Econometrica, Vol. 90, No. 6 (November, 2022), 2567–2602

ROBUST EMPIRICAL BAYES CONFIDENCE INTERVALS

TIMOTHY B. ARMSTRONG
Department of Economics, University of Southern California

MICHAL KOLESÁR
Department of Economics, Princeton University

MIKKEL PLAGBORG-MØLLER
Department of Economics, Princeton University

We construct robust empirical Bayes confidence intervals (EBCIs) in a normal means problem. The intervals are centered at the usual linear empirical Bayes estimator, but use a critical value accounting for shrinkage. Parametric EBCIs that assume a normal distribution for the means (Morris (1983b)) may substantially undercover when this assumption is violated. In contrast, our EBCIs control coverage regardless of the means distribution, while remaining close in length to the parametric EBCIs when the means are indeed Gaussian. If the means are treated as fixed, our EBCIs have an average coverage guarantee: the coverage probability is at least $1-\alpha$ on average across the n EBCIs for each of the means. Our empirical application considers the effects of U.S. neighborhoods on intergenerational mobility.

KEYWORDS: Average coverage, empirical Bayes, confidence interval, shrinkage.

1. INTRODUCTION

EMPIRICAL RESEARCHERS IN ECONOMICS are often interested in estimating effects for many individuals or units, such as estimating teacher quality for teachers in a given geographic area. In such problems, it is common to shrink unbiased but noisy preliminary estimates of these effects toward baseline values, say the average effect for teachers with the same experience. In addition to estimating teacher quality (Kane and Staiger (2008), Jacob and Lefgren (2008), Chetty, Friedman, and Rockoff (2014)), shrinkage techniques are used in a wide range of applications including estimating school quality (Angrist, Hull, Pathak, and Walters (2017)), hospital quality (Hull (2020)), the effects of neighborhoods on intergenerational mobility (Chetty and Hendren (2018)), and patient risk scores across regional health care markets (Finkelstein, Gentzkow, Hull, and Williams (2017)).

The shrinkage estimators used in these applications can be motivated by an empirical Bayes (EB) approach. One imposes a working assumption that the individual effects are drawn from a normal distribution (or, more generally, a known family of distributions). The mean squared error (MSE) optimal point estimator then has the form of a Bayesian

Timothy B. Armstrong: timothy.armstrong@usc.edu
Michal Kolesár: mkolesar@princeton.edu
Mikkel Plagborg-Møller: mikkelpm@princeton.edu

This paper is dedicated to the memory of Gary Chamberlain, who had a profound influence on our thinking about decision problems in econometrics, and empirical Bayes methods in particular. Luther Yap provided excellent research assistance. We received helpful comments from four anonymous referees, Otávio Bartalotti, Toru Kitagawa, Laura Liu, Ulrich Müller, Stefan Wager, Mark Watson, Martin Weidner, and numerous seminar participants. We are especially indebted to Bruce Hansen and Roger Koenker for inspiring our simulation study. Armstrong acknowledges support by the National Science Foundation Grant SES-2049765. Kolesár acknowledges support by the Sloan Research Fellowship and by the National Science Foundation Grant SES-22049356. Plagborg-Møller acknowledges support by the National Science Foundation Grant SES-1851665.

posterior mean, treating this distribution as a prior distribution. Rather than specifying the unknown parameters in the prior distribution ex ante, the EB estimator replaces them with consistent estimates, just as in random effects models. This approach is attractive because one does not need to assume that the effects are in fact normally distributed, or even take a "Bayesian" or "random effects" view: the EB estimators have lower MSE (averaged across units) than the unshrunk unbiased estimators, even when the individual effects are treated as nonrandom (James and Stein (1961)).

In spite of the popularity of EB methods, it is currently not known how to provide uncertainty assessments to accompany the point estimates without imposing strong parametric assumptions on the effect distribution. Indeed, Hansen (2016, p. 116) described inference in shrinkage settings as an open problem in econometrics. The natural EB version of a confidence interval (CI) takes the form of a Bayesian credible interval, again using the postulated effect distribution as a prior (Morris (1983b)). If the distribution is correctly specified, this *parametric empirical Bayes confidence interval (EBCI)* will cover 95%, say, of the true effect parameters, under repeated sampling of the observed data *and* of the effect parameters. We refer to this notion of coverage as "EB coverage," following the terminology in Morris (1983b). Unfortunately, we show that, in the context of a normal means model, the parametric EBCI with nominal level 95% can have actual EB coverage as low as 74% for certain non-normal effect distributions. The potential undercoverage is increasing in the degree of shrinkage, and we derive a simple "rule of thumb" for gauging the potential coverage distortion.

To allow easy uncertainty assessment in EB applications that is reliable irrespective of the degree of shrinkage, we construct novel *robust EBCIs* that take a simple form and control EB coverage *regardless* of the true effect distribution. Our baseline model is an (approximate) normal means problem $Y_i \sim N(\theta_i, \sigma_i^2)$, $i = 1, \dots, n$. In applications, Y_i represents a preliminary estimate of the effect θ_i for unit i. Like the parametric EBCI that assumes a normal distribution for θ_i, the robust EBCI we propose is centered at the normality-based EB point estimate $\hat{\theta}_i$ that shrinks Y_i toward some baseline value, but it uses a larger critical value to account for bias due to shrinkage.[1] EB coverage is controlled in the class of all distributions for θ_i that satisfy certain moment bounds, which we estimate consistently from the data (similarly to the parametric EBCI, which uses the second moment). We show that the baseline implementation of our robust EBCI is "adaptive": its length is close to that of the parametric EBCI when the θ_i's are in fact normally distributed. Thus, little efficiency is lost from using the robust EBCI in place of the non-robust parametric one.[2]

In addition to controlling EB coverage, the robust EBCIs with level $1-\alpha$ have a frequentist *average coverage* property: If the means $\theta_1, \dots, \theta_n$ are treated as *fixed*, the coverage probability, averaged across the n parameters θ_i, is at least $1-\alpha$. In fact, under mild conditions, at least a fraction $1-\alpha$ of the n EBCIs will contain their respective parameters (with high probability as $n \to \infty$). This weakening of the usual requirement of coverage for *each* parameter θ_i allows our robust EBCI to be shorter than the usual CI centered at the unshrunk estimate Y_i, and often substantially so.[3] Intuitively, the average coverage criterion only requires us to guard against the *average* coverage distortion induced by the

[1]Our methods are implemented in the Stata package `ebreg`, R package `ebci`, and Matlab package `ebci_matlab`, which are available at SSC, CRAN, and GitHub, respectively.

[2]If the θ_i's are not normally distributed, our robust EBCIs are valid but may leave room for greater efficiency improvement, as we discuss in Section 5.3.

[3]Relaxing the usual notion of coverage in some way is necessary to obtain intervals that reflect the efficiency improvement of the empirical Bayes approach. In particular, the results in Pratt (1961) imply that for CIs

biases of the individual shrinkage estimators $\hat{\theta}_i$, and the data are quite informative about whether *most* of these biases are large, even though individual biases are difficult to estimate. To complement the frequentist properties, our EBCIs can be viewed as Bayesian credible sets that are robust to the prior on θ_i, in terms of ex ante coverage.

The average coverage criterion has the same motivation as the usual frequentist justification of the EB *point estimator*: the EB point estimator achieves lower MSE on average across units at the expense of potentially worse performance for some individual units (see, e.g., Efron (2010, Chapter 1.3)). Thus, researchers who use EB estimators instead of the unshrunk Y_i's prioritize favorable group performance over protecting individual performance. Our average coverage intervals make an analogous tradeoff: they guarantee coverage and achieve short length on average across units at the expense of giving up on a coverage guarantee for every individual unit. We examine this tradeoff in more detail in Section 5.

We caution, however, that the average coverage criterion is typically inappropriate in applications where shrinkage point estimation is unattractive. This includes settings where one is interested in the magnitude or the identity of the largest θ_i, or the true effect for the largest observed Y_i (as in, e.g., Hung and Fithian (2019), or Andrews, Kitagawa, and McCloskey (2021)).[4] It also includes settings where a particular effect, say θ_1, is of primary interest, or, more generally, settings where the effects are not exchangeable, and their ordering is relevant (Greenshtein and Ritov (2019)). Our methods are also not applicable if one is interested in functionals of the random effects *distribution* (as in Bonhomme and Weidner (2021), or Ignatiadis and Wager (2021)), rather than in the effects themselves. Finally, the justification for our methods is asymptotic in the number of parameters n. In our Monte Carlo simulations, we find that our EBCIs have close to nominal coverage over a range of data generating processes (DGPs) once n is greater than 100.

We illustrate our results by computing EBCIs for the causal effects of growing up in different U.S. neighborhoods (specifically commuting zones) on intergenerational mobility. We follow Chetty and Hendren (2018), who applied EB shrinkage to initial fixed effects estimates. Depending on the specification, we find that the robust EBCIs are on average 12–25% as long as the unshrunk CIs.

Our underlying ideas extend to other linear and nonlinear shrinkage settings with possibly non-Gaussian data. For example, our techniques allow for the construction of robust EBCIs that contain (nonlinear) soft thresholding estimators, as well as average coverage confidence bands for nonparametric regression functions.

The average coverage criterion was originally introduced in the literature on nonparametric regression (Wahba (1983), Nychka (1988), Wasserman (2006, Chapter 5.8)). Cai, Low, and Ma (2014) constructed adaptive average coverage confidence bands. These procedures are challenging to implement in our EB setting, and lack a clear finite-sample justification, unlike our procedure. Liu, Moon, and Schorfheide (2022) constructed forecast intervals in a dynamic panel data model that guarantee average coverage in a Bayesian sense (for a fixed prior). We discuss alternative approaches to inference in EB settings in Section 5.

with coverage 95%, one cannot achieve expected length improvements greater than 15% relative to the usual unshrunk CIs, even if one happens to optimize length for the true parameter vector $(\theta_1, \dots, \theta_n)$. See, for example, Corollary 3.3 in Armstrong and Kolesár (2018) and the discussion following it.

[4] As we show in Section 6.2, our methods do extend to settings where we keep a subset of units i that exceed a given cutoff. However, we do not allow this cutoff to diverge with the sample size, such as when one focuses on the unit i with the single largest observed Y_i.

The rest of this paper is organized as follows. Section 2 illustrates our methods in the context of a simple homoscedastic Gaussian model. Section 3 presents our recommended baseline procedure and discusses practical implementation issues. Section 4 presents our main results on the coverage and efficiency of the robust EBCI, and on the coverage distortions of the parametric EBCI; we also verify the finite-sample coverage accuracy of the robust EBCI through extensive simulations. Section 5 compares our EBCI with other inference approaches. Section 6 discusses extensions of the basic framework. Section 7 contains an empirical application to inference on neighborhood effects. Appendices A to C give details on finite-sample corrections, computational details, and formal asymptotic coverage results. The Supplemental Material (Armstrong, Kolesár, and Plagborg-Møller (2022)) contains proofs as well as further technical results. Applied readers are encouraged to focus on Sections 2, 3, and 7.

2. SIMPLE EXAMPLE

This section illustrates the construction of the robust EBCIs that we propose in a simplified setting with no covariates and with known, homoscedastic errors. Section 3 relaxes these restrictions, and discusses other empirically relevant extensions of the basic framework, as well as implementation issues.

We observe n estimates Y_i of elements of the parameter vector $\theta = (\theta_1, \dots, \theta_n)'$. Each estimate is normally distributed with common, known variance σ^2,

$$Y_i | \theta \sim N(\theta_i, \sigma^2), \quad i = 1, \dots, n. \tag{1}$$

In many applications, the Y_i's arise as preliminary least squares estimates of the parameters θ_i. For instance, they may correspond to fixed effect estimates of teacher or school value added, neighborhood effects, or firm and worker effects. In such cases, Y_i will only be *approximately* normal in large samples by the central limit theorem (CLT); we take this explicitly into account in the theory in Appendix C.

A popular approach to estimation that substantially improves upon the raw estimator $\hat{\theta}_i = Y_i$ under the compound MSE $\sum_{i=1}^n E[(\hat{\theta}_i - \theta_i)^2]$ is based on empirical Bayes (EB) shrinkage. In particular, suppose that the θ_i's are themselves normally distributed,

$$\theta_i \sim N(0, \mu_2). \tag{2}$$

Our discussion below applies if Eq. (2) is viewed as a subjective Bayesian prior distribution for a single parameter θ_i, but for concreteness we will think of Eq. (2) as a "random effects" sampling distribution for the n mean parameters $\theta_1, \dots, \theta_n$. Under Eq. (2), it is optimal to estimate θ_i using the posterior mean $\hat{\theta}_i = w_{\text{EB}} Y_i$, where $w_{\text{EB}} = 1 - \sigma^2/(\sigma^2 + \mu_2)$. To avoid having to specify the variance μ_2, the EB approach treats it as an unknown parameter, and replaces the marginal precision of Y_i, $1/(\sigma^2 + \mu_2)$, with a method of moments estimate $n/\sum_{i=1}^n Y_i^2$, or the degrees-of-freedom adjusted estimate $(n-2)/\sum_{i=1}^n Y_i^2$. The latter leads to the classic estimator of James and Stein (1961), $\hat{w}_{\text{EB}} = 1 - \sigma^2(n-2)/\sum_{i=1}^n Y_i^2$.

One can also use Eq. (2) to construct CIs for the θ_i's. In particular, since the marginal distribution of $w_{\text{EB}} Y_i - \theta_i$ is normal with mean zero and variance $(1 - w_{\text{EB}})^2 \mu_2 + w_{\text{EB}}^2 \sigma^2 = w_{\text{EB}} \sigma^2$, this leads to the $1 - \alpha$ CI

$$w_{\text{EB}} Y_i \pm z_{1-\alpha/2} w_{\text{EB}}^{1/2} \sigma, \tag{3}$$

where z_α is the α quantile of the standard normal distribution. Since the form of the interval is motivated by the parametric assumption (2), we refer to it as a parametric EBCI. With μ_2 unknown, one can replace w_{EB} by $\hat{w}_{\text{EB}}$.[5] This is asymptotically equivalent to (3) as $n \to \infty$.

The coverage of the parametric EBCI in (3) is $1-\alpha$ under repeated sampling of (Y_i, θ_i) according to Eqs. (1) and (2). To distinguish this notion of coverage from the case with fixed θ, we refer to coverage under repeated sampling of (Y_i, θ_i) as "empirical Bayes coverage." This follows the definition of an empirical Bayes confidence interval (EBCI) in Morris (1983b, Eq. 3.6) and Carlin and Louis (2000, Chapter 3.5). Unfortunately, this coverage property relies heavily on the parametric assumption (2). We show in Section 4.3 that the actual EB coverage of the nominal $1-\alpha$ parametric EBCI can be as low as $1 - 1/\max\{z_{1-\alpha/2}, 1\}$ for certain non-normal distributions of θ_i with variance μ_2; for 95% EBCIs, this evaluates to 74%. This contrasts with existing results on estimation: although the EB estimator is motivated by the parametric assumption (2), it performs well even if this assumption is dropped, with low MSE even if we treat θ as fixed.

This paper constructs an EBCI with a similar robustness property: the interval will be close in length to the parametric EBCI when Eq. (2) holds, but its EB coverage is at least $1-\alpha$ without any parametric assumptions on the distribution of θ_i. To describe the construction, suppose that all that is known is that θ_i is sampled from a distribution with second moment given by μ_2 (in practice, we can replace μ_2 by the consistent estimate $n^{-1}\sum_{i=1}^n Y_i^2 - \sigma^2$). Conditional on θ_i, the estimator $w_{\text{EB}} Y_i$ has bias $(w_{\text{EB}} - 1)\theta_i$ and variance $w_{\text{EB}}^2 \sigma^2$, so that the t-statistic $(w_{\text{EB}} Y_i - \theta_i)/w_{\text{EB}}\sigma$ is normally distributed with mean $b_i = (1 - 1/w_{\text{EB}})\theta_i/\sigma$ and variance 1. Therefore, if we use a critical value χ, the non-coverage of the CI $w_{\text{EB}} Y_i \pm \chi w_{\text{EB}}\sigma$, conditional on θ_i, will be given by the probability

$$r(b_i, \chi) = P\big(|Z - b_i| \geq \chi | \theta_i\big) = \Phi(-\chi - b_i) + \Phi(-\chi + b_i), \tag{4}$$

where Z denotes a standard normal random variable, and Φ denotes its cdf. Thus, by iterated expectations, under repeated sampling of θ_i, the non-coverage is bounded by

$$\rho\big(\sigma^2/\mu_2, \chi\big) = \sup_F E_F\big[r(b, \chi)\big] \quad \text{s.t. } E_F\big[b^2\big] = \frac{(1 - 1/w_{\text{EB}})^2}{\sigma^2}\mu_2 = \frac{\sigma^2}{\mu_2}, \tag{5}$$

where E_F denotes expectation under $b \sim F$. Although this is an infinite-dimensional optimization problem over the space of distributions, it turns out that it admits a simple closed-form solution, which we give in Proposition B.1 in Appendix B. Moreover, because the optimization is a linear program, it can be solved even in the more general settings of applied relevance that we consider in Section 3.

Set $\chi = \text{cva}_\alpha(\sigma^2/\mu_2)$, where $\text{cva}_\alpha(t) = \rho^{-1}(t, \alpha)$, and the inverse is with respect to the second argument. Then the resulting interval

$$w_{\text{EB}} Y_i \pm \text{cva}_\alpha\big(\sigma^2/\mu_2\big) w_{\text{EB}}\sigma \tag{6}$$

will maintain coverage $1-\alpha$ among all distributions of θ_i with $E[\theta_i^2] = \mu_2$ (recall that we estimate μ_2 consistently from the data). For this reason, we refer to it as a robust EBCI. Figure 1 in Section 3.1 gives a plot of the critical values for $\alpha = 0.05$. We show in Section 4.2 below that by also imposing a constraint on the fourth moment of θ_i, in

[5] Alternatively, to account for estimation error in $\hat{w}_{\text{EB}}$, Morris (1983b) suggested adjusting the variance estimate $\hat{w}_{\text{EB}}\sigma^2$ to $\hat{w}_{\text{EB}}\sigma^2 + 2Y_i^2(1 - \hat{w}_{\text{EB}})^2/(n-2)$. The adjustment does not matter asymptotically.

