expanding on our second motivation, the (univariate) normal location problem is of intrinsic interest, but it is also essential in the theory of pretest, shrinkage, and model averaging estimators, which can all be viewed as estimating the location parameters in the diagonal form of a Gaussian sequence

model (Johnstone 2019, Chapter 2). For example, the pretest estimator corresponds to a two-step procedure where one first 'tests' whether $|\theta| > c$, and then chooses between the usual and the silly estimator based on this preliminary 'test'. This procedure mimics the usual practice in applied regression analysis, where the analyst first performs a $t$-test and then chooses the model (and hence the estimator) based on the outcome of this preliminary $t$-test. Other examples include the empirical Bayes thresholding estimator of Johnstone and Silverman (2004), the penalized least squares estimators analyzed by Pötscher and Leeb (2009), the weighted-average least squares (WALS) estimator of Magnus et al. (2010), the Bayesian model averaging estimator of Lee and Oh (2013), and the wavelet shrinkage estimators discussed in Johnstone (2019, Chapter 7).

Consider the linear regression model

$$y = X_1 \beta_1 + X_2 \beta_2 + u, \qquad u \sim \mathrm{N}(0, \sigma^2 I_n), \tag{5}$$

where $X_1$ $(n \times k_1)$ is a matrix of 'focus' variables that are required to be in the model on theoretical or other grounds, while there is doubt whether the additional 'auxiliary' variables in $X_2$ $(n \times k_2)$ should be in the model or not. Thus, there are $2^{k_2}$ models to consider. If $\hat{\beta}_{1(i)}$ denotes the estimator of $\beta_1$ in model $i$, then

$$\tilde{\beta}_1 = \sum_{i=1}^{2^{k_2}} \omega_i \hat{\beta}_{1(i)} \tag{6}$$

is the model-averaging estimator of $\beta_1$. When the auxiliary regressors in $X_2$ can be made orthogonal using some preliminary data transformation, the equivalence theorem of Magnus and Durbin (1999) and Danilov and Magnus (2004) shows that the choice of model weights $\omega_i$ in (6) is equivalent to the choice of $\lambda(x)$ in a shrinkage estimator for the location parameter $\theta$. In particular, the theory developed for our Bayesian shrinkage estimator of $\theta$ has direct implications on the sampling properties of the WALS estimator; see Magnus and De Luca (2016) for a review of the WALS approach and De Luca et al. (2018, 2022, 2023) for additional generalizations and developments of this model averaging procedure.

The present paper considers a small but important variant of model (1), which allows us to study the sampling properties of the posterior mean in large samples. This variant is explored in Section 2, where we distinguish between two Bayesian approaches to the normal location problem: the standard approach and the one we propose. Section 3 discusses regularity conditions

placed on the prior to ensure that the posterior mean resulting from the proposed Bayesian approach satisfies certain desirable properties, and Section 4 specifies four priors which satisfy these conditions. Section 5 studies the finite-sample distribution of the posterior mean. In Section 6 we investigate uniform $\sqrt{n}$-consistency and in Section 7 we investigate the asymptotic distribution of the posterior mean using a general moving-parameter setup that includes both the fixed-parameter and the local-parameter settings as special cases. Section 8 concludes. Proofs of all propositions are in the Appendix.

## 2  Two Bayesian approaches

In model (1) there is only one observation, and hence the sample size plays no role and any discussion of asymptotics is meaningless. Things change, however, if we replace $x$ with a sample statistic $x_n$ (e.g., a sample mean) that depends on the sample size $n$ and satisfies

$$x_n \sim \mathrm{N}\left(\theta, \frac{\sigma^2}{n}\right),\tag{7}$$

where $\sigma^2 > 0$ is assumed to be known. The assumption that $\sigma^2$ is known is typically motivated by the wish to derive simple and easy-to-interpret approximations that can prove helpful for the asymptotic analysis; see, e.g., Pötscher 2006, Pötscher and Leeb 2009, and DasGupta and Johnstone 2014. The assumption can be relaxed if $\sigma^2$ is estimated consistently by an estimator that is independent of $x_n$; see, e.g., Leeb and Pötscher 2003. The normality assumption, on the other hand, is employed to derive our Bayesian estimator of $\theta$ and its finite-sample distribution, and it will be weakened in the large sample theory of Sections 6 and 7.

The obvious estimator of $\theta$ in model (7) is $x_n$ (the 'usual' estimator), which is unbiased and consistent. The mean square error of the usual estimator is $\mathrm{MSE}(x_n) = \sigma^2/n$, while the silly estimator has $\mathrm{MSE}(0) = \theta^2$. So we prefer the silly estimator (in the MSE sense) if and only if $|\theta| < \sigma/\sqrt{n}$. Defining

$$x_n^* = \frac{x_n}{\sigma/\sqrt{n}}, \qquad \theta_n^* = \frac{\theta}{\sigma/\sqrt{n}},\tag{8}$$

we can write (7) equivalently as

$$x_n^* \sim \mathrm{N}(\theta_n^*, 1).\tag{9}$$

With $\sigma$ known, the transformed random variable $x_n^*$ is the $t$-ratio for testing the hypothesis $\theta = 0$, so its mean $\theta_n^*$ may be viewed as the 'population $t$-ratio'. There is no essential difference between the approach via (7) and the approach via (9). However, if we add a prior, then it *does* make a difference whether we place the prior on $\theta$ or on $\theta_n^*$.

The standard Bayesian approach places a prior on $\theta$. This prior does not depend on the sample size, so the resulting posterior mean of $\theta$ given $x_n$ is consistent for $\theta$ (van der Vaart 1998, Chapter 10), because, as $n$ increases, the data information becomes increasingly important and dominates the prior information which remains constant.

Since $\mathrm{MSE}(0) < \mathrm{MSE}(x_n)$ if and only if $|\theta_n^*| < 1$ and, more generally, since model selection and model averaging typically depend on diagnostics (such as $t$-ratios) rather than on parameter estimates, it makes sense in our context to place a prior on $\theta_n^*$ rather than on $\theta$. This approach, first suggested by Huntsberger (1955) and Hjort (1986), plays a key role in the Bayesian shrinkage step of the WALS estimator of Magnus et al. (2010). An important issue is whether this non-standard approach leads to a consistent estimator of $\theta$. For example, suppose the prior distribution of $\theta_n^*$ is $\mathrm{N}(0, \tau^2)$, where $\tau$ does not depend on $n$. Then the posterior distribution of $\theta_n^*$ is $\mathrm{N}(\xi\, x_n^*, \xi)$, with $\xi = \tau^2/(1 + \tau^2)$. Under quadratic loss, the Bayes estimator of $\theta_n^*$ is the posterior mean $m_n^* = \mathbb{E}(\theta_n^*|x_n^*) = \xi\, x_n^*$, while $\tilde{\theta}_n = \sigma\, m_n^*/\sqrt{n} = \xi\, x_n$ is the implied estimator of $\theta$. From a frequentist perspective, the bias and variance of $\tilde{\theta}_n$ are $\mathbb{E}(\tilde{\theta}_n) - \theta = (\xi - 1)\theta$ and $\mathbb{V}(\tilde{\theta}_n) = \xi^2 \sigma^2/n$, so the variance of $\tilde{\theta}_n$ vanishes as $n \to \infty$ but the bias does not. Hence, with a normal prior for $\theta_n^*$, $\tilde{\theta}_n$ is not a consistent estimator of $\theta$.

