# Robust Empirical Bayes Confidence Intervals*

Timothy B. Armstrong[†] Yale University

Michal Kolesár[‡] Princeton University

Mikkel Plagborg-Møller[§] Princeton University

April 7, 2020

**Abstract**

We construct robust empirical Bayes confidence intervals (EBCIs) in a normal means problem. The intervals are centered at the usual empirical Bayes estimator, but use a larger critical value to account for the effect of shrinkage. We show that in this setting, parametric EBCIs based on the assumption that the means are normally distributed (Morris, 1983) can have coverage substantially below the nominal level when the normality assumption is violated, and we derive a simple rule of thumb for gauging the potential coverage distortion. In contrast, while our EBCIs remain close in length to the parametric EBCIs when the means are indeed normally distributed, they achieve correct coverage regardless of the means distribution. 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. We illustrate our methods with applications to effects of U.S. neighborhoods on intergenerational mobility, and structural changes in factor loadings in a large dynamic factor model for the Eurozone. Our approach generalizes to the construction of intervals with average coverage guarantees in other regularized estimation settings.

*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.

[†]email: timothy.armstrong@yale.edu

[‡]email: mkolesar@princeton.edu

[§]email: mikkelpm@princeton.edu

# 1 Introduction

Empirical researchers in economics are often interested in estimating effects for a large number of individuals or units, such as estimating teacher quality for teachers in a given geographic area. In such problems, it has become common to shrink unbiased but noisy preliminary estimates of these effects toward baseline values, say the average fixed effect for teachers with the same experience. In addition to estimating teacher quality (Kane and Staiger, 2008; Jacob and Lefgren, 2008; Chetty et al., 2014), shrinkage techniques have been used recently in a wide range of applications including estimating school quality (Angrist et al., 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 et al., 2017). Fessler and Kasy (2019) propose using economic theory to guide the shrinkage direction.

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 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 effects distribution. Indeed, Hansen (2016, p. 116) describes 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 effects distribution as a prior (Morris, 1983). 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" (Carlin and Louis, 2000). 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 distributions. On the other hand, if the degree of shrinkage is small, the coverage distortion is limited, and

we derive a simple "rule of thumb", in the form of a universal cut-off value on the degree of shrinkage, ensuring that the coverage distortion of the parametric EBCI is limited.

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 effects distribution. Our baseline model is an (approximate) normal means problem $Y_i \sim N(\theta_i, \sigma_i^2)$, $i = 1, \ldots, n$. In applications, $Y_i$ represents a preliminary asymptotically unbiased estimate of the effect $\theta_i$ for unit $i$. Like the parametric EBCI that assumes a normal distribution for $\theta_i$, the robust EBCI we propose is centered at the normality-based EB point estimate $\hat{\theta}_i$, but it uses a larger critical value to take into account the bias due to shrinkage. For convenient practical implementation, we provide software implementing our methods. EB coverage is controlled in the class of all distributions for $\theta_i$ that satisfy certain moment bounds, which we estimate consistently from the data (similarly to the parametric EBCI, which uses the second moment). Importantly, we show that the baseline implementation of our robust EBCI is "adaptive", in the sense that its length is very close to the length of the parametric EBCI when the $\theta_i$'s are in fact normally distributed. Thus, little efficiency is lost from using the robust EBCI in place of the non-robust parametric one.

In addition to controlling EB coverage, we show that the robust $1-\alpha$ EBCIs have a frequentist *average coverage* property: If the mean parameters $\theta_1, \ldots, \theta_n$ are treated as *fixed*, the coverage probability—averaged across the $n$ parameters $\theta_i$—is at least $1-\alpha$. Although the usual CI centered at the unshrunk estimate $Y_i$ clearly also achieves average coverage, the unshrunk CIs are wider than our robust EBCIs, and often substantially so. This improvement in CI length is possible because the average coverage criterion relaxes the usual notion of coverage, which would be imposed separately for each $\theta_i$.[1] Intuitively, the average coverage criterion only requires us to guard against the *average* coverage distortion induced by the biases of the individual estimators $\hat{\theta}_i$, and the data is quite informative about whether *most* of these biases are large, even though individual biases are difficult to estimate. While one might hope to achieve similar improvements in CI length while satisfying the usual notion of coverage, it can be shown formally that this is not possible, regardless of how one forms the CI.[2]

We also show how the underlying ideas may be translated to other shrinkage settings,

[1] This stands in contrast to the requirement of *simultaneous* coverage, which strengthens the usual notion of (pointwise) coverage.

[2] In particular, it follows from the results in Pratt (1961) that for CIs with nominal 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 $(\theta_1, \ldots, \theta_n)$. See, for example, Corollary 3.3 in Armstrong and Kolesár (2018) and the discussion following it.

not just the normal means model. Our CI construction generalizes naturally to settings in which one has available approximately normal, but biased estimates of parameters $\theta_i$, and one can consistently estimate moments of the bias normalized by the standard error. This includes classic nonparametric estimation problems, such as estimating the conditional mean function using local polynomials or regression trees. Here $\theta_i$ corresponds to the conditional mean given covariates of observation $i$, and the resulting CIs can be interpreted as an average coverage confidence band for the regression function.

We illustrate our results through two empirical applications. The first application considers the effect of growing up in different U.S. neighborhoods (specifically commuting zones) on intergenerational mobility. We follow Chetty and Hendren (2018), who apply EB shrinkage to initial fixed effects estimates. Depending on the specification, we find that the robust EBCIs are on average 12–52% as long as the unshrunk CIs. Our second application estimates the extent of structural change in a dynamic factor model (DFM) of the Eurozone. Employing a large panel of macroeconomic time series for the 19 Eurozone countries, we construct robust EBCIs for the breaks in the factor loadings following the Great Recession. We shrink the loading breaks towards zero to reduce the influence of estimation error due to the short sample. Our robust EBCIs for the loading breaks are on average 77–86% as long as the unshrunk CIs.

The robust EBCI we develop can also be viewed as a (pure) Bayesian 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 $\theta_i$, for any prior distribution that satisfies the moment bounds. In contrast, *conditional* gamma-minimax credible intervals, discussed recently by Kitagawa et al. (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, and it would lead to reporting the entire parameter space (up to the moment bounds).

The average coverage criterion was originally introduced in the literature on nonparametric regression (Wahba, 1983; Nychka, 1988; Wasserman, 2006, Chapter 5.8). Cai et al. (2014) construct rate-optimal adaptive confidence bands that achieve average coverage. These procedures for nonparametric regression are challenging to implement in our EB setting and do not have a clear finite-sample justification, unlike our procedure. Outside the nonparametric regression context, Liu et al. (2019) construct forecast intervals that guarantee average coverage in a Bayesian sense (for a fixed prior). Bonhomme and Weidner (2020) and Ignatiadis and Wager (2019) consider robust estimation and inference on functionals of the effects $\theta_i$, rather than the effects themselves.

The rest of this paper is organized as follows. Section 2 illustrates our methods in the context of a simple homoskedastic 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. Section 5 discusses extensions of the basic framework. Section 6 contains two empirical applications: (i) inference on neighborhood effects and (ii) inference on structural breaks in a DFM. Appendix A gives computational details, and Appendix B formal proofs of our coverage results. The Online Supplement provides additional empirical results. Applied readers are encouraged to focus on Sections 2, 3 and 6.

# 2 Simple example

This section illustrates the construction of the robust EBCIs that we propose in a simplified setting with homoskedastic errors. In the next section, we show how to generalize these results when the variances of the $Y_i$'s are heteroskedastic along with several other empirically relevant extensions of the basic framework, and we discuss implementation issues.

We observe $n$ independent, normally distributed estimates

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

of the parameter vector $\theta = (\theta_1, \dots, \theta_n)'$. In many applications, the $Y_i$'s arise as preliminary least squares estimates of the parameters $\theta_i$. 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, the normality in Eq. (1) is only approximate, and justified by large-sample arguments; for simplicity, we assume here that it is exact. We also assume that the variance $\sigma^2$ is known.

A popular approach to estimation that substantially improves upon the raw estimator $\hat{\theta} = Y$ 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 $\theta_i$'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 $\theta_i$, 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 this normal sampling distribution, it is optimal to estimate $\theta_i$ using the posterior mean $\hat{\theta}_i = w_{EB} Y_i$, where $w_{EB} = (1 - \sigma^2/(\sigma^2 + \mu_2))$. To avoid having to specify the variance $\mu_2$ of the distribution of $\theta_i$,

the EB approach treats it as an unknown parameter, and uses the data to estimate this posterior, replacing 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 unbiased estimate $(n-2)/\sum_{i=1}^n Y_i^2$. The latter leads to $\hat{w}_{EB} = (1 - \sigma^2(n-2)/\sum_{i=1}^n Y_i^2)$, which is the classic estimator of James and Stein (1961).

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

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

where $z_\alpha$ is the $\alpha$ 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 $\mu_2$ unknown, one can replace $w_{EB}$ by $\hat{w}_{EB}$.[3] 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 $(Y_i, \theta_i)$ according to Eqs. (1) and (2). To distinguish this notion of coverage from the case with fixed $\theta$, we refer to coverage under repeated sampling of $(Y_i, \theta_i)$ as "empirical Bayes coverage". This follows the definition in Carlin and Louis (2000, Chapter 3.5), who refer to an interval with coverage $1-\alpha$ under repeated sampling of $(Y_i, \theta_i)$ as an (unconditional) empirical Bayes confidence interval (EBCI). 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 95% parametric EBCI can be as low as 74% for certain non-normal distributions of $\theta_i$ with variance $\mu_2$; more generally, for a nominal $1-\alpha$ confidence level, it can be as low as $1 - 1/\max\{z_{1-\alpha/2}, 1\}$. This contrasts with existing results on estimation: Although the empirical Bayes estimator is motivated by the parametric assumption (2), it performs well even if this assumption is dropped, with low MSE even if we treat $\theta$ as fixed.

In this paper, we construct 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 will remain $1-\alpha$ without making any parametric assumptions on the distribution of $\theta_i$. To describe how we construct an EBCI with such a robustness property, suppose that all that is known is that $\theta_i$ is sampled i.i.d. from a distribution with second moment given by $\mu_2$ (in practice, we can replace it by the consistent estimate $n^{-1}\sum_{i=1}^n Y_i^2 - \sigma^2$). Conditional on $\theta_i$, the estimator $w_{EB}Y_i$ has bias $(w_{EB}-1)\theta_i$ and variance $w_{EB}^2\sigma^2$, so that the $t$-statistic $(w_{EB}Y_i - \theta_i)/w_{EB}\sigma$ is normally distributed with mean $b_i = (1 - 1/w_{EB})\theta_i/\sigma$ and variance 1. Therefore, if we use a critical value $\chi$, the non-coverage of the CI $w_{EB}Y_i \pm \chi w_{EB}\sigma$ conditional on $\theta_i$ will be given by the probability $r(b_i, \chi) = P(|Z - b_i| \geq \chi \mid \theta_i) = \Phi(-\chi - b_i) + \Phi(-\chi + b_i)$, where $Z$

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

denotes a standard normal random variable, and $\Phi$ denotes the standard normal cdf. Thus, by iterated expectations, under repeated sampling of $\theta_i$, the non-coverage is bounded by

$$\rho(\sigma^2/\mu_2, \chi) = \sup_F E_F[r(b, \chi)] \quad \text{s.t.} \quad E_F[b^2] = \frac{(1 - 1/w_{EB})^2}{\sigma^2}\mu_2 = \frac{\sigma^2}{\mu_2}, \tag{4}$$

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.[4] 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 $\text{cva}_\alpha(t) = \rho^{-1}(t, \alpha)$, where the inverse is with respect to the second argument. Then the resulting interval

$$w_{EB}Y_i \pm \text{cva}_\alpha(\sigma^2/\mu_2)w_{EB}\sigma \tag{5}$$

will maintain coverage $1 - \alpha$ among all distributions of $\theta_i$ with $E[\theta_i^2] = \mu_2$ (recall that we estimate $\mu_2$ consistently from the data). For this reason, we refer to it as a robust EBCI. Figure 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 $\theta_i$, in addition to a second moment constraint, one can construct robust EBCIs that "adapt" to the Gaussian case in the sense that if these moment constraints are compatible with a normal distribution, its length will be close to the parametric EBCI in Eq. (3).

Instead of considering EB coverage, one may alternatively wish to assess uncertainty associated with the estimates $w_{EB}Y_i$ when $\theta$ is treated as fixed. In this case, the EBCI in Eq. (5) has an average coverage guarantee that

$$\frac{1}{n}\sum_{i=1}^{n} P\big(\theta_i \in [w_{EB}Y_i \pm \text{cva}_\alpha(\sigma^2/\mu_2)w_{EB}\sigma] \;\big|\; \theta\big) \geq 1 - \alpha, \tag{6}$$

provided that the moment constraint can be interpreted as a constraint on the empirical second moment on the $\theta_i$'s, $n^{-1}\sum_{i=1}^n \theta_i^2 = \mu_2$. In other words, if we condition on $\theta$, then the coverage is at least $1-\alpha$ on average across the $n$ EBCIs for $\theta_1, \ldots, \theta_n$. To see this, note that the average non-coverage of the intervals is bounded by (4), except that the supremum is only

[4] Specifically, Proposition A.1 in Appendix A shows that

$$\rho(t, \chi) = \begin{cases} r(0, \chi) + \frac{t}{t_0}(r(t_0^{1/2}, \chi) - r(0, \chi)) & \text{if } |t| < t_0, \\ r(t^{1/2}, \chi) & \text{otherwise.} \end{cases}$$

Here $t_0$ solves $r(t^{1/2}, \chi) - t\frac{\partial}{\partial t}r_0(t^{1/2}, \chi) = r(0, \chi)$. The solution is unique if $\chi \geq \sqrt{3}$; if $\chi < \sqrt{3}$, put $t_0 = 0$.

taken over possible empirical distributions for $\theta_1, \ldots, \theta_n$ 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 $1 - \alpha$.[5]

The usual CIs $Y_i \pm z_{1-\alpha/2}\sigma$ also of course achieve average coverage $1-\alpha$. The robust EBCI in Eq. (5) will however be shorter, especially when $\mu_2$ is small relative to $\sigma^2$—see Figure 4 below: by weakening the requirement that each CI covers the true parameter $1 - \alpha$ percent of the time to the requirement that the coverage is $1 - \alpha$ on average across the CIs, we can substantially shorten the CI length. It may seem surprising at first that we can achieve this by centering the CI at the shrinkage estimates $w_{EB}Y_i$. The intuition for this is that the shrinkage reduces the variability of the estimates. This comes at the expense of introducing bias in the estimates. However, we can on average control the resulting coverage loss by using the 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 $\theta_i$, 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 towards more general regression estimates rather than towards 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 Model and robust EBCI

In applied settings, the standard errors for the unshrunk estimates $Y_i$ will often vary. Furthermore, rather than shrinking towards zero, it is common to shrink toward an estimate of $\theta_i$ 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 the heteroskedastic model

$$Y_i \mid \theta_i, X_i, \sigma_i \sim N(\theta_i, \sigma_i^2). \tag{7}$$

[5]This link between average risk of separable decision rules (here coverage of CIs, each of which depends only on $Y_i$) when the parameters $\theta_1, \ldots, \theta_n$ are treated as fixed, and the risk of a single decision rule when these parameters are i.i.d. is a special case of what Jiang and Zhang (2009) call the fundamental theorem of compound decisions, which goes back to Robbins (1951).