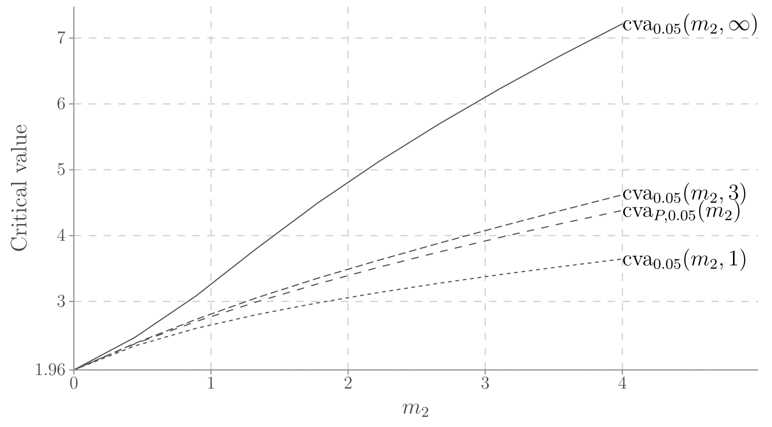

FIGURE 1.—Function $\text{cva}_\alpha(m_2, \kappa)$ for $\alpha = 0.05$ and selected values of κ. The function $\text{cva}_\alpha(m_2)$, defined in Section 2, that only imposes a constraint on the second moment, corresponds to $\text{cva}_\alpha(m_2, \infty)$. The function $\text{cva}_{P,\alpha}(m_2) = z_{1-\alpha/2}\sqrt{1+m_2}$ corresponds to the critical value under the assumption that θ_i is normally distributed.

addition to the second moment constraint, one can construct a robust EBCI that "adapts" to the Gaussian case in the sense that its length will be close to that of the parametric EBCI in Eq. (3) if these moment constraints are compatible with a normal distribution.

Instead of considering EB coverage, one may alternatively wish to assess uncertainty associated with the estimates $\hat{\theta}_i = w_{\text{EB}} Y_i$ when θ is treated as fixed. In this case, the EBCI in Eq. (6) has an average coverage guarantee that

$$\frac{1}{n}\sum_{i=1}^{n} P\big(\theta_i \in \big[w_{\text{EB}} Y_i \pm \text{cva}_\alpha\big(\sigma^2/\mu_2\big) w_{\text{EB}}\sigma\big] \,|\, \theta\big) \geq 1-\alpha, \tag{7}$$

provided that the moment constraint can be interpreted as a constraint on the empirical second moment on the θ_i's, $n^{-1}\sum_{i=1}^{n}\theta_i^2 = \mu_2$. In other words, if we condition on θ, then the coverage is at least $1-\alpha$ on average across the n EBCIs for $\theta_1, \dots, \theta_n$. To see this, note that the average non-coverage of the intervals is bounded by (5), except that the supremum is only taken over possible empirical distributions for $\theta_1, \dots, \theta_n$ satisfying the moment constraint. Since this supremum is necessarily smaller than $\rho(\sigma^2/\mu_2, \chi)$, it follows that the average coverage is at least $1-\alpha$. In fact, if the Y_i's exhibit limited dependence across i, a stronger property holds: the probability that at least a fraction $1-\alpha$ of the n EBCIs contain their respective true parameters converges to 1 as $n \to \infty$; cf. Remark 4.1 below.

The usual CIs $Y_i \pm z_{1-\alpha/2}\sigma$ also of course achieve average coverage $1-\alpha$. The robust EBCI in Eq. (6) will, however, be shorter, especially when μ_2 is small relative to σ^2—see Figure 3 below. The reduction in length is achieved by weakening the requirement that each CI covers its true parameter $1-\alpha$ percent of the time to the requirement that the coverage probability equal $1-\alpha$ on average across the CIs. It may seem surprising that we can construct a narrower CI by centering it at the shrinkage estimates $w_{\text{EB}} Y_i$. The intuition for this is that the shrinkage reduces the variability of the estimates, at the expense of introducing bias in the estimates. The bias necessitates the use of a larger critical value $\text{cva}_\alpha(\sigma^2/\mu_2)$. Because under the average coverage criterion we only need to control the bias *on average* across i, rather than for each individual θ_i, this increase in the critical value is smaller than the reduction in the standard error.

3. PRACTICAL IMPLEMENTATION

We now describe how to compute a robust EBCI that allows for heteroscedasticity, shrinks toward more general regression estimates rather than toward zero, and exploits higher moments of the bias to yield a narrower interval. In Section 3.1, we describe the empirical Bayes model that motivates our baseline approach. Section 3.2 describes the practical implementation of our baseline approach.

3.1. *Motivating Model and Robust EBCI*

In applied settings, the unshrunk estimates Y_i will typically have heteroscedastic variances. Furthermore, rather than shrinking toward zero, it is common to shrink toward an estimate of θ_i based on some covariates X_i, such as a regression estimate $X_i'\hat{\delta}$. We now describe how to adapt the ideas in Section 2 to such settings.

Consider a generalization of the model in Eq. (1) that allows for heteroscedasticity and covariates,

$$Y_i|\theta_i, X_i, \sigma_i \sim N\big(\theta_i, \sigma_i^2\big), \quad i = 1, \dots, n. \tag{8}$$

The covariate vector X_i may contain just the intercept, and it may also contain (functions of) σ_i. To construct an EB estimator of θ_i, consider the working assumption that the sampling distribution of the θ_i's is conditionally normal:

$$\theta_i|X_i, \sigma_i \sim N(\mu_{1,i}, \mu_2), \quad \text{where } \mu_{1,i} = X_i'\delta. \tag{9}$$

The hierarchical model (8)–(9) leads to the Bayes estimate $\hat{\theta}_i = \mu_{1,i} + w_{\text{EB},i}(Y_i - \mu_{1,i})$, where $w_{\text{EB},i} = \frac{\mu_2}{\mu_2+\sigma_i^2}$. This estimate shrinks the unrestricted estimate Y_i of θ_i toward $\mu_{1,i} = X_i'\delta$. In contrast to (8), the normality assumption (9) typically cannot be justified simply by appealing to the CLT; the linearity of the conditional mean $\mu_{1,i} = X_i'\delta$ may also be suspect. Our robust EBCI will therefore be constructed so that it achieves valid EB coverage even if assumption (9) fails. To obtain a narrow robust EBCI, we augment the second moment restriction used to compute the critical value in Eq. (5) with restrictions on higher moments of the bias of $\hat{\theta}_i$. In our baseline specification, we add a restriction on the fourth moment.

In particular, we replace assumption (9) with the much weaker requirement that the conditional second moment and kurtosis of $\varepsilon_i = \theta_i - X_i'\delta$ do not depend on (X_i, σ_i):

$$E\big[\big(\theta_i - X_i'\delta\big)^2|X_i, \sigma_i\big] = \mu_2, \qquad E\big[\big(\theta_i - X_i'\delta\big)^4|X_i, \sigma_i\big]/\mu_2^2 = \kappa, \tag{10}$$

where δ is defined as the probability limit of the regression estimate $\hat{\delta}$.[6] We discuss this requirement further in Remark 3.1 below, and we relax it in Remark 3.2 below.

We now apply analysis analogous to that in Section 2. Let us suppose for simplicity that δ, μ_2, κ, and σ_i are known; we relax this assumption in Section 3.2 below, and in the theory in Section 4. Denote the conditional bias of $\hat{\theta}_i$ normalized by the standard error by $b_i = (w_{\text{EB},i} - 1)\varepsilon_i/(w_{\text{EB},i}\sigma_i) = -\sigma_i\varepsilon_i/\mu_2$. Under repeated sampling of θ_i, the non-coverage of the CI $\hat{\theta}_i \pm \chi w_{\text{EB},i}\sigma$, conditional on (X_i, σ_i), depends on the distribution of

[6]Our framework can be modified to let (X_i, σ_i) be fixed, in which case δ depends on n. See the discussion following Theorem 4.1 below.

the normalized bias b_i, as in Section 2. Given the moments μ_2 and κ, the *maximal* non-coverage is given by

$$\rho(m_{2,i}, \kappa, \chi) = \sup_F E_F\big[r(b, \chi)\big] \quad \text{s.t. } E_F\big[b^2\big] = m_{2,i},\ E_F\big[b^4\big] = \kappa m_{2,i}^2, \tag{11}$$

where b is distributed according to the distribution F. Here $m_{2,i} = E[b_i^2 | X_i, \sigma_i] = \sigma_i^2/\mu_2$. Observe that the kurtosis of b_i matches that of ε_i. Appendix B shows that the infinite-dimensional linear program (11) can be reduced to two nested *univariate* optimizations. We also show that the least favorable distribution—the distribution F maximizing (11)—is a discrete distribution with up to four support points (see Remark B.1).

Define the critical value $\text{cva}_\alpha(m_{2,i}, \kappa) = \rho^{-1}(m_{2,i}, \kappa, \alpha)$, where the inverse is in the last argument. Figure 1 plots this function for $\alpha = 0.05$ and selected values of κ. This leads to the robust EBCI

$$\hat{\theta}_i \pm \text{cva}_\alpha(m_{2,i}, \kappa) w_{\text{EB},i} \sigma_i, \tag{12}$$

which, by construction, has coverage at least $1 - \alpha$ under repeated sampling of (Y_i, θ_i), conditional on (X_i, σ_i), so long as Eq. (10) holds; it is not required that (9) holds. Note that both the critical value and the CI length are increasing in σ_i.

3.2. *Baseline Implementation*

Our baseline implementation of the robust EBCI plugs in consistent estimates of the unknown quantities in Eq. (12), based on the data $\{Y_i, X_i, \hat{\sigma}_i\}_{i=1}^n$, where $\hat{\sigma}_i$ is a consistent estimate of σ_i (such as the standard error of the preliminary estimate Y_i), and X_i is a vector of covariates that are thought to help predict θ_i.

1. Regress Y_i on X_i to obtain the fitted values $X_i'\hat{\delta}$, with $\hat{\delta} = (\sum_{i=1}^n \omega_i X_i X_i')^{-1} \sum_{i=1}^n \omega_i \times X_i Y_i$ denoting the weighted least squares estimate with precision weights ω_i. Two natural choices are setting $\omega_i = \hat{\sigma}_i^{-2}$, or setting $\omega_i = 1/n$ for unweighted estimates; see Appendix A.2 for further discussion. Let $\hat{\mu}_2 = \max\{\frac{\sum_{i=1}^n \omega_i(\hat{\varepsilon}_i^2 - \hat{\sigma}_i^2)}{\sum_{i=1}^n \omega_i}, \frac{2\sum_{i=1}^n \omega_i^2 \hat{\sigma}_i^4}{\sum_{i=1}^n \omega_i \cdot \sum_{i=1}^n \omega_i \hat{\sigma}_i^2}\}$, and $\hat{\kappa} = \max\{\frac{\sum_{i=1}^n \omega_i(\hat{\varepsilon}_i^4 - 6\hat{\sigma}_i^2\hat{\varepsilon}_i^2 + 3\hat{\sigma}_i^4)}{\hat{\mu}_2^2 \sum_{i=1}^n \omega_i}, 1 + \frac{32\sum_{i=1}^n \omega_i^2 \hat{\sigma}_i^8}{\hat{\mu}_2^2 \sum_{i=1}^n \omega_i \cdot \sum_{i=1}^n \omega_i \hat{\sigma}_i^4}\}$, where $\hat{\varepsilon}_i = Y_i - X_i'\hat{\delta}$.
2. Form the EB estimate

$$\hat{\theta}_i = X_i'\hat{\delta} + \hat{w}_{\text{EB},i}\big(Y_i - X_i'\hat{\delta}\big), \quad \text{where } \hat{w}_{\text{EB},i} = \frac{\hat{\mu}_2}{\hat{\mu}_2 + \hat{\sigma}_i^2}.$$

3. Compute the critical value $\text{cva}_\alpha(\hat{\sigma}_i^2/\hat{\mu}_2, \hat{\kappa})$ defined below Eq. (11).
4. Report the robust EBCI

$$\hat{\theta}_i \pm \text{cva}_\alpha\big(\hat{\sigma}_i^2/\hat{\mu}_2, \hat{\kappa}\big)\hat{w}_{\text{EB},i}\hat{\sigma}_i. \tag{13}$$

We provide fast and stable software packages that automate these steps (see footnote 1). We now discuss the assumptions needed for validity of the robust EBCI.

REMARK 3.1—Conditional EB coverage and moment independence: A potential concern about EB coverage in a heteroscedastic setting is that in order to reduce the length of the CI on average, one could choose to overcover parameters θ_i with small σ_i and undercover parameters θ_i with large σ_i. Our robust EBCI ensures that this does not happen by requiring EB coverage to hold conditional on (X_i, σ_i). This also avoids analogous coverage concerns as a result of the value of X_i.

The key to ensuring this property is assumption (10) that the conditional second moment and kurtosis of $\varepsilon_i = \theta_i - X_i'\delta$ do not depend on (X_i, σ_i). Conditional moment independence assumptions of this form are common in the literature. For instance, it is imposed in the analysis of neighborhood effects in Chetty and Hendren (2018) (their approach requires independence of the second moment), which is the basis for our empirical application in Section 7. Nonetheless, such conditions may be strong in some settings, as argued by Xie, Kou, and Brown (2012) in the context of EB point estimation. In Remark 3.2 below, we drop condition (10) entirely by replacing $\hat{\mu}_2$ and $\hat{\kappa}$ with nonparametric estimates of these conditional moments; alternatively, one could relax it by using a flexible parametric specification.[7]

REMARK 3.2—Nonparametric moment estimates: As a robustness check to guard against failure of the moment independence assumption (10), one may replace the critical value in Eq. (13) with $\mathrm{cva}_\alpha((1-1/\hat{w}_{\mathrm{EB},i})^2\hat{\mu}_{2i}/\hat{\sigma}_i^2, \hat{\kappa}_i)$, where $\hat{\mu}_{2i}$ and $\hat{\kappa}_i$ are consistent nonparametric estimates of $\mu_{2i} = E[(\theta_i - X_i'\delta)^2|X_i, \sigma_i]$ and $\kappa_i = E[(\theta_i - X_i'\delta)^4|X_i, \sigma_i]/\mu_{2i}^2$. The resulting CI will be asymptotically equivalent to the CI in the baseline implementation if Eq. (10) holds, but it will achieve valid EB coverage even if this assumption fails. In our empirical application, we use nearest-neighbor estimates, as described in Appendix A.1. As a simple diagnostic to gauge how much the second moment of $\theta_i - X_i'\delta$ varies with (X_i, σ_i), one can report the R^2 gain in predicting $\hat{\varepsilon}_i^2 - \hat{\sigma}_i^2$ using $\hat{\mu}_{2i}$ rather than the baseline estimate $\hat{\mu}_2$, as we illustrate in our empirical application.

REMARK 3.3—Average coverage and non-independent sampling: We show in Section 4 that the robust EBCI satisfies an average coverage criterion of the form (7) when the parameters $\theta = (\theta_1, \dots, \theta_n)$ are considered fixed, in addition to achieving valid EB coverage when the θ_i's are viewed as random draws from some underlying distribution. To guarantee average coverage or EB coverage, we do not need to assume that the Y_i's and θ_i's are drawn independently across i. This is because the average coverage and EB coverage criteria only depend on the marginal distribution of (Y_i, θ_i), not the joint distribution. Indeed, in deriving the infeasible CI in Eq. (12), we made no assumptions about the dependence structure of (Y_i, θ_i) across i. Consequently, to guarantee asymptotic coverage of the feasible interval in Eq. (13) as $n \to \infty$, we only need to ensure that the estimates $\hat{\mu}_2$, $\hat{\kappa}$, $\hat{\delta}$, $\hat{\sigma}_i$ are consistent for μ_2, κ, δ, σ_i, which is the case under many forms of weak dependence or clustering. Furthermore, our baseline implementation above does not require the researcher to take an explicit stand on the dependence of the data; for example, in the case of clustering, the researcher does not need to take an explicit stand on how the clusters are defined.

REMARK 3.4—Estimating moments of the distribution of θ_i: The estimators $\hat{\mu}_2$ and $\hat{\kappa}$ in step 1 of our baseline implementation above are based on the moment conditions $E[(Y_i - X_i'\delta)^2 - \sigma_i^2|X_i, \sigma_i] = \mu_2$ and $E[(Y_i - X_i'\delta)^4 + 3\sigma_i^4 - 6\sigma_i^2(Y_i - X_i'\delta)^2|X_i, \sigma_i] = \kappa\mu_2^2$, replacing population expectations by weighted sample averages. In addition, to avoid small-sample coverage issues when μ_2 and κ are near their theoretical lower bounds of 0 and 1, respectively, these estimates incorporate truncation on $\hat{\mu}_2$ and $\hat{\kappa}$. These truncated

[7]Another way to drop condition (10) is to base shrinkage on the t-statistics Y_i/σ_i, applying the baseline implementation above with $Y_i/\hat{\sigma}_i$ in place of Y_i and 1 in place of $\hat{\sigma}_i$. Then the homoscedastic analysis in Section 2 applies, leading to valid EBCIs without any assumptions about independence of the moments. See Remark 3.8 and Appendix D.1 in Armstrong, Kolesár, and Plagborg-Møller (2020) for further discussion.

estimates approximate the Bayesian posterior means under a flat prior on μ_2 and κ, as in Morris (1983a,b). Although the resulting EBCIs do not directly account for estimation uncertainty in μ_2 and κ, we verify their small-sample coverage accuracy via extensive simulations in Section 4.4. Appendix A.1 discusses the choice of the moment estimates, as well as other ways of performing truncation.

REMARK 3.5—Using higher moments and other forms of shrinkage: In addition to using the second and fourth moment of bias, one may augment (11) with restrictions on higher moments of the bias in order to further tighten the critical value. In Section 4.2, we show that using other moments in addition to the second and fourth moment does not substantially decrease the critical value in the case where θ_i is normally distributed. Thus, the CI in our baseline implementation is robust to failure of the normality assumption (9), while being near-optimal when this assumption does hold. Section 4.2 also shows that further efficiency gains are possible if one uses the linear estimator $\tilde{\theta}_i = \mu_{1,i} + w_i(Y_i - \mu_{1,i})$ with the shrinkage coefficient w_i chosen to optimize CI length, instead of using the MSE-optimal shrinkage $w_{\text{EB},i}$. For efficiency under a non-normal distribution of θ_i, one needs to consider nonlinear shrinkage; we discuss this extension in Section 6.1.

4. MAIN RESULTS

This section provides formal statements of the coverage properties of the CIs presented in Sections 2 and 3. Furthermore, we show that the CIs presented in Sections 2 and 3 are highly efficient when the mean parameters are in fact normally distributed. Next, we calculate the maximal coverage distortion of the parametric EBCI, and derive a rule of thumb for gauging the potential coverage distortion. Finally, we present a comprehensive simulation study of the finite-sample performance of the robust EBCI. Applied readers interested primarily in implementation issues may skip ahead to the empirical application in Section 7.