We thus need conditions on the prior to ensure that $\tilde{\theta}_n$ enjoys attractive sampling properties as an estimator of $\theta$. Henceforth we shall write the posterior mean of $\theta_n^*$ as $m_n^* = m(x_n^*)$ and the posterior variance as $v_n^{*2} = v^2(x_n^*)$, where the functions $m\colon \mathbb{R} \to \mathbb{R}$ and $v^2\colon \mathbb{R} \to \mathbb{R}_+$ have properties that depend on those of the prior density $\pi$ of $\theta_n^*$.

# 3 Regularity conditions on the prior

Our first three conditions on the prior $\pi$ are just mild regularity conditions:

(C1) $\pi$ is symmetric around zero;
(C2) $\pi$ is positive and non-increasing on $(0, \infty)$; and

(C3) $\pi$ is differentiable, except possibly at 0.

Kumar and Magnus (2013) show that, under (C1)–(C3), the posterior mean function $m$ satisfies the following properties: $m(-x) = -m(x)$, $m(0) = 0$, $m'(x) > 0$, $0 < m(x) < x$ for $x > 0$, and $m(x) \to \infty$ as $x \to \infty$. Since $m$ is continuous and strictly increasing, its inverse function $\ell\colon \mathbb{R} \to \mathbb{R}$ exists, is continuous, strictly increasing, and satisfies $\ell(-t) = -\ell(t)$, $\ell(0) = 0$, and $\ell(t) \to \infty$ as $t \to \infty$. Because $m$ is a shrinkage rule, there exists a continuous and symmetric shrinkage function $w(x) = m(x)/x$ for $x \neq 0$, satisfying $w(-x) = w(x)$, $0 < w(x) < 1$, and $w(x) \to v^2(0)$ as $x \to 0$. The latter property follows immediately by l'Hôpital's rule and the Brown–Tweedie formula (Brown 1971; Pericchi and Smith 1992): $\lim_{x \to 0} m(x)/x = \lim_{x \to 0} m'(x) = v^2(0)$.

An important requirement for posterior inference is that, when the data information is sufficiently strong, the prior should have bounded influence on $m(x)$ (Sansó and Pericchi 1992). If this is not the case, then the MSE of $m_n^*$ as an estimator of $\theta_n^*$ is not bounded in $\theta_n^*$ (Brown 1971). Although $m(x)$ is bounded from above by $x$ (for $x > 0$), conditions (C1)–(C3) are not enough to imply this additional property.

To see this, consider again the example of a normal prior for $\theta_n^*$ and let us introduce the following two functions:

$$g(x) = x - m(x) = (1 - w(x))x, \qquad \psi(\theta) = -\frac{d \ln \pi(\theta)}{d\theta} = -\frac{\pi'(\theta)}{\pi(\theta)}. \tag{10}$$

Since $g(x)$ measures the deviation between $x$ and $m(x)$, we call it the 'discrepancy function'. In our example, $g(x) = (1 - \xi)x$, $w(x) = \xi$, and $\psi(\theta) = \theta/\tau^2$. Hence, the bias of $m_n^*$ is equal to $-g(\theta_n^*)$ and its MSE is equal to $\xi^2 + g(\theta_n^*)^2$, both of which are unbounded in $\theta_n^*$ because the function $g$ is unbounded. Since $g(\theta_n^*) = \xi \psi(\theta_n^*)$, it follows that the bias and the MSE of $m_n^*$ are unbounded in $\theta_n^*$ because the function $\psi$ is unbounded.

To avoid this problem, we therefore impose the following additional condition:

(C4) $\psi(\theta) \to \psi_0$ as $\theta \to \infty$, where $\psi_0 \geqslant 0$ is some finite constant.

Kumar and Magnus (2013, Theorem 4.1) prove that $g(x) \to 0$ if and only if $\psi_0 = 0$. A straightforward extension of their theorem shows that, under

(C1)–(C4), $g(x) \to \psi_0 < \infty$ as $x \to \infty$.[1] This implies that the estimator $m_n^*$ of $\theta_n^*$ has bounded MSE and is admissible. When $\psi_0 = 0$, it also implies the stronger property, sometimes called 'Bayesian robustness', that $x - m(x) \to 0$ and $v^2(x) \to 1$ as $x \to \infty$, so that prior information is essentially ignored when $x$ is sufficiently large. For robust priors (i.e., when $\psi_0 = 0$), we also have $w(x) \to 1$ as $x \to \infty$.

In a Bayesian analysis one often needs to formalize the concept of prior ignorance, and a flat (improper) prior is then typically used for computational convenience. In our context, a flat prior fails to capture a notion of prior ignorance which we call 'neutrality', namely not knowing whether or not $|\theta_n^*| < 1$, that is, whether or not the silly estimator of $\theta$ has a lower MSE than the usual estimator. Our final condition on $\pi$ ensures that neutrality is satisfied:

(C5) $\pi$ satisfies $\int_0^1 \pi(\theta) \, d\theta = \int_1^\infty \pi(\theta) \, d\theta = 1/4$.

Together with condition (C1), condition (C5) implies that the events $|\theta_n^*| < 1$ and $|\theta_n^*| > 1$ are equally likely a priori.