The covariate vector $X_i$ may contain just the intercept, and it may also contain (functions of) $\sigma_i$. $Y_i$ will typically be some preliminary unrestricted estimate of $\theta_i$ that is only *approximately* normal in large samples by the central limit theorem (CLT), a feature that we will explicitly take into account in the theory in Appendix B. To construct an EB estimator of $\theta_i$, consider the working assumption that the sampling distribution of the $\theta_i$'s is conditionally normal:

$$\theta_i \mid X_i, \sigma_i \sim N(\mu_{1,i}, \mu_2), \quad \text{where} \quad \mu_{1,i} = X_i'\delta. \tag{8}$$

The hierarchical model (7)–(8) leads to the Bayes estimate $\hat{\theta}_i = \mu_{1,i} + w_{EB,i}(Y_i - \mu_{1,i})$, where $w_{EB,i} = \frac{\mu_2}{\mu_2+\sigma_i^2}$. This estimate shrinks the unrestricted estimate $Y_i$ of $\theta_i$ toward $\mu_{1,i} = X_i'\delta$. Although convenient, the normality assumption (8) typically cannot be justified simply by appealing to the CLT, and 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 (8) fails. To obtain a narrow robust EBCI, we augment the second moment restriction used to compute the critical value in Eq. (4) with restrictions on higher moments of the bias of $\hat{\theta}_i$. In our baseline specification, we add a restriction on the fourth moment.

We replace assumption (8) with the much weaker requirement that the conditional second moment and kurtosis of $\varepsilon_i = \theta_i - X_i'\delta$ do not depend on $(X_i, \sigma_i)$:

$$E[(\theta_i - X_i'\delta)^2 \mid X_i, \sigma_i] = \mu_2, \quad E[(\theta_i - X_i'\delta)^4 \mid X_i, \sigma_i]/\mu_2^2 = \kappa, \tag{9}$$

where, to relax the assumption that the conditional mean is linear, $\delta = \operatorname{plim} \hat{\delta}$ is defined as the probability limit of the ordinary least squares (OLS) estimates $\hat{\delta}$ in a regression of $Y_i$ on $X_i$.[6] We discuss this requirement further in Remark 3.2 below, and we relax it in Remark 3.8 below.

We now apply analysis analogous to that in Section 2. Let us suppose for simplicity that $\delta$, $\mu_2$, and $\kappa$ are known; we discuss practical implementation in Section 3.2 below. Denote the ratio of the conditional bias and standard deviation of $\hat{\theta}_i$ by $b_i = (1 - w_{EB,i})\varepsilon_i/(w_{EB,i}\sigma_i) = (1/w_{EB,i} - 1)\varepsilon_i/\sigma_i$. Under repeated sampling of $\theta_i$, the non-coverage of the CI $w_{EB,i}Y_i \pm \chi w_{EB,i}\sigma$, conditional on $(X_i, \sigma_i)$, depends on the distribution of the standardized bias $b_i$, as in Section 2. Given the known moments $\mu_2$ and $\kappa$, the *maximal* non-coverage is given by

$$\rho(m_{2,i}, \kappa, \chi) = \sup_F E_F[r(b, \chi)] \quad \text{s.t.} \quad E_F[b^2] = m_{2,i},\ E_F[b^4] = \kappa m_{2,i}^2, \tag{10}$$

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

where $b$ is distributed according to the distribution $F$. Here $m_{2,i} = E[b_i^2 \mid X_i, \sigma_i] = (1 - 1/w_{EB,i})^2 \mu_2/\sigma_i^2 = \sigma_i^2/\mu_2$. Observe that the kurtosis of $b_i$ matches that of $\varepsilon_i$. Appendix A shows that the infinite-dimensional linear program (10) can be reduced to two nested *univariate* optimizations. We also show that the least favorable distribution—the distribution $F$ maximizing (10)—is a discrete distribution with up to 4 support points (see Remark A.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 $\kappa$. This leads to the robust EBCI

$$\hat{\theta}_i \pm \text{cva}_\alpha(m_{2,i}, \kappa) w_{EB,i} \sigma_i. \tag{11}$$

By construction, this CI has coverage at least $1 - \alpha$ under repeated sampling of $(Y_i, \theta_i)$, conditional on $(X_i, \sigma_i)$, so long as Eq. (9) holds; it is not required that the conditional distribution of $\theta_i$ be normal with a linear conditional mean.

## 3.2 Baseline implementation

Our baseline implementation of the robust EBCI simply plugs in consistent estimates of the unknown quantities in Eq. (11):

1. Let $Y_i$ be an estimate of $\theta_i$ with standard error $\hat{\sigma}_i$, and let $X_i$ be covariates that are thought to help predict $\theta_i$.

2. Regress $Y_i$ on $X_i$ to obtain the fitted values $X_i'\hat{\delta}$, where $\hat{\delta} = (\sum_{i=1}^n X_i X_i')^{-1} \sum_{i=1}^n X_i Y_i$. Let $\hat{\varepsilon}_i = Y_i - X_i'\hat{\delta}$ be the residuals from this regression. Let $\hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^n (\hat{\varepsilon}_i^2 - \hat{\sigma}_i^2)$, and $\hat{\kappa} = \frac{1}{n}\sum_{i=1}^n (\hat{\varepsilon}_i^4 - 6\hat{\sigma}_i^2\hat{\varepsilon}_i^2 + 3\hat{\sigma}_i^4)/\hat{\mu}_2^2$.

3. Form the EB estimate

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

4. Compute the critical value $\text{cva}_\alpha(\hat{\sigma}_i^2/\hat{\mu}_2, \hat{\kappa})$ defined in (10). We provide a fast and stable numerical algorithm in our software package.[7]

5. Report the robust EBCI

$$\hat{\theta}_i \pm \text{cva}_\alpha(\hat{\sigma}_i^2/\hat{\mu}_2, \hat{\kappa}) \hat{w}_{EB,i} \hat{\sigma}_i. \tag{12}$$

[7] Software implementing these critical values is available at https://github.com/kolesarm/ebci

We discuss the assumptions needed for validity of the robust EBCI in Remarks 3.2, 3.6 and 3.7 below.

**Remark 3.1** (Rule of thumb for when to use parametric EBCI)**.** If we take the normality assumption (8) seriously, we may use the parametric EBCI

$$\hat{\theta}_i \pm z_{1-\alpha/2}\hat{w}_{EB,i}^{1/2}\hat{\sigma}_i, \tag{13}$$

which is an EB version of a Bayesian credible interval that treats (8) as a prior. We show in Section 4.3 that for significance levels $\alpha = 0.05$ or $0.10$, if we drop the normality assumption (8), then the parametric EBCI has a maximum coverage distortion of at most 5 percentage points, provided that the shrinkage factor $\hat{w}_{EB,i} \geq 0.3$. Hence, if moderate coverage distortions can be tolerated, a simple rule of thumb would be to report the parametric EBCI unless $\hat{w}_{EB,i}$ falls below this threshold. Importantly, however, Section 4.2 below will show that the robust EBCI (11) is almost as narrow as the parametric EBCI if the normality assumption (8) in fact holds, so little is lost by always reporting the robust EBCI.

**Remark 3.2** (Conditional EB coverage and moment independence)**.** A potential concern about the EB coverage criterion in a heteroskedastic setting is that in order to reduce the length of the CI on average, one could overcover parameters $\theta_i$ with small $\sigma_i$ and give up entirely on covering parameters $\theta_i$ for which the standard error $\sigma_i$ is large. Our robust EBCI avoids these issues by requiring EB coverage to hold conditional on $(X_i, \sigma_i)$. This also prevents similar conditional coverage issues arising depending on the value of $X_i$.

The key assumption we need to ensure this property is the assumption in (9) that the conditional second moment and kurtosis of $\varepsilon_i = \theta_i - X_i'\delta$ doesn't depend on $(X_i, \sigma_i)$. 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 6.1. Nonetheless, such conditions may be strong in some settings, as argued by Xie et al. (2012) in the context of EB point estimation. As discussed in Remark 3.7 below, the condition (9) can be avoided entirely by replacing $\hat{\mu}_2$ and $\hat{\kappa}$ with nonparametric estimates of these conditional moments, or relaxed using a flexible parametric specification.

**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 (6) when the parameters $\theta = (\theta_1, \ldots, \theta_n)$ are considered fixed, in addition to achieving valid EB coverage when the $\theta_i$'s are viewed as random draws from some underlying distribution. To guarantee average

coverage, we do not need to assume that the $Y_i$'s and $\theta_i$'s are drawn independently across $i$. This is because the average coverage criterion (6) only depends on the marginal distribution of $(Y_i, \theta_i)$, not the joint distribution. We only require that the estimates $\hat{\mu}_2, \hat{\kappa}, \hat{\delta}, \hat{\sigma}_i$ are consistent for $\mu_2, \kappa, \delta, \sigma_i$, which is the case under many forms of weak dependence or clustering. Notice that 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 doesn't need to take an explicit stand on how the clusters are defined.

**Remark 3.4** (Length-optimal shrinkage)**.** The shrinkage coefficient $w_{EB,i} = \mu_2/(\mu_2 + \sigma_i^2)$ is designed to optimize MSE of the point estimator $\hat{\theta}_i$. If an EBCI is directly of interest rather than a point estimate, it may be desirable to optimize shrinkage to minimize the length of the robust EBCI. The length of the EBCI based on the estimator $\mu_{1,i} + w_i(Y_i - \mu_{1,i})$ is $\text{cva}_\alpha((1 - 1/w_i)^2\mu_2/\sigma_i^2, \kappa_i)w_i$. This expression can be numerically minimized as a function of $w_i$ to find the EBCI length-optimal shrinkage $w_{opt,i} = w_{opt}(\mu_2/\sigma_i^2, \kappa, \alpha)$ given $\mu_2/\sigma_i^2$ and $\kappa$. We show theoretically in Section 4.2 and empirically in Section 6 that the efficiency gains from using length-optimal shrinkage relative to MSE-optimal shrinkage are only substantial if the distribution of $\theta_i$ is not close to the normal distribution.

**Remark 3.5** (Using higher moments)**.** In addition to using the second and fourth moment of bias, one may augment (10) 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 substantively decrease the critical value in the case where the normal distribution for $\theta_i$ is correct. Thus, the CI in our baseline implementation is robust to failure of the normality assumption (8), while being near-optimal when this assumption does hold.

If greater efficiency is desired under departures from a normal distribution for $\theta_i$, one can add other moment restrictions to the optimization problem (10) and plug in consistent estimates of these moments. For example, if the distribution of $\theta_i$ is substantially skewed, then adding a restriction on the third moment may shrink the CI. Note, however, that our approach restricts attention to linear (or, at least, affine) shrinkage estimators: the estimator takes the form $\hat{\theta}_i = a_i + w_i(Y_i - a_i)$ for some constants $a_i$ and $w_i$ that can depend on the covariates $X_i$ but cannot depend on $Y_i$ directly. To obtain a fully efficient EBCI when the distribution of $\theta_i$ is not normal, one will need to consider estimators $\hat{\theta}_i$ that are nonlinear functions of $Y_i$.

In finite samples, one may also be concerned that inaccurate estimates of $\hat{\kappa}$ may affect coverage. To address this concern, one may impose only a constraint on the second moment, which is equivalent to setting $\kappa = \infty$ in Eq. (10). One may then also weaken assumption (9)

by only imposing independence of the second moment. Alternatively, since the solution to Eq. (10) remains unchanged when the fourth-moment constraint is imposed as an inequality $E_F[b^4] \leq \kappa m_{2,i}^2$, one can replace $\hat{\kappa}$ with a conservative estimate.

**Remark 3.6** (Estimating moments of the distribution of $\theta_i$)**.** The estimates $\hat{\mu}_2$ and $\hat{\kappa}$ in step 2 of our baseline implementation above is based on the moment conditions $E[(Y_i - X_i'\delta)^2 - \sigma_i^2 \mid X_i, \sigma_i] = \mu_2$ and $E[(Y_i - X_i'\delta)^4 + 3\sigma_i^4 - 6\sigma_i^2(Y_i - X_i'\delta)^4 \mid X_i, \sigma_i] = \kappa\mu_2^2$, replacing population expectations by sample averages. However, since our theory only requires that estimates be consistent, one may use alternative estimates. For example, to increase precision of the estimates, one can use precision weights when forming sample averages. If one has access to the original data used to compute the estimates $Y_i$, and the estimates can be written as sample means $Y_i = T^{-1}\sum_{t=1}^T W_{it}$, and $W_{it}$ is independent across $t$ conditional on $\theta_i$, one can use the unbiased jackknife estimate $\frac{2}{nT(T-1)}\sum_{i=1}^n\sum_{t=2}^T\sum_{s=1}^{t-1} W_{it}W_{is}$ of $\mu_2$, and an analogous jackknife estimate for $\kappa$.

**Remark 3.7** (Nonparametric moment estimates)**.** If conditional EB coverage is desired, but the moment independence assumption (9) is implausible, it is straightforward in principle to allow the conditional moments of $\varepsilon_i$ to depend nonparametrically on $(X_i, \sigma_i)$, and use kernel or series estimators $\hat{\mu}_{2i}$ and $\hat{\kappa}_i$ of $\mu_2(X_i, \sigma_i) = E[(Y_i - X_i'\delta)^2 \mid X_i, \sigma_i]$ and $\kappa(X_i, \sigma_i) = E[(Y_i - X_i'\delta)^4 \mid X_i, \sigma_i]/\mu_2(X_i, \sigma_i)^2$. If these estimates are consistent, and one replaces the critical value in Eq. (12) with $\text{cva}_\alpha((1/\hat{w}_{EB,i} - 1)^2\hat{\mu}_{2i}/\hat{\sigma}_i^2, \hat{\kappa}_i)$, the resulting CI achieves valid EB coverage with assumption (9) dropped. Similarly, one can replace $X_i'\delta$ in the definition of $w_{EB,i}$ and $\varepsilon_i$ with a non-parametric estimate of the conditional mean $E[Y_i \mid X_i, \sigma_i] = E[\theta_i \mid X_i, \sigma_i]$.

**Remark 3.8** ($t$-statistic shrinkage)**.** Another way to avoid the moment independence condition (9) is to base shrinkage on the $t$-statistics $Y_i/\sigma_i$. Since these have constant variance equal to 1 by construction, we can apply 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 homoskedastic analysis in Section 2 applies, leading to valid EBCIs without any assumptions about independence of the moments. We discuss this approach further in Appendix B.5, and illustrate it in the empirical applications in Section 6. A disadvantage of this approach is that, while the resulting intervals satisfy the EB coverage property unconditionally, they do not satisfy the conditional coverage property discussed in Remark 3.2.

# 4 Main results

This section provides formal statements of the coverage properties of the CIs presented in Sections 2 and 3. In addition, we show that the CIs presented in Sections 2 and 3 are highly efficient when the mean parameters are in fact normally distributed. Finally, we calculate the maximal coverage distortion of the parametric EBCI. Applied readers interested primarily in implementation issues may skip ahead to the empirical applications in Section 6.