4.1. *Coverage Under Baseline Implementation*

In order to state the formal result, let us first carefully define the notions of coverage that we consider. Consider intervals $CI_1, \dots, CI_n$ for elements of the parameter vector $\theta = (\theta_1, \dots, \theta_n)'$. The probability measure P denotes the joint distribution of θ and $CI_1, \dots, CI_n$. Following Morris (1983b, Eq. 3.6) and Carlin and Louis (2000, Chapter 3.5), we say that the interval CI_i is an (asymptotic) $1-\alpha$ empirical Bayes confidence interval (EBCI) if

$$\liminf_{n\to\infty} P(\theta_i \in CI_i) \geq 1 - \alpha. \tag{14}$$

We say that the intervals CI_i are (asymptotic) $1-\alpha$ average coverage intervals (ACIs) under the parameter sequence $\theta_1, \dots, \theta_n$ if

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} P(\theta_i \in CI_i | \theta) \geq 1 - \alpha. \tag{15}$$

The average coverage property (15) is a property of the distribution of the data conditional on θ and therefore does not require that we view the θ_i's as random (as in a Bayesian or "random effects" analysis). To maintain consistent notation, we nonetheless use the conditional notation $P(\cdot|\theta)$ when considering average coverage. See Appendix C for a formulation with θ treated as nonrandom.

Observe that under the exchangeability condition that $P(\theta_i \in CI_i) = P(\theta_j \in CI_j)$ for all i, j, if the ACI property (15) holds almost surely, then the EBCI property (14) holds, since then

$$P(\theta_i \in CI_i) = \frac{1}{n}\sum_{j=1}^{n} P(\theta_j \in CI_j) \geq 1-\alpha+o(1) \quad \text{for all } i.$$

We now provide coverage results for the baseline implementation described in Section 3.2. To keep the statements in the main text as simple as possible, we (i) maintain the assumption that the unshrunk estimates Y_i follow an exact normal distribution conditional on the parameter θ_i, (ii) state the results only for the homoscedastic case where the variance σ_i of the unshrunk estimate Y_i does not vary across i, and (iii) consider only unconditional coverage statements of the form (14) and (15). In Appendix C, we allow the estimates Y_i to be only approximately normally distributed and allow σ_i to vary, and we verify that our assumptions hold in a linear fixed effects panel data model. We also formalize the statements about conditional coverage made in Remark 3.1.

THEOREM 4.1: *Suppose $Y_i|\theta \sim N(\theta_i, \sigma^2)$. Let $\mu_{j,n} = \frac{1}{n}\sum_{i=1}^{n}(\theta_i - X_i'\delta)^j$ and let $\kappa_n = \mu_{4,n}/\mu_{2,n}^2$. Suppose the sequence $\theta = \theta_1, \dots, \theta_n$ and the conditional distribution $P(\cdot|\theta)$ satisfy the following conditions with probability* 1:

1. *$\mu_{2,n} \to \mu_2$ and $\mu_{4,n}/\mu_{2,n}^2 \to \kappa$ for some $\mu_2 \in (0, \infty)$ and $\kappa \in (1, \infty)$.*
2. *Conditional on θ, $(\hat{\delta}, \hat{\sigma}, \hat{\mu}_2, \hat{\kappa})$ converges in probability to $(\delta, \sigma, \mu_2, \kappa)$.*

Then the CIs in Eq. (13) *with $\hat{\sigma}_i = \hat{\sigma}$ satisfy the ACI property* (15) *with probability* 1. *Furthermore, if $\theta_1, \dots, \theta_n$ follow an exchangeable distribution and the estimators $\hat{\delta}$, $\hat{\sigma}$, $\hat{\mu}_2$, and $\hat{\kappa}$ are exchangeable functions of the data $(X_1', Y_1)', \dots, (X_n', Y_n)'$, then these CIs satisfy the EB coverage property* (14).

Theorem 4.1 follows immediately from Theorem C.2 in Appendix C. In order to cover both the EB coverage condition (14) and the average coverage condition (15), Theorem 4.1 considers a random sequence of parameters $\theta_1, \dots, \theta_n$, and shows average coverage conditional on these parameters. See Appendix C for a formulation with θ treated as nonrandom.

The condition on the moments μ_2 and κ avoids degenerate cases such as when $\mu_2 = 0$, in which case the EB point estimator $\hat{\theta}_i$ shrinks each preliminary estimate Y_i all the way to $X_i'\hat{\delta}$. Note also that the theorem does not require that $\hat{\delta}$ be the ordinary least squares (OLS) estimate in a regression of Y_i onto X_i, and that δ be the population analog; one can define δ in other ways, the theorem only requires that $\hat{\delta}$ be a consistent estimate of it. The definition of δ does, however, affect the plausibility of the moment independence assumption in Eq. (10) needed for conditional coverage results stated in Appendix C.[8]

REMARK 4.1: As shown in Appendix C, if CIs satisfy the average coverage condition (15) given $\theta_1, \dots, \theta_n$, they will typically also satisfy the stronger condition

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{I}\{\theta_i \in CI_i\} \geq 1-\alpha+o_{P(\cdot|\theta)}(1), \tag{16}$$

[8]The specification of $\mu_{1i} = X_i'\delta$ also affects the EBCI width through its effect on μ_2 and κ.

where $o_{P(\cdot|\theta)}(1)$ denotes a sequence that converges in probability to zero conditional on θ (Eq. (16) implies Eq. (15) since the left-hand side is uniformly bounded). That is, at least a fraction $1-\alpha$ of the n CIs contain their respective true parameters, asymptotically. This is analogous to the result that for estimation, the difference between the squared error $\frac{1}{n}\sum_{i=1}^n(\hat{\theta}_i-\theta_i)^2$ and the MSE $\frac{1}{n}\sum_{i=1}^n E[(\hat{\theta}_i-\theta_i)^2|\theta]$ typically converges to zero.

4.2. *Relative Efficiency*

The robust EBCI in Eq. (12), unlike the parametric EBCI $\hat{\theta}_i \pm z_{1-\alpha/2}\sigma_i\sqrt{w_{\text{EB},i}}$, does not rely on the normality assumption in Eq. (9) for its validity. We now show that this robustness does not come at a high cost in terms of efficiency: if the normality assumption (9) in fact holds, the efficiency loss is limited unless the signal-to-noise ratio μ_2/σ_i^2 is very small.

There are two reasons for the inefficiency of the robust EBCI. First, the robust EBCI only makes use of the second and fourth moment of the conditional distribution of $\theta_i - X_i'\delta$, rather than its full distribution. Second, if we only have knowledge of these two moments, it is no longer optimal to center the EBCI at the estimator $\hat{\theta}_i$: one may need to consider other, perhaps nonlinear, shrinkage estimators, as we do below in Section 6.1.

We decompose the sources of inefficiency by studying the relative length of the robust EBCI relative to the EBCI that picks the amount of shrinkage optimally. For the latter, we maintain assumption (10), and consider a more general class of estimators $\tilde{\theta}(w_i)=\mu_{1,i}+w_i(Y_i-\mu_{1,i})$. For tractability, we focus on fixed-length CIs based on linear shrinkage estimators, but allow the amount of shrinkage w_i to be optimally determined. The normalized bias of $\tilde{\theta}(w_i)$ is given by $b_i=(1/w_i-1)\varepsilon_i/\sigma_i$, which leads to the EBCI

$$\mu_{1,i}+w_i(Y_i-\mu_{1,i})\pm \text{cva}_\alpha\big((1-1/w_i)^2\mu_2/\sigma_i^2,\kappa\big)w_i\sigma_i.$$

The half-length of this EBCI, $\text{cva}_\alpha((1-1/w_i)^2\mu_2/\sigma_i^2,\kappa)w_i\sigma_i$, can be numerically minimized as a function of w_i to find the EBCI length-optimal shrinkage. Denote the minimizer by $w_{\text{opt}}(\mu_2/\sigma_i^2,\kappa,\alpha)$. Like $w_{\text{EB},i}$, the optimal shrinkage depends on μ_2 and σ_i^2 only through the signal-to-noise ratio μ_2/σ_i^2. Numerically evaluating the minimizer shows that $w_{\text{opt}}(\cdot,\kappa,\alpha)\geq w_{\text{EB},i}$ for $\kappa\geq 3$ and $\alpha\in\{0.05,0.1\}$. The resulting EBCI is optimal among all fixed-length EBCIs centered at linear estimators under (10), and we call it the optimal robust EBCI.[9]

Figure 2 plots the ratio of lengths of the optimal robust EBCI and robust EBCI relative to the parametric EBCI, for $\alpha=0.05$. The figure shows that to maintain efficiency relative to the normal benchmark, it is important to impose the fourth moment constraint. If this constraint is imposed, the efficiency loss of the robust EBCI is modest unless the signal-to-noise ratio is very small: if $w_{\text{EB},i}\geq 0.1$ (which is equivalent to $\mu_2/\sigma_i^2\geq 1/9$), the efficiency loss is at most 11.4% for $\alpha=0.05$; up to half of the efficiency loss is due to not using the optimal shrinkage. For $\alpha=0.1$ (not plotted), the results are very similar; in particular, if $w_{\text{EB},i}\geq 0.1$, the efficiency loss is at most 12.9%.

When the signal-to-noise ratio is very small, so that $w_{\text{EB},i}<0.1$, the efficiency loss of the robust EBCI is higher (up to 39% for $\alpha=0.05$ or 0.1). Using the optimal robust EBCI ensures that the efficiency loss is below 20%, irrespective of the signal-to-noise ratio. On the other hand, when the signal-to-noise ratio is small, any of these CIs will be

[9]Since the optimal robust EBCI is always shorter than the robust EBCI in Eq. (12), the former is preferable on efficiency grounds. It may not contain the MSE-optimal point estimator $\hat{\theta}_i$, however.

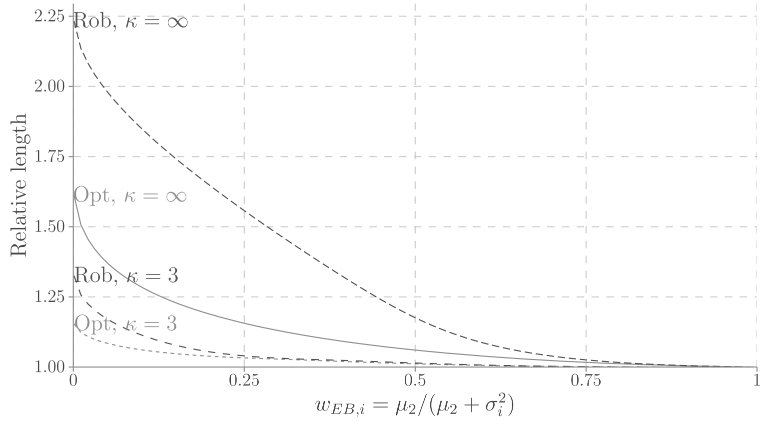

FIGURE 2.—Relative efficiency of robust EBCI (Rob) and optimal robust EBCI (Opt) relative to the normal benchmark, for $\alpha = 0.05$. The figure plots ratios of Rob length, $2\,\mathrm{cva}_\alpha(\sigma_i^2/\mu_2, \kappa) \cdot \sigma_i\mu_2/(\mu_2 + \sigma_i^2)$, and Opt length, $2\,\mathrm{cva}_\alpha((1 - 1/w_{\mathrm{opt}}(\mu_2/\sigma_i^2, \kappa, \alpha))^2\mu_2/\sigma_i^2, \kappa) \cdot \sigma_i w_{\mathrm{opt}}(\mu_2/\sigma_i^2, \kappa, \alpha)$, relative to the parametric EBCI length $2z_{1-\alpha/2}\sqrt{\mu_2/(\mu_2 + \sigma_i^2)}\sigma_i$ as a function of the shrinkage factor $w_{\mathrm{EB},i} = \mu_2/(\mu_2 + \sigma_i^2)$, which maps the signal-to-noise ratio μ_2/σ_i^2 to the interval [0, 1].

significantly tighter than the unshrunk CI $Y_i \pm z_{1-\alpha/2}\sigma_i$. To illustrate this point, Figure 3 plots the efficiency of the robust EBCI that imposes the second moment constraint only, relative to this unshrunk CI. It can be seen from the figure that shrinkage methods allow us to tighten the CI by 44% or more when $\mu_2/\sigma_i^2 \leq 0.1$.

4.3. *Undercoverage of Parametric EBCI*

The parametric EBCI $\hat{\theta}_i \pm z_{1-\alpha/2}w_{\mathrm{EB},i}^{1/2}\sigma_i$ is an EB version of a Bayesian credible interval that treats (9) as a prior. We now assess its potential undercoverage when Eq. (9) is violated.

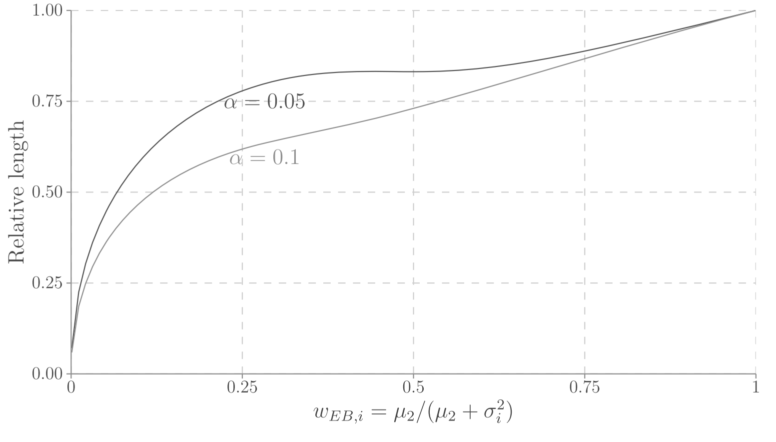

FIGURE 3.—Efficiency of robust EBCI $\hat{\theta}_i \pm \mathrm{cva}_\alpha(\sigma_i^2/\mu_2, \kappa = \infty) \cdot \sigma\mu_2/(\mu_2 + \sigma_i^2)$ relative to the unshrunk CI $Y_i \pm z_{1-\alpha/2}\sigma_i$. The figure plots the ratio of the length of the robust EBCI relative to the unshrunk CI as a function of the shrinkage factor $w_{\mathrm{EB},i} = \mu_2/(\mu_2 + \sigma_i^2)$.

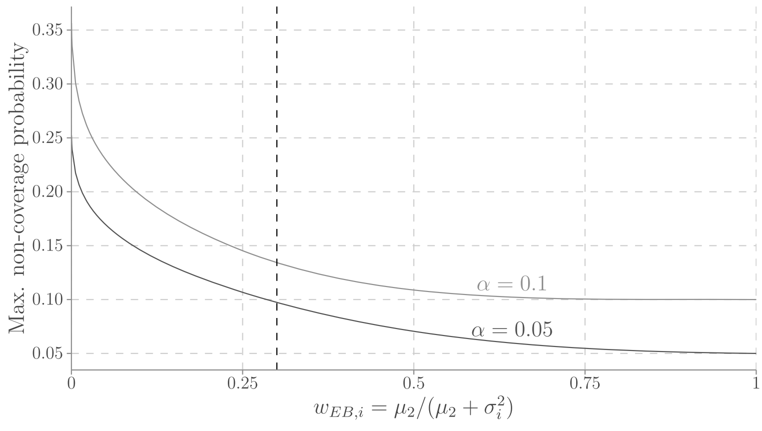

FIGURE 4.—Maximal non-coverage probability of parametric EBCI, $\alpha \in \{0.05, 0.10\}$. The vertical line marks the "rule of thumb" value $w_{\mathrm{EB},i} = 0.3$, above which the maximal coverage distortion is less than 5 percentage points for these two values of α.

Given knowledge of only the second moment μ_2 of $\varepsilon_i = Y_i - X_i'\delta$, the maximal under-coverage of this interval is given by

$$\rho\big(1/w_{\mathrm{EB},i} - 1, z_{1-\alpha/2}/\sqrt{w_{\mathrm{EB},i}}\big), \tag{17}$$

since $w_{\mathrm{EB},i} = \mu_2/(\mu_2 + \sigma_i^2)$. Here ρ is the non-coverage function defined in Eq. (5). Figure 4 plots the maximal non-coverage probability as a function of $w_{\mathrm{EB},i}$, for significance levels $\alpha = 0.05$ and $\alpha = 0.10$. The figure suggests a simple "rule of thumb": if $w_{\mathrm{EB},i} \geq 0.3$, the maximal coverage distortion is less than 5 percentage points for these values of α.

The following lemma confirms that the maximal non-coverage is decreasing in $w_{\mathrm{EB},i}$, as suggested by the figure. It also gives an expression for the maximal non-coverage across all values of $w_{\mathrm{EB},i}$ (which is achieved in the limit $w_{\mathrm{EB},i} \to 0$).

LEMMA 4.1: *The non-coverage probability* (17) *of the parametric EBCI is weakly decreasing as a function of* $w_{\mathrm{EB},i}$, *with the supremum given by* $1/\max\{z_{1-\alpha}^2, 1\}$.

The maximal non-coverage probability $1/\max\{z_{1-\alpha/2}^2, 1\}$ equals 0.260 for $\alpha = 0.05$ and 0.370 for $\alpha = 0.10$. For $\alpha > 2\Phi(-1) \approx 0.317$, the maximal non-coverage probability is 1.

If we additionally impose knowledge of the kurtosis of ε_i, the maximal non-coverage of the parametric EBCI can be similarly computed using Eq. (11), as illustrated in the application in Section 7.

4.4. *Monte Carlo Simulations*

Here we show through simulations that the robust EBCI achieves accurate average coverage in finite samples.

4.4.1. *Design*

The DGP is a simple linear fixed effects panel data model. We first draw θ_i, $i = 1, \dots, n$, i.i.d. from a random effects distribution specified below. Then we simulate panel data from the model

$$W_{it} = \theta_i + U_{it}, \quad i = 1, \dots, n, t = 1, \dots, T,$$

where the errors U_{it} are mean zero and i.i.d. across (i,t) and independent of the θ_i's. The unshrunk estimator of θ_i is the sample average of W_{it} for unit i, with standard error obtained from the usual unbiased variance estimator:

$$Y_i = \frac{1}{T}\sum_{t=1}^{T} W_{it}, \qquad \hat{\sigma}_i = \sqrt{\frac{1}{T(T-1)}\sum_{t=1}^{T}(W_{it}-Y_i)^2}.$$

We draw U_{it} from one of two distributions: (1) a normal distribution and (2) a (shifted) chi-squared distribution with 3 degrees of freedom. In case (1), Y_i is exactly normal conditional on θ_i, but $\hat{\sigma}_i^2$ does not exactly equal $\operatorname{var}(Y_i|\theta_i)$ for finite T. In case (2), Y_i is non-normal and positively skewed (conditional on θ_i) for finite T.