# 4    Specification of the prior

Many priors satisfy conditions (C1)–(C5). We shall study four. The first two priors are based on the gamma function

$$\Gamma(r) = \int_0^\infty t^{r-1} e^{-t} \, dt \qquad (r > 0). \tag{11}$$

If we substitute $t = c\theta^b$ with $dt = bc\theta^{b-1} \, d\theta$, and let $r = (1-a)/b$, we obtain the class of reflected gamma-type priors

$$\pi(\theta) = \frac{bc^{(1-a)/b}}{2\Gamma((1-a)/b)} |\theta|^{-a} e^{-c|\theta|^b} \qquad (0 \leqslant a < 1, \, b > 0, \, c > 0), \tag{12}$$

where we have imposed the additional restriction $a \geqslant 0$, because otherwise the prior will be increasing rather than decreasing for small values of $\theta > 0$. Special cases are the Weibull prior ($a + b = 1$):

$$\pi(\theta) = \frac{bc}{2} |\theta|^{-(1-b)} e^{-c|\theta|^b} \qquad (0 < b \leqslant 1, \, c > 0), \tag{13}$$

---

[1] A proof of this result, omitted for brevity, is available from the authors upon request.

and the Laplace prior ($a = 0$ and $b = 1$), which is a special case of the Weibull prior:

$$\pi(\theta) = \frac{c}{2} e^{-c|\theta|} \qquad (c > 0). \tag{14}$$

The Weibull (and hence also the Laplace) prior satisfies regularity conditions (C1)–(C4). As shown in Kumar and Magnus (2013) and Magnus and De Luca (2016), a reflected generalized gamma prior is robust if and only if $0 < b < 1$ and is neutral if and only if

$$\int_0^c t^{(1-a-b)/b} e^{-t} \, dt = \frac{\Gamma((1-a)/b)}{2}.$$

Thus, the Weibull prior with $b < 1$ is robust, but the Laplace prior is not (although it has bounded influence since $\psi_0 = c > 0$). Neutrality leads to $c = \ln 2$ for both priors. For the Weibull prior, we fix the prior parameter $b$ by the minimax regret criterion for $m_n^*$, with regret defined as the difference between the MSE of $m_n^*$ and the minimum MSE in estimating $\theta_n^*$. Based on this criterion, Magnus and De Luca (2016) find that the 'optimal' neutral and robust Weibull prior is achieved for $b \approx 0.8876$ and $c = \ln 2$.

The other two priors are based on the beta function

$$B(r, s) = \frac{\Gamma(r)\Gamma(s)}{\Gamma(r+s)} = \int_0^1 t^{r-1}(1-t)^{s-1} \, dt \qquad (r > 0, \, s > 0). \tag{15}$$

Substituting $t = (1 + c\theta^b)^{-1}$ with $dt = -bc(1 + c\theta^b)^{-2}\theta^{b-1}d\theta$, and letting $r = 1/a - 1/b$ and $s = 1/b$, we obtain the class of reflected beta-type priors

$$\pi(\theta) = \frac{c^{1/b}b}{2 \, B(1/a - 1/b, 1/b)}(1 + c|\theta|^b)^{-1/a} \qquad (0 < a < b, \, c > 0), \tag{16}$$

special cases of which are the Pareto prior ($b = 1$):

$$\pi(\theta) = \frac{c(1-a)}{2a}(1 + c|\theta|)^{-1/a} \qquad (0 < a < 1, \, c > 0) \tag{17}$$

and the Cauchy prior ($a = 1$, $b = 2$, $c = 1$):

$$\pi(\theta) = \frac{1}{\pi} \cdot \frac{1}{1 + \theta^2}, \tag{18}$$

where $\pi(\theta)$ denotes the prior function and $\pi \approx 3.14$ denotes Archimedes' constant. The Pareto and Cauchy priors also satisfy conditions (C1)–(C4).

10

A reflected beta-type prior is robust when $b > a > 0$, and hence both the Pareto and the Cauchy priors are robust. The reflected beta-type prior is neutral when

$$\int_{1/(c+1)}^{1} t^{r-1}(1-t)^{s-1}\, dt = \frac{B(r,s)}{2},$$

where $r = 1/a - 1/b$ and $s = 1/b$. Specifically, the Pareto prior is neutral for $c = 2^{a/(1-a)} - 1$, while the Cauchy prior is always neutral. For the Pareto prior, we also fix the prior parameter $a$ by the minimax regret criterion for $m_n^*$. The 'optimal' neutral and robust Pareto prior is achieved for $a \approx 0.0862$ and $c \approx 0.0676$. The maximum regret for $m_n^*$ is equal to 0.4546 under the Weibull prior, 0.4959 under the Pareto prior, 0.5127 under the Laplace prior, and 0.6332 under the Cauchy prior.

Figure 1: Posterior mean and discrepancy functions
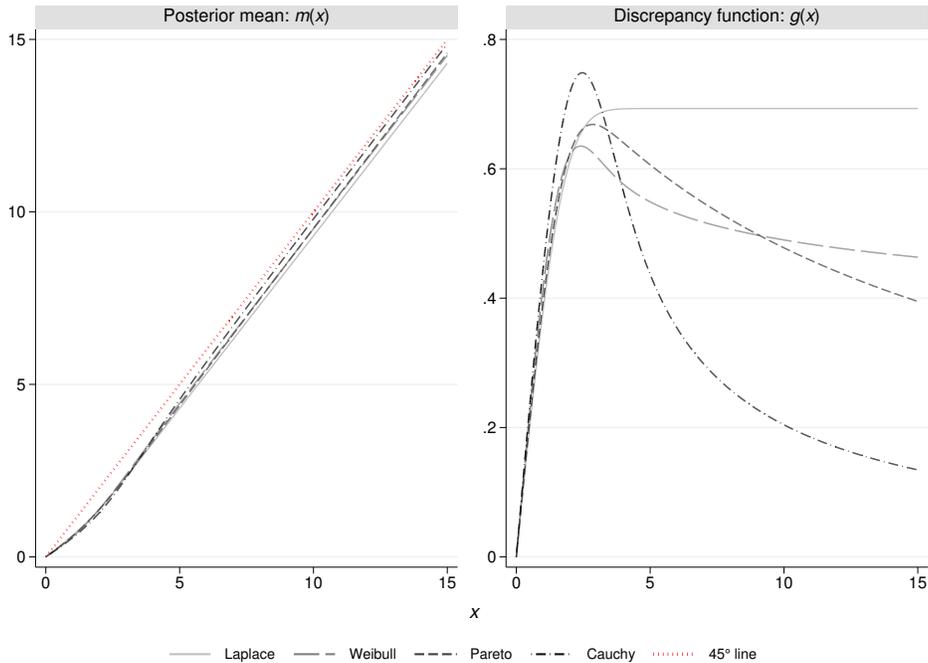under four neutral priors



Figure 1 plots the posterior mean and discrepancy functions of the four priors after imposing neutrality and minimax regret optimality. For large $\theta_n^*$, the Laplace prior has thin tails that decay exponentially, while the Cauchy

11

prior has much thicker tails that decay at the rate $(\theta_n^*)^{-2}$. The Weibull and Pareto priors are in-between the Laplace and Cauchy priors. This implies that, as $x$ increases, the Cauchy posterior mean function $m(x)$ curves back towards the $45°$ line much faster than the other three priors. Turning to the discrepancy functions in the right panel, we see that the Laplace prior is not robust since its discrepancy function converges to the prior parameter $c = \ln 2$. The Weibull and Pareto priors are robust since $g(x) \to 0$ as $x \to \infty$, but convergence is slow, especially for the Weibull prior. In contrast, the Cauchy discrepancy function converges to zero very quickly.