## 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 $CI_1, \ldots, CI_n$ for elements of the parameter vector $\theta = (\theta_1, \ldots, \theta_n)'$. We use the probability measure $P$ to denote the joint distribution of $\theta$ and $CI_1, \ldots, CI_n$. We say that the interval $CI_i$ is an (asymptotic) $1-\alpha$ empirical Bayes CI 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 $\theta_1, \ldots, \theta_n$ if

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

Note that the average coverage property (15) is a property of the distribution of the data conditional on the parameter $\theta$ and therefore does not require that we view the $\theta_i$'s as random (as in a Bayesian or "random effects" analysis). We nonetheless maintain the conditioning notation $P(\cdot \mid \theta)$ when stating results on average coverage, in order to maintain consistent notation.

Under an exchangeability condition, the ACI property (15) implies the EBCI property (14). Suppose that the average coverage property (15) holds almost surely and that the marginal distribution of $\{\theta_i, CI_i\}_{i=1}^n$ is exchangeable in the sense that

$$P(\theta_i \in CI_i) = P(\theta_j \in CI_j) \quad \text{for all } i, j.$$

Then, the EBCI property (14) holds since, for all $j$,

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

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 $\theta_i$, (ii) state the results only for the homoskedastic case where the variance $\sigma_i$ of the unshrunk estimate $Y_i$ does not vary across $i$, and (iii) we consider only unconditional coverage statements of the form (14) and (15). In Theorem B.2 in Appendix B, we allow the estimates $Y_i$ to be only approximately normally distributed and allow $\sigma_i$ to vary, and we formalize the statements about conditional coverage made in Remark 3.2. The following theorem is a special case of this result.

**Theorem 4.1.** *Suppose* $Y_i \mid \theta \sim N(\theta_i, \sigma)$. *Let* $\mu_{j,n} = \frac{1}{n}\sum_{i=1}^{n}(\theta_i - X_i'\delta)^j$ *and let* $\kappa_n = \mu_{4,n}/\mu_{2,n}^2$. *Let* $\theta_1, \ldots, \theta_n$ *be a sequence such that* $\mu_{2,n} \to \mu_2$ *and* $\mu_{4,n}/\mu_{2,n}^2 \to \kappa$ *for some* $\mu_2$ *and* $\kappa$ *such that* $(\mu_2, \kappa\mu_2^2)'$ *is in the interior of the set of values of* $E_F[(x^2, x^4)']$ *with* $F$ *ranging over all probability distributions. Suppose that, conditional on* $\theta$, $(\hat{\delta}, \hat{\sigma}, \hat{\mu}_2, \hat{\kappa})$ *converges in probability to* $(\delta, \sigma, \mu_2, \kappa)$. *Then the CIs in Eq.* (12) *with* $\hat{\sigma}_i = \hat{\sigma}$ *satisfy the ACI property* (15). *Furthermore, if, these conditions hold for* $\theta$ *in a probability one set,* $\theta_1, \ldots, \theta_n$ *follow an exchangeable distribution and the estimators* $\hat{\delta}$, $\hat{\sigma}$, $\hat{\mu}_2$ *and* $\hat{\kappa}$ *are exchangeable functions of the data* $(X_1', Y_1)', \ldots, (X_n', Y_n)'$, *then these CIs satisfy the empirical Bayes coverage property* (14).

The requirement that the moments $(\mu_2, \kappa\mu_2^2)'$ be in the interior of the set of feasible moments is needed to avoid degenerate cases such as when $\mu_2 = 0$, in which case the EBCI shrinks each estimate all the way to $X_i'\hat{\delta}$. Note also that the theorem doesn't require that $\hat{\delta}$ be the OLS estimate in a regression of $Y_i$ onto $X_i$, and that $\delta$ be the population analog; one can define $\delta$ in other ways, the theorem only requires that $\hat{\delta}$ be a consistent estimate of it. The definition of $\delta$ does, however, affect the plausibility of the moment independence assumption in Eq. (9) needed for conditional coverage results.

**Remark 4.1.** Typically, if CIs satisfy the average coverage condition (15) given $\theta_1, \ldots, \theta_n$, they will also satisfy the stronger condition

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

where $o_{P(\cdot|\theta)}(1)$ denotes a sequence that converges in probability to zero conditional on $\theta$ (Eq. (16) implies Eq. (15) since the left-hand side is uniformly bounded). This is analogous to the result in the empirical Bayes estimation setting that the difference between the squared error $\frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)^2$ and the MSE $E[\frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)^2 \mid \theta]$ typically converges to zero.

**Remark 4.2.** In the homoskedastic setting in Section 2, the CI asymptotically takes the form $\{\hat{\theta}_i \pm \zeta\}$ where $\hat{\theta}_i = w_{EB} Y_i$ and $\zeta = \chi w_{EB} \sigma$. Thus, Eq. (15) can be written as a bound on $\frac{1}{n}\sum_{i=1}^n P(|\hat{\theta}_i - \theta_i| > \zeta \mid \theta)$. This can be interpreted as the risk of the estimator $\hat{\theta}$ with compound loss defined using the 0-1 loss function $\ell(\theta_i, \hat{\theta}_i) = \mathrm{I}\{|\hat{\theta}_i - \theta_i| > \zeta\}$. The average coverage criterion states that the risk of the estimator $\hat{\theta}$ is bounded by $\alpha$ under this loss function. In the heteroskedastic setting in Section 3, a similar statement holds, but with $\zeta_i$ varying over $i$ so that the loss function varies with $i$.

## 4.2 Relative efficiency

The robust EBCI in Eq. (11) is inefficient relative to the parametric EBCI $\hat{\theta}_i \pm z_{1-\alpha/2}\sigma_i\sqrt{w_{EB,i}}$ when in fact the normality assumption (8) holds. We now quantify this inefficiency and show, in particular, that the amount of inefficiency is small unless the signal-to-noise ratio $\mu_2/\sigma_i^2$ is very small.

There are two reasons for the inefficiency relative to this normal benchmark. 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 shrinkage estimators, and perhaps relax the restriction that the shrinkage is linear.

We now quantify the inefficiency of the robust EBCI relative to the normal benchmark. We also 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, as discussed in Remark 3.4, we maintain assumption (9), and consider a more general class of estimators $\tilde{\theta}(w_i) = \mu_{1,i} + w_i(Y_i - \mu_{1i})$: we impose the requirement that the shrinkage is linear for tractability, but allow the amount of shrinkage $w_i$ to be optimally determined. The normalized bias is then given by $b_i = (1/w_i - 1)\varepsilon_i/\sigma_i$, which leads to the EBCI

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

The optimal amount of shrinkage $w_i$ minimizes the half-length $\mathrm{cva}_\alpha((1-1/w_i)^2\mu_2/\sigma_i^2, \kappa)w_i\sigma_i$ of this EBCI. Denote the minimizer by $w_{opt}(\mu_2/\sigma_i^2, \kappa, \alpha)$. Like $w_{EB,i}$, the optimal shrinkage depends on $\mu_2$ and $\sigma_i^2$ only through the signal-to-noise ratio $\mu_2/\sigma_i^2$. The resulting EBCI is optimal among all EBCIs based on linear estimators under (9), and we refer to it as the optimal robust EBCI.

Figure 2 plots the optimal shrinkage for $\kappa = \infty$ (which corresponds to not imposing any constraints on the fourth moment of $\varepsilon_i$), and $\kappa = 3$ (which is the case under the normal

benchmark). It is clear from the figure that relative to the normal benchmark, it is optimal to employ less shrinkage, and that the optimal amount of shrinkage depends on the coverage level $1-\alpha$ as well as moments of $\theta_i - \mu_{1,i}$.

Figure 3 plots the ratio of lengths of the optimal robust EBCI and robust EBCI relative to the parametric EBCI. The figure shows that for efficiency relative to the normal benchmark, for significance levels $\alpha = 0.1$ and $\alpha = 0.05$, it is relatively more important to impose the fourth moment constraint than to use the optimal amount of shrinkage (and only impose the second moment constraint). Furthermore, it is clear from the figure that the efficiency loss of the robust EBCI is modest unless the signal-to-noise ratio is very small: if $\mu_2/\sigma_i^2 \geq 0.1$, the efficiency loss is at most 12.3% for $\alpha = 0.05$, and 13.6% for $\alpha = 0.1$; up to half of the efficiency loss is due to not using the optimal shrinkage.

When the signal-to-noise ratio is very small, $\mu_2/\sigma_i^2 < 0.1$, the efficiency loss of the robust EBCI is higher (up to 39% for these significance levels). To keep the EBCIs short relative to the normal benchmark, it then becomes more important to use the optimal shrinkage $w_{opt}$, which 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 significantly tighter than the unshrunk CI $Y_i \pm z_{1-\alpha/2}\sigma_i$. To illustrate this point, Figure 4 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 maximal non-coverage probability of the parametric EBCI (13), given knowledge of only the second moment $\mu_2$ of $\varepsilon_i = Y_i - X_i'\delta$, is given by

$$\rho(\sigma_i^2/\mu_2, z_{1-\alpha/2}/\sqrt{w_{EB,i}}),$$

where $w_{EB,i} = \mu_2/(\mu_2 + \sigma_i^2)$. Here $\rho$ is the non-coverage function defined in Eq. (4), and for simplicity we pretend that $\mu_2$ and $\sigma_i$ are known. Note that the non-coverage probability depends on $\mu_2$ and $\sigma_i^2$ only through the signal-to-noise ratio $\mu_2/\sigma_i^2 = w_{EB,i}/(1-w_{EB,i})$.

Figure 5 plots the maximal non-coverage probability as a function of $w_{EB}$, for significance levels $\alpha = 0.05$ and $\alpha = 0.10$. If $w_{EB} \geq 0.3$, the maximal coverage distortion is less than 5 percentage points for these two significance levels. This justifies the "rule of thumb" proposed in Remark 3.1. The following lemma confirms that the maximal non-coverage is decreasing in $w_{EB}$, as suggested by the figure. Moreover, the lemma gives an expression for the maximal non-coverage across all values of $w_{EB}$ (which is achieved in the limit $w_{EB} \to 0$).

**Lemma 4.1.** *Define, for any $z > 0$, the function $\tilde{\rho}\colon (0,1] \to [0,1]$ given by*

$$\tilde{\rho}(w) = \rho(1/w - 1, z/\sqrt{w}), \quad 0 < w \leq 1.$$

*This function is weakly decreasing, and* $\sup_{w\in(0,1]} \tilde{\rho}(w) = 1/\max\{z^2, 1\}$.

Thus, for any significance level $\alpha \leq 2\Phi(-1) \approx 0.317$, the maximal non-coverage probability of the parametric EBCI across all possible distributions of $\varepsilon_i$ (with any second moment) is $1/z^2_{1-\alpha/2}$. This number equals 0.260 for $\alpha = 0.05$ and 0.370 for $\alpha = 0.10$. For $\alpha > 2\Phi(-1)$, the maximal non-coverage probability across all distributions is 1.

If we additionally impose knowledge of the kurtosis of $\varepsilon_i$, the maximal non-coverage of the parametric EBCI can be similarly computed using the function (10), as illustrated in the applications in Section 6.

# 5 Extensions: general shrinkage estimators

The ideas in Sections 2 and 3 go through for any shrinkage estimators $\hat{\theta}_i$ that follow an approximate normal distribution conditional on $\theta_i$. For simplicity, we consider in the main text the case where this holds exactly:

$$\frac{\hat{\theta}_i - \theta_i}{\text{se}_i} \,\Big|\, \theta \sim N(b_i, 1), \tag{17}$$

where $\text{se}_i$ is the standard error of the shrinkage estimator $\hat{\theta}_i$ and $b_i$ is the normalized bias. We relax the normality assumption in Appendix B. In our baseline implementation for the empirical Bayes setting, we used estimates of the second and fourth moments of the bias. More generally, letting $g : \mathbb{R} \to \mathbb{R}^p$ be some vector of moment functions, we can use estimates $\hat{m}$ of the empirical moments $m_n = \frac{1}{n}\sum_{i=1}^n g(b_i)$ of the normalized bias. This leads to the critical value $\text{cva}_{\alpha,g}(m) = \inf\{\chi : \rho_g(m,\chi) \leq \alpha\}$ where

$$\rho_g(m,\chi) = \sup_F E_F[r(b,\chi)] \quad \text{s.t.} \quad E_F[g(b)] = m. \tag{18}$$

This leads to the interval $\hat{\theta}_i \pm \text{cva}_{\alpha,g}(\hat{m})\text{se}_i$. The program (18) is an infinite-dimensional linear programming problem. Even with several constraints, its solution can be computed to high degree of precision by discretizing the support of $b$ and applying efficient finite-dimensional linear programming solution algorithms. See Appendix A for details.

More generally, we can condition the entire analysis on covariates (which could include $\text{se}_i$) when estimating the moments, as discussed in Remark 3.2, and we allow for this possibility

in our general results in Appendix B. The following theorem is a special case of Theorem B.1 in Appendix B.

**Theorem 5.1.** *Suppose that (17) holds and that $m_n \to m$ and $\hat{m}$ converges in probability to $m$ conditional on $\theta$, where $m$ is in the interior of the set of values of $E_F[g(b)]$ with $F$ ranging over all probability distributions. Suppose also that, for some $j$, $\lim_{b\to\infty} g_j(b) = \lim_{b\to-\infty} g_j(b) = \infty$ and $g_j(b) \geq 0$. Then the average coverage property (15) holds for the CIs $\hat{\theta}_i \pm \text{cva}_{\alpha,g}(\hat{m})se_i$ conditional on $\theta$.*

The assumption that $\lim_{b\to\infty} g_j(b) = \lim_{b\to-\infty} g_j(b) = \infty$ and $g_j(b) \geq 0$ for some $j$ is made so that the conditions on the empirical moments of the bias $\frac{1}{n}\sum_{i=1}^n g(b_i)$ place a strong enough bound on the bias so that the critical value is finite.

The normality assumption (17) will hold exactly if $\hat{\theta}_i$ is a linear function of jointly normal observations $W_1, \ldots, W_N$:

$$\hat{\theta}_i = \sum_{j=1}^{N} k_{ij} W_j \quad \text{for some deterministic weights } k_{ij}. \tag{19}$$

This holds for the shrinkage estimator $\hat{\theta}_i = w_{EB} Y_i$ when $Y_i \mid \theta_i \sim N(\theta_i, \sigma^2)$ as in Section 2. Another setting where such estimators are of interest is in nonparametric or regularized regression, in which $\theta_i = f(X_i)$ and $W_i$ follows the regression model $W_i = f(X_i) + \epsilon_i$ with $\epsilon_i \mid X_i \sim N(0, \sigma_i^2)$ (here we condition on the $X_i$'s so that the weights $k_{ij}$ may depend on $X_1, \ldots, X_n$). The Nadaraya-Watson estimator takes this form with $k_{ij} = \tilde{k}((X_i - X_j)/h_n) / \sum_{j=1}^n \tilde{k}((X_i - X_j)/h_n)$ for a kernel function $\tilde{k}$ and bandwidth $h_n$.

If (19) holds but $W_1, \ldots, W_N$ does not follow a normal distribution, then the normality condition (17) will not hold exactly but will hold approximately so long as the weights $k_{ij}$ satisfy a Lindeberg condition. A further complication is that the weights may depend on the data $W_1, \ldots, W_n$ through a preliminary estimate of a tuning parameter, as with the James and Stein (1961) estimate $\hat{w}_{EB} = (1 - (n-2)/\sum_{i=1}^n Y_i^2)$ of the mean squared error optimal weight $w_{EB}$ described in Section 2. In Appendix B, we provide high level conditions that allow for such complications, and we verify them for the empirical Bayes setting in Section 3.