We consider six random effects distributions for θ_i (see Supplemental Material Appendix E.1 (Armstrong, Kolesár, and Plagborg-Møller (2022)) for detailed definitions): (i) normal (kurtosis $\kappa = 3$); (ii) scaled chi-squared with 1 degree of freedom ($\kappa = 15$); (iii) two-point distribution ($\kappa \approx 8.11$); (iv) three-point distribution ($\kappa = 2$); (v) the least favorable distribution for the robust EBCI that exploits only second moments (κ depends on μ_2, see Appendix B); and (vi) the least favorable distribution for the parametric EBCI.

Given T, we scale the θ_i distribution to match one of four signal-to-noise ratios $\mu_2/\operatorname{var}(Y_i|\theta_i) \in \{0.1, 0.5, 1, 2\}$, for a total of $6 \times 4 = 24$ DGPs for each distribution of U_{it}. We shrink toward the grand mean ($X_i = 1$ for all i). We construct the robust EBCIs following the baseline implementation in Section 3.2 (with $\omega_i = 1/n$), as well as a version that does not impose constraints on the kurtosis.

As $T \to \infty$, we recover the idealized setting in Section 2, with $(Y_i - \theta_i)/\sqrt{\operatorname{var}(Y_i|\theta_i)}$ converging in distribution to a standard normal (conditional on θ_i), and $\hat{\sigma}_i^2/\operatorname{var}(Y_i|\theta_i)$ converging in probability to 1, for each i.

4.4.2. *Results*

Table I shows that the 95% robust EBCIs achieve good average coverage when the panel errors U_{it} are normally distributed. This is true for all DGPs, panel dimensions n

TABLE I

MONTE CARLO SIMULATION RESULTS, PANEL DATA WITH NORMAL ERRORS.

	Robust, μ_2 only				Robust, μ_2 & κ				Parametric			
T	10	20	∞	ora	10	20	∞	ora	10	20	∞	ora
Panel A: Average coverage (%), minimum across 24 DGPs												
$n=100$	92.1	93.7	94.0	95.0	91.8	93.2	93.2	94.6	79.2	79.7	79.3	86.9
$n=200$	91.9	93.4	92.9	95.0	91.8	93.3	92.9	94.8	80.7	80.3	81.0	86.3
$n=500$	91.9	93.6	94.8	95.0	91.9	93.5	94.3	94.9	84.2	85.1	85.1	85.6
Panel B: Relative average length, average across 24 DGPs												
$n=100$	1.09	1.10	1.11	1.16	1.03	1.02	1.02	1.00	0.81	0.82	0.83	0.86
$n=200$	1.09	1.10	1.12	1.16	1.02	1.02	1.01	1.00	0.81	0.82	0.84	0.86
$n=500$	1.10	1.11	1.13	1.16	1.04	1.03	1.01	1.00	0.82	0.83	0.84	0.86

Note: Normally distributed errors. Nominal average confidence level $1-\alpha = 95\%$. All EBCI procedures use baseline estimate of $\hat{\mu}_2$ and (if applicable) $\hat{\kappa}$, except columns labeled "ora," which use oracle values of μ_2 and κ. Columns $T = \infty$ and "ora" use oracle standard errors σ_i. For each DGP, "average coverage" and "average length" refer to averages across units $i = 1, \dots, n$ and across 2000 Monte Carlo repetitions. Average CI length is measured relative to the robust EBCI that exploits the oracle values of μ_2, κ, and σ_i (but not of the grand mean $\delta = E[\theta]$).

TABLE II

MONTE CARLO SIMULATION RESULTS, PANEL DATA WITH CHI-SQUARED ERRORS.

	Robust, μ_2 only				Robust, μ_2 & κ				Parametric			
T	10	20	50	ora	10	20	50	ora	10	20	50	ora
					Panel A: Average coverage (%), minimum across 24 DGPs							
$n = 100$	87.9	90.9	93.1	95.0	87.8	90.8	92.6	94.7	79.9	79.3	79.3	87.0
$n = 200$	87.9	90.8	93.0	94.9	87.8	90.8	92.8	94.8	77.8	79.8	80.3	86.2
$n = 500$	87.8	90.8	93.0	95.0	87.8	90.7	92.9	94.9	82.0	84.1	84.8	85.6
					Panel B: Relative average length, average across 24 DGPs							
$n = 100$	1.05	1.08	1.10	1.16	1.01	1.02	1.02	1.00	0.79	0.81	0.82	0.86
$n = 200$	1.04	1.08	1.10	1.16	0.99	1.00	1.00	1.00	0.78	0.81	0.82	0.86
$n = 500$	1.05	1.09	1.11	1.16	0.99	1.00	1.00	1.00	0.79	0.82	0.83	0.86

Note: Chi-squared distributed errors. See caption for Table I. Results for $T = \infty$ are by definition the same as in Table I.

and T, and whether we exploit one or both of the (estimated) moments μ_2 and κ. When the time dimension T equals 10, the maximal coverage distortion across all DGPs and all cross-sectional dimensions $n \in \{100, 200, 500\}$ is 3.2 percentage points. For $T \geq 20$, the coverage distortion of the robust EBCIs is always below 2.1 percentage points.

Table II shows that coverage distortions are somewhat larger when the panel errors U_{it} are chi-squared distributed and T is small. The robust EBCIs undercover by up to 7.2 percentage points when $T = 10$ due to the pronounced non-normality of Y_i given θ_i. However, the distortion is at most 4.3 percentage points when $T = 20$, and at most 2.4 percentage points when $T \geq 50$. The coverage distortion due to non-normality when T is small is similar to the coverage distortion of the usual unshrunk CI (not reported).

Importantly, in all cases considered in Tables I and II, the worst-case coverage distortion of the parametric EBCI substantially exceeds that of the corresponding robust EBCIs, sometimes by more than 10 percentage points. Nevertheless, the cost of robustness in terms of extra CI length is modest and consistent with the theoretical results in Section 4.2.

Both the estimation of the standard errors σ_i and the estimation of the moments μ_2 and κ contribute to the finite-sample coverage distortions. The "ora" columns in Table I exploit the oracle (true) values of μ_2, κ, and $\sigma_i = \sqrt{\operatorname{var}(Y_i|\theta_i)}$, while the $T = \infty$ columns use oracle standard errors but not oracle moments. By comparing these columns, we see that estimation of μ_2 and κ is responsible for modest coverage distortions when $n = 100$ or 200. However, estimation of the standard errors σ_i also contributes to the distortions, as can be seen by comparing the $T = 10$ and $T = \infty$ columns.

In Supplemental Material Appendix E.2, we show that the robust EBCI also has good coverage in a heteroscedastic design calibrated to the empirical application in Section 7 below.

5. COMPARISON WITH OTHER APPROACHES

Here we compare our EBCI procedure with other approaches to confidence interval construction in the normal means model. We also discuss other related inference problems.

5.1. *Average Coverage versus Alternative Coverage Concepts*

The average coverage requirement in Eq. (15) is less stringent than the usual (pointwise) notion of frequentist coverage that $P(\theta_i \in CI_i|\theta) \geq 1-\alpha$ for all i. An even stronger coverage requirement is that of simultaneous coverage: $P(\forall i\colon\ \theta_i \in CI_i|\theta) \geq 1-\alpha$. As outlined in footnote 3, under the pointwise coverage criterion, one cannot achieve substantial reductions in length relative to the unshrunk CI. Under the simultaneous coverage criterion, it is likewise impossible to substantially improve upon the usual sup-t confidence band based on the unshrunk estimates (Cai, Low, and Ma (2014)). Thus, undercoverage for some θ_i's must be tolerated if one wants to use shrinkage to improve CI length.

The fact that our EBCIs achieve improvements in average length at the expense of undercovering for certain units i is analogous to well-known properties of EB point estimators. We now show that the units i for which our EBCI undercovers are quantitatively similar to the units for which the shrinkage estimator $\hat{\theta}_i$ has higher MSE than the unshrunk estimator Y_i. Let $\varepsilon_i = \theta_i - X_i'\delta$ be the "shrinkage error" defined in Section 3.1. The pointwise coverage of our EBCI is decreasing in the normalized shrinkage error $|\varepsilon_i|/\sqrt{\mu_2}$, for a fixed signal-to-noise ratio μ_2/σ_i^2.[10] Hence, the units i for which our EBCI undercovers are those whose covariate-predicted value $X_i'\delta$ fails to approximate their true effect θ_i well. The MSE of the shrinkage estimator (for an individual unit i), normalized by the MSE of the unshrunk estimator, is similarly increasing in $|\varepsilon_i|/\sqrt{\mu_2}$.[11]

Figure 5 shows that the knife-edge value of $|\varepsilon_i|/\sqrt{\mu_2}$ for which the pointwise coverage of our EBCI equals $1-\alpha$ is quantitatively close to the value of $|\varepsilon_i|/\sqrt{\mu_2}$ for which the MSE of the shrinkage estimator equals that of the unshrunk estimator. In other words, to the extent that one worries about undercoverage for certain types of θ_i values, one should simultaneously worry about the relative performance of the shrinkage point estimator for those same values.

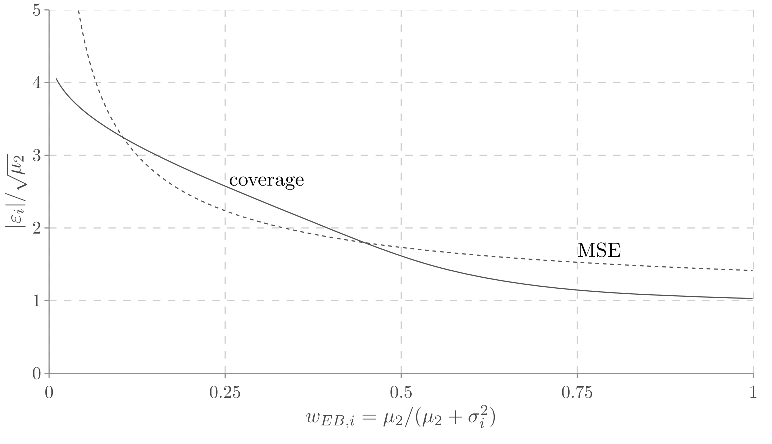

FIGURE 5.—Value of $|\varepsilon_i|/\sqrt{\mu_2}$, as a function of $w_{\mathrm{EB},i}$, such that the MSE of the shrinkage point estimator equals that of the unshrunk estimator (MSE), and such that the coverage of the robust EBCI with $\kappa=\infty$ equals the nominal average coverage $1-\alpha$ (coverage), for $\alpha = 0.05$.

[10]The pointwise coverage (conditional on X_i) equals $1 - r(\sqrt{1/w_{\mathrm{EB},i} - 1}\cdot|\varepsilon_i|/\sqrt{\mu_2}, \mathrm{cva}_\alpha(1/w_{\mathrm{EB},i}-1,\kappa))$, with r defined in Eq. (4) and $w_{\mathrm{EB},i} = \mu_2/(\mu_2+\sigma_i^2)$.

[11]The ratio of MSEs equals $E[(\hat{\theta}_i-\theta_i)^2|\theta_i, X_i]/\sigma_i^2 = w_{\mathrm{EB},i}^2 + (1-w_{\mathrm{EB},i})w_{\mathrm{EB},i}\cdot|\varepsilon_i|/\sqrt{\mu_2}$.

We stress that the pointwise coverage depends on the unobservable shrinkage error ε_i, which cannot be gauged directly from the observables (Y_i, X_i). If one wishes to avoid systematic differences in coverage across units i with different genders, say (i.e., one is worried that ε_i correlates with gender), one can simply add gender to the set of covariates X_i: the baseline procedure in Section 3.2 ensures control of average coverage conditional on the covariates X_i. In Section 6.2, we show how to adapt our EBCIs to settings where one focuses the analysis on a subset of units i based on the values of their unshrunk estimates Y_i (e.g., keeping only the estimates that exceed a given threshold).

From a Bayesian point of view, our robust EBCI can be viewed as an uncertainty interval that is robust to the choice of prior distribution in the *unconditional* gamma-minimax sense: the coverage probability of this CI is at least $1-\alpha$ when averaged over the distribution of the data and over the prior distribution for θ_i, for any prior distribution that satisfies the moment bounds. This follows directly from the derivations in Section 2, reinterpreting the random effects distribution for θ_i as a prior distribution. In contrast, *conditional* gamma-minimax credible intervals, discussed recently by Giacomini, Kitagawa, and Uhlig (2019, p. 6), are too stringent in our setting. This notion requires that the posterior credibility of the interval be at least $1-\alpha$ regardless of the choice of prior, in any data sample, which would require reporting the entire parameter space (up to the moment bounds).

5.2. *Finite-Sample versus Asymptotic Coverage*

Our procedures are asymptotically valid as $n \to \infty$, as proved in Section 4.1. These asymptotics do not capture the impact of estimation error in the "hyper-parameters" $\hat{\sigma}_i$, $\hat{\delta}$, $\hat{\mu}_2$, and $\hat{\kappa}$, or the impact of lack of exact normality of the Y_i's, on the finite-sample performance of the EBCIs. As detailed in Section 3.2 and Appendix A, we do apply a finite-sample adjustment to the moments $\hat{\mu}_2$ and $\hat{\kappa}$, which is motivated by the same heuristic arguments that Morris (1983a,b) used to motivate finite-sample adjustments to the parametric EBCI.[12] The promising simulation results in Section 4.4 notwithstanding, these adjustments do not ensure exact average coverage control in finite samples.[13]

Our results are thus analogous to standard results on coverage of Eicker–Huber–White CIs in cross-sectional OLS: asymptotic validity follows by consistency of the OLS variance estimate and asymptotic normality of the outcomes, while adjustments to account for finite-sample issues (such as the HC2 or HC3 variance estimators studied in MacKinnon and White (1985)) are justified heuristically. Deriving EBCIs with finite-sample coverage guarantees is an interesting problem that we leave for future research; the problem appears to be challenging even in the context of constructing parametric EBCIs.

5.3. *Local versus Global Optimality*

Our EBCIs are designed to provide uncertainty assessments to accompany linear shrinkage estimates that, as the Introduction argues, have been popular in applied work.

[12]An alternative approach would be to adapt the bootstrap adjustment proposed by Carlin and Louis (2000, Chapter 3.5.3) in the context of parametric EBCI construction (see also Efron (2019)). As with the Morris (1983a,b) adjustment, we are not aware of a formal result justifying it.

[13]One could account for hyper-parameter uncertainty by computing the critical value $\sup_{\tilde{\sigma}_i,\tilde{\mu}_2,\tilde{\kappa}\in\hat{\mathcal{C}}_i} \mathrm{cva}_\alpha(\tilde{\sigma}_i^2/\tilde{\mu}_2, \tilde{\kappa})$ over an initial confidence set $\hat{\mathcal{C}}_i$ for the hyper-parameters, coupled with a Bonferroni adjustment of the confidence level $1-\alpha$. This approach appears to be highly conservative in practice.

Our procedure's global validity, as well as local near-optimality when the θ_i's are normal (cf. Section 4.2), is analogous to Eicker–Huber–White CIs for OLS estimators: these CIs are optimal under normal homoscedastic regression errors, but remain valid when this assumption is dropped.

Similarly to the Eicker–Huber–White CIs, our EBCIs are not globally efficient: when the θ_i's are not Gaussian, it is generally inefficient to restrict attention to CIs that are centered at a linear point estimator and have fixed width. While we expect our EBCIs to remain near-efficient under mild departures from normality, substantial efficiency gains may be possible if the effect size distribution is, for example, heavy-tailed or bimodal.[14] Section 6.1 shows how our method can be adapted to construct EBCIs that are locally near-optimal under non-normal baseline priors using nonlinear shrinkage, such as soft thresholding. Since the distribution of θ_i is nonparametrically identified under the normal model (8), it is in principle possible to construct EBCIs that are globally efficient using nonparametric methods. In the context of the homoscedastic model with no covariates in Eq. (1), various approaches to nonparametric point estimation of the θ_i's have been proposed, including kernels (Brown and Greenshtein (2009)), splines (Efron (2019)), or nonparametric maximum likelihood (Kiefer and Wolfowitz (1956), Jiang and Zhang (2009), Koenker and Mizera (2014)). An interesting problem for future research is to adapt these methods to EBCI construction, while ensuring asymptotic validity, good finite-sample performance, and allowing for covariates, heteroscedasticity, and possible dependence across i.

5.4. *Other Inference Problems*

A number of alternative inference procedures have been proposed in the context of the normal means model. Efron (2015) developed a formula for the frequentist standard error of EB estimators, but this cannot be used to construct CIs without a corresponding estimate of the bias. There is a substantial literature on shrinkage confidence balls, that is, confidence sets of the form $\{\theta\colon \sum_{i=1}^{n}(\theta_i - \hat{\theta}_i)^2 \leq \hat{c}\}$ (see Casella and Hwang (2012), for a review). While theoretically interesting, these sets can be difficult to visualize and report in practice.[15]

Finally, while we focus on CI length in our relative efficiency comparisons, our approach can be fruitfully applied when the goal of CI construction is to discern non-null effects, rather than to construct short CIs. In particular, suppose one forms a test of the null hypothesis $H_{0,i}: \theta_i = \theta_0$ for some null value θ_0 by rejecting when $\theta_0 \notin CI_i$, where CI_i is our robust EBCI given in (12). In Supplemental Material Appendix F, we show that the test based on our EBCI has higher average power than the usual z-test based on the unshrunk estimate when $X_i'\delta$ (the regression line toward which we shrink) is far enough from the null value θ_0, and that these power gains can be substantial. Furthermore, such tests can be combined with corrections from the multiple testing literature to form procedures that asymptotically control the false discovery rate (FDR), a commonly used criterion for multiple testing.[16]

[14]Indeed, if the true effect distribution puts mass $1/2$ on $\theta_i = K$ and $\theta_i = -K$, then, as K gets large, our EBCIs become arbitrarily conservative relative to an oracle that reports the highest posterior density set under this prior.

[15]Confidence balls can be translated into average coverage intervals using Chebyshev's inequality (see Wasserman (2006, Chapter 5.8)). However, such intervals are very conservative compared to the ones we construct.

[16]In particular, Storey (2002) showed that the Benjamini and Hochberg (1995) procedure asymptotically controls the FDR so long as the p-values do not exhibit too much statistical dependence and the proportion of

6. EXTENSIONS

We now discuss two extensions of our method: adapting our intervals to general, possibly nonlinear shrinkage, and constructing intervals that achieve coverage conditional on Y_i falling into a pre-specified interval.