Figure 2: Posterior variance and shrinkage functions
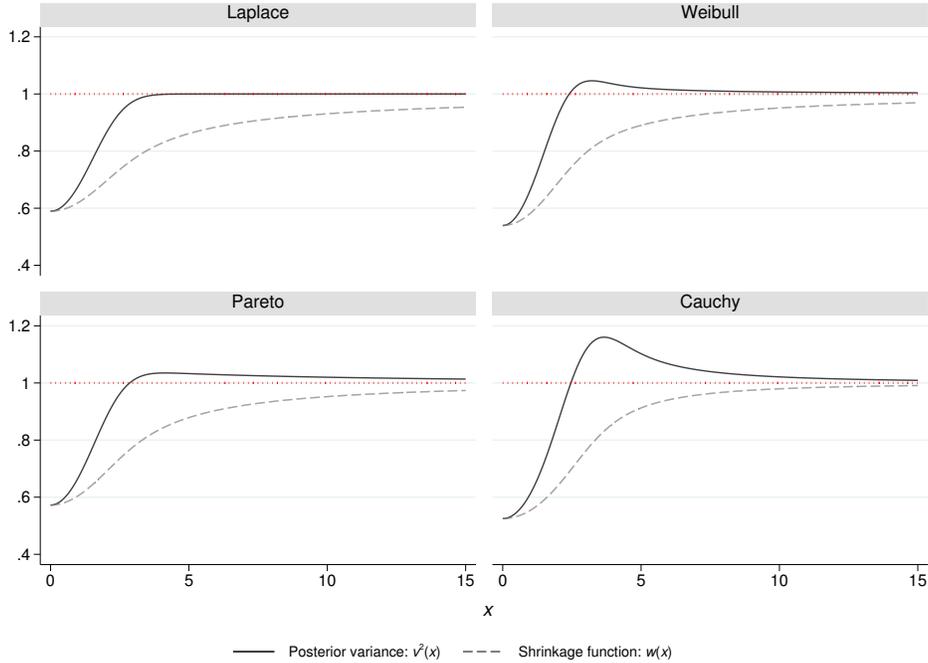under four neutral priors



Figure 2 plots the shrinkage and the posterior variance functions resulting from our four neutral priors. For small $x$, the Cauchy prior implies slightly more shrinkage than the other three priors but, for all of them, $w(x) \to v^2(0)$ as $x \to 0$ and $v^2(x) \to 1$ as $x \to \infty$. In agreement with Mitchell (1994), the Laplace posterior variance is non-decreasing and converges to one from below as $x \to \infty$. Since $v^2(x) = m'(x)$, the Laplace posterior mean is convex

12

for $x > 0$. The Weibull, Pareto, and Cauchy posterior variances reach a maximum greater than one before approaching one from above as $x \to \infty$. Moreover, for $x > 0$, their shrinkage functions are non-decreasing because $v^2(x) \geqslant w(x)$. In the next three sections, we show that the properties of the functions $m$, $\ell$, $g$, $v^2$, and $w$ play a crucial role in determining the finite-sample and asymptotic properties of our shrinkage estimator of $\theta$.

# 5    Finite-sample properties

In this section we study the sampling properties of the shrinkage estimator $\tilde{\theta}_n$ of $\theta$, given by $\tilde{\theta}_n = (\sigma/\sqrt{n})m_n^*$, as developed in Section 2. Extending the results of De Luca et al. (2022) on the bias and variance of $m_n^*$, we establish the finite-sample distribution of $t_n = \sqrt{n}(\tilde{\theta}_n - \theta)/\sigma = m_n^* - \theta_n^*$ under model (9).

PROPOSITION 1  *Let $x_n^* \sim \mathrm{N}(\theta_n^*, 1)$.  Then,*

$$t_n = \frac{\tilde{\theta}_n - \theta}{\sigma/\sqrt{n}} = w(\theta_n^* + z_n)z_n - (1 - w(\theta_n^* + z_n))\,\theta_n^*,$$

*where $w$ is the shrinkage function and $z_n \sim \mathrm{N}(0, 1)$.  Under conditions (C1)–(C3), the density of $t_n$ at a point $t$ is given by*

$$f(t; \theta_n^*) = \frac{\phi\left(\ell(t + \theta_n^*) - \theta_n^*\right)}{v^2(\ell(t + \theta_n^*))},$$

*where $v^2$ is the posterior variance function, $\ell$ is the inverse of the posterior mean function, and $\phi$ is the standard-normal density.*

Let us consider again the case of a $\mathrm{N}(0, \tau^2)$ prior for $\theta_n^*$, where $m_n^* = \xi\,x_n^*$, $\ell(t + \theta_n^*) = (t + \theta_n^*)/\xi$, and $v^2(\ell(t + \theta_n^*)) = \xi$. Proposition 1 then implies that

$$f(t; \theta_n^*) = \frac{1}{\sqrt{2\pi\xi}}\exp-\frac{1}{2}\left(\frac{t - (\xi - 1)\theta_n^*}{\xi}\right)^2,$$

that is, $t_n \sim \mathrm{N}((\xi - 1)\theta_n^*, \xi^2)$. As in Section 2, we have $\mathbb{E}(t_n) = \mathbb{E}(m_n^* - \theta_n^*) = (\xi - 1)\theta_n^*$, so the bias of our estimator of $\theta$ is $\mathbb{E}(\tilde{\theta}_n - \theta) = (\xi - 1)(\sigma\theta_n^*/\sqrt{n}) = (\xi - 1)\theta$. The result on the sampling variance of $t_n$ is also in agreement with the surprising findings of Efron (2015, Theorem 1) and De Luca et al. (2022,

13

Proposition 2), namely that, for any positive and bounded prior density, the posterior *variance* represents a first-order approximation to the sampling *standard deviation* (not the sampling variance) of $m_n^*$. For the normal prior, the posterior mean is linear and hence the approximation is exact.

More generally, Proposition 1 shows that, except in the case of a normal prior, the finite-sample distribution of $t_n$ is the same as that of a non-linear function of $z_n \sim \mathrm{N}(0, 1)$, and is therefore non-normal in general. This distribution depends on the sample size $n$ only through $\theta_n^*$, and the continuity of $\ell$ and $v^2$ ensures that the density $f$ is continuous in both its arguments.