More generally, our approach could be applied to other estimators that use shrinkage or regularization, so long as they can be expressed in the linear form (19) and so long as one can deal with the dependence of $k_{ij}$ on any data-driven tuning parameters. For example, regression trees take the linear form (19) with $k_{ij}$ depending on the choice of "leaves," which are typically chosen using data-driven methods such as cross-validation. In the regression trees setting and other more complicated settings, it may be difficult to characterize how the

linear weights $k_{ij}$ depend on the data, and methods such as sample splitting may provide a promising approach.

A substantive restriction of the normality condition (17) (or versions of this condition that require only approximate normality) is that it rules out estimators where non-linearity plays an essential form in shrinkage, rather than just through tuning parameters. For example, our approach rules out nonlinear estimators in the empirical Bayes setting of the form $\hat{\theta}_i = g(Y_i)$ for a nonlinear function $g(\cdot)$, such as the hard thresholding estimator $\hat{\theta}_i = Y_i \, \mathrm{I}\{|Y_i| > \varrho\}$ for some threshold $\varrho$.

# 6 Empirical applications

We illustrate our methods through two empirical applications: estimating (i) the effects of neighborhoods on intergenerational mobility, and (ii) the extent of structural changes in a large dynamic factor model (DFM) of the Eurozone economies.

## 6.1 Neighborhood effects

Our first application is based on the data and model in Chetty and Hendren (2018), who are interested in the effect of neighborhoods on intergenerational mobility. We adopt their main specification, which focuses on two definitions of a "neighborhood effect" $\theta_i$. The first defines it as the effect of spending one 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, but for children with parents at the 75th percentile. 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) exploit variation in the age at which children move between CZs to obtain preliminary fixed effect estimates $Y_i$ of $\theta_i$.

Since these preliminary estimates are measured with noise, to predict $\theta_i$, Chetty and Hendren (2018) shrink $Y_i$ towards 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 accuracy of these forecasts, Chetty and Hendren (2018) report estimates of the unconditional MSE associated with these forecasts (i.e. treating $\theta_i$ as random), under the implicit assumption that the moment independence assumption in Eq. (12) holds. Here we complement their analysis by constructing robust EBCIs associated with these forecasts.

### 6.1.1 Framework

Our sample consists of 595 U.S. CZs, with population over 25,000 in the 2000 census, which is the set of CZs for which Chetty and Hendren (2018) report baseline fixed effect estimates $Y_i$ of the effects $\theta_i$. These baseline estimates are normalized so that their population-weighted mean is zero. Thus, we may interpret the effects $\theta_i$ as being 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 $1/\hat{\sigma}_i^2$ when constructing the estimates $\hat{\delta}$, $\hat{\mu}_2$ and $\hat{\kappa}$ (see Remark 3.6). For comparison, we also report results based on shrinking the $t$-statistic, following Remark 3.8. Here we do not use precision weights for simplicity.

### 6.1.2 Results

Table 1 summarizes the main estimation and efficiency results. The shrinkage magnitude and relative efficiency results are similar for children with parents at the 25th and 75th percentiles of the income distribution, but there are interesting differences depending on whether we follow the baseline implementation, or shrink the $t$-statistics. In all four specifications reported in Table 1, the estimate of the kurtosis $\kappa$ is large enough so that it doesn't affect the critical values or the form of the optimal shrinkage: specifications that only impose constraints on the second moment yield identical results. In line with this finding, Supplemental Appendix C.1 gives a plot of the $t$-statistics showing that they exhibit a fat lower tail.

The robust EBCIs based on EB estimates are 47.9–87.7% 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 robust EBCIs in column (1) implies CIs of the form ±\$160 on average, while the unshrunk CIs are of the form ±\$643 on average. These large gains are a consequence of a low ratio of signal-to-noise $\mu_2/\sigma_i^2$ ($\mu_2$ in the case of $t$-statistic shrinkage) in this application. Consequently, in the specifications in columns (1) and (2), the shrinkage coefficient $w_{EB,i}$ falls below the threshold of 0.3 in our "rule of thumb" in Remark 3.1 for over 90% of the CIs, and for $t$-statistic shrinkage, the coefficient is also below the rule of thumb threshold (here the shrinkage doesn't vary across observations). Because the shrinkage magnitude is so large on average, the tail behavior of the bias matters, and since the kurtosis estimates suggests these tails are fat, it is important to use the robust critical value: the parametric EBCI exhibits average potential size distortions of 8.1–17.8 percentage points.

Because the precision of the fixed effect estimates $Y_i$ depends on the number of movers between CZs, there is considerable heterogeneity in the precision of these preliminary estimates: at the 25th (75th) percentile, the standard errors range between 0.042 and 6.57 (0.045 and 5.78). As a result, imposing assumption (9), and following our baseline implementation yields considerably shorter EBCIs on average relative to shrinking the $t$-statistic, with a large degree of shrinkage for smaller CZs, and the EBCIs being considerably tighter than the usual unshrunk CIs. Figure 6 plots the unshrunk CIs based on the preliminary estimates, as well as robust EBCIs based on EB estimates for New York for children with parents at the 25th percentile to illustrate this result. While the EBCIs for large CZs like New York City or Buffalo are similar to the unshrunk CIs, they are considerably tighter for smaller CZs like Plattsburgh or Watertown, with point estimates that shrink the preliminary estimates $Y_i$ considerably toward the regression line $X_i'\hat{\delta}$. See Supplemental Appendix C.1 for an analogous plot for the 75th percentile.

Due to the low signal-to-noise ratio and high kurtosis of the effects distribution, there is substantial gain from optimizing the degree of shrinkage to minimize the length of the robust EBCIs, as discussed in Remark 3.4. The robust EBCIs centered at the MSE-optimal degree of shrinkage $w_{EB}$ are on average 26.1–35.2% longer than the robust EBCIs that use the length-optimal degree of shrinkage $w_{opt}$.

In summary, using shrinkage allows us considerably tighten the CIs based on the preliminary estimates, even if we don't impose assumption (9) and shrink the $t$-statistics. This is true in spite of the fact that the CIs only effectively use second moment constraints—imposing constraints on the kurtosis does not affect the critical values.

## 6.2 Structural change in the Eurozone

Our second application constructs robust EBCIs for structural breaks in the factor loadings of a DFM. Specifically, we estimate a DFM on a large quarterly data set of several economic variables pertaining to each of the 19 current Eurozone countries. By estimating the model separately on the pre- and post-2009 samples and differencing, we are able to estimate the structural break, if any, in the loadings of each individual series on a common Eurozone-wide real activity factor. Our goal is to gauge whether the pattern of Eurozone co-movements changed substantially following the financial crisis of 2008–2009, and if so, which countries or economic variables tend to exhibit more breaks.

Because the break magnitudes are estimated on short samples, it is appealing to apply EB shrinkage to the unrestricted point estimates of the breaks.[8] We construct robust EBCIs to

[8] Koopman and Mesters (2017) show that EB methods can improve the accuracy of point estimators in the DFM context, although they do not consider the case of structural breaks. Fully Bayesian inference

complement the shrinkage point estimates. Based on these confidence intervals, we conclude that breaks have occurred in the loadings of financial series in several countries, whereas real business cycle indicators have not exhibited marked breaks in their relationships with the overall Eurozone real activity cycle. There is no tendency for "core" or "periphery" countries to experience a higher frequency of structural breaks. Our analysis is purely descriptive and is merely intended to illustrate the potential benefits of adopting an EB approach when doing inference on multiple structural breaks, each of which may be of substantive interest.

### 6.2.1 Data

We construct a quarterly data set of 13 economic variables for each of the 19 current Eurozone countries, spanning the years 1999–2018. The 13 variables fall into several categories, including familiar real business cycle variables, the current account, consumer confidence, goods and house prices, wages, asset prices, and credit aggregates. We supplement with aggregate data on oil prices (Brent), Eurozone short-term interest rates, and euro exchange rates versus each of five major currencies. To enhance cross-country comparability, we use only data sets maintained by Eurostat, the BIS, and the OECD, rather than supplementing with data from national statistical agencies. We discard any time series that are available for fewer than 15 years.

The resulting data set features 221 time series, 8 of which are Eurozone-wide. There are at least 7 country-specific variables available for every Eurozone country. We transform all variables to stationarity following similar conventions as in the rich U.S. data set constructed by Stock and Watson (2016). We impute a small number of missing observations by iterating on an initial preliminary factor model. The detailed list of countries, variables, and their construction is given in Supplemental Appendix C.2.

### 6.2.2 Model, estimation, and inference

We assume that the $n$ observed times series $z_{i,t}$ are driven by a small number $r$ of common factors $f_t = (f_{1,t}, \ldots, f_{r,t})'$, where $i = 1, \ldots, n$ and $t = 1, \ldots, T$. The data $z_{i,t}$ are given by the $n = 221$ country-specific or Eurozone-wide series described above. Specifically, we consider a Dynamic Factor model with a potential break in the factor loadings at a known date $t = T_0 + 1$ (= 2009q1):

$$z_{i,t} = \begin{cases} \lambda_i^{(0)\prime} f_t + \epsilon_{i,t} & \text{if } t = 1, \ldots, T_0, \\ \lambda_i^{(1)\prime} f_t + \epsilon_{i,t} & \text{if } t = T_0 + 1, \ldots, T. \end{cases}$$

methods are often applied to DFMs, see for example Bai and Wang (2015) and references therein.

Here $\lambda_i^{(0)}, \lambda_i^{(1)} \in \mathbb{R}^r$ are the pre- and post-break factor loadings, respectively. In our data, $T_0 = T - T_0 = 40$ quarters. We assume that the idiosyncratic errors $\epsilon_{i,t}$ are weakly correlated across $t$ and across $i$, as well as weakly correlated with the common factors $f_t$ (see Bai and Ng, 2008, for standard assumptions). We do not require $\epsilon_{i,t}$ to be stationary across the break date, and its variance, for example, is allowed to change after time $T_0$.

We normalize the first factor as being the latent factor driving Eurozone-wide GDP growth (the "named factor" normalization, cf. Stock and Watson, 2016). Let the scalar time series $s_t$ denote real aggregate GDP growth in the 19 current Eurozone countries. Then

$$s_t = f_{1,t} + u_t, \quad t = 1, \ldots, T, \tag{20}$$

where $u_t$ is weakly stationary within the two subsamples and uncorrelated with all factors $f_t$ and idiosyncratic errors $\epsilon_{i,t}$. That is, we identify Eurozone-wide GDP growth $s_t$ as being driven solely (and one-for-one) by the first latent factor $f_{1,t}$, which we thus interpret as an Eurozone-wide real activity factor. Because we are only interested in the loadings on this factor, we do not need further normalizations, except that we impose the conventional assumption that the $r$ factors are mutually uncorrelated.

Our parameters of interest are the structural breaks in the loadings of each series on the Eurozone-wide real activity factor $f_{1,t}$,

$$\theta_i = \lambda_{i,1}^{(1)} - \lambda_{i,1}^{(0)}, \quad i = 1, \ldots, n.$$

Following conventional practice (Stock and Watson, 2016), before analysis, all series $\{s_t\}$ and $\{z_{i,t}\}_t$, $i = 1, \ldots, n$, have been standardized to each have sample mean 0 and sample variance 1. Hence, the magnitudes of $\theta_i$ can be meaningfully compared across different series $i$. We estimate $\theta_i$ as follows:

1. Estimate the DFM separately on the two subsamples (before and after $T_0$) by applying principal components to the data $\{z_{i,t}\}_{i,t}$, cf. Stock and Watson (2016).[9] Let $\hat{f}_t^{(j)}$ denote the principal component factor estimates from subsamples $j = 0, 1$.

2. For each series $i = 1, \ldots, n$ and each subsample $j = 0, 1$, estimate $\lambda_i^{(j)}$ by running a two-stage least squares (2SLS) regression of $z_{i,t}$ on $s_t$, with the $r$ instruments given by $\hat{f}_t$. Call the coefficient estimate $\hat{\lambda}_i^{(j)}$.

3. Compute the preliminary estimator $Y_i = \hat{\lambda}_{i,1}^{(1)} - \hat{\lambda}_{i,1}^{(0)}$, $i = 1, \ldots, n$.

[9] We choose the number $r$ of factors using the "$IC_{p2}$" information criterion of Bai and Ng (2002). This criterion selects 5 and 4 factors on the early and late subsample, respectively, although the scree plot is flat around the optimum. Thus, we conservatively set $r = 6$ on both subsamples.

This estimator is consistent as $n, T_0, (T - T_0) \to \infty$ under conditions similar to those stated in Bai and Ng (2008), since (i) the principal components $\hat{f}_t$ consistently estimate the linear space spanned by the true factors $f_t$, and (ii) the fitted value from the first stage of the 2SLS regression is a consistent estimate of $f_{1,t}$ by the normalization (20). We compute standard errors for $\hat{\lambda}_{i,1}^{(j)}$ using the usual 2SLS formula, with a Newey-West correction for serial correlation of $u_t$ (bandwidth = 8 lags). The standard errors $\hat{\sigma}_i$ for $Y_i$ are obtained by assuming independence of the two subsamples (weak dependence would suffice in practice).

Due to the small sample size—10 years of quarterly data on each subsample—and because we are interested in inspecting the individual break magnitudes, we shrink the estimated breaks $Y_i$ toward 0 using EB methods. We consider the parametric and robust EB procedures, using the baseline implementation in Section 3.2 (with $X_i'\hat{\delta} = 0$). We choose to shrink toward 0 because the no-break case appears to be an economically relevant focal point. In contrast, shrinkage towards the grand mean implicitly rests on the prior belief that most loadings exhibit breaks in the same direction (positive or negative), which may not be plausible given our heterogeneous collection of variables and countries.[10] In addition to shrinking $Y_i$, which relies on the moment independence assumption (9), we also consider shrinking the $t$-statistic $Y_i/\hat{\sigma}_i$ as described in Remark 3.8.

We are not aware of any pre-existing work that uses EB methods for structural break testing in settings with many time series. Although fully Bayesian approaches to modeling structural breaks in DFMs exist, two advantages of our approach is (i) prior-robustness and (ii) frequentist average coverage control, as discussed earlier. Stock and Watson (2016, section 2.5) review classical and Bayesian approaches to estimating and testing for breaks in DFMs.

### 6.2.3 Results

Table 2 shows that in this application the robust EBCIs are close to the parametric EBCI, especially if both the estimated 2nd and 4th moments of the bias are exploited. Throughout we use a nominal coverage level of 95%.

Under our baseline shrinkage specification in Section 3.2, which imposes the moment independence assumption (9), the maximal coverage distortion of the parametric EBCI (averaged across series $i$) is at most 0.2 percentage points based on the estimated 2nd and 4th moments of the break distribution. Hence, if we impose both 2nd and 4th moments, the robust EBCIs (whether optimizing the length or not) only need to be very marginally wider than the parametric one to ensure the desired average coverage. This is consistent with

[10] This issue could be ameliorated by shrinking toward variable- or country-specific means, but these means would be estimated with less precision.

the "rule of thumb" mentioned in Remark 3.1, since the shrinkage factor $w_{EB}$ exceeds 0.3. In fact, the estimated kurtosis $\kappa$ of the loading break distribution is 2.994, consistent with normality.[11] Imposing the second and fourth moments of the break magnitude distribution leads to a non-trivial 10.0% reduction in the length of the robust EBCI, relative to only imposing the second moment.