6.1. *General Shrinkage*

Our method can be generalized to cover general, possibly nonlinear shrinkage based on possibly non-Gaussian data. Let $\mathcal{S}(y; \chi, \tilde{X}_i) \subseteq \mathbb{R}$ be a family of candidate confidence sets for a parameter θ_i, which depends on the data $Y_i = y$, a tuning parameter $\chi \in \mathbb{R}$ to be selected below, and covariates $\tilde{X}_i$ (that include any known nuisance parameters) that we treat as fixed. We assume that $\mathcal{S}$ is increasing in χ, in the sense of set containment, and that the non-coverage probability conditional on θ satisfies

$$P\big(\theta_i \notin \mathcal{S}(Y_i; \chi, \tilde{X}_i) | \theta, \tilde{X}^{(n)}\big) = \tilde{r}(a_i, \chi), \tag{18}$$

where a_i is some function of θ_i, $\tilde{X}^{(n)} = (\tilde{X}_1, \dots, \tilde{X}_n)$, and $\tilde{r}$ is a known function (perhaps computed numerically or through simulation). Similarly to linear shrinkage in the normal means model, Eq. (18) may only hold approximately if the set $\mathcal{S}$ depends on estimated parameters (such as standard error estimates or tuning parameters), or if we use a large-sample approximation to the distribution of Y_i. We assume that a_i satisfies the moment constraints $E_F[g(a_i)|\tilde{X}^{(n)}] = m$, where g is a p-vector of moment functions, and the expectation is over the conditional distribution F of a_i conditional on $\tilde{X}^{(n)}$.[17] To guarantee EB coverage, we compute the maximal non-coverage

$$\rho_g(m, \chi) = \sup_F E_F\big[\tilde{r}(a, \chi)\big], \quad E_F\big[g(a)\big] = m, \tag{19}$$

analogously to Eq. (11). This is a linear program, which can be computed numerically to a high degree of precision even with several constraints; see Appendix B for details. Given an estimate $\hat{m}$ of the moment vector m, we form a robust EBCI as

$$\mathcal{S}(Y_i; \hat{\chi}, \tilde{X}_i), \quad \text{where } \hat{\chi} = \inf\big\{\chi : \rho_g(\hat{m}, \chi) \le \alpha\big\}. \tag{20}$$

EXAMPLE 6.1—Linear shrinkage in the normal model: The setting in Section 3.1 obtains if we set $\tilde{X}_i = (X_i, \sigma_i)$ and $\mathcal{S}(y; \chi, \tilde{X}_i) = \{(1 - w_{\mathrm{EB},i})X_i'\delta + w_{\mathrm{EB},i}Y_i \pm \chi w_{\mathrm{EB},i}\sigma_i\}$. Here a_i is given by the normalized bias $b_i = (1/w_{\mathrm{EB},i} - 1)(\theta_i - X_i'\delta)/\sigma_i$, and the function $\tilde{r}$ is given by the function $r(b, \chi)$ defined in (4). Our baseline implementation uses constraints on the second and fourth moments, $g(a_i) = (a_i^2, a_i^4)$.

rejected null hypotheses does not converge too quickly to zero. While Storey (2002) assumed that the uncorrected tests control size in the classical sense, the argument goes through essentially unchanged so long as the tests invert CIs that satisfy (16), which holds so long as the CIs do not exhibit too much statistical dependence, as discussed in Remark 4.1. We note, however, that this does not hold for modifications of the Benjamini and Hochberg (1995) procedure that use initial estimates of the proportion of true null hypotheses.

[17]The moment functions g need not be simple moments, and could incorporate constraints used for selection of hyper-parameters, such as constraints on the marginal data distribution or, if an unbiased risk criterion is used, the constraint that the derivative of the risk equals zero at the selected prior hyper-parameters.

EXAMPLE 6.2—Nonlinear soft thresholding: Consider for simplicity the homoscedastic normal model $Y_i|\theta_i \sim N(\theta_i, \sigma^2)$ without covariates. A popular alternative to linear estimators is the soft thresholding estimator $\hat{\theta}_{\mathrm{ST},i} = \operatorname{sign}(Y_i)\max\{|Y_i| - \sqrt{2\sigma^2/\mu_2}, 0\}$ (e.g., Abadie and Kasy (2019)). It equals the posterior mode corresponding to a baseline Laplace prior with second moment μ_2, which has density $\pi_0(\theta) = \frac{1}{\sqrt{2\mu_2}}\exp(-|\theta|\sqrt{2/\mu_2})$ (Johnstone (2019, Example 2.5)). To construct a robust EBCI that always contains the soft thresholding estimator, we calibrate the corresponding highest posterior density set:

$$\mathcal{S}(Y_i;\chi) = \left\{t \in \mathbb{R}\colon \log \frac{\sigma^{-1}\phi\big((Y_i - t)/\sigma\big)\pi_0(t)}{\int_{-\infty}^{\infty} \sigma^{-1}\phi\big((Y_i - \tilde{\theta})/\sigma\big)\pi_0(\tilde{\theta})\, d\tilde{\theta}} + \chi \geq 0\right\}, \tag{21}$$

where ϕ is the standard normal density. This set is available in closed form and takes the form of an interval (see Supplemental Material Appendix G.1). Here $a_i = \theta_i$, and the function $\tilde{r}(a, \chi)$ in (18) can be computed via numerical integration.

In contrast to the EBCIs in Example 6.1 (which may be viewed as calibrating the highest posterior density set under a *normal* prior), the Laplace prior π_0 leads to nonlinear shrinkage and an EBCI whose length depends on the data Y_i. This reflects the suboptimality of linear shrinkage and fixed-length intervals under the Laplace prior.

In Supplemental Material Appendix G.1, we show that the resulting robust EBCI that imposes the constraint $E[\theta_i^2] = \mu_2$ not only has robust EB coverage (by definition), it also achieves substantial expected length improvements when the θ_i's are in fact Laplace distributed. For $\alpha = 0.05$ and $\mu_2/\sigma^2 \leq 0.2$, the expected length under the Laplace distribution of the soft thresholding EBCI is at least 49% smaller than the length of the unshrunk CI. This exceeds the length reduction achieved by the linear robust EBCI shown in Figure 3.

EXAMPLE 6.3—Poisson shrinkage: Supplemental Material Appendix G.2 constructs a robust EBCI for the rate parameter θ_i in a Poisson model $Y_i|\theta_i \sim \mathrm{Poisson}(\theta_i)$. This example demonstrates that our general approach does not require normality of the data.

EXAMPLE 6.4—Linear estimators in other settings: While our focus has been on EB shrinkage, our approach applies to other settings in which an estimator $\hat{\theta}_i$ is approximately normally distributed with non-negligible bias. In particular, suppose $(\hat{\theta}_i - \theta_i)/\mathrm{se}_i$ is distributed $N(a_i, 1)$, where se_i is the standard deviation of the estimate $\hat{\theta}_i$, which for simplicity we take to be known. This holds whenever $\hat{\theta}_i$ is a linear function of jointly normal observations $W_1, \ldots, W_N$, that is, $\hat{\theta}_i = \sum_{j=1}^N k_{ij} W_j$ for some deterministic weights k_{ij}. Examples include series, kernel, or local polynomial estimators in a nonparametric regression with fixed covariates and normal errors. We can construct a confidence interval for θ_i as $\hat{\theta}_i \pm \chi \cdot \mathrm{se}_i$, in which case Eq. (18) holds with $\tilde{r} = r$ given in Eq. (4). It follows from Theorem C.1 in Appendix C that if the moment constraints m on the normalized bias in Eq. (19) are replaced by consistent estimates, the resulting robust EBCI will satisfy the average coverage property (15) in large samples. We leave a full treatment of these applications for future research.

6.2. *Coverage After Selection*

In some applications, researchers may be primarily interested in parameters corresponding to those units i whose initial estimates Y_i fall in a given interval $[\iota_1, \iota_2]$, where

$-\infty \le \iota_1 < \iota_2 \le \infty$. For example, in a teacher value added application, we may only be interested in the ability θ_i of those teachers i whose fixed effects estimates Y_i are positive, corresponding to setting $\iota_1 = 0$ and $\iota_2 = \infty$. Because of the selection on outcomes, naïvely applying our baseline EBCI procedure to the selected sample $\{i\colon Y_i \in [\iota_1, \iota_2]\}$ does not yield the desired average coverage across the selected units i. We now show how to correct for the selection bias in the simple homoscedastic model $Y_i|\theta_i \sim N(\theta_i, \sigma^2)$ without covariates from Section 2 (reintroducing the extra model features in Section 3.1 only complicates notation).

We seek a critical value χ such that the average coverage of the CI $[\hat{\theta}_i \pm \chi w_{\text{EB}}\sigma]$ is at least $1-\alpha$ *conditional* on the sample selection, that is,

$$P\big(\theta_i \in \hat{\theta}_i \pm \chi w_{\text{EB}}\sigma \,|\, Y_i \in [\iota_1, \iota_2]\big) \ge 1-\alpha \tag{22}$$

under repeated sampling of (Y_i, θ_i), regardless of the distribution for θ_i (we maintain focus on linear shrinkage for simplicity, but our approach extends to nonlinear shrinkage using the ideas in Section 6.1). Straightforward calculations show that the non-coverage, conditional on θ_i and on selection, equals

$$\begin{aligned}
&\tilde{r}_{\iota_1,\iota_2}(\theta_i, \chi) \\
&\quad = P\big(\theta_i \notin \hat{\theta}_i \pm \chi w_{\text{EB}}\sigma \,|\, Y_i \in [\iota_1, \iota_2], \theta_i\big) \\
&\quad = \min\left\{1 - \frac{\Phi\big(\min\{\chi - b_i, (\iota_2 - \theta_i)/\sigma\}\big) - \Phi\big(\max\{-\chi - b_i, (\iota_1 - \theta_i)/\sigma\}\big)}{\Phi\big((\iota_2 - \theta_i)/\sigma\big) - \Phi\big((\iota_1 - \theta_i)/\sigma\big)}, 1\right\},
\end{aligned}$$

where $b_i = (1 - 1/w_{\text{EB}})\theta_i/\sigma$ as in Section 2. Among all distributions for θ_i consistent with the conditional moment $\tilde{\mu}_{2,\iota_1,\iota_2} = E[\theta_i^2 | Y_i \in [\iota_1, \iota_2]]$, the worst-case non-coverage probability, conditional on selection, is given by

$$\tilde{\rho}_{\iota_1,\iota_2}(\tilde{\mu}_{2,\iota_1,\iota_2}, \chi) \equiv \sup_F E_F\big[\tilde{r}_{\iota_1,\iota_2}(\theta_i, \chi)\big] \quad \text{s.t. } E_F\big[\theta_i^2\big] = \tilde{\mu}_{2,\iota_1,\iota_2},$$

where E_F denotes expectation under $\theta_i \sim F$. This is an infinite-dimensional linear program that can be solved numerically to a high degree of accuracy; cf. Appendix B. To achieve robust conditional coverage, we solve numerically for the χ such that $\tilde{\rho}_{\iota_1,\iota_2}(\tilde{\mu}_{2,\iota_1,\iota_2}, \chi) = \alpha$.

We can estimate the conditional second moment $\tilde{\mu}_{2,\iota_1,\iota_2}$ as follows. Denote the log marginal density of Y_i by $\ell(y) \equiv \log \int \phi(y-\theta)\, d\Gamma_0(\theta)$, where Γ_0 is the true distribution of θ_i. Tweedie's formulas (e.g., Efron (2019, Eq. (26))) imply

$$\tilde{\mu}_{2,\iota_1,\iota_2} = E\big[\theta_i^2 | Y_i \in [\iota_1, \iota_2]\big] = 1 + E\big[\big(Y_i + \ell'(Y_i)\big)^2 + \ell''(Y_i) \,|\, Y_i \in [\iota_1, \iota_2]\big]. \tag{23}$$

Let $\hat{\ell}(y)$ be a kernel estimate of the log marginal density function of the data $Y_1, \dots, Y_n$. Then the estimate

$$\widehat{\tilde{\mu}}_{2,\iota_1,\iota_2} \equiv 1 + \frac{\sum_{i\colon Y_i \in [\iota_1,\iota_2]} \big\{\big(Y_i + \hat{\ell}'(Y_i)\big)^2 + \hat{\ell}''(Y_i)\big\}}{\#\big\{i\colon Y_i \in [\iota_1, \iota_2]\big\}}$$

will be consistent as $n \to \infty$ for $\tilde{\mu}_{2,\iota_1,\iota_2}$ in (23) under mild regularity conditions.

The criterion (22) can be viewed as the EB analogue of the criterion $P(\theta_i \in CI_i | Y_i \in [\iota_1, \iota_2], \theta) \geq 1 - \alpha$, which requires frequentist coverage conditional on the event $\{Y_i \in [\iota_1, \iota_2]\}$. The latter criterion has been considered in the recent "selective inference" literature (Benjamini and Yekutieli (2005), Lee, Sun, Sun, and Taylor (2016), Hung and Fithian (2019), Andrews, Kitagawa, and McCloskey (2021)). In contrast to this literature, we cannot allow ι_1 to be given by the maximum of the initial estimates (as in Andrews, Kitagawa, and McCloskey (2021)), as we require ι_1 and ι_2 to converge in probability to distinct nonrandom limits. On the other hand, weakening the notion of frequentist coverage to EB (or average) coverage allows for improvements in the length of the intervals, similar to the analysis in Section 4.2 in the absence of selection.

7. EMPIRICAL APPLICATION

We illustrate our methods using the data and model in Chetty and Hendren (2018), who were interested in the effect of neighborhoods on intergenerational mobility.

7.1. *Framework*

We follow Chetty and Hendren (2018) in using two definitions of a "neighborhood effect" θ_i. The first focuses on effects for children growing up in low-income families, and defines θ_i as the effect of spending an additional year of childhood in commuting zone (CZ) i on children's rank in the income distribution at age 26, for children with parents at the 25th percentile of the national income distribution. The second definition is analogous, except it focuses on children growing up in high-income families, and consequently conditions on children with parents at the 75th percentile. Chetty and Hendren (2018) argued that these definitions approximately capture the mean rank effects for children in below-median and above-median income families. Using de-identified tax returns for all children born between 1980 and 1986 who move across CZs exactly once as children, Chetty and Hendren (2018) exploited variation in the age at which children move between CZs to obtain preliminary fixed effects estimates Y_i of θ_i.

Since these preliminary estimates are measured with noise, to predict θ_i, Chetty and Hendren (2018) shrunk Y_i toward average outcomes of permanent residents of CZ i (children with parents at the same percentile of the income distribution who spent all of their childhood in the CZ). To give a sense of the accuracy of these forecasts, Chetty and Hendren (2018) reported estimates of their unconditional MSE (i.e., treating θ_i as random), under the implicit assumption that the moment independence assumption in Eq. (10) holds. Here we complement their analysis by constructing robust EBCIs associated with these forecasts.

Our sample consists of 595 U.S. CZs, with population over 25,000 in the 2000 census: this is the sample for which Chetty and Hendren (2018) reported baseline estimates Y_i of the effects θ_i. These baseline estimates are normalized so that their population-weighted mean is zero. We may therefore interpret θ_i as the effect relative to an "average" CZ. We follow the baseline implementation from Section 3.2 with standard errors $\hat{\sigma}_i$ reported by Chetty and Hendren (2018), and covariates X_i corresponding to a constant and the average outcomes for permanent residents. In line with the original analysis, we use precision weights $\omega_i = 1/\hat{\sigma}_i^2$ when constructing the estimates $\hat{\delta}$, $\hat{\mu}_2$, and $\hat{\kappa}$.

7.2. *Results*

Columns (1) and (2) in Table III summarize the main estimation and efficiency results. The shrinkage magnitude and relative efficiency results are similar for children with

TABLE III

STATISTICS FOR 90% EBCIS FOR NEIGHBORHOOD EFFECTS.

	Baseline		Nonparametric	
Percentile	(1) 25th	(2) 75th	(3) 25th	(4) 75th
Panel A: Summary statistics				
$E[\sqrt{\mu_{2,i}}]$	0.079	0.044	0.076	0.042
$E[\kappa_i]$	778.5	5948.6	1624.9	43,009.9
$E[\mu_{2i}/\sigma_i^2]$	0.142	0.040	0.139	0.072
$\hat{\delta}_{\text{intercept}}$	−1.441	−2.162	−1.441	−2.162
$\hat{\delta}_{\text{perm. resident}}$	0.032	0.038	0.032	0.038
$E[w_{\text{EB},i}]$	0.093	0.033	0.093	0.033
$E[w_{\text{opt},i}]$	0.191	0.100	0.191	0.100
E[non-cov of parametric EBCI$_i$]	0.227	0.278	0.210	0.292
Panel B: E[half-length$_i$]				
Robust EBCI	0.195	0.122	0.186	0.116
Optimal robust EBCI	0.149	0.090	0.145	0.094
Parametric EBCI	0.123	0.070	0.123	0.070
Unshrunk CI	0.786	0.993	0.786	0.993
Panel C: Efficiency relative to robust EBCI				
Optimal robust EBCI	1.312	1.352	1.289	1.238
Parametric EBCI	1.582	1.731	1.509	1.648
Unshrunk CI	0.248	0.123	0.237	0.117

Note: Columns (1) and (2) correspond to shrinking Y_i as in the baseline implementation that imposes Eq. (10), so that $\mu_{2i} = E[(\theta_i - X_i'\delta)^2|X_i, \sigma_i]$ and $\kappa_i = E[(\theta_i - X_i'\delta)^4|X_i, \sigma_i]/\mu_{2i}^2$ do not vary with i. Columns (3) and (4) use nonparametric estimates of μ_{2i} and κ_i, using the nearest neighbor estimator described in Appendix A.1. The number of nearest neighbors $J = 422$ (column (3)) and $J = 525$ (column (4)) is selected using cross-validation. For all columns, $\hat{\delta} = (\hat{\delta}_{\text{intercept}}, \hat{\delta}_{\text{perm. resident}})$ is computed by regressing Y_i onto a constant and outcomes for permanent residents. "Optimal Robust EBCI" refers to a robust EBCI based on length-optimal shrinkage $w_{\text{opt},i}$, described in Section 4.2. "E[non-cov of parametric EBCI$_i$]": average of maximal non-coverage probability of parametric EBCI, given the estimated moments.

parents at the 25th and 75th percentiles of the income distribution. In both columns, the estimate of the kurtosis κ is large enough so that it does not affect the critical values or the form of the optimal shrinkage: specifications that only impose constraints on the second moment yield identical results.[18] In line with this finding, a density plot of the t-statistics (reported as Figure S2 in Armstrong, Kolesár, and Plagborg-Møller (2020)) exhibits a fat lower tail. As a robustness check, columns (3) and (4) show that results based on nonparametric moment estimates (see Remark 3.2 and Appendix A.1) are very similar to our baseline specification. Indeed, the R^2 gain in predicting $\hat{\varepsilon}_i^2 - \hat{\sigma}_i^2$ using $\hat{\mu}_{2i}$ is less than 0.001 in both specifications, indicating that there is little evidence in the data against the moment independence assumption.