At $\theta_n^* = 0$, the finite-sample distribution of $t_n$ is the same as the distribution of $w(z_n)z_n$, which does not depend on $n$ and is symmetric around the origin. All its odd moments thus vanish. Further, since $\mathbb{E}[(w(z_n))^{2p} z_n^{2p}] < \mathbb{E}(z_n^{2p})$, all its even moments are smaller than the corresponding moments of the standard-normal distribution. Setting $p = 1$ shows that, at $\theta_n^* = 0$, our shrinkage estimator is unbiased and more efficient than the usual estimator $x_n^*$.

Figure 3: Density of $t_n = \sqrt{n}(\tilde{\theta}_n - \theta)/\sigma$
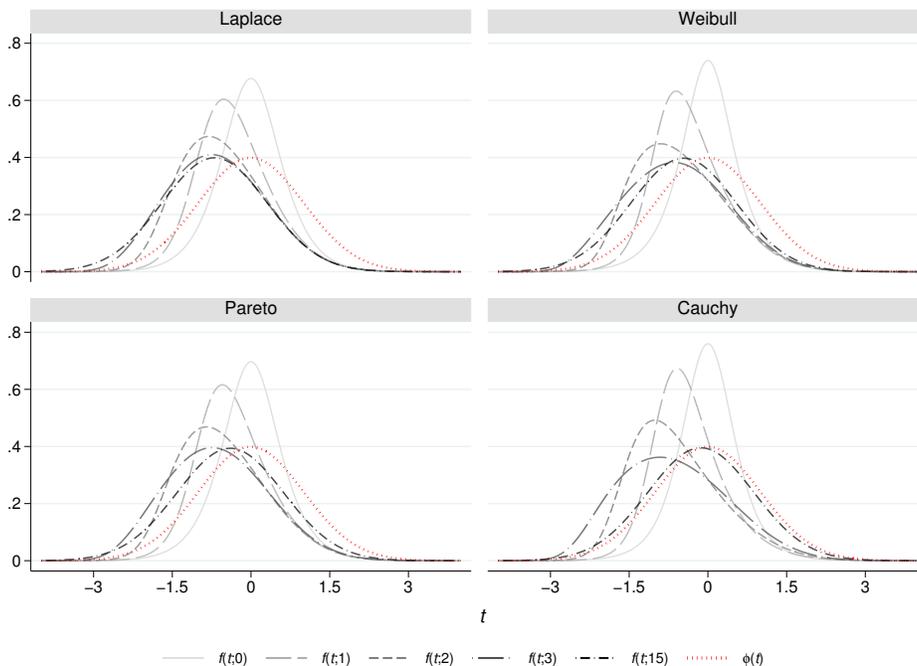under four neutral priors and five values of $\theta_n^*$



14

Figure 3 plots the finite-sample densities of $t_n$ under our four neutral priors for five values of $\theta_n^*$. These densities are all non-normal but, unlike many post-selection estimators (see, e.g., Figure 2 in Leeb and Pötscher 2005) and many thresholding estimators (see, e.g., Figures 1–3 in Pötscher and Leeb 2009), they are all unimodal and smooth. This is because Bayesian shrinkage estimators are smooth functions of the data. At $\theta_n^* = 0$, all densities are symmetric and centered at the origin. They are also more concentrated than the standard-normal density, and leptokurtic. For small non-zero values of $\theta_n^*$, there is a bias which has opposite sign as $\theta_n^*$, while the skewness has the same sign as $\theta_n^*$. For large enough values of $\theta_n^*$, all densities tend to have a normal shape but they are not necessarily centered at zero because of the bias. In Proposition 3 we shall prove that the finite-sample distribution of $t_n$ indeed converges to a normal distribution as $\theta_n^* \to \infty$.

Figure 4: Bias, variance, skewness and excess kurtosis of $m_n^*$ under four neutral priors
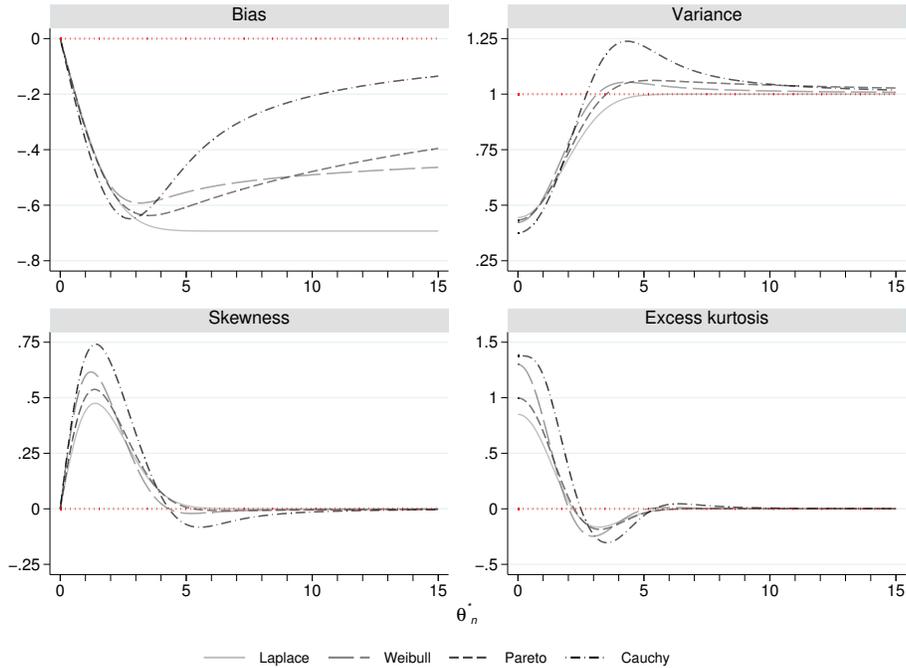


Figure 4 plots the bias, variance, skewness, and excess kurtosis of $m_n^*$ under our four neutral priors. As $\theta_n^*$ increases, the bias of $m_n^*$ converges to

zero much faster for the Cauchy and Pareto priors than for the Weibull prior, which is closely related to the behavior of the discrepancy function shown in the right panel of Figure 1. When $\theta_n^*$ is small, the Cauchy posterior mean has a relatively small variance. However, as $\theta_n^*$ increases, its variance first reaches a peak of 1.24 at $\theta_n^* = 4.33$ and then approaches 1 monotonically. Although there is no uniform improvement in the MSE, the Cauchy prior may be preferred to the other priors because of its greater robustness. However, robustness is not the only criterion, and the Pareto prior offers, in our view, the best compromise between robustness and minimax regret. The skewness reaches a positive maximum around $\theta_n^* = 1.4$ and the peak is largest for the Cauchy prior. As $\theta_n^*$ increases, the skewness converges to zero monotonically for the Laplace prior, while for the other priors it first attains a negative minimum around $\theta_n^* = 5$ before approaching zero from below. In contrast, the excess kurtosis attains a global maximum at $\theta_n^* = 0$, the peak being again largest for the Cauchy prior. As $\theta_n^*$ increases, the excess kurtosis first decreases, attains a negative minimum between $\theta_n^* = 3$ and $\theta_n^* = 4$ for all priors, and then approaches zero, although not necessarily monotonically and not necessarily from below.