If we instead use $t$-statistic shrinkage as discussed in Remark 3.8, the robust EBCIs are again close to the parametric ones. However, in this case exploiting information about the fourth moment of the break magnitude distribution leads to very little reduction in CI length.

On average, the unshrunk confidence intervals are 17.0–30.1% longer than the robust EBCIs that exploit fourth moments. The gain over the unshrunk interval is particularly large when we impose the moment independence assumption.

Figure 7 plots the shrinkage-estimated loading breaks and associated robust EBCIs. For clarity, we focus on three series: real GDP growth (GDP), changes in the 10-year government bond spread vis-à-vis the 3-month Eurozone interest rate (GOVBOND), and stock price index growth (STOCKP). Results for the remaining series are reported in Supplemental Appendix C.2, where we also explain the 2-letter country codes used in the figure. Recall that an estimated break magnitude of 0.5, say, means that the variable in question responds by 0.5 standard deviation units less to a one unit increase in the Eurozone-wide real activity factor in the post-2009 sample than in the pre-2009 sample. Because the substantive results are similar for baseline and $t$-statistic shrinkage, we show only the former here and relegate the latter to Supplemental Appendix C.2. We also focus on the robust EBCI that uses MSE-optimal weights and exploits 2nd and 4th moments.

While only one country (Malta) experiences a significant break in its real GDP loading, many countries experience breaks in the loadings on the two financial series. The government bond spread exhibits statistically significant breaks in 10 countries (in the sense that the EBCI excludes 0). Since all but one estimated break is negative, and the estimated pre-2009 loadings were negative in all countries, these spreads have become even more negatively related to the Eurozone-wide real activity factor following the financial crisis. Stock price indices exhibit significant breaks in 9 countries, but in this case the tendency has been for national indices to become less strongly (positively) correlated with Eurozone real activity. In Supplemental Appendix C.2, we show that CPI inflation has similarly become less positively correlated with Eurozone real activity. Moreover, some largest-in-magnitude *point estimates* of breaks occur for credit aggregates (especially credit to households, but also to non-financial businesses) in periphery countries; yet, many of these breaks are imprecisely

[11] Supplemental Appendix C.2 shows that the cross-sectional histogram of the $t$-statistics for $\hat{\theta}_i$ also appears to be consistent with approximate normality.

estimated according to the robust EBCI. As is the case for GDP growth, other traditional real business cycle indicators such as real consumption growth, capacity utilization, wage growth, and the unemployment rate do not undergo significant breaks in most countries.[12]

Moreover, we find that no particular group of countries exhibits especially many or few breaks following the financial crisis of 2008–2009. Each country has at least one significant break, and no country has more than six breaks (which is the case for Italy; Luxembourg, Netherlands, and Spain have five). In terms of the frequency of breaks, there does not appear to be any clear systematic differences between "periphery" versus "core" Eurozone countries, or between newer and older Eurozone members.

We conclude that the financial crisis of 2008–2009 was not associated with substantial breaks in the differential co-movement patterns of core and periphery Eurozone countries, but several financial variables do exhibit markedly different behavior post-crisis. Traditional real business cycle indicators have by and large not undergone significant breaks in terms of their relationship with the overall Eurozone real activity cycle. Although our analysis is purely descriptive, the results are consistent with chronologies of the Eurozone crisis that attach special importance to the effects of the European Central Bank's extraordinary interventions in financial markets.

## 6.3 Comparison of the two applications

The two applications demonstrate the usefulness of robust EBCIs in both cross-sectional and time series settings. In addition, the applications illustrate the range of potential shrinkage settings that can be encountered in practice, as we now discuss.

In the neighborhood effects application, the degree of shrinkage is high, so that our "rule of thumb" in Remark 3.1 indicates that the parametric EBCI could have severely distorted coverage. This is because the estimated signal-to-noise ratio is small. In the Eurozone DFM application, the estimated signal-to-noise ratio is somewhat higher. Hence, in this case, the rule of thumb indicates that the parametric EBCI has at worst very mild coverage distortions; however, it follows that using the robust EBCI costs very little in terms of efficiency, so applied researchers do not need to make tough decisions about when to use the parametric EBCI. As shown in Section 4.2, the relatively high signal-to-noise ratio in

[12] One possible spurious reason why financial series might seem to exhibit more breaks than non-financial series could be that these two different types of economic variables are fundamentally different and should not be pooled together in the shrinkage estimation. However, in the case of $t$-statistic shrinkage, the estimated values of $\sqrt{\mu_2} = 2.26$ and $\kappa = 2.47$ for the subset of 65 financial time series in our data set (excluding house prices) are similar to the full-sample estimates reported in Table 2. The estimated moments for financial series would lead to an increase in $w_{EB}$ of 14.7% (i.e., less shrinkage toward zero) and a modest 6.6% increase in the length of the robust EBCI (based on $w_{EB}$), relative to the full-sample estimates. The number of robust EBCIs that exclude 0 would increase from 30 to 31.

the Eurozone application implies that there is less to gain from optimizing the degree of shrinkage to minimize CI length than in the neighborhood effects application.

In the neighborhood effects application, imposing the estimated fourth moment of the effects distribution leads to no reduction in the critical value, relative to only using the estimated second moment. This is because the effects distribution appears to exhibit a fat lower tail and thus very high kurtosis. In contrast, in the Eurozone DFM application, the estimate of the kurtosis is consistent with near-normality of the effects distribution, so that imposing the fourth moment (in addition to the second) leads to a non-trivial reduction in the critical value, at least in the case of moment independence.

The difference between robust EBCIs that impose the moment independence assumption (9) and those that use $t$-statistic shrinkage is only pronounced in the neighborhood effects application. This is because the standard errors $\hat{\sigma}_i$ are much less heterogeneous in the Eurozone DFM application, so the moment independence assumption has less bite.

Most importantly, in both applications, the robust EBCI is easy to implement and is substantially shorter than the conventional unshrunk CI. The reduction in average length is as large as 87.7% in some specifications in the neighborhood effects application, and as large as 23.1% in the Eurozone DFM application.

# A Computational details

As in the main text, let $r(b,\chi) = \Phi(-\chi - b) + \Phi(-\chi + b)$. To simplify the statement of the results below, let $r_0(b,\chi) = r(\sqrt{b},\chi)$. We state and discuss the results first, with proofs relegated to the end of this appendix.

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. (4) has a simple solution:

**Proposition A.1.** *The solution to the problem*

$$\rho(m_2,\chi) = \sup_F E_F[r(b,\chi)] \quad s.t. \quad E_F[b^2] = m_2 \tag{21}$$

*is given by* $\rho(m_2,\chi) = \sup_{u\geq m_2}\{(1-m_2/u)r_0(0,\chi) + \frac{m_2}{u}r_0(u,\chi)\}$. *Let* $t_0 = 0$ *if* $\chi \leq \sqrt{3}$, *and otherwise let* $t_0 > 0$ *denote the solution to* $r_0(0,\chi) - r_0(u,\chi) + u\frac{\partial}{\partial u}r_0(u,\chi) = 0$. *This solution is unique, and the optimal* $u$ *satisfies* $u = m_2$ *for* $m_2 > t_0$ *and* $u = t_0$ *otherwise.*

The proof of Proposition A.1 shows that $\rho(m_2,\chi)$ is given by the least concave majorant of the function $r_0$. This majorant function can be computed via a univariate optimization problem given in the statement of Proposition A.1.

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

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

*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 A.1. If $m_2 < t_0$, then the solution is given by*

$$\inf_{0<x_0\leq t_0} \left\{ r_0(x_0, \chi) + (m_2 - x_0)\frac{\partial r_0(x_0, \chi)}{\partial x_0} + ((x_0 - m_2)^2 + (\kappa - 1)m_2^2) \sup_{0\leq x\leq t_0} \delta(x; x_0) \right\}, \tag{22}$$

*where $\delta(x; x_0) = \frac{r_0(x,\chi) - r_0(x_0,\chi) - (x-x_0)\frac{\partial r_0(x_0,\chi)}{\partial x_0}}{(x-x_0)^2}$ 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 doesn't help to reduce the maximal non-coverage probability, and $\rho(m_2, \kappa, \chi) = \rho(m_2, \chi)$.

**Remark A.1** (Least favorable distributions)**.** It follows from the proof of these propositions that distributions maximizing Eq. (21)—the least favorable distributions for the standardized 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, the probabilities are not uniquely determined: any distribution with these two support points maximizes Eq. (21). 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. (22).

**Remark A.2** (Certificate of optimality)**.** Since the optimization problem is a linear program, we can computationally verify that this solution is correct using duality theory, and we do this in our software implementation. In particular, the solution in the statement of Proposition A.2 is based on the solution to the dual. By the duality theorem, the value of the dual is necessarily greater than the value of the primal. Therefore, if the implied least favorable distribution discussed in Remark A.1 satisfies the primal constraints on the moments of $b$, and the implied non-coverage rate equals $\rho(m_2, \kappa, \chi)$, it follows that the value of the primal equals the value of the dual and the solution is correct. Alternatively, we can solve the primal directly by discretizing the support of $F$ on $[0, t_0]$ (in the proof of

Proposition A.2, we show that the solution is supported on this interval) using $K$ support points, for some large $K$. This turns the primal into a finite-dimensional linear program. Since discretizing the support can only lower the value of the primal, if the solution is numerically close to Eq. (22) (using some small numerical tolerance), it follows that this solution must be numerically close to correct.

Finally, the characterization of the solution to the general program in Eq. (18) depends on the form of the constraint function $g$. To solve the program numerically, one can 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

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

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

## A.1 Proof of Proposition A.1

Let $r_0(t,\chi) = r(\sqrt{t},\chi)$. Since $r(b,\chi)$ is symmetric in $b$, Eq. (21) is equivalent to maximizing $E_F[r_0(t,\chi)]$ over distributions $F$ of $t$ with $E_F[t] = m_2$. Let $\bar{r}(t,\chi)$ denote the least concave majorant of $r_0(t,\chi)$. We first show that $\rho(m_2,\chi) = \bar{r}(m_2,\chi)$.

Observe that $\rho(m_2,\chi) \leq \bar{\rho}(m_2,\chi)$, where $\bar{\rho}(m_2,\chi)$ denotes the value of the problem

$$\bar{\rho}(m_2,\chi) = \sup_F E_F[\bar{r}(t,\chi)] \quad \text{s.t.} \quad E_F[t] = m_2.$$

Furthermore, since $\bar{r}$ is concave, by Jensen's inequality, the optimal solution $F^*$ to this problem puts point mass on $m_2$, so that $\bar{\rho}(m_2,\chi) = \bar{r}(m_2,\chi)$, and hence $\rho(m_2,\chi) \leq \bar{r}(m_2,\chi)$.

Next, we show that the reverse inequality holds, $\rho(m_2,\chi) \geq \bar{r}(m_2,\chi)$. By Corollary 17.1.4 on page 157 in Rockafellar (1970), the majorant can be written as

$$\bar{r}(t,\chi) = \sup\{\lambda r_0(x_1,\chi) + (1-\lambda) r_0(x_2,\chi) \colon \lambda x_1 + (1-\lambda) x_2 = t,\ 0 \leq x_1 \leq x_2, \lambda \in [0,1]\}, \quad (23)$$

which corresponds to the problem in Eq. (21), with the distribution $F$ constrained to be a discrete distribution with two support points. Since imposing this additional constraint on $F$ must weakly decrease the value of the solution, it follows that $\rho(m_2,\chi) \geq \bar{r}(m_2,\chi)$. Thus, $\rho(m_2,\chi) = \bar{r}(m_2,\chi)$. The proposition then follows by Lemma A.2 below.

**Lemma A.1.** *Let $r_0(t,\chi) = r(\sqrt{t},\chi)$. If $\chi \leq \sqrt{3}$, then $r_0$ is concave in $t$. If $\chi > \sqrt{3}$, then its second derivative is positive for $t$ small enough, negative for $t$ large enough, and crosses*

*zero exactly once, at some* $t_1 \in [\chi^2 - 3, (\chi - 1/\chi)^2]$.

*Proof.* Letting $\phi$ denote the standard normal density, the first and second derivative of $r_0(t) = r_0(t, \chi)$ are given by

$$\begin{aligned} r_0'(t) &= \frac{1}{2\sqrt{t}}\Big[\phi(\sqrt{t} - \chi) - \phi(\sqrt{t} + \chi)\Big] \geq 0, \\ r_0''(t) &= \frac{\phi(\chi - \sqrt{t})(\chi\sqrt{t} - t - 1) + \phi(\chi + \sqrt{t})(\chi\sqrt{t} + t + 1)}{4t^{3/2}} \\ &= \frac{\phi(\chi + \sqrt{t})}{4t^{3/2}}\Big[e^{2\chi\sqrt{t}}(\chi\sqrt{t} - t - 1) + (\chi\sqrt{t} + t + 1)\Big] = \frac{\phi(\chi + \sqrt{t})}{4t^{3/2}} f(\sqrt{t}), \end{aligned}$$

where the last line uses $\phi(a + b)e^{-2ab} = \phi(a - b)$, and

$$f(u) = (\chi u + u^2 + 1) - e^{2\chi u}(u^2 - \chi u + 1).$$

Thus, the sign of $r_0''(t)$ corresponds to that of $f(\sqrt{t})$, with $r_0''(t) = 0$ if and only if $f(\sqrt{t}) = 0$. Observe $f(0) = 0$, and $f(u) < 0$ is negative for $u$ large enough, since the term $-u^2 e^{2\chi u}$ dominates. Furthermore,

$$\begin{aligned} f'(u) &= 2u + \chi - e^{2\chi u}(2\chi(u^2 - \chi u + 1) + 2u - \chi) & f'(0) &= 0 \\ f''(u) &= e^{2\chi u}(4\chi^3 u - 4\chi^2 u^2 - 8\chi u - 2) + 2 & f''(0) &= 0 \\ f^{(3)}(u) &= 4\chi e^{2\chi u}(2\chi^3 u + \chi^2(1 - 2u^2) - 6\chi u - 3) & f^{(3)}(0) &= 4\chi(\chi^2 - 3). \end{aligned}$$

Therefore for $u > 0$ small enough, $f(u)$, and hence $r_0''(u^2)$ is positive if $\chi^2 \geq 3$, and negative otherwise.

Now suppose that $f(u_0) = 0$ for some $u_0 > 0$, so that

$$\chi u_0 + u_0^2 + 1 = e^{2\chi u_0}(u_0^2 - \chi u_0 + 1) \tag{24}$$

Since $\chi u + u^2 + 1$ is strictly positive, it must be the case that $u_0^2 - \chi u_0 + 1 > 0$. Multiplying and dividing the expression for $f'(u)$ above by $u_0^2 - \chi u_0 + 1$ and plugging in the identity in Eq. (24) and simplifying the expression yields

$$\begin{aligned} f'(u_0) &= \frac{(u_0^2 - \chi u_0 + 1)(2u_0 + \chi) - (\chi u_0 + u_0^2 + 1)(2\chi(u_0^2 - \chi u_0 + 1) + 2u_0 - \chi)}{u_0^2 - \chi u_0 + 1} \\ &= \frac{2u_0^2\chi(\chi^2 - 3 - u_0^2)}{u_0^2 - \chi u_0 + 1}. \end{aligned} \tag{25}$$

Suppose $\chi^2 < 3$. Then $f'(u_0) < 0$ at all positive roots $u_0$ by Eq. (25). But if $\chi^2 < 3$, then

$f(u)$ is initially negative, so by continuity it must be that $f'(u_1) \geq 0$ at the first positive root $u_1$. Therefore, if $\chi^2 \leq 3$, $f$, and hence $r_0''$, cannot have any positive roots. Thus, if $\chi^2 \leq 3$, $r_0$ is concave as claimed.