The baseline robust 90% EBCIs are 75.2–87.7% shorter than the usual unshrunk CIs $Y_i \pm z_{1-\alpha/2}\hat{\sigma}_i$. To interpret these gains in dollar terms, for children with parents at the 25th percentile of the income distribution, a percentile gain corresponds to an annual income gain of \$818 (Chetty and Hendren (2018, p. 1183)). Thus, the average half-length of the baseline robust EBCIs in column (1) implies CIs of the form ±\$160 on average, while the

[18]The truncation in the $\hat{\kappa}$ formula in our baseline algorithm in Section 3.2 binds in columns (1) and (2), although the non-truncated estimates 345.3 and 5024.9 are similarly large; using these non-truncated estimates yields identical results.

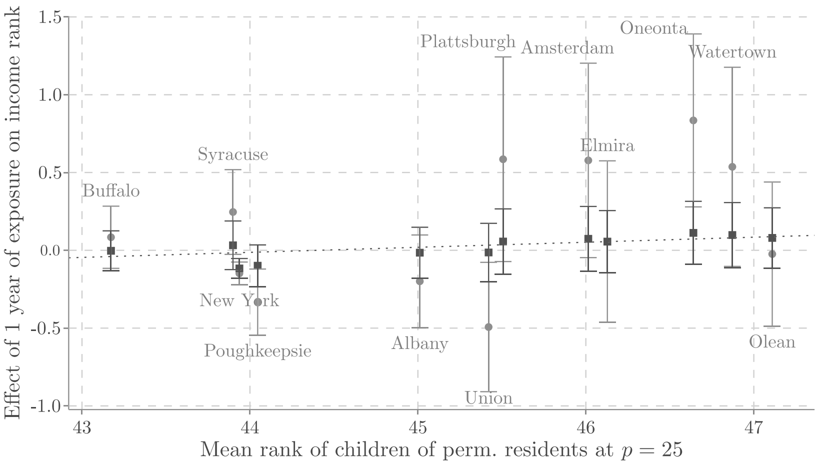

FIGURE 6.—Neighborhood effects for New York and 90% robust EBCIs for children with parents at the $p = 25$ percentile of the national income distribution, plotted against mean outcomes of permanent residents. Gray lines correspond to CIs based on unshrunk estimates represented by circles, and black lines correspond to robust EBCIs based on EB estimates represented by squares that shrink toward a dotted regression line based on permanent residents' outcomes. Baseline implementation as in Section 3.2.

unshrunk CIs are of the form ±\$643 on average. These large gains are a consequence of a low signal-to-noise ratio μ_2/σ_i^2 in this application. Because the shrinkage magnitude is so large on average, the tail behavior of the bias matters, and since the kurtosis estimates suggest these tails are fat, it is important to use the robust critical value: the parametric EBCI exhibits average potential size distortions of 12.7–17.8 percentage points. Indeed, for over 90% of the CIs in the specifications in columns (1) and (2), the shrinkage coefficient $w_{\text{EB},i}$ falls below the "rule of thumb" threshold of 0.3 derived in Section 4.3.

To visualize these results, Figure 6 plots the unshrunk 90% CIs based on the preliminary estimates, as well as robust EBCIs based on EB estimates for cities in the state of New York for children with parents at the 25th percentile. While the EBCIs for large CZs like New York City or Buffalo are similar to the unshrunk CIs, they are much tighter for smaller CZs like Plattsburgh or Watertown, with point estimates that shrink the preliminary estimates Y_i most of the way toward the regression line $X_i'\hat{\delta}$.

In summary, shrinkage allows us to considerably tighten the CIs based on preliminary estimates. This is true even though the CIs effectively only use second moment constraints—imposing kurtosis constraints does not affect the critical values in this application.

APPENDIX A: MOMENT ESTIMATES

The EBCI in our baseline implementation has valid EB coverage asymptotically as $n \to \infty$, so long as the estimates $\hat{\mu}_2$ and $\hat{\kappa}$ are consistent. While the particular choice of the estimates $\hat{\mu}_2$ and $\hat{\kappa}$ does not affect the CI asymptotically, finite-sample considerations can be important for small to moderate values of n. In particular, unrestricted moment-based estimates of μ_2 and κ may fall below their theoretical lower bounds of 0 and 1, in which case it is not clear how to define the EBCI.[19] To address this issue, in analogy to finite-

[19]Formally, our results are asymptotic and require $\mu_2 > 0$ and $\kappa > 1$, so that these issues do not occur when n is large enough. We discuss the difficulty of providing finite-sample coverage guarantees in Section 5.

sample corrections to parametric EBCIs proposed in Morris (1983a,b), Appendix A.1 derives two finite-sample corrections to the unrestricted estimates that approximate a Bayesian estimate under a flat hyperprior on (μ_2, κ). We verify that these corrections give good coverage in an extensive set of Monte Carlo designs in Section 4.4. We also discuss implementation of nonparametric moment estimates. Appendix A.2 discusses the choice of weights ω_i.

A.1. *Finite n Corrections and Nonparametric Moment Estimates*

To derive our estimates of μ_2 and κ, we first consider unrestricted estimation under the moment independence condition (10). For μ_2, this condition implies the moment condition $E[(Y_i - X_i'\delta)^2 - \sigma_i^2 | X_i, \sigma_i] = \mu_2$. Replacing $Y_i - X_i'\delta$ with the residual $\hat{\varepsilon}_i = Y_i - X_i'\hat{\delta}$ yields the estimate

$$\hat{\mu}_{2,\mathrm{UC}} = \frac{\sum_{i=1}^{n} \omega_i \mathcal{W}_{2i}}{\sum_{i=1}^{n} \omega_i}, \quad \mathcal{W}_{2i} = \hat{\varepsilon}_i^2 - \hat{\sigma}_i^2, \tag{24}$$

for any weights $\omega_i = \omega_i(X_i, \hat{\sigma}_i)$. Here, UC stands for "unconstrained," since the estimate $\hat{\mu}_{2,\mathrm{UC}}$ can be negative. To incorporate the constraint $\mu_2 > 0$, we use an approximation to a Bayesian approach with a flat prior on the set $[0, \infty)$. A full Bayesian approach to estimating μ_2 would place a hyperprior on possible joint distributions of X_i, σ_i, θ_i, which could potentially lead to using complicated functions of the data to estimate μ_2. For simplicity, we compute the posterior mean given $\hat{\mu}_{2,\mathrm{UC}}$, and we use a normal approximation to the likelihood. Since the posterior distribution only uses knowledge of $\hat{\mu}_{2,\mathrm{UC}}$, we refer to this as a flat prior limited information Bayes (FPLIB) approach.

To derive this formula, first note that, if $\hat{m}$ is an estimate of a parameter m with $\hat{m}|m \sim N(m, V)$, then under a flat prior for m on $[0, \infty)$, the posterior mean of m is given by

$$b(\hat{m}, V) = \hat{m} + \sqrt{V}\phi(\hat{m}/\sqrt{V})/\Phi(\hat{m}/\sqrt{V}),$$

where ϕ and Φ are the standard normal pdf and cdf, respectively. Furthermore, if $\hat{m} = \sum_{i=1}^n \omega_i Z_i / \sum_{i=1}^n \omega_i$, where the Z_i's are independent with mean m conditional on the weights $\omega = (\omega_1, \dots, \omega_n)'$, then an unbiased estimate of the variance of $\hat{m}$ given ω is given by

$$V(Z, \omega) = \frac{\sum_{i=1}^{n} \omega_i^2 (Z_i^2 - \hat{m}^2)}{\left(\sum_{i=1}^{n} \omega_i\right)^2 - \sum_{i=1}^{n} \omega_i^2}.$$

Conditioning on the X_i's and σ_i's (and ignoring sampling variation in $\hat{\delta}$ and the $\hat{\sigma}_i$'s), we can then apply this formula to $\hat{\mu}_{2,\mathrm{UC}}$, with $Z_i = \mathcal{W}_{2i}$, where $\mathcal{W}_{2i}$ is given in (24). This gives the FPLIB estimate for μ_2:

$$\hat{\mu}_{2,\mathrm{FPLIB}} = b\big(\hat{\mu}_{2,\mathrm{UC}}, V(\mathcal{W}_2, \omega)\big).$$

To derive the FPLIB estimate for κ, we begin with an unconstrained estimate of $\mu_4 = E[(\theta_i - X_i'\delta)^4]$. The moment independence condition (10) delivers the moment condition $\mu_4 = E[(Y_i - X_i'\delta)^4 + 3\sigma_i^4 - 6\sigma_i^2(Y_i - X_i'\delta)^2 | X_i, \sigma_i]$, which leads to the unconstrained estimate

$$\hat{\mu}_{4,\mathrm{UC}} = \frac{\sum_{i=1}^{n} \omega_i \mathcal{W}_{4i}}{\sum_{i=1}^{n} \omega_i}, \quad \mathcal{W}_{4i} = \hat{\varepsilon}_i^4 - 6\hat{\sigma}_i^2\hat{\varepsilon}_i^2 + 3\hat{\sigma}_i^4.$$

To avoid issues with small values of estimates of μ_2 in the denominator, we apply the FPLIB approach to an estimate of $\mu_4 - \mu_2^2$, using a flat prior on the parameter space $[0, \infty)$. Using the delta method leads to approximating the variance of $\hat{\mu}_{4,\mathrm{UC}} - \hat{\mu}_{2,\mathrm{UC}}^2$ with the variance of $\sum_{i=1}^n \omega_i(\mathcal{W}_{4i} - 2\mu_2\mathcal{W}_{2i}) / \sum_{i=1}^n \omega_i$, so that the FPLIB estimate of $\mu_4 - \mu_2^2$ is $b(\hat{\mu}_{4,\mathrm{UC}} - \hat{\mu}_{2,\mathrm{UC}}^2, V(\mathcal{W}_4 - 2\hat{\mu}_{2,\mathrm{FPLIB}}\mathcal{W}_2, \omega))$, and the FPLIB estimate of κ is

$$\hat{\kappa}_{\mathrm{FPLIB}} = 1 + \frac{b\big(\hat{\mu}_{4,\mathrm{UC}} - \hat{\mu}_{2,\mathrm{UC}}^2, V(\mathcal{W}_4 - 2\hat{\mu}_{2,\mathrm{FPLIB}}\mathcal{W}_2, \omega)\big)}{\hat{\mu}_{2,\mathrm{FPLIB}}^2}.$$

As a further simplification, we derive approximations in which the posterior mean formula $b(\hat{m}, V)$ is replaced by a simple truncation formula. We refer to this approach as posterior mean trimming (PMT). In particular, suppose we apply the formula $b(\hat{m}, V)$ to an estimator $\hat{m}$ such that $\hat{m} \geq m_0$ and $V \geq V_0$ by construction, where $m_0 < 0$. Then the posterior mean satisfies $b(\hat{m}, V) \geq b(m_0, V_0)$ (Pinelis (2002, Proposition 1.2)). Thus, a simple approximation to the FPLIB estimator is to truncate $\hat{m}$ from below at $b(m_0, V_0)$. To obtain an even simpler formula, we use the approximation $b(m_0, V_0) = -V_0/m_0 + O(V_0^{3/2})$ (Pinelis (2002, Proposition 1.3)), which holds as $V_0 \to 0$ (or, equivalently, as $n \to \infty$, provided the estimator $\hat{m}$ is consistent). The variance of $\hat{\mu}_{2,\mathrm{UC}}$ conditional on (X_i, σ_i) is bounded below by $2\sum_{i=1}^n \omega_i^2\sigma_i^4 / (\sum_{i=1}^n \omega_i)^2$, and $\hat{\mu}_{2,\mathrm{UC}} \geq -\sum_{i=1}^n \omega_i\sigma_i^2 / \sum_{i=1}^n \omega_i$, so we can use $V_0/m_0 = -\frac{2\sum_{i=1}^n \omega_i^2\sigma_i^4}{\sum_{i=1}^n \omega_i\sigma_i^2 \cdot \sum_{i=1}^n \omega_i}$, which gives the PMT estimator

$$\hat{\mu}_{2,\mathrm{PMT}} = \max\left\{\hat{\mu}_{2,\mathrm{UC}}, \frac{2\sum_{i=1}^{n} \omega_i^2\sigma_i^4}{\sum_{i=1}^{n} \omega_i\sigma_i^2 \cdot \sum_{i=1}^{n} \omega_i}\right\}.$$

For κ, we simplify our approach to deriving a trimming rule by treating μ_2 as known, and considering the variance of the infeasible estimate $\hat{\kappa}_{\mathrm{UC}}^* = \frac{\sum_{i=1}^n \omega_i(\hat{\varepsilon}_i^4 - 6\hat{\sigma}_i^2\mu_2 - 3\hat{\sigma}_i^4)}{\mu_2^2 \sum_{i=1}^n \omega_i}$. Using the above truncation formula for $\hat{\kappa}_{\mathrm{UC}}^* - 1$ along with the fact that $\hat{\kappa}_{\mathrm{UC}}^* \geq \frac{\sum_{i=1}^n \omega_i(-6\hat{\sigma}_i^2\mu_2 - 3\hat{\sigma}_i^4)}{\mu_2^2 \sum_{i=1}^n \omega_i}$ and the lower bound $8\sum_i \omega_i^2(2\mu_2^3\sigma_i^2 + 21\mu_2^2\sigma_i^4 + 48\mu_2\sigma_i^6 + 12\sigma_i^8)/\mu_2^4(\sum_i \omega_i)^2$ on the variance yields $V_0/m_0 = -\frac{8\sum_i \omega_i^2(2\mu_2^3\sigma_i^2 + 21\mu_2^2\sigma_i^4 + 48\mu_2\sigma_i^6 + 12\sigma_i^8)}{\mu_2^2(\sum_i \omega_i)\sum_{i=1}^n \omega_i(\mu_2^2 + 6\hat{\sigma}_i^2\mu_2 + 3\hat{\sigma}_i^4)}$. To simplify the trimming rule even further, we only use the leading term of V_0/m_0 as $\mu_2 \to 0$, $V_0/m_0 = -\frac{32\sum_i \omega_i^2\sigma_i^8}{\mu_2^2(\sum_i \omega_i)\sum_{i=1}^n \omega_i\hat{\sigma}_i^4} + o(1/\mu_2^2)$.

Plugging in $\hat{\mu}_{2,\text{PMT}}$ in place of the unknown μ_2 then gives the PMT estimator

$$\hat{\kappa}_{\text{PMT}} = \max\left\{\frac{\hat{\mu}_{4,\text{UC}}}{\hat{\mu}_{2,\text{PMT}}^2}, 1 + \frac{32\sum_{i=1}^{n}\omega_i^2\hat{\sigma}_i^8}{\hat{\mu}_{2,\text{PMT}}^2\sum_{i=1}^{n}\omega_i \cdot \sum_{i=1}^{n}\omega_i\hat{\sigma}_i^4}\right\}.$$

The estimators in step 1 of our baseline implementation in Section 3.2 correspond to $\hat{\mu}_{2,\text{PMT}}$ and $\hat{\kappa}_{\text{PMT}}$, due to their slightly simpler form relative to the FPLIB estimators. In unreported simulations based on the designs described in Section 4.4 and Supplemental Material Appendix E.2, we find that EBCIs based on FPLIB lead to even smaller finite-sample coverage distortions than those based on the baseline implementation that uses PMT, at the expense of slightly longer average length.

To implement the nonparametric estimates $\hat{\kappa}_i$ and $\hat{\mu}_{2i}$ in Remark 3.2, we use the nearest-neighbor estimator that, for each i, computes the PMT estimates $\hat{\mu}_{2,\text{PMT}}$ and $\hat{\kappa}_{\text{PMT}}$ described above, using only the J observations closest to i, rather than the full sample of n observations. We define distance as a Euclidean distance on (X_i, σ_i), after scaling elements of this vector by their standard deviations. Under regularity conditions, the resulting estimates will be consistent for μ_{2i} and κ_i, so long as $J \to \infty$ and $J/n \to 0$. We select J using leave-one-out cross-validation, using the squared prediction error in predicting $\mathcal{W}_{2i}$ as the criterion. For simplicity, we use the same J for estimating the kurtosis as that used for estimating the second moment.

A.2. *Choice of Weighting*

Under condition (10), the weights ω_i used to estimate μ_2 and κ can be any function of X_i, σ_i. Furthermore, while $\hat{\delta}$ can be essentially arbitrary as long as it converges in probability to some δ such that Eq. (10) holds, that equation will often be motivated by the assumption that the conditional mean of θ_i is linear in X_i,

$$E\left[\theta_i - X_i'\delta \mid X_i, \sigma_i\right] = 0. \tag{25}$$

Under this condition, the weights ω_i used to estimate δ can also be any function of X_i, σ_i.

Thus, under conditions (10) and (25), the choice of weighting can be guided by efficiency concerns. In general, the optimal weights are different for each of the three estimates of δ, μ_2, and κ, and implementing them requires first-stage estimates of the variances of Y_i, $\mathcal{W}_{2i}$, and $\mathcal{W}_{4i}$, conditional on (X_i, σ_i) (with $\mathcal{W}_{2i}$ and $\mathcal{W}_{4i}$ defined in Appendix A.1). To avoid estimation of these variances, consider the limiting case where the signal-to-noise ratio goes to 0, $\mu_2/\min_i \sigma_i^2 \to 0$. The resulting weights will be near-optimal under a low signal-to-noise ratio, when precise estimation of these parameters is relatively more important for accurate coverage (under a high signal-to-noise ratio, shrinkage is limited, and estimation error in these parameters has little effect on coverage). Let us also ignore estimation error in δ for simplicity, and suppose that the Y_i's are independent conditional on $(\theta_i, X_i, \sigma_i)$. Then, as $\mu_2/\min_i \sigma_i^2 \to 0$, the weights $\hat{\sigma}_i^{-2}$, $\hat{\sigma}_i^{-4}$, and $\hat{\sigma}_i^{-8}$, for estimating δ, μ_2, and μ_4, respectively, become optimal. For simplicity, the baseline implementation in Section 3.2 uses the same weights ω_i for each of the estimates; the choice $\omega_i = \hat{\sigma}_i^{-2}$ targets optimal estimation of δ. However, one could relax this constraint, and use the weights $\hat{\sigma}_i^{-4}$ and $\hat{\sigma}_i^{-8}$ for estimating μ_2 and μ_4 instead. The choice $\omega_i = 1/n$ has the advantage of simplicity; one may also motivate it by robustness concerns when Eq. (10) fails, though our preferred robustness check is to use nonparametric moment estimates, as outlined in Remark 3.2.