# 6 Consistency

We next investigate the sampling properties of our shrinkage estimator of $\theta$ as $n \to \infty$, relaxing the assumption that $x_n$ is normally distributed. In the current section we establish (uniform) consistency, in the next section the asymptotic distribution. The uniform $\sqrt{n}$-consistency of $\tilde{\theta}_n$ will be proved under the mild assumption that $\sqrt{n}(x_n - \theta)/\sigma \xrightarrow{d} \mathrm{N}(0,1)$.

Formally, $\tilde{\theta}_n$ is (pointwise) $\sqrt{n}$-consistent for $\theta$ if, given any $\theta \in \mathbb{R}$ and $\varepsilon > 0$, there exist $M$ and $N$ (depending on both $\theta$ and $\varepsilon$) such that

$$n > N \implies \Pr\left[n^{1/2}|\tilde{\theta}_n - \theta| \geqslant M\right] \leqslant \varepsilon, \tag{19}$$

implying that $\lim_{n\to\infty} \Pr\left[n^\delta|\tilde{\theta}_n - \theta| \geqslant \varepsilon\right] = 0$ for any $\theta \in \mathbb{R}$, $\varepsilon > 0$, and $0 \leqslant \delta < 1/2$. If the same $N$ and $M$ work equally well for every $\theta$, the consistency is said to be uniform. That is, $\tilde{\theta}_n$ is *uniformly* $\sqrt{n}$-consistent for $\theta$ if, for any $\varepsilon > 0$, there exist $M$ and $N$ such that

$$n > N \implies \sup_{\theta\in\mathbb{R}} \Pr\left[n^{1/2}|\tilde{\theta}_n - \theta| \geqslant M\right] \leqslant \varepsilon. \tag{20}$$

16

PROPOSITION **2** *Suppose that $\sqrt{n}(x_n - \theta)/\sigma \xrightarrow{d} \mathrm{N}(0,1)$ and that the prior density of $\theta_n^*$ satisfies conditions (C1)–(C4). Then $\tilde{\theta}_n$ is uniformly $\sqrt{n}$-consistent for $\theta$, and hence*

$$\lim_{n\to\infty} \sup_{\theta\in\mathbb{R}} \Pr\left[n^\delta|\tilde{\theta}_n - \theta| \geqslant \varepsilon\right] = 0$$

*for every $0 \leqslant \delta < 1/2$ and every $\varepsilon > 0$.*

The proof of this result exploits the fact that, under (C1)–(C4), $g(x_n^*)$ is bounded, and so $|\tilde{\theta}_n - x_n|$ is bounded by a deterministic sequence proportional to $\sigma/\sqrt{n}$. Unlike Proposition 1, condition (C4) rules out a conjugate normal prior for $\theta_n^*$ because the corresponding discrepancy function would be unbounded (see Section 3). The uniform convergence rate of $\tilde{\theta}_n$ is similar to that of conservative post-selection estimators, in contrast to consistent post-selection and thresholding estimators which typically achieve a convergence rate that is slower and depends on the underlying tuning parameters (Pötscher and Leeb 2009).

# 7    Asymptotic distribution

When establishing the asymptotic distribution of $\tilde{\theta}_n$, we allow explicitly for the possibility that $\theta$ depends on $n$. To structure this dependence, we set $\theta_n = n^{-\delta}\theta_0$ with $\delta \geqslant 0$ and $\theta_0 \in \mathbb{R}$. The most important special cases are $\delta = 0$ and $\delta = 1/2$. When $\delta = 0$ we have $\theta_n = \theta_0$ (the 'fixed-parameter' case), and when $\delta = 1/2$ we have $\theta_n = n^{-1/2}\theta_0$ (the 'local-parameter' case). For $\theta_0 = 0$ we have $\theta_n^* = 0$, but for $\theta_0 \neq 0$ our parameter of interest becomes

$$\theta_n^* = \frac{n^{1/2}\theta_n}{\sigma} = \frac{n^{1/2-\delta}\theta_0}{\sigma} \to \begin{cases} \pm\infty & \text{if } 0 \leqslant \delta < 1/2, \\ \theta_0/\sigma & \text{if } \delta = 1/2, \\ 0 & \text{if } \delta > 1/2, \end{cases} \tag{21}$$

as $n \to \infty$. What matters for the asymptotic distribution of $\tilde{\theta}_n$ is whether $\theta_n^*$ diverges or converges to a finite constant, in which case its speed of convergence also matters.

PROPOSITION **3** *Suppose the conditions of Proposition 2 hold and let $t_n = \sqrt{n}(\tilde{\theta}_n - \theta_n)/\sigma$, where $\theta_n = n^{-\delta}\theta_0$ with $\delta \geqslant 0$ and $\theta_0 \in \mathbb{R}$. Let $w$ be the*

*shrinkage function and $z \sim \mathrm{N}(0,1)$. Then, $t_n \xrightarrow{d} t$, where the distribution of $t$ depends on $\delta$ and $\theta_0$ as follows:*
*(a) If $0 \leqslant \delta < 1/2$, then*

$$t = \begin{cases} w(z)z & \text{if } \theta_0 = 0, \\ z - \mathrm{sign}(\theta_0)\,\psi_0 & \text{if } \theta_0 \neq 0, \end{cases}$$

*where $\psi_0$ is defined in condition (C4). If the prior is robust then $\psi_0 = 0$.*
*(b) If $\delta = 1/2$, then $t = w(\zeta + z)z - [1 - w(\zeta + z)]\zeta$ with density*

$$f(t;\zeta) = \frac{\phi\left(\ell(t+\zeta) - \zeta\right)}{v^2(\ell(t+\zeta))},$$

*where $\zeta = \theta_0/\sigma$ and the functions $\phi$, $v^2$, and $\ell$ are defined as in Proposition 1.*
*(c) If $\delta > 1/2$, then $t = w(z)z$ with density $f(t) = \phi(\ell(t))/v^2(\ell(t))$.*

In contrast to Proposition 1, condition (C4) is needed in Proposition 3 to obtain the distribution of $t$ when $0 \leqslant \delta < 1/2$ and $\theta_0 \neq 0$. At $\theta_0 = 0$, the value of $\delta$ does not matter and the distribution of $t_n$ always converges to the distribution of $t = w(z)z$. When $\theta_0 \neq 0$ there are two relevant intervals for $\delta$, namely $0 \leqslant \delta < 1/2$ and $\delta \geqslant 1/2$, though in practice the most important values are $\delta = 0$ and $\delta = 1/2$, which we have called the fixed-parameter and the local-parameter case, respectively.