Now suppose that $\chi^2 \geq 3$, so that $f(u)$ is initially positive. By continuity, this implies that $f'(u_1) \leq 0$ at its first positive root $u_1$. By Eq. (25), this implies $u_1 \geq \sqrt{\chi^2 - 3}$. As a result, again by Eq. (25), $f(u_i) \leq 0$ for all remaining positive roots. But since by continuity, the signs of $f'$ must alternate at the roots of $f$, this implies that $f$ has at most a single positive root. Since $f$ is initially positive, and negative for large enough $u$, it follows that it has a single positive root $u_1 \geq \sqrt{\chi^2 - 3}$. Finally, to obtain an upper bound for $t_1 = u_1^2$, observe that if $f(u_1) = 0$, then, by Taylor expansion of the exponential function,

$$1 + \frac{2\chi u_1}{\chi u_1 + u_1^2 + 1} = e^{2\chi u_1} \geq 1 + 2\chi u_1 + 2(\chi u_1)^2,$$

which implies that $1 \geq (1 + \chi u_1)(\chi u_1 + u_1^2 + 1)$, so that $u_1 \leq \chi - 1/\chi$. □

**Lemma A.2.** *The problem in Eq. (23) can be written as*

$$\bar{r}(t, \chi) = \sup_{u \geq t} \{(1 - t/u) r_0(0, \chi) + \frac{t}{u} r_0(u, \chi)\}. \tag{26}$$

*Let $t_0 = 0$ if $\chi \leq \sqrt{3}$, and otherwise let $t_0 > 0$ denote the solution to $r_0(0, \chi) - r_0(u, \chi) + u\frac{\partial}{\partial u} r_0(u, \chi) = 0$. This solution is unique, and the optimal $u$ solving Eq. (26) satisfies $u = t$ for $t > t_0$ and $u = t_0$ otherwise.*

*Proof.* If the optimization problem in Eq. (23), the constraint on $x_2$ binds, or either constraint on $\lambda$ binds, then the optimum is achieved at $r_0(t) = r_0(t, \chi)$, with $x_1 = t$ and $\lambda = 1$ and $x_2$ arbitrary; $x_2 = t$ and $\lambda = 0$ and $x_1$ arbitrary; or else $x_1 = x_2$ and $\lambda$ arbitrary. In any of these cases $\bar{r}$ takes the form in Eq. (26) as claimed. If, on the other hand, these constraints do not bind, then $x_2 > t > x_1$, and substituting $\lambda = (x_2 - t)/(x_2 - x_1)$ into the objective function yields the first-order conditions

$$r_0(x_2) - (x_2 - x_1) r_0'(x_1) - r_0(x_1) = \mu \frac{(x_2 - x_1)^2}{(x_2 - t)}, \tag{27}$$

$$r_0(x_2) + (x_1 - x_2) r_0'(x_2) - r_0(x_1) = 0, \tag{28}$$

where $\mu \geq 0$ is the Lagrange multiplier on the constraint that $x_1 \geq 0$. Subtracting Eq. (28) from Eq. (27) and applying the fundamental theorem of calculus then yields

$$\mu \frac{x_2 - x_1}{(x_2 - t)} = r_0'(x_2) - r_0'(x_1) = \int_{x_1}^{x_2} r_0''(t)\, dt > 0, \tag{29}$$

which implies that $\mu > 0$. Here the last inequality follows because by Taylor's theorem, Eq. (28) implies that $\int_{x_1}^{x_2} r_0''(t)(t - x_1)\,dt = 0$. Since $r_0''$ is positive for $t \le t_1$ and negative for $t \ge t_1$ by Lemma A.1, it follows that $x_1 \le t_1 \le x_2$, and hence that

$$0 = \int_{x_1}^{t_1} r_0''(t)(t - x_1)\,dt + \int_{t_1}^{x_2} r_0''(t)(t - x_1)\,dt$$
$$< (t_1 - x_1)\int_{x_1}^{t_1} r_0''(t)\,dt + (t_1 - x_1)\int_{t_1}^{x_2} r_0''(t)\,dt = (t_1 - x_1)\int_{x_1}^{x_2} r_0''(t)\,dt.$$

Finally Eq. (29) implies that $\mu > 0$, so that $x_1 = 0$ at the optimum. Consequently, the problem in Eq. (23) takes the form in Eq. (26) as claimed.

To show the second part of the Lemma A.2, note that by Lemma A.1, if $\chi \le \sqrt{3}$, $r_0$ is concave, so that we can put $u = t$ in Eq. (26). Otherwise, $\mu \ge 0$ denote the Lagrange multiplier associated with the constraint $u \ge t$ in the optimization problem in Eq. (26). The first-order condition is then given by

$$r_0(0) - r_0(u) + ur_0'(u) = \frac{-\mu u^2}{t}.$$

Let $f(u) = r_0(0) - r_0(u) + ur_0'(u)$. Since $f'(u) = ur_0''(u)$, it follows from Lemma A.1 that $f(u)$ is increasing for $u \le t_1$ and decreasing for $u \ge t_1$. Since $f(0) = 0$ and $\lim_{u\to\infty} f(u) < r_0(0) - 1 < 0$, it follows that $f(u)$ has exactly one positive zero, at some $t_0 > t_1$. Thus, if $t < t_0$, $u = t_0$ is the unique solution to the first-order condition. If $t > t_0$, $u = t$ is the unique solution. □

## A.2 Proof of Proposition A.2

Since $r(b, \chi)$ is symmetric in $b$, letting $t = b^2$, we can equivalently write the optimization problem as

$$\rho(m_2, \kappa, \chi) = \sup_F E_F[r_0(t, \chi)] \quad \text{s.t.} \quad E_F[t] = m_2,\ E_F[t^2] = \kappa m_2^2, \tag{30}$$

where $r_0(t, \chi) = r(\sqrt{t}, \chi)$, and the supremum is over all distributions supported on the positive part of the real line. The dual of this problem is

$$\min_{\lambda_0,\lambda_1,\lambda_2} \lambda_0 + \lambda_1 m_2 + \lambda_2 \kappa m_2^2 \qquad \text{s.t.} \quad \lambda_0 + \lambda_1 t + \lambda_2 t^2 \ge r_0(t), \quad 0 \le t < \infty,$$

where $\lambda_0$ the Lagrange multiplier associated with the implicit constraint that $E_F[1] = 1$, and $r_0(t) = r_0(t, \chi)$. So long as $\kappa > 1$ and $m_2 > 0$, so that the moments $(m_2, \kappa m_2^2)$ lie in the

interior of the space of possible moments of $F$, by the duality theorem in Smith (1995), the duality gap is zero, and if $F^*$ and $\lambda^* = (\lambda_0^*, \lambda_1^*, \lambda_2^*)$ are optimal solutions to the primal and dual problems, then $F^*$ has mass points only at those $t$ with $\lambda_0^* + \lambda_1^* t + \lambda_2^* t^2 = r(\sqrt{t}, \chi)$.

Define $t_0$ as in Lemma A.2. First, we claim that if $m_2 \geq t_0$, then $\rho(m_2, \kappa, \chi) = \rho(m_2, \chi)$, the value of the objective function in Proposition A.1. The reason that adding the constraint $E_F[t^2] = \kappa m_2^2$ doesn't change the optimum is that it follows from the proof of Proposition A.1 that the distribution achieving the rejection probability $\rho(m_2, \chi)$ is a point mass on $m_2$. Consider adding another support point $x_2 = \sqrt{n}$ with probability $\kappa m_2^2/n$, with the remaining probability on the support point $m_2$. Then, as $n \to \infty$, the mean of this distribution converges to $m_2$, and its second moment converges to $\kappa m_2^2$, so that the constraints in Eq. (30) are satisfied, while the rejection probability converges to $\rho(m_2, \chi)$. Since imposing the additional constraint $E_F[t^2] = \kappa m_2^2$ cannot increase optimum, the claim follows.

Suppose that $m_2 < t_0$. At optimum, the majorant $g(x) = \lambda_0 + \lambda_1 t + \lambda_2 t^2$ in the dual constraint must satisfy $g(x_0) = r_0(x_0)$ for at least one $x_0 > 0$. Otherwise, if the constraint never binds, we could lower the value of the objective function by decreasing $\lambda_0$; furthermore, $x_0 = 0$ cannot be the unique point at which the constraint binds, since by the duality theorem, this would imply that the distribution that puts point mass on 0 maximizes the primal, which cannot be the case.

At such $x_0$, we must also have $g'(x_0) = r_0'(x_0)$, otherwise the constraint would be locally violated. Using this fact together with the equality $g(x_0) = r_0(x_0)$, we therefore have that $\lambda_0 = r_0(x_0) - \lambda_1 x_0 - \lambda_2 x_0^2$ and $\lambda_1 = r_0'(x_0) - 2\lambda_2 x_0$, so that the dual problem may be written as

$$\min_{x_0 > 0, \lambda_2} r_0(x_0) + r_0'(x_0)(m_2 - x_0) + \lambda_2((x_0 - m_2)^2 + (\kappa - 1)m_2^2)$$
$$\text{s.t.} \quad r_0(x_0) + r_0'(x_0)(x - x_0) + \lambda_2(x - x_0)^2 \geq r_0(x). \quad (31)$$

Since $\kappa > 1$, the objective is increasing in $\lambda_2$. Therefore, given $x_0$, the optimal value of $\lambda_2$ is as small as possible while still satisfying the constraint,

$$\lambda_2 = \sup_{x > 0} \delta(x; x_0), \qquad \delta(x; x_0) = \frac{r_0(x) - r_0(x_0) - r_0'(x_0)(x - x_0)}{(x - x_0)^2}.$$

Next, we claim that the dual constraint cannot bind for $x_0 > t_0$. Observe that $\lambda_2 \geq 0$, otherwise the constraint would be violated for $t$ large enough. However, setting $\lambda_2 = 0$ still satisfies the constraint. This is because the function $h(x) = r_0(x_0) + r_0'(x_0)(x - x_0) - r_0(x)$ is minimized at $x = x_0$, with its value equal to 0. To see this, note that its derivative the equals zero if $r_0'(x_0) = r'(x)$. By Lemma A.1, $r_0'(t)$ is increasing for $t \leq t_0$ and decreasing for $t > t_0$.

Therefore, if $r_0'(x_0) < r_0'(0)$, $h'(x) = 0$ has a unique solution, $x = x_0$. If $r_0'(x_0) > r_0'(0)$, there is another solution at some $x_1 \in [0, t_0]$. However, $h''(x_1) = -r_0''(x_1) < 0$, so $h(x)$ achieves a local maximum here. Since $h(0) > 0$ by arguments in the proof of Lemma A.1, it follows that the maximum of $h(x)$ occurs at $x = x_0$, and equals 0. However, Eq. (31) cannot be maximized at $(x_0, 0)$, since by Proposition A.1, setting $(x_2, \lambda_2) = (t_0, 0)$ achieves a lower value of the objective function, which proves the claim.

Therefore, Eq. (31) can be written as

$$\min_{0<x_0\leq t_0} r_0(x_0) + r_0'(x_0)(m_2 - x_0) + ((x_0 - m_2)^2 + (\kappa - 1)m_2^2) \sup_{x\geq 0} \delta(x; x_0),$$

To finish the proof of the proposition, it remains to show that $\delta$ cannot be maximized at $x > t_0$. This follows from observing that the dual constraint in Eq. (31) binds at any $x$ that maximizes $\delta$. However, by the claim above, the constraint cannot bind for $x > t_0$.

# B Coverage results

This appendix provides coverage results that generalize Theorems 4.1 and 5.1, along with proofs of these results. In addition, we provide details and formal results for the $t$-statistic shrinkage approach described in Remark 3.8, and proof of Lemma 4.1. We start by this proof in Appendix B.1. Appendix B.2 generalizes the setup in the main text. Appendix B.3 provides results for general shrinkage estimators, from which Theorem 5.1 in the main text follows. Appendix B.4 considers a generalization of our baseline specification in the empirical Bayes setting, and proves a generalization of Theorem 4.1. Appendix B.5 provides details and formal results for the $t$-statistic shrinkage approach described in Remark 3.8. Appendix B.7 contains technical lemmas.

## B.1 Proof of Lemma 4.1

**Part (i).** Let $\Gamma(m)$ denote the space of probability measures on $\mathbb{R}$ with second moment bounded above by $m > 0$. By definition of the maximal non-coverage probability,

$$\tilde{\rho}(w) = \sup_{F\in\Gamma(1/w-1)} E_{b\sim F}\left[P(|b - Z| > z/\sqrt{w} \mid b)\right] = \sup_{F\in\Gamma(1/w-1)} P_{b\sim F}\big(\sqrt{w}|b - Z| > z\big), \quad (32)$$

where $Z$ denotes a $N(0, 1)$ variable that is independent of $b$.

Consider any $w_0, w_1$ such that $0 < w_0 \leq w_1 < 1$. Let $F_1^* \in \Gamma(1/w_1 - 1)$ denote the least-favorable distribution—i.e., the distribution that achieves the supremum (32)—when

$w = w_1$. (Proposition A.1 implies that the supremum is in fact attained at a particular discrete distribution.) Let $\tilde{F}_0$ denote the distribution of the linear combination

$$\sqrt{\frac{w_1}{w_0}}b - \sqrt{\frac{w_1 - w_0}{w_0}}Z$$

when $b \sim F_1^*$ and $Z \sim N(0,1)$ are independent. Note that the second moment of this distribution is $\frac{w_1}{w_0} \times \frac{1-w_1}{w_1} + \frac{w_1-w_0}{w_0} = \frac{1-w_0}{w_0}$, so $\tilde{F}_0 \in \Gamma(1/w_0 - 1)$. Thus, if we let $\tilde{Z}$ denote another $N(0,1)$ variable that is independent of $(b, Z)$, then

$$\begin{aligned}
\tilde{\rho}(w_0) &\geq P_{b\sim\tilde{F}_0}\left(\sqrt{w_0}|b - Z| > z\right) \\
&= P_{b\sim F_1^*}\left(\sqrt{w_0}\left|\sqrt{\frac{w_1}{w_0}}b - \sqrt{\frac{w_1 - w_0}{w_0}}\tilde{Z} - Z\right| > z\right) \\
&= P_{b\sim F_1^*}\left(\left|\sqrt{w_1}b - \underbrace{(\sqrt{w_1 - w_0}\tilde{Z} + \sqrt{w_0}Z)}_{\sim N(0,w_1)}\right| > z\right) \\
&= P_{b\sim F_1^*}\left(\sqrt{w_1}|b - Z| > z\right) \\
&= \tilde{\rho}(w_1).
\end{aligned}$$

**Part (ii).** It follows from Proposition A.1 that, if we define $r(b,\chi) = \Phi(-\chi - b) + \Phi(-\chi + b)$, then

$$\rho(t,\chi) = \sup_{0\leq\lambda\leq 1}(1-\lambda)r(0,\chi) + \lambda r((t/\lambda)^{1/2},\chi).$$

Note that $r(0, z/\sqrt{w}) \to 0$ as $w \to 0$. Thus,

$$\lim_{w\to 0}\tilde{\rho}(w) = \lim_{w\to 0}\rho\left(1/w - 1, z/\sqrt{w}\right) = \lim_{w\to 0}\sup_{0\leq\lambda\leq 1}\lambda r\left(\lambda^{-1/2}(1/w-1)^{1/2}, zw^{-1/2}\right),$$

provided the latter limit exists. We will first show that the supremum above is bounded below by an expression that tends to $1/\max\{z^2, 1\}$. Then we will show that the supremum is bounded above by an expression that tends to $1/z^2$ (and the supremum is obviously also bounded above by 1).