APPENDIX B: COMPUTATIONAL DETAILS

To simplify the statement of the results below, let $r_0(b, \chi) = r(\sqrt{b}, \chi)$, and put $m_2 = \sigma^2/\mu_2$. The next proposition shows that, if only a second moment constraint is imposed, the maximal non-coverage probability $\rho(m_2, \chi)$, defined in Eq. (5), has a simple solution:

PROPOSITION B.1: *Consider the problem in Eq*. (5). *The solution is given by*

$$\rho(m_2, \chi) = \begin{cases} r_0(0, \chi) + \dfrac{m_2}{t_0}\big(r_0(t_0, \chi) - r_0(0, \chi)\big) & \text{if } m_2 < t_0, \\ r_0(m_2, \chi) & \text{otherwise}. \end{cases}$$

Here $t_0 = 0$ *if* $\chi < \sqrt{3}$; *otherwise* t_0 *is the unique solution to* $r_0(t, \chi) + u\frac{\partial}{\partial u} r_0(u, \chi) = r_0(u, \chi)$.

The proof of Proposition B.1 shows that $\rho(m_2, \chi)$ corresponds to the least concave majorant of the function r_0.

The next result shows that, if in addition to a second moment constraint, we impose a constraint on the kurtosis, the maximal non-coverage probability can be computed as a solution to two nested univariate optimizations:

PROPOSITION B.2: *Suppose* $\kappa > 1$ *and* $m_2 > 0$. *Then the solution to the problem*

$$\rho(m_2, \kappa, \chi) = \sup_F E_F\big[r(b, \chi)\big] \quad \text{s.t. } E_F\big[b^2\big] = m_2, E_F\big[b^4\big] = \kappa m_2^2,$$

is given by $\rho(m_2, \kappa, \chi) = r_0(m_2, \chi)$ *if* $m_2 \geq t_0$, *with* t_0 *defined in Proposition* B.1. *If* $m_2 < t_0$, *then the solution is given by*

$$\inf_{0 < x_0 \leq t_0} \Big\{ r_0(x_0, \chi) + (m_2 - x_0) r_0'(x_0, \chi) + \big((x_0 - m_2)^2 + (\kappa - 1)m_2^2\big) \sup_{0 \leq x \leq t_0} \delta(x; x_0) \Big\}, \tag{26}$$

where $r_0'(x_0, \chi) = \partial r_0(x_0, \chi)/\partial x_0$, $\delta(x; x_0) = \frac{r_0(x,\chi) - r_0(x_0,\chi) - (x - x_0) r_0'(x_0,\chi)}{(x - x_0)^2}$ *if* $x \neq x_0$, *and* $\delta(x_0; x_0) = \lim_{x \to x_0} \delta(x; x_0) = \frac{1}{2}\frac{\partial^2}{\partial x_0^2} r_0(x_0, \chi)$.

If $m_2 \geq t_0$, then imposing a constraint on the kurtosis does not help to reduce the maximal non-coverage probability, and $\rho(m_2, \kappa, \chi) = \rho(m_2, \chi)$.

REMARK B.1—Least favorable distributions: It follows from the proof of these propositions that distributions maximizing Eq. (5)—the least favorable distributions for the normalized bias b—have two support points if $m_2 \geq t_0$, namely, $-\sqrt{m_2}$ and $\sqrt{m_2}$ (since the rejection probability $r(b, \chi)$ depends on b only through its absolute value, any distribution with these two support points maximizes Eq. (5)). If $m_2 < t_0$, there are three support points, $b = 0$ with probability $1 - m_2/t_0$ and $b = \pm\sqrt{t_0}$ with total probability m_2/t_0 (again, only the sum of the probabilities is uniquely determined). If the kurtosis constraint is also imposed, then there are four support points, $\pm\sqrt{x_0}$ and $\pm\sqrt{x}$, where x and x_0 optimize Eq. (26).

Finally, the characterization of the solution to the general program in Eq. (19) depends on the form of the constraint function g. To solve the program numerically, discretize the support of F to turn the problem into a finite-dimensional linear program, which can be solved using a standard linear solver. In particular, we solve the problem

$$\rho_g(m,\chi)=\sup_{p_1,\dots,p_K}\sum_{k=1}^{K}p_k r(x_k,\chi)\quad \text{s.t.}\ \sum_{k=1}^{K}p_k g(x_k)=m,\ \sum_{k=1}^{K}p_k=1,\ p_k\geq 0.$$

Here $x_1,\dots,x_K$ denote the support points of b, with p_k denoting the associated probabilities.

APPENDIX C: COVERAGE RESULTS

This appendix provides coverage results that generalize Theorem 4.1. Appendix C.1 introduces the general setup. Appendix C.2 provides results for general shrinkage estimators that satisfy an approximate normality assumption. Appendix C.3 considers a generalization of our baseline specification in the EB setting, and states a generalization of Theorem 4.1.

C.1. *General Setup and Notation*

Let $\hat{\theta}_1,\dots,\hat{\theta}_n$ be estimates of $\theta_1,\dots,\theta_n$, with standard errors $\mathrm{se}_1,\dots,\mathrm{se}_n$. The standard errors may be random variables that depend on the data. We are interested in coverage properties of the intervals $CI_i=\{\hat{\theta}_i\pm\mathrm{se}_i\cdot\chi_i\}$ for some $\chi_1,\dots,\chi_n$, which may be chosen based on the data. In some cases, we will condition on some variable $\tilde{X}_i$ when defining EB coverage or average coverage. Let $\tilde{X}^{(n)}=(\tilde{X}_1,\dots,\tilde{X}_n)'$ and let $\chi^{(n)}=(\chi_1,\dots,\chi_n)'$.

As discussed in Section 4.1, the average coverage criterion does not require thinking of θ as random. To save on notation, we will state most of our average coverage results and conditions in terms of a general sequence of probability measures $\tilde{P}=\tilde{P}^{(n)}$ and triangular arrays θ and $\tilde{X}^{(n)}$. We will use $E_{\tilde{P}}$ to denote expectation under $\tilde{P}$. We can then obtain EB coverage statements by considering a distribution P for the data and θ, $\tilde{X}^{(n)}$ and an additional variable ν such that these conditions hold for the measure $\tilde{P}(\cdot)=P(\cdot|\theta,\nu,\tilde{X}^{(n)})$ for θ, ν, $\tilde{X}^{(n)}$ in a probability 1 set. The variable ν is allowed to depend on n, and can include nuisance parameters as well as additional variables.

It will be useful to formulate a conditional version of the average coverage criterion (15), to complement the conditional version of EB coverage discussed in the main text. Due to discreteness of the empirical measure of the $\tilde{X}_i$'s, we consider coverage conditional on each set in some family $\mathcal{A}$ of sets. To formalize this, let $\mathcal{I}_{\mathcal{X},n}=\{i\in\{1,\dots,n\}\colon \tilde{X}_i\in\mathcal{X}\}$, and let $N_{\mathcal{X},n}=\#\mathcal{I}_{\mathcal{X},n}$. The sample average non-coverage on the set $\mathcal{X}$ is then given by

$$ANC_n\big(\chi^{(n)};\mathcal{X}\big)=\frac{1}{N_{\mathcal{X},n}}\sum_{i\in\mathcal{I}_{\mathcal{X},n}}\mathrm{I}\big\{\theta_i\notin\{\hat{\theta}\pm\mathrm{se}_i\cdot\chi_i\}\big\}=\frac{1}{N_{\mathcal{X},n}}\sum_{i\in\mathcal{I}_{\mathcal{X},n}}\mathrm{I}\big\{|Z_i|>\chi_i\big\},$$

where $Z_i=(\hat{\theta}_i-\theta_i)/\mathrm{se}_i$. We consider two notions of average coverage control, conditional on the set $\mathcal{X}\in\mathcal{A}$:

$$ANC_n(\chi;\mathcal{X})\leq\alpha+o_{\tilde{P}}(1), \tag{27}$$

and

$$\limsup_n E_{\tilde{P}}\big[ANC_n(\chi; \mathcal{X})\big] = \limsup_n \frac{1}{N_{\mathcal{X},n}} \sum_{i \in \mathcal{I}_{\mathcal{X},n}} \tilde{P}\big(|Z_i| > \chi_i\big) \le \alpha. \tag{28}$$

Since $ANC_n(\chi; \mathcal{X})$ is uniformly bounded, (27) implies (28). Furthermore, if we integrate with respect to some distribution on ν, $\tilde{X}^{(n)}$ such that (28) holds with $\tilde{P}(\cdot) = P(\cdot|\theta, \nu, \tilde{X}^{(n)})$ almost surely, we get (again by uniform boundedness) $\limsup_n E[ANC_n(\chi; \mathcal{X})|\theta] \le \alpha$, which, if $\mathcal{X}$ contains all $\tilde{X}_i$'s with probability 1, is condition (15) from the main text.

Now consider EB coverage, as defined in Eq. (14) in the main text, but conditioning on $\tilde{X}_i$. We consider EB coverage under a distribution P for the data, $\tilde{X}^{(n)}$, θ, and ν, where ν includes additional nuisance parameters and covariates, and where the average coverage condition (28) holds with $P(\cdot|\theta, \nu, \tilde{X}^{(n)})$ playing the role of $\tilde{P}$ with probability 1. Suppose $\tilde{X}_i$ is discretely distributed under P, and that the exchangeability condition

$$P(\theta_i \in CI_i|\mathcal{I}_{\{\tilde{x}\},n}) = P(\theta_j \in CI_j|\mathcal{I}_{\{\tilde{x}\},n}) \quad \text{for all } i, j \in \mathcal{I}_{\{\tilde{x}\},n} \tag{29}$$

holds with probability 1. Then, for each j,

$$\begin{aligned} &P(\theta_j \in CI_j|\tilde{X}_j = \tilde{x}) \\ &\quad = P(\theta_j \in CI_j|j \in \mathcal{I}_{\{\tilde{x}\},n}) = E\big[P(\theta_j \in CI_j|\mathcal{I}_{\{\tilde{x}\},n})|j \in \mathcal{I}_{\{\tilde{x}\},n}\big] \\ &\quad = E\left[\frac{1}{\mathcal{N}_{\{\tilde{x}\},n}} \sum_{i \in \mathcal{I}_{\{\tilde{x}\}}} P(\theta_i \in CI_i|\mathcal{I}_{\{\tilde{x}\}}) \,|\, j \in \mathcal{I}_{\{\tilde{x}\},n}\right]. \end{aligned}$$

Plugging in $P(\cdot|\theta, \nu, \tilde{X}^{(n)})$ for $\tilde{P}$ in the coverage condition (28), taking the expectation conditional on $\mathcal{I}_{\{\tilde{x}\},n}$, and using uniform boundedness, it follows that the lim inf of the term in the conditional expectation is no less than $1 - \alpha$. Then, by uniform boundedness of this term,

$$\liminf_{n \to \infty} P(\theta_j \in CI_j|\tilde{X}_j = \tilde{x}) \ge 1 - \alpha. \tag{30}$$

This is a conditional version of the EB coverage condition (14) from the main text.

C.2. *Results for General Shrinkage Estimators*

We assume that $Z_i = (\hat{\theta}_i - \theta_i)/\text{se}_i$ is approximately normal with variance 1 and mean b_i under the sequence of probability measures $\tilde{P} = \tilde{P}^{(n)}$. To formalize this, we consider a triangular array of distributions satisfying the following conditions.

ASSUMPTION C.1: *For some random variables $\tilde{b}_i$ and constants $b_{i,n}$, $Z_i - \tilde{b}_i$ satisfies*

$$\lim_{n \to \infty} \max_{1 \le i \le n} \big|\tilde{P}(Z_i - \tilde{b}_i \le t) - \Phi(t)\big| = 0$$

for all $t \in \mathbb{R}$ and, for all $\mathcal{X} \in \mathcal{A}$ and any $\varepsilon > 0$, $\frac{1}{N_{\mathcal{X},n}} \sum_{i \in \mathcal{I}_{\mathcal{X},n}} \tilde{P}(|\tilde{b}_i - b_{i,n}| \ge \varepsilon) \to 0$.

Note that, when applying the results with $\tilde{P}(\cdot)$ given by the sequence of measures $P(\cdot|\theta, \nu, \tilde{X}^{(n)})$, the constants $b_{i,n}$ will be allowed to depend on θ, ν, $\tilde{X}^{(n)}$.

Let $g\colon \mathbb{R} \to \mathbb{R}^p$ be a vector of moment functions. We consider critical values $\hat{\chi}^{(n)} = (\hat{\chi}_1, \dots, \hat{\chi}_n)$ based on an estimate of the conditional expectation of $g(b_{i,n})$ given $\tilde{X}_i$, where the expectation is taken with respect to the empirical distribution of $\tilde{X}_i$, $b_{i,n}$. Due to the discreteness of this measure, we consider the behavior of this estimate on average over sets $\mathcal{X} \in \mathcal{A}$. We assume that there exists a function $m : \mathcal{X} \to \mathbb{R}^p$ that plays the role of the conditional expectation of $g(b_{i,n})$ given $\tilde{X}_i$, along with estimates $\hat{m}_i$ of $m(\tilde{X}_i)$, which satisfy the following assumptions.

ASSUMPTION C.2: *For all $\mathcal{X} \in \mathcal{A}$, $N_{\mathcal{X},n} \to \infty$, $\frac{1}{N_{\mathcal{X},n}} \sum_{i \in \mathcal{I}_{\mathcal{X},n}} (g(b_{n,i}) - m(\tilde{X}_i)) \to 0$, and, for all $\varepsilon > 0$, $\frac{1}{N_{\mathcal{X},n}} \sum_{i \in \mathcal{I}_{\mathcal{X},n}} \tilde{P}(\|\hat{m}_i - m(\tilde{X}_i)\| \geq \varepsilon) \to 0$.*

ASSUMPTION C.3: *For every $\mathcal{X} \in \mathcal{A}$ and every $\varepsilon > 0$, there is a partition $\mathcal{X}_1, \dots, \mathcal{X}_J \in \mathcal{A}$ of $\mathcal{X}$ and $m_1, \dots, m_J$ such that, for each j and all $x \in \mathcal{X}_j$, $m(x) \in B_\varepsilon(m_j)$, where $B_\varepsilon(m) = \{\tilde{m} : \|\tilde{m} - m\| \leq \varepsilon\}$.*

ASSUMPTION C.4: *For some compact set M in the interior of the set of values of $\int g(b)dF(b)$ where F ranges over all probability measures on $\mathbb{R}$, we have $m(x) \in M$ for all x.*

Let $\rho_g(m, \chi)$ and $\mathrm{cva}_{\alpha,g}(m)$ be defined as in Section 6,

$$\mathrm{cva}_{\alpha,g}(m) = \inf\{\chi : \rho_g(m, \chi) \leq \alpha\} \quad \text{where } \rho_g(m, \chi) = \sup_F E_F[r(b, \chi)] \text{s.t.} E_F[g(b)] = m.$$

Let $\hat{\chi}_i = \mathrm{cva}_{\alpha,g}(\hat{m}_i)$. We will consider the average non-coverage $ANC_n(\hat{\chi}^{(n)}; \mathcal{X})$ of the collection of intervals $\{\hat{\theta}_i \pm \mathrm{se}_i \cdot \hat{\chi}_i\}$.

THEOREM C.1: *Suppose that Assumptions C.1, C.2, C.3, and C.4 hold, and that, for some j, $\lim_{b\to\infty} g_j(b) = \lim_{b\to-\infty} g_j(b) = \infty$ and $\inf_b g_j(b) \geq 0$. Then, for all $\mathcal{X} \in \mathcal{A}$,*

$$E_{\tilde{P}} ANC_n(\hat{\chi}^{(n)}; \mathcal{X}) \leq \alpha + o(1).$$

If, in addition, $Z_i - \tilde{b}_i$ is independent over i under $\tilde{P}$, then $ANC_n(\hat{\chi}^{(n)}; \mathcal{X}) \leq \alpha + o_{\tilde{P}}(1)$.

C.3. *Empirical Bayes Shrinkage Toward Regression Estimate*

We now apply the general results in Appendix C.2 to the EB setting. As in Section 3, we consider unshrunk estimates $Y_1, \dots, Y_n$ of parameters $\theta = (\theta_1, \dots, \theta_n)'$, along with regressors $X^{(n)} = (X_1, \dots, X_n)$ and variables $\tilde{X}^{(n)} = (\tilde{X}_1, \dots, \tilde{X}_n)'$, which include σ_i, and which play the role of the conditioning variables (the setting in Section 3 obtains as a special case with $\tilde{X}_i = (X_i, \sigma_i))$. The initial estimate Y_i has standard deviation σ_i, and we observe an estimate $\hat{\sigma}_i$. We obtain average coverage results by considering a triangular array of probability distributions $\tilde{P} = \tilde{P}^{(n)}$, in which the X_i's, σ_i's, and θ_i's are fixed. EB coverage can then be obtained for a distribution P of the data, θ, and some nuisance parameter $\tilde{\nu}$ such that these conditions hold almost surely with $P(\cdot|\theta, \tilde{\nu}, \tilde{X}^{(n)}, X^{(n)})$ playing the role of $\tilde{P}$.