Part (a) of Proposition 3 shows that, when $\delta = 0$, the asymptotic distribution of $\tilde{\theta}_n$ is normal (standard-normal if the prior is robust) for all $\theta_0 \neq 0$, but not for $\theta_0 = 0$. Thus, for $\theta_0$ not equal but close to 0, the asymptotic approximation could misrepresent important features of the finite-sample distribution of $\tilde{\theta}_n$. Moreover, like for the classical Hodges estimator (see, e.g., van der Vaart 1998, pp. 109–110), the fixed-parameter case gives the misleading impression that $\tilde{\theta}_n$ is superefficient at the origin, as it is asymptotically equivalent to $x_n$ when $\theta_0 \neq 0$ but asymptotically more efficient when $\theta_0 = 0$. This overoptimistic conclusion reflects the lack of uniformity in the convergence of the finite-sample distribution to the asymptotic distribution.[2]

Part (b) of Proposition 3 shows that, when $\delta = 1/2$, the asymptotic density of $t_n = \sqrt{n}(\tilde{\theta}_n - \theta_n)/\sigma$ is the same as the finite-sample density in Proposition 1, which justifies using the local-parameter framework to study

---

[2] Related results are discussed in detail in Leeb and Pötscher (2005, 2006a,b, 2008), Pötscher (2006), and Pötscher and Leeb (2009).

18

the sampling properties of $\tilde{\theta}_n$. In a neighborhood of $\zeta = 0$, the asymptotic distribution of $t_n$ is characterized by sizeable departures from normality, both in terms of skewness and excess kurtosis. This is also the region of the parameter space where $\tilde{\theta}_n$ has smaller asymptotic risk than the usual estimator $x_n$, although the risk improvements are not uniform. However, for sufficiently large values of $\zeta$, the asymptotic distribution of $t_n$ is normal.

Although one can construct consistent estimators of the finite-sample distribution function of $\sqrt{n}(\tilde{\theta}_n - \theta)$ for any given value of $\theta$, these estimators are not uniformly close to the true distribution function. In fact, Leeb and Pötscher (2006b, Lemmas 3.1 and 3.5) show that no uniformly consistent estimator of the distribution of shrinkage-type estimators can exist. This 'impossibility result' is a general feature of a large class of estimators of $\theta$ that includes post-selection, thresholding, and model-averaging estimators. In our case, it means that one should be careful with inference based on the estimated distribution of $\tilde{\theta}_n$, as there is no guarantee that it will be close to the true one.

# 8    Conclusions

We have investigated the sampling properties of a Bayesian estimator of $\theta$ in the normal local model $x_n \sim \mathrm{N}\left(\theta, \sigma^2/n\right)$ with $\sigma^2$ known. Unlike the standard Bayesian approach, which places a prior on $\theta$ that does not depend on the sample size $n$, we place a prior on the 'population $t$-ratio' $\theta_n^* = \sqrt{n}\theta/\sigma$. This non-standard approach is motivated by the fact that model selection and model averaging estimators typically depend on diagnostics such as $t$-ratios. Moreover, since MSE comparisons between alternative estimators of $\theta$ depend crucially on $\theta_n^*$, it leads to a transparent notion of prior ignorance about $\theta_n^*$ that is formalized in the neutrality condition (C5).

Results on the large sample properties of standard Bayesian estimators of $\theta$ do not directly extend to our approach. In addition to the mild regularity conditions (C1)–(C3), we require the prior on $\theta_n^*$ to have a bounded influence on the posterior mean function. The additional condition (C4) rules out a conjugate normal prior for $\theta_n^*$ and is related to the notion of Bayesian robustness.

Our main results are as follows. First, we extend earlier results on the first two sampling moments of our estimator $\tilde{\theta}_n = n^{-1/2}\sigma\,\mathbb{E}(\theta_n^*|x_n)$ and derive its finite-sample distribution under the assumption that $\sigma$ is known. We show

that this distribution displays sizeable departures from normality in terms of skewness and excess kurtosis. Second, the choice of prior affects both the speed at which the estimation bias tends to zero and the speed at which the sampling variance tends to one. This bias-precision trade-off depends on the thickness of the prior tails. Third, $\tilde{\theta}_n$ is a uniformly $\sqrt{n}$-consistent estimator of $\theta$ under very mild conditions on $x_n$. Fourth, we derive the asymptotic distribution of $\tilde{\theta}_n$ under a moving-parameter setup that encompasses both the fixed-parameter and the local-parameter settings.

These results have direct implications for the WALS estimator of Magnus et al. (2010). They may also be helpful for other model selection or model averaging procedures.

# Appendix: Proofs

*Proof of Proposition 1:* Write $x_n^* = \theta_n^* + z_n$. Then, since $m_n^* = w(x_n^*)x_n^*$, the first result follows from

$$\sqrt{n}(\tilde{\theta}_n - \theta)/\sigma = m_n^* - \theta_n^* = w(\theta_n^* + z_n)z_n - [1 - w(\theta_n^* + z_n)]\theta_n^*.$$

Since $t_n = m_n^* - \theta_n^* = m(x_n^*) - \theta_n^*$ is a continuous one-to-one transformation of $x_n^*$ with inverse function $x_n^* = \ell(t_n + \theta_n^*)$, its density at a point $t$ is

$$f(t; \theta_n^*) = |\ell'(t + \theta_n^*)| \, \phi(\ell(t + \theta_n^*) - \theta_n^*),$$

where $\ell'(t + \theta_n^*) = [m'(x_n^*)]^{-1}$. The second result then follows from the Brown–Tweedie formula, using the fact that $m'(x) = v^2(x) > 0$.

*Proof of Proposition 2:* Since the function $g$ is bounded, there exists a finite constant $G > 0$ such that $0 < g(x) \leqslant G$ for all $x$. The triangle inequality gives

$$|\tilde{\theta}_n - \theta| \leqslant |x_n - \theta| + |\tilde{\theta}_n - x_n|,$$

from which we obtain

$$\begin{aligned}
\Pr\left(n^{1/2}|\tilde{\theta}_n - \theta| \geqslant M\right) &\leqslant \Pr\left(n^{1/2}|x_n - \theta| + n^{1/2}|\tilde{\theta}_n - x_n| \geqslant M\right) \\
&\leqslant \Pr\left(n^{1/2}|x_n - \theta| \geqslant M/2\right) + \Pr\left(n^{1/2}|\tilde{\theta}_n - x_n| \geqslant M/2\right) \\
&\leqslant \frac{4\sigma^2}{M^2} + \Pr\left(n^{1/2}|\tilde{\theta}_n - x_n| \geqslant \sigma G\right),
\end{aligned}$$

where the first term on the last line uses Chebyshev's inequality, and the second term follows by choosing $M > 2\sigma G$. Since

$$n^{1/2}|\tilde{\theta}_n - x_n| = \sigma|m(x_n^*) - x_n^*| = \sigma|g(x_n^*)| \leqslant \sigma G$$

almost surely, the result follows by setting $M = n^{1/2-\delta}\varepsilon$ and choosing $n$ sufficiently large, so that $M > 2\sigma G$.