Let $\varepsilon(w) \geq 0$ be any function of $w$ such that $\varepsilon(w) \to 0$ and $\varepsilon(w)(1/w-1)^{1/2} \to \infty$ as $w \to 0$. Let $\tilde{z} = \max\{z, 1\}$. Note first that, by setting $\lambda = (\tilde{z}(1-w)^{-1/2} + \varepsilon(w))^{-2} \in [0,1]$,

$$\sup_{0\leq\lambda\leq 1}\lambda r\left(\lambda^{-1/2}(1/w-1)^{1/2}, zw^{-1/2}\right) \geq \frac{r\left((\tilde{z}(1-w)^{-1/2} + \varepsilon(w))(1/w-1)^{1/2}, zw^{-1/2}\right)}{(\tilde{z}(1-w)^{-1/2} + \varepsilon(w))^2} \to \frac{1}{\tilde{z}^2}$$

as $w \to 0$, since $r(b, \chi) \to 1$ when $(b - \chi) \to \infty$, and

$$(\tilde{z}(1-w)^{-1/2} + \varepsilon(w))(1/w - 1)^{1/2} - zw^{-1/2} \geq (z(1-w)^{-1/2} + \varepsilon(w))(1/w - 1)^{1/2} - zw^{-1/2}$$
$$= \varepsilon(w)(1/w - 1)^{1/2} \to \infty.$$

Second,

$$\sup_{0 \leq \lambda \leq 1} \lambda r\left(\lambda^{-1/2}(1/w - 1)^{1/2}, zw^{-1/2}\right)$$
$$\leq \Phi\left(-zw^{-1/2}\right) + \sup_{0 \leq \lambda \leq 1} \lambda \Phi\left(\lambda^{-1/2}(1/w - 1)^{1/2} - zw^{-1/2}\right).$$

The first term above tends to 0 as $w \to 0$. The second term above equals

$$\max\left\{ \sup_{0 \leq \lambda \leq (z - \varepsilon(w))^{-2}} \lambda \Phi\left(\lambda^{-1/2}(1/w - 1)^{1/2} - zw^{-1/2}\right), \right.$$
$$\left. \sup_{(z - \varepsilon(w))^{-2} < \lambda \leq 1} \lambda \Phi\left(\lambda^{-1/2}(1/w - 1)^{1/2} - zw^{-1/2}\right) \right\}. \quad (33)$$

The first argument to the maximum above is bounded above by

$$\sup_{0 \leq \lambda \leq (z - \varepsilon(w))^{-2}} \lambda = (z - \varepsilon(w))^{-2} \to \frac{1}{z^2}.$$

The second argument to the maximum in (33) tends to 0 as $w \to 0$, since

$$\lambda^{-1/2}(1/w - 1)^{1/2} - zw^{-1/2} \leq (\lambda^{-1/2} - z)(1/w - 1)^{1/2} \leq -\varepsilon(w)(1/w - 1)^{1/2}$$

for all $\lambda > (z - \varepsilon(w))^{-2}$, and the far right-hand side above tends to $-\infty$ as $w \to 0$.

## B.2 General setup and notation

Let $\hat{\theta}_1, \ldots, \hat{\theta}_n$ be estimates of parameters $\theta_1, \ldots, \theta_n$, with standard errors $\mathrm{se}_1, \ldots, \mathrm{se}_n$. 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 $\chi_1, \ldots, \chi_n$, which may be chosen based on the data. In some cases, we will condition on a variable $\tilde{X}_i$ when defining empirical Bayes coverage or average coverage. Let $\tilde{X}^{(n)} = (X_1, \ldots, X_n)'$.

As discussed in Section 4.1, the average coverage criterion does not require thinking of $\theta$ 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 $\theta$ and $\tilde{X}^{(n)}$. We will use $E_{\tilde{P}}$ to denote expectation under the measure $\tilde{P}$. We can then obtain empirical Bayes coverage statements by considering a distribution $P$ for the data and $\theta$, $\tilde{X}^{(n)}$ and an additional variable $\nu$ such that these conditions hold for the measure $\tilde{P}(\cdot) = P(\cdot \mid \theta, \nu, \tilde{X}^{(n)})$ for $\theta, \nu, \tilde{X}^{(n)}$ in a probability one set. The variable $\nu$ 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 empirical Bayes 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 $\mathcal{I}_{\mathcal{X},n} = \{i \in \{1, \ldots, n\} \colon \tilde{X}_i \in \mathcal{X}\}$, and let $N_{\mathcal{X},n} = \#\mathcal{I}_{\mathcal{X},n}$. Let $\tilde{X}^{(n)} = (\tilde{X}_1, \ldots, \tilde{X}_n)'$ and let $\chi^{(n)} = (\chi_1, \ldots, \chi^{(n)})'$. The sample average non-coverage on the set $\mathcal{X}$ is then given by

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

where $Z_i = (\hat{\theta}_i - \theta_i)/\mathrm{se}_i$.

We consider the following notions of average coverage control, conditional on the set $\mathcal{X} \in \mathcal{A}$:

$$ANC(\chi; \mathcal{X}) \leq \alpha + o_{\tilde{P}}(1), \tag{34}$$

and

$$\limsup_n E_{\tilde{P}}\left[ANC_n(\chi; \mathcal{X})\right] = \limsup_n \frac{1}{N_{\mathcal{X},n}} \sum_{i \in \mathcal{I}_{\mathcal{X},n}} \tilde{P}(|Z_i| > \chi_i) \leq \alpha. \tag{35}$$

Note that (34) implies (35), since $ANC_n(\chi; \mathcal{X})$ is uniformly bounded. Furthermore, if we integrate with respect to some distribution on $\nu, \tilde{X}^{(n)}$ such that (35) holds with $\tilde{P}(\cdot) = P(\cdot \mid \theta, \nu, \tilde{X}^{(n)})$ almost surely, we get (again by uniform boundedness)

$$\limsup_n E\left[ANC_n(\chi; \mathcal{X}) \mid \theta\right] \leq \alpha,$$

which, in the case where $\mathcal{X}$ contains all $\tilde{X}_i$'s with probability one, is condition (15) from the main text.

Now consider empirical Bayes coverage, as defined in (14) in the main text, but conditioning on the variable $\tilde{X}_i$. We consider empirical Bayes coverage under a distribution $P$ for the data, $\tilde{X}^{(n)}$, $\theta$ and $\nu$, where $\nu$ includes additional nuisance parameters and covariates, and

where the average coverage condition (35) holds with $P(\cdot \mid \theta, \nu, \tilde{X}^{(n)})$ playing the role of $\tilde{P}$ with probability one. Consider the case where $\tilde{X}_i$ is discretely distributed under $P$. Suppose that the exchangeability condition

$$P(\theta_i \in CI_i \mid \mathcal{I}_{\{\tilde{x}\},n}) = P(\theta_j \in CI_j \mid \mathcal{I}_{\{\tilde{x}\},n}) \text{ for all } i, j \in \mathcal{I}_{\{\tilde{x}\},n} \tag{36}$$

holds with probability one. Then, for each $j$,

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

Plugging in $P(\cdot \mid \theta, \nu, \tilde{X}^{(n)})$ for $\tilde{P}$ in the coverage condition (35), 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 $1 - \alpha$. By uniform boundedness of this term, it then follows that

$$\liminf_{n \to \infty} P(\theta_j \in CI_j \mid \tilde{X}_j = \tilde{x}) \geq 1 - \alpha. \tag{37}$$

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

## B.3 Results for general shrinkage estimators

We assume that the $Z_i$'s are approximately normal with variance one 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. We use $\Phi(\cdot)$ to denote the standard normal cdf.

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

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

*for all $t \in \mathbb{R}$ and, for all $\mathcal{X} \in \mathcal{A}$ and any $\varepsilon > 0$,*

$$\frac{1}{N_{\mathcal{X},n}} \sum_{i \in \mathcal{I}_{\mathcal{X},n}} \tilde{P}\left(|\tilde{b}_i - b_{i,n}| \geq \varepsilon\right) \to 0.$$

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

As in the main text, define $r(b,\chi) = \Phi(-\chi - b) + 1 - \Phi(\chi - b) = \Phi(-\chi - b) + \Phi(-\chi + b)$.

**Lemma B.1.** *Under Assumption B.1, we have, for any deterministic $\chi_1, \ldots, \chi_n$, and any $\mathcal{X} \in \mathcal{A}$ with $N_{\mathcal{X},n} \to \infty$,*

$$\lim_{n\to\infty} \frac{1}{N_{\mathcal{X},n}} \sum_{i \in \mathcal{I}_{\mathcal{X},n}} \tilde{P}\left(|Z_i| > \chi_i\right) - \frac{1}{N_{\mathcal{X},n}} \sum_{i \in \mathcal{I}_{\mathcal{X},n}} r(b_{i,n}, \chi_i) = 0.$$

*Furthermore, if $Z_i - \tilde{b}_i$ is independent over $i$ under $\tilde{P}$, then*

$$\frac{1}{N_{\mathcal{X},n}} \sum_{i \in \mathcal{I}_{\mathcal{X},n}} I\left(|Z_i| > \chi_i\right) - \frac{1}{N_{\mathcal{X},n}} \sum_{i \in \mathcal{I}_{\mathcal{X},n}} r(b_{i,n}, \chi_i) = o_{\tilde{P}}(1).$$

*Proof.* For any $\varepsilon > 0$, $\frac{1}{N_{\mathcal{X},n}} \sum_{i \in \mathcal{I}_{\mathcal{X},n}} I\left(|Z_i| > \chi_i\right)$ is bounded from above by

$$\frac{1}{N_{\mathcal{X},n}} \sum_{i \in \mathcal{I}_{\mathcal{X},n}} I\left(|Z_i - \tilde{b}_i + b_{i,n}| > \chi_i - \varepsilon\right) + \frac{1}{N_{\mathcal{X},n}} \sum_{i \in \mathcal{I}_{\mathcal{X},n}} I\left(|\tilde{b}_i - b_{i,n}| \geq \varepsilon\right).$$

The expectation under $\tilde{P}$ of the second term converges to zero by Assumption B.1. The expectation under $\tilde{P}$ of the first term is $\frac{1}{N_{\mathcal{X},n}} \sum_{i \in \mathcal{I}_{\mathcal{X},n}} \tilde{r}_{i,n}(b_{i,n}, \chi_i - \varepsilon)$ where $\tilde{r}_{i,n}(b,\chi) = \tilde{P}(Z_i - \tilde{b}_i < -\chi - b) + 1 - \tilde{P}(Z_i - \tilde{b}_i \leq \chi - b)$. Note that $r_{i,n}(b,\chi)$ converges to $r(b,\chi)$ uniformly over $b, \chi$ under Assumption B.1, using the fact that the convergence in Assumption B.1 is uniform in $t$ by Lemma 2.11 in van der Vaart (1998), and the fact that $\tilde{P}(Z_i - \tilde{b}_i < -\chi - b) = \lim_{t \uparrow -\chi - b} P(Z_i - \tilde{b}_i \leq t)$. It follows that the expectation of the above display under $\tilde{P}$ is bounded by $\frac{1}{N_{\mathcal{X},n}} \sum_{i \in \mathcal{I}_{\mathcal{X},n}} \tilde{r}(b_{i,n}, \chi_i - \varepsilon) + o(1)$. If $Z_i - \tilde{b}_i$ is independent over $i$, the variance of each term in the above display converges to zero, so that the above display equals $\frac{1}{N_{\mathcal{X},n}} \sum_{i \in \mathcal{I}_{\mathcal{X},n}} \tilde{r}(b_{i,n}, \chi_i - \varepsilon) + o_{\tilde{P}}(1)$. Taking $\varepsilon \to 0$ and noting that $r(b,\chi)$ is uniformly continuous in both arguments, and using an analogous argument with a lower bound, gives the result. □

Let $g : \mathbb{R} \to \mathbb{R}^p$ be a vector of moment functions. We consider critical values $\hat{\chi}^{(n)} = (\hat{\chi}_1, \ldots, \hat{\chi}_n)$ 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 B.2.** *For all $\mathcal{X} \in \mathcal{A}$, $N_{\mathcal{X},n} \to \infty$ and*

$$\frac{1}{N_{\mathcal{X},n}} \sum_{i \in \mathcal{I}_{\mathcal{X},n}} g(b_{n,i}) - \frac{1}{N_{\mathcal{X},n}} \sum_{i \in \mathcal{I}_{\mathcal{X},n}} m(\tilde{X}_i) \to 0$$

*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 B.3.** *For every $\mathcal{X} \in \mathcal{A}$ and every $\varepsilon > 0$, there is a partition $\mathcal{X}_1, \ldots, \mathcal{X}_J \in \mathcal{A}$ of $\mathcal{X}$ and $m_1, \ldots, m_J$ 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 B.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 $\text{cva}_{\alpha,g}(m)$ be defined as in Section 5:

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

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

**Theorem B.1.** *Suppose that Assumptions B.1, B.2, B.3 and B.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 $ANC_n(\hat{\chi}^{(n)}; \mathcal{X}) \leq \alpha + o_{\tilde{P}}(1)$.*

To prove this theorem, we begin with some lemmas regarding $\rho_g$.

**Lemma B.2.** *$\rho_g(\chi; m)$ is continuous in $\chi$. Furthermore, for any $m^*$ in the interior of the set of values of $\int g(b)\, dF(b)$, where $F$ ranges over all probability measures on $\mathbb{R}$, $\rho_g(\chi; m)$ is continuous with respect to $m$ at $m^*$.*

*Proof.* To show continuity with respect to $\chi$, note that

$$|\rho_g(\chi; m) - \rho_g(\tilde{\chi}; m)| \leq \sup_F \left| \int [r(b, \chi) - r(b, \tilde{\chi})]\, dF(b) \right| \quad \text{s.t.} \int g(b)\, dF(b) = m,$$

where we use the fact that the difference between suprema of two functions over the same constraint set is bounded by the supremum of the absolute difference of the two functions.

The above display is bounded by $\sup_b |r(b,\chi) - r(b,\tilde{\chi})|$, which is bounded by a constant times $|\tilde{\chi} - \chi|$ by uniform continuity of the standard normal CDF.