We generalize the baseline specification in the main text, and consider

$$\hat{\theta}_i = \hat{X}_i'\hat{\delta} + w(\hat{\gamma}, \hat{\sigma}_i)(Y_i - \hat{X}_i'\hat{\delta}),$$

where $\hat{X}_i$ is an estimate of X_i (this allows some elements of X_i to be estimated rather than observed directly, such as when σ_i is included in X_i), $\hat{\delta}$ is any random vector that depends on the data (such as the OLS estimator in a regression of Y_i on X_i), and $\hat{\gamma}$ is a tuning parameter that determines shrinkage and may depend on the data. This leads to the standard error $\text{se}_i = w(\hat{\gamma}, \hat{\sigma}_i)\hat{\sigma}_i$ so that the t-statistic is

$$Z_i = \frac{\hat{\theta}_i - \theta_i}{\text{se}_i} = \frac{\hat{X}_i'\hat{\delta} + w(\hat{\gamma}, \hat{\sigma}_i)\big(Y_i - \hat{X}_i'\hat{\delta}\big) - \theta_i}{w(\hat{\gamma}, \hat{\sigma}_i)\hat{\sigma}_i} = \frac{Y_i - \theta_i}{\hat{\sigma}_i} + \frac{\big[w(\hat{\gamma}, \hat{\sigma}_i) - 1\big]\big(\theta_i - \hat{X}_i'\hat{\delta}\big)}{w(\hat{\gamma}, \hat{\sigma}_i)\hat{\sigma}_i}.$$

We use estimates of moments of the bias of positive integer order $\ell_1 < \cdots < \ell_p$. Let $\hat{\mu}_\ell$ be an estimate of the ℓth moment of $\theta_i - X_i'\delta$, and suppose that this moment is independent of σ_i in a sense formalized below. Then an estimate of the ℓ_jth moment of the bias is $\hat{m}_{i,j} = \frac{[w(\hat{\gamma}, \hat{\sigma}_i) - 1]^{\ell_j}\hat{\mu}_{\ell_j}}{w(\hat{\gamma}, \hat{\sigma}_i)^{\ell_j}\hat{\sigma}_i^{\ell_j}}$. Let $\hat{m}_i = (\hat{m}_1, \ldots, \hat{m}_p)'$. The EBCI is then given by $\hat{\theta}_i \pm w(\hat{\gamma}, \hat{\sigma}_i)\hat{\sigma}_i \cdot \text{cva}_{\alpha,g}(\hat{m}_i)$, where $g_j(b) = b^{\ell_j}$. We obtain the baseline specification in Section 3.2 when $p = 2$, $\ell_1 = 2$, $\ell_2 = 4$, $\hat{\gamma} = \hat{\mu}_2$, and $w(\hat{\mu}_2, \hat{\sigma}_i) = \hat{\mu}_2/(\hat{\mu}_2 + \hat{\sigma}_i^2)$.

We make the following assumptions.

ASSUMPTION C.5: $\lim_{n\to\infty} \max_{1\le i\le n} |\tilde{P}((Y_i - \theta_i)/\hat{\sigma}_i \le t) - \Phi(t)| = 0$.

Supplemental Material Appendix D.1 gives primitive conditions for Assumption C.5, and verifies them in a linear fixed effects panel data model. These conditions involve considering a triangular array of parameter values such that sampling error and empirical moments of the parameter value sequence are of the same order of magnitude, and defining θ_i to be a scaled version of the corresponding parameter.

ASSUMPTION C.6: *The standard deviations σ_i are bounded away from zero. In addition, for some δ and γ, $\hat{\delta}$ and $\hat{\gamma}$ converge to δ and γ under $\tilde{P}$, and, for any $\varepsilon > 0$,*

$$\lim_{n\to\infty} \max_{1\le i\le n} \tilde{P}\big(|\hat{\sigma}_i - \sigma_i| \ge \varepsilon\big) = 0 \quad \textit{and} \quad \lim_{n\to\infty} \max_{1\le i\le n} \tilde{P}\big(|\hat{X}_i - X_i| \ge \varepsilon\big) = 0.$$

ASSUMPTION C.7: *The variable $\tilde{X}_i$ takes values in $\mathcal{S}_1 \times \cdots \times \mathcal{S}_s$ where, for each k, either $\mathcal{S}_k = [\underline{x}_k, \overline{x}_k]$ (with $-\infty < \underline{x}_k < \overline{x}_k < \infty$) or $\mathcal{S}_k$ is a finitely discrete set with minimum element $\underline{x}_k$ and maximum element $\overline{x}_k$. In addition, $\tilde{X}_{i1} = \sigma_i$ (the first element of $\tilde{X}_i$ is given by σ_i). Furthermore, for some μ_0 such that $(\mu_{0,\ell_1}, \ldots, \mu_{0,\ell_p})$ is in the interior of the set of values of $\int g(b)\,dF(b)$ where F ranges over probability measures on $\mathbb{R}$ where $g_j(b) = b^{\ell_j}$ and some constant K, the following holds. Let $\mathcal{A}$ denote the collection of sets $\widetilde{\mathcal{S}}_1 \times \cdots \times \widetilde{\mathcal{S}}_s$ where $\widetilde{\mathcal{S}}_k$ is a positive Lebesgue measure interval contained in $[\underline{x}_k, \overline{x}_k]$ in the case where $\mathcal{S}_k = [\underline{x}_k, \overline{x}_k]$, and $\widetilde{\mathcal{S}}_k$ is a nonempty subset of $\mathcal{S}_k$ in the case where $\mathcal{S}_k$ is finitely discrete. For any $\mathcal{X} \in \mathcal{A}$, $N_{\mathcal{X},n} \to \infty$ and*

$$\frac{1}{N_{\mathcal{X},n}} \sum_{i\in\mathcal{I}_{\mathcal{X},n}} \big(\theta_i - X_i'\delta\big)^{\ell_j} \to \mu_{0,\ell_j}, \qquad \frac{1}{N_{\mathcal{X},n}} \sum_{i\in\mathcal{I}_{\mathcal{X},n}} |\theta_i|^{\ell_j} \le K, \quad \textit{and} \quad \frac{1}{N_{\mathcal{X},n}} \sum_{i\in\mathcal{I}_{\mathcal{X},n}} \|X_i\|^{\ell_j} \le K.$$

In addition, the estimate $\hat{\mu}_{\ell_j}$ converges in probability to μ_{0,ℓ_j} under $\tilde{P}$ for each j.

THEOREM C.2: *Let $\hat{\theta}_i$ and se_i be given above and let $\hat{\chi}_i = \mathrm{cva}_{\alpha,g}(\hat{m}_i)$ where $\hat{m}_i$ is given above and $g(b) = (b^{\ell_1}, \dots, b^{\ell_p})$ for some positive integers $\ell_1, \dots, \ell_p$, at least one of which is even. Suppose that Assumptions C.5, C.6, and C.7 hold, and that $w()$ is continuous in an open set containing $\{\gamma\} \times \mathcal{S}_1$ and is bounded away from zero on this set. Let $\mathcal{A}$ be as given in Assumption C.7. Then, for all $\mathcal{X} \in \mathcal{A}$, $E_{\tilde{P}} ANC_n(\hat{\chi}^{(n)}; \mathcal{X}) \leq \alpha + o(1)$. If, in addition, $(Y_i, \hat{\sigma}_i)$ is independent over i under $\tilde{P}$, then $ANC_n(\hat{\chi}^{(n)}; \mathcal{X}) \leq \alpha + o_{\tilde{P}}(1)$.*

As a consequence of Theorem C.2, we obtain, under the exchangeability condition (29), conditional EB coverage, as defined in Eq. (30), for any distribution P of the data and θ, $\tilde{\nu}$ such that the conditions of Theorem C.2 hold with probability 1 with the sequence of probability measures $P(\cdot|\theta, \tilde{\nu}, X^{(n)}, \tilde{X}^{(n)})$ playing the role of $\tilde{P}$. This follows from the arguments in Appendix C.1.

COROLLARY C.1: *Let θ, ν, $X^{(n)}$, $\tilde{X}^{(n)}$, Y_i follow a sequence of distributions P such that the conditions of Theorem C.2 hold with $\tilde{X}_i$ taking on finitely many values, and $P(\cdot|\theta, \nu, X^{(n)}, \tilde{X}^{(n)})$ playing the role of $\tilde{P}$ with probability 1, and such that the exchangeability condition (29) holds. Then the intervals $CI_i = \{\hat{\theta}_i \pm w(\hat{\gamma}, \hat{\sigma}_i)\hat{\sigma}_i \cdot \mathrm{cva}_{\alpha,g}(\hat{m}_i)\}$ satisfy the conditional EB coverage condition (30).*

The first part of Theorem 4.1 (average coverage) follows by applying Theorem C.2 with the conditional distribution $P(\cdot|\theta)$ playing the role of $\tilde{P}$. The second part (EB coverage) follows immediately from Corollary C.1.

REFERENCES

ABADIE, ALBERTO, AND MAXIMILIAN KASY (2019): "Choosing Among Regularized Estimators in Empirical Economics: The Risk of Machine Learning," *The Review of Economics and Statistics*, 101 (5), 743–762. https://doi.org/10.1162/rest_a_00812. [2587]

ANDREWS, ISAIAH, TORU KITAGAWA, AND ADAM MCCLOSKEY (2021): "Inference on Winners," Unpublished manuscript, Harvard University. [2569,2589]

ANGRIST, JOSHUA D., PETER D. HULL, PARAG A. PATHAK, AND CHRISTOPHER R. WALTERS (2017): "Leveraging Lotteries for School Value-Added: Testing and Estimation," *The Quarterly Journal of Economics*, 132 (2), 871–919. https://doi.org/10.1093/qje/qjx001. [2567]

ARMSTRONG, TIMOTHY B., AND MICHAL KOLESÁR (2018): "Optimal Inference in a Class of Regression Models," *Econometrica*, 86 (2), 655–683. https://doi.org/10.3982/ECTA14434. [2569]

ARMSTRONG, TIMOTHY B., MICHAL KOLESÁR, AND MIKKEL PLAGBORG-MØLLER (2020): "Robust Empirical Bayes Confidence Intervals," arXiv:2004.03448v2. [2575,2590]

——— (2022): "Supplement to 'Robust Empirical Bayes Confidence Intervals'," *Econometrica Supplemental Material*, 90, https://doi.org/10.3982/ECTA18597. [2570,2581]

BENJAMINI, YOAV, AND YOSEF HOCHBERG (1995): "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, 57 (1), 289–300. [2585,2586]

BENJAMINI, YOAV, AND DANIEL YEKUTIELI (2005): "False Discovery Rate–Adjusted Multiple Confidence Intervals for Selected Parameters," *Journal of the American Statistical Association*, 100 (469), 71–81. https://doi.org/10.1198/016214504000001907. [2589]

BONHOMME, STÉPHANE, AND MARTIN WEIDNER (2021): "Posterior Average Effects," Journal of Business & Economic Statistics, forthcoming. https://doi.org/10.1080/07350015.2021.1984928. [2569]

BROWN, LAWRENCE D., AND EITAN GREENSHTEIN (2009): "Nonparametric Empirical Bayes and Compound Decision Approaches to Estimation of a High-Dimensional Vector of Normal Means," *The Annals of Statistics*, 37 (4), 1685–1704. https://doi.org/10.1214/08-AOS630. [2585]

CAI, T. TONY, MARK LOW, AND ZONGMING MA (2014): "Adaptive Confidence Bands for Nonparametric Regression Functions," *Journal of the American Statistical Association*, 109 (507), 1054–1070. https://doi.org/10.1080/01621459.2013.879260. [2569,2583]

CARLIN, BRADLEY P., AND THOMAS A. LOUIS (2000): *Bayes and Empirical Bayes Methods for Data Analysis* (second Ed.). New York, NY: Chapman & Hall/CRC. [2571,2576,2584]

CASELLA, GEORGE, AND J. T. GENE HWANG (2012): "Shrinkage Confidence Procedures," *Statistical Science*, 27 (1), 51–60. https://doi.org/10.1214/10-STS319. [2585]

CHETTY, RAJ, AND NATHANIEL HENDREN (2018): "The Impacts of Neighborhoods on Intergenerational Mobility II: County-Level Estimates," *The Quarterly Journal of Economics*, 133 (3), 1163–1228. https://doi.org/10.1093/qje/qjy006. [2567,2569,2575,2589,2590]

CHETTY, RAJ, JOHN N. FRIEDMAN, AND JONAH E. ROCKOFF (2014): "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates," *American Economic Review*, 104 (9), 2593–2632. https://doi.org/10.1257/aer.104.9.2593. [2567]

EFRON, BRADLEY (2010): *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. New York, NY: Cambridge University Press. https://doi.org/10.1017/CBO9780511761362. [2569]

——— (2015): "Frequentist Accuracy of Bayesian Estimates," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77 (3), 617–646. https://doi.org/10.1111/rssb.12080. [2585]

——— (2019): "Bayes, Oracle Bayes and Empirical Bayes," *Statistical Science*, 34 (2), 177–201. https://doi.org/10.1214/18-STS674. [2584,2585,2588]

FINKELSTEIN, AMY, MATTHEW GENTZKOW, PETER HULL, AND HEIDI WILLIAMS (2017): "Adjusting Risk Adjustment—Accounting for Variation in Diagnostic Intensity," *New England Journal of Medicine*, 376 (7), 608–610. https://doi.org/10.1056/NEJMp1613238. [2567]

GIACOMINI, RAFFAELLA, TORU KITAGAWA, AND HARALD UHLIG (2019): "Estimation Under Ambiguity," Cemmap Working Paper 24/19. [2584]

GREENSHTEIN, EITAN, AND YA'ACOV RITOV (2019): "Comment: Empirical Bayes, Compound Decisions and Exchangeability," *Statistical Science*, 34 (2), 224–228. https://doi.org/10.1214/19-STS709. [2569]

HANSEN, BRUCE E. (2016): "Efficient Shrinkage in Parametric Models," *Journal of Econometrics*, 190 (1), 115–132. https://doi.org/10.1016/j.jeconom.2015.09.003. [2568]

HULL, PETER (2020): "Estimating Hospital Quality With Quasi-Experimental Data," Unpublished manuscript, University of Chicago. [2567]

HUNG, KENNETH, AND WILLIAM FITHIAN (2019): "Rank Verification for Exponential Families," *The Annals of Statistics*, 47 (2), 758–782. https://doi.org/10.1214/17-AOS1634. [2569,2589]

IGNATIADIS, NIKOLAOS, AND STEFAN WAGER (2021): "Confidence Intervals for Nonparametric Empirical Bayes Analysis," Journal of the American Statistical Association, forthcoming. https://doi.org/10.1080/01621459.2021.2008403. [2569]

JACOB, BRIAN A., AND LARS LEFGREN (2008): "Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education," *Journal of Labor Economics*, 26 (1), 101–136. https://doi.org/10.1086/522974. [2567]

JAMES, WILLARD, AND CHARLES M. STEIN (1961): "Estimation With Quadratic Loss," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, ed. by Jerzy Neyman. Berkeley, CA: University of California Press, 361–379, https://projecteuclid.org/euclid.bsmsp/1200512173. [2568,2570]

JIANG, WENHUA, AND CUN-HUI ZHANG (2009): "General Maximum Likelihood Empirical Bayes Estimation of Normal Means," *The Annals of Statistics*, 37 (4), 1647–1684. https://doi.org/10.1214/08-AOS638. [2585]

JOHNSTONE, IAIN M. (2019): *Gaussian Estimation: Sequence and Multiresolution Models*. Book draft, https://statweb.stanford.edu/~imj/GE_09_16_19.pdf. [2587]

KANE, THOMAS, AND DOUGLAS STAIGER (2008): "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation,". Technical Report 14607, National Bureau of Economic Research, Cambridge, MA. https://doi.org/10.3386/w14607. [2567]

KIEFER, JACK, AND JACOB WOLFOWITZ (1956): "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters," *The Annals of Mathematical Statistics*, 27 (4), 887–906. https://doi.org/10.1214/aoms/1177728066. [2585]

KOENKER, ROGER, AND IVAN MIZERA (2014): "Convex Optimization, Shape Constraints, Compound Decisions, and Empirical Bayes Rules," *Journal of the American Statistical Association*, 109 (506), 674–685. https://doi.org/10.1080/01621459.2013.869224. [2585]

LEE, JASON D., DENNIS L. SUN, YUEKAI SUN, AND JONATHAN E. TAYLOR (2016): "Exact Post-Selection Inference, With Application to the Lasso," *The Annals of Statistics*, 44 (3), 907–927. https://doi.org/10.1214/15-AOS1371. [2589]

LIU, LAURA, HYUNGSIK ROGER MOON, AND FRANK SCHORFHEIDE (2022): "Forecasting With a Panel Tobit Model," arXiv:2110.14117. [2569]

MACKINNON, JAMES G., AND HALBERT WHITE (1985): "Some Heteroskedasticity-Consistent Covariance Matrix Estimators With Improved Finite Sample Properties," *Journal of Econometrics*, 29 (3), 305–325. https://doi.org/10.1016/0304-4076(85)90158-7. [2584]

MORRIS, CARL N. (1983a): "Parametric Empirical Bayes Confidence Intervals," in *Scientific Inference, Data Analysis, and Robustness*, ed. by George E. P. Box, Tom Leonard, and Chien-Fu Wu. New York, NY: Academic Press, 25–50. https://doi.org/10.1016/B978-0-12-121160-8.50008-9. [2576,2584,2592]

——— (1983b): "Parametric Empirical Bayes Inference: Theory and Applications," *Journal of the American Statistical Association*, 78 (381), 47–55. https://doi.org/10.1080/01621459.1983.10477920. [2567,2568,2571, 2576,2584,2592]

NYCHKA, DOUGLAS (1988): "Bayesian Confidence Intervals for Smoothing Splines," *Journal of the American Statistical Association*, 83 (404), 1134–1143. https://doi.org/10.1080/01621459.1988.10478711. [2569]

PINELIS, IOSIF (2002): "Monotonicity Properties of the Relative Error of a Padé Approximation for Mills' Ratio," Journal of Inequalities in Pure & Applied Mathematics, 3 (2), http://eudml.org/doc/122588. [2593]

PRATT, JOHN W. (1961): "Length of Confidence Intervals," *Journal of the American Statistical Association*, 56 (295), 549–567. https://doi.org/10.2307/2282079. [2568]

STOREY, JOHN D. (2002): "A Direct Approach to False Discovery Rates," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64 (3), 479–498. https://doi.org/10.1111/1467-9868.00346. [2585,2586]

WAHBA, GRACE (1983): "Bayesian "Confidence Intervals" for the Cross-Validated Smoothing Spline," *Journal of the Royal Statistical Society: Series B (Methodological)*, 45 (1), 133–150. https://doi.org/10.1111/j.2517-6161.1983.tb01239.x. [2569]

WASSERMAN, LARRY (2006): *All of Nonparametric Statistics*. New York, NY: Springer. https://doi.org/10.1007/0-387-30623-4. [2569,2585]

XIE, XIANCHAO, SAMUEL KOU, AND LAWRENCE D. BROWN (2012): "SURE Estimates for a Heteroscedastic Hierarchical Model," *Journal of the American Statistical Association*, 107 (500), 1465–1479. https://doi.org/10.1080/01621459.2012.728154. [2575]

Guest Co-editor Alberto Abadie handled this manuscript.

Manuscript received 15 June, 2020; final version accepted 9 June, 2022; available online 15 July, 2022.