*Proof of Proposition 3:* Cases (b) and (c), where $\delta \geqslant 1/2$, follow directly from Proposition 1 and the assumption that

$$\sqrt{n}(x_n - \theta)/\sigma = x_n^* - \theta_n^* = z_n \xrightarrow{d} z \sim \mathrm{N}(0, 1).$$

Note that $\zeta = 0$ in case (c).

Let us prove case (a), where $0 \leqslant \delta < 1/2$. We first rewrite $t_n = \sqrt{n}(\tilde{\theta}_n - \theta_n)/\sigma$ as

$$t_n = m_n^* - \theta_n^* = z_n - g(x_n^*),$$

with $m_n^* = w(x_n^*)x_n^*$, $\theta_n^* = n^{1/2-\delta}\theta_0/\sigma$, and $g(x_n^*) = x_n^* - m_n^*$. Since $z_n \xrightarrow{d} z$, we only need to study the asymptotic behavior of $g(x_n^*)$. If $\theta_0 > 0$, then $\theta_n^* \to \infty$ as $n \to \infty$. Hence,

$$\Pr(x_n^* \geqslant M) = \Pr(z_n \geqslant M - \theta_n^*) \to 1$$

for every $M > 0$, which we write as $\operatorname{plim} x_n^* = \infty$. By the (generalized) continuous mapping theorem (see Theorem 18.11 and Example 18.4 in van der Vaart 1998), we then find

$$\operatorname{plim} g(x_n^*) = g(\operatorname{plim} x_n^*) = g(\infty) = \psi_0.$$

Similarly, if $\theta_0 < 0$, then $\theta_n^* \to -\infty$ as $n \to \infty$ so that

$$\Pr(x_n^* \leqslant -M) = \Pr(z_n \leqslant -M - \theta_n^*) \to 1$$

for every $M > 0$, which we write as $\operatorname{plim} x_n^* = -\infty$. Since $m$ is an odd function, we then find $\operatorname{plim} g(x_n^*) = -\psi_0$. Finally, if $\theta_0 = 0$, then $x_n^* = z_n$ so that $m_n^* = w(z_n)z_n$, $g(x_n^*) = (1 - w(z_n))z_n$, and $t_n = w(z_n)z_n \xrightarrow{d} w(z)z$.

21

## Acknowledgements

# References

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd edition. Springer, New York.

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

Carlin, B. P., and Louis, T. A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd edition. Chapman and Hall–CRC, Boca Raton.

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

Danilov, D., and Magnus, J. R. (2004). On the harm that ignoring pretesting can cause. *Journal of Econometrics*, 122, 27–46.

DasGupta, A., and Johnstone, I. (2014). Asymptotic risk and Bayes risk of thresholding and superefficient estimates and optimal thresholding. In: Lahiri, S., Schick, A., SenGupta, A., Sriram, T. (eds), *Contemporary Developments in Statistical Theory*. Springer Proceedings in Mathematics & Statistics, 68. Springer, Cham, Chapter 4.

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

De Luca, G., Magnus, J. R., and Peracchi, F. (2022). Sampling properties of the Bayesian posterior mean with an application to WALS estimation. *Journal of Econometrics*, 230, 299–317.

De Luca, G., Magnus, J. R., and Peracchi, F. (2023). Weighted-average least squares (WALS): Confidence and prediction intervals. *Computational Economics*, 61, 1637–1664.

Efron, B. (2012). Bayesian inference and the parametric bootstrap. *The Annals of Applied Statistics*, 6, 1971–1997.

Efron, B. (2014). Estimation and accuracy after model selection. *Journal of the American Statistical Association*, 109, 991–1007.

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

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

Huntsberger, D. V. (1955). A generalization of a preliminary testing procedure for pooling data. *Annals of Mathematical Statistics*, 26, 734–743.

Joanes, D. N., and Peers, H. W. (1974). On the sampling properties of Bayesian point estimators, *Journal of the American Statistical Association*, 69, 560–564.

Johnstone, I. M. (2019). *Gaussian Estimation: Sequence and Wavelet Models*. Draft version, available from statweb.stanford.edu.

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

Judge, G. G., and Bock, M. E. (1983). Biased estimation. In: Griliches, Z., and Intriligator, M. D. (eds), *Handbook of Econometrics*, Vol. I. North-Holland, Amsterdam, Chapter 10.

Kabaila, P. (1995). The effect of model selection on confidence regions and prediction regions. *Econometric Theory*, 11, 537–549.

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

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

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

Leeb, H., and Pötscher, B. M. (2003). The finite-sample distribution of post-model-selection estimators and uniform versus non-uniform approximations. *Econometric Theory*, 19, 100–142.

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

Leeb, H., and Pötscher, B. M. (2006a). Can one estimate the conditional distribution of post-model-selection estimators? *The Annals of Statistics*, 34, 2554–2591.

Leeb, H., and Pötscher, B. M. (2006b). Performance limits for estimators of the risk or distribution of shrinkage-type estimators, and some general lower risk-bound results. *Econometric Theory*, 22, 69–97.

Leeb, H., and Pötscher, B. M. (2008). Sparse estimators and the oracle property, or the return of Hodges' estimator. *Journal of Econometrics*, 142, 201–211.

Magnus, J. R. (1999). The traditional pretest estimator. *Theory of Probability and Its Applications*, 44, 293–308.

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

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.

Mitchell, F. S. (1994). A note on posterior moments for a normal mean with double-exponential prior. *Journal of the Royal Statistical Society, Series B*, 56, 605–610.

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

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

Pötscher, B. M. (2006). The distribution of model averaging estimators and an impossibility result regarding its estimation. *IMS Lecture Notes — Monograph Series Time Series and Related Topics*, 52, 113–129.

Pötscher, B. M., and Leeb, H. (2009). On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding. *Journal of Multivariate Analysis*, 100, 2065–2082.

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

Strawderman, W. E., and Cohen, A. (1971). Admissibility of estimators of the mean vector of a multivariate normal distribution with quadratic loss. *Annals of Mathematical Statistics*, 42: 270–296.

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