To show continuity with respect to $m$, note that, by Lemma B.5, the conditions for the Duality Theorem in Smith (1995, p. 812) hold for $m$ in a small enough neighborhood of $m^*$, so that

$$\rho_g(\chi; m) = \inf_{\lambda_0, \lambda} \lambda_0 + \lambda' m \text{ s.t. } \lambda_0 + \lambda' g(b) \geq r(b, \chi) \text{ all } b \in \mathbb{R}$$

and the above optimization problem has a finite solution. Thus, for $m$ in this neighborhood of $m^*$, $\rho_g(\chi; m)$ is the infimum of a collection of affine functions of $m$, which implies that it is concave function of $m$ (Boyd and Vandenberghe, 2004, p. 81). By concavity, $\rho_g(\chi; m)$ is also continuous as a function of $m$ in this neighborhood of $m^*$. □

**Lemma B.3.** *Let $M$ be any compact subset of the interior of the set of values of $\int g(b)\, dF(b)$, where $F$ ranges over all measures on $\mathbb{R}$ with the Borel $\sigma$-algebra. Suppose that $\lim_{b\to\infty} g_j(b) = \lim_{b\to-\infty} g_j(b) = \infty$ and $\inf_b g_j(b) \geq 0$ for some $j$. Then $\lim_{\chi\to\infty} \sup_{m\in M} \rho_g(\chi; m) = 0$ and $\rho_g(\chi; m)$ is uniformly continuous with respect to $(\chi, m')'$ on the set $[0, \infty) \times M$.*

*Proof.* The first claim (that $\lim_{\chi\to\infty} \sup_{m\in M} \rho_g(\chi; m) = 0$) follows by Markov's inequality and compactness of $M$. Given $\varepsilon > 0$, let $\overline{\chi}$ be large enough so that $\rho_g(\chi; m) < \varepsilon$ for all $\chi \in [\overline{\chi}, \infty)$ and all $m \in M$. By Lemma B.2, $\rho_g(\chi; m)$ is continuous on $[0, \overline{\chi} + 1] \times M$, so, since $[0, \overline{\chi} + 1] \times M$ is compact, it is uniformly continuous on this set. Thus, there exists $\delta$ such that, for any $\chi, m$ and $\tilde{\chi}, \tilde{m}$ with $\chi, \tilde{\chi} \leq \overline{\chi} + 1$ and $\|(\tilde{\chi}, \tilde{m}')' - (\chi, m')'\| \leq \delta$, we have $|\rho_g(\chi; m) - \rho_g(\tilde{\chi}; \tilde{m})| < \varepsilon$. If we also set $\delta < 1$, then, if either $\chi \geq \overline{\chi} + 1$ or $\tilde{\chi} \geq \overline{\chi} + 1$ we must have both $\chi \geq \overline{\chi}$ and $\tilde{\chi} \geq \overline{\chi}$, so that $\rho_g(\tilde{\chi}; \tilde{m}) < \varepsilon$ and $\rho_g(\chi; m) < \varepsilon$, which also implies $|\rho_g(\chi; m) - \rho_g(\tilde{\chi}; \tilde{m})| < \varepsilon$. This completes the proof. □

For any $\varepsilon > 0$, let

$$\overline{\rho}_g(\chi; m, \varepsilon) = \sup_{\tilde{m} \in B_\varepsilon(m)} \rho_g(\chi; \tilde{m})$$

and

$$\underline{\rho}_g(\chi; m, \varepsilon) = \inf_{\tilde{m} \in B_\varepsilon(m)} \rho_g(\chi; \tilde{m}).$$

**Lemma B.4.** *Let $M$ be any compact subset of the interior of the set of values of $\int g(b)\, dF(b)$, where $F$ ranges over all measures on $\mathbb{R}$ with the Borel $\sigma$-algebra and suppose $\lim_{b\to\infty} g_j(b) = \lim_{b\to-\infty} g_j(b) = \infty$ and $\inf_b g_j(b) \geq 0$ for some $j$. Then, for $\varepsilon$ smaller than a constant that depends only on $M$, the functions $\overline{\rho}_g(\chi; m, \varepsilon)$ and $\underline{\rho}_g(\chi; m, \varepsilon)$ are continuous in $\chi$. Furthermore, we have $\lim_{\varepsilon\to 0} \sup_{\chi\in[0,\infty), m\in M} [\overline{\rho}_g(\chi; m, \varepsilon) - \underline{\rho}_g(\chi; m, \varepsilon)] = 0$.*

*Proof.* For $\varepsilon$ smaller than a constant that depends only on $M$, the set $\cup_{m\in M} B_\varepsilon(m)$ is contained in another compact subset of the interior of the set of values of $\int g(b)\, dF(b)$, where

$F$ ranges over all measures on $\mathbb{R}$ with the Borel $\sigma$-algebra. The result then follows from Lemma B.3, where, for the first claim, we use the fact that $|\overline{\rho}_g(\chi; m, \varepsilon) - \overline{\rho}_g(\tilde{\chi}; m, \varepsilon)| \leq \sup_{\tilde{m} \in B_\varepsilon(m)} |\rho_g(\chi; \tilde{m}) - \rho_g(\tilde{\chi}; \tilde{m})|$ and similarly for $\underline{\rho}_g$. □

Given $\mathcal{X} \in \mathcal{A}$ and $\varepsilon > 0$, let $m_1, \ldots, m_J$ and $\mathcal{X}_1, \ldots, \mathcal{X}_J$ be as in Assumption B.3. Let $\underline{\chi}_j = \min\{\chi \colon \underline{\rho}_g(\chi; m_j, 2\varepsilon) \leq \alpha\}$. For $\hat{m}_i \in B_{2\varepsilon}(m_j)$, we have $\underline{\rho}_g(\chi; m_j, 2\varepsilon) \leq \rho_g(\chi; \hat{m}_i)$ for all $\chi$, so that, using the fact that $\underline{\rho}_g(\chi; m_j, 2\varepsilon)$ and $\rho_g(\chi; \hat{m}_i)$ are weakly decreasing in $\chi$, we have $\underline{\chi}_j \leq \hat{\chi}_i$. Thus, letting $\underline{\tilde{\chi}}^{(n)}$ denote the sequence with $i$th element equal to $\underline{\chi}_j$ when $\tilde{X}_i \in \mathcal{X}_j$, we have

$$\begin{aligned} ANC_n(\hat{\chi}^{(n)}; \mathcal{X}) &\leq \max_{1 \leq j \leq J} ANC_n(\underline{\tilde{\chi}}^{(n)}; \mathcal{X}_j) \\ &\leq \max_{1 \leq j \leq J} \left[ \frac{1}{N_{\mathcal{X}_j, n}} \sum_{i \in \mathcal{I}_{\mathcal{X}_j, n}} \mathrm{I}\{\hat{m}_i \notin B_{2\varepsilon}(m_j)\} + \frac{1}{N_{\mathcal{X}_j, n}} \sum_{i \in \mathcal{I}_{\mathcal{X}_j, n}} \mathrm{I}\{|Z_i| > \underline{\chi}_j\} \right]. \end{aligned}$$

The first term is bounded by $\frac{1}{N_{\mathcal{X}_j, n}} \sum_{i \in \mathcal{I}_{\mathcal{X}_j, n}} \mathrm{I}\{\|\hat{m}_i - m(\tilde{X}_i)\| > \varepsilon\}$ since, for $i \in \mathcal{I}_{\mathcal{X}_j, n}$, we have $\|\hat{m}_i - m_j\| \leq \varepsilon + \|\hat{m}_i - m(\tilde{X}_i)\|$. This converges in probability (and expectation) to zero under $\tilde{P}$ by Assumption B.2. By Lemma B.1, the second term is equal to, letting $F_{j,n}$ denote the empirical distribution of the $b_{i,n}$'s for $i$ with $x_i \in \mathcal{X}_j$,

$$\int r(b, \underline{\chi}_j) \, dF_{j,n}(b) + R_n \leq \overline{\rho}_g(\underline{\chi}_j; \mu_j, 2\varepsilon) + R_n$$

where $R_n$ is a term such that $E_{\tilde{P}} R_n \to 0$ and such that, if $Z_i - \tilde{b}_i$ is independent over $i$ under $\tilde{P}$, then $R_n$ converges in probability to zero under $\tilde{P}$. The result will now follow if we can show that $\max_{1 \leq j \leq J} [\overline{\rho}_g(\underline{\chi}_j; \mu_j, 2\varepsilon) - \alpha]$ can be made arbitrarily small by making $\varepsilon$ small. This holds by Lemma B.4 and the fact that $\underline{\rho}_g(\underline{\chi}_j; \mu_j, 2\varepsilon) \leq \alpha$ by construction.

## B.4 Empirical Bayes shrinkage toward regression estimate

We now apply the general results in Appendix B.3 to the empirical Bayes setting. As in Section 3, we consider unshrunk estimates $Y_1, \ldots, Y_n$ of parameters $\theta = (\theta_1, \ldots, \theta_n)'$, along with regressors $X^{(n)} = (X_1, \ldots, X_n)$ and variables $\tilde{X}^{(n)} = (\tilde{X}_1, \ldots, \tilde{X}_n)'$, which include $\sigma_i$ and which play the role of the conditioning variables. (While Section 3 uses $X_i, \sigma_i$ as the conditioning variable $\tilde{X}_i$, here we generalize the results by allowing the conditioning variables to differ from $X_i$.) The initial estimate $Y_i$ has variance $\sigma_i$, 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, $\sigma_i$'s and $\theta_i$'s are fixed. Empirical Bayes coverage can then be

obtained for a distribution $P$ of the data, $\theta$ and some nuisance parameter $\tilde{\nu}$ such that these conditions hold almost surely with $P(\cdot \mid \theta, \tilde{\nu}, \tilde{X}^{(n)}, X^{(n)})$ playing the role of $\tilde{P}$.

We consider the following generalization of the baseline specification considered in the main text. Let

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

where $\hat{X}_i$ is an estimate of $X_i$ (we allow for the possibility that some elements of $X_i$ are estimated rather than observed directly, which will be the case, for example, when $\sigma_i$ 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 $\mathrm{se}_i = w(\hat{\gamma}, \hat{\sigma}_i)\hat{\sigma}_i$ so that the $t$-statistic is

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

We use estimates of moments of order $\ell_1 < \cdots < \ell_p$ of the bias, where $\ell_1 < \cdots < \ell_p$ are positive integers. Let $\hat{\mu}_\ell$ be an estimate of the $\ell$th moment of $(\theta_i - X_i'\delta)$, and suppose that this moment is independent of $\sigma_i$ in a sense formalized below. Then an estimate of the $\ell_j$th 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 empirical Bayes CI is then given by $\hat{\theta}_i \pm w(\hat{\gamma}, \hat{\sigma}_i)\hat{\sigma}_i \cdot \mathrm{cva}_{\alpha,g}(\hat{m}_i)$ 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 $w(\hat{\mu}_2, \hat{\sigma}_i) = \hat{\mu}_2/(\hat{\mu}_2 + \hat{\sigma}_i)$.

We make the following assumptions.

**Assumption B.5.**

$$\lim_{n\to\infty} \max_{1\le i\le n} \left| \tilde{P}\left( \frac{Y_i - \theta_i}{\hat{\sigma}_i} \le t \right) - \Phi(t) \right| = 0.$$

We give primitive conditions for Assumption B.5 in Appendix B.6. This involves 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 $\theta_i$ to be a scaled version of the corresponding parameter.

**Assumption B.6.** *The standard deviations $\sigma_i$ are bounded away from zero. In addition, for some $\delta$ and $\gamma$, $\hat{\delta}$ and $\hat{\gamma}$ converge to $\delta$ and $\gamma$ under $\tilde{P}$, and, for any $\varepsilon > 0$,*

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

**Assumption B.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* $\sigma_i$*). Furthermore, for some* $\mu_0$ *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*

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

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

**Theorem B.2.** *Let* $\hat{\theta}_i$ *and* $se_i$ *be given above and let* $\hat{\chi}_i = \text{cva}_{\alpha,g}(\hat{m}_i)$ *where* $\hat{m}_i$ *is given above and* $g(b) = (b^{\ell_1}, \ldots, b^{\ell_p})$ *for some positive integers* $\ell_1, \ldots, \ell_p$*, at least one of which is even. Suppose that Assumptions B.5, B.6 and B.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 B.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* $ANC_n(\hat{\chi}^{(n)}; \mathcal{X}) \leq \alpha + o_{\tilde{P}}(1)$*.*

To prove Theorem B.2, we will verify the conditions of Theorem B.1 with $\mathcal{A}$ given above, $m_j(\tilde{X}_i) = c(\gamma, \sigma_i)^{\ell_j} \mu_{0,\ell_j}$ and $\tilde{b}_i = c(\hat{\gamma}, \hat{\sigma}_i)(\theta_i - \hat{X}_i'\hat{\delta})$ and $b_{i,n} = c(\gamma, \sigma_i)(\theta_i - \hat{X}_i'\delta)$ where $c(\gamma, \sigma) = \frac{w(\gamma,\sigma)-1}{w(\gamma,\sigma)\sigma}$. The first part of Assumption B.1 is immediate from Assumption B.5 since $Z_i - \tilde{b}_i = (Y_i - \theta_i)/\hat{\sigma}_i$. For the second part, we have

$$\begin{aligned} \tilde{b}_i - b_{i,n} &= c(\hat{\gamma}, \hat{\sigma}_i)(\theta_i - \hat{X}_i'\hat{\delta}) - c(\gamma, \sigma_i)(\theta_i - X_i'\delta) \\ &= [c(\hat{\gamma}, \hat{\sigma}_i) - c(\gamma, \sigma_i)](\theta_i - X_i'\delta) + c(\hat{\gamma}, \hat{\sigma}_i) \cdot [(\hat{X}_i - X_i)'\hat{\delta} - X_i'(\delta - \hat{\delta})]. \end{aligned}$$

For $\|\theta_i\| + \|X_i\| \leq C$, the above expression is bounded by

$$[c(\hat{\gamma}, \hat{\sigma}_i) - c(\gamma, \sigma_i)] \cdot (\|\delta\| + 1) \cdot C + c(\hat{\gamma}, \hat{\sigma}_i) \left[ \|\hat{\delta} - \delta\| \cdot C + \|\hat{X}_i - X_i\| \cdot (C + \|\hat{\delta} - \delta\|) \right].$$

By uniform continuity of $c()$ on an open set containing $\{\gamma\} \times \mathcal{S}_1$, for every $\varepsilon > 0$ there exists $\eta > 0$ such that $\|(\hat{\sigma}_i - \sigma_i, \hat{\gamma} - \gamma, \hat{\delta}' - \delta', \hat{X}_i' - X_i')'\| \leq \eta$ implies that the absolute value of the above display is less than $\varepsilon$. Thus, for any $\mathcal{X} \in \mathcal{A}$,

$$\lim_{n \to \infty} \frac{1}{N_{\mathcal{X},n}} \sum_{i \in \mathcal{I}_{\mathcal{X},n}} \tilde{P}(|\tilde{b}_i - b_{i,n}| \geq \varepsilon)$$

$$\leq \lim_{n\to\infty} \frac{1}{N_{\mathcal{X},n}} \sum_{i\in\mathcal{I}_{\mathcal{X},n}} \tilde{P}(\|(\hat{\sigma}_i - \sigma_i, \hat{\gamma} - \gamma, \hat{\delta}' - \delta', \hat{X}_i' - X_i')'\| > \eta)\, \mathrm{I}\{\|\theta_i\| + \|X_i\| \leq C\}$$
$$+ \limsup_{n\to\infty} \frac{1}{N_{\mathcal{X},n}} \sum_{i\in\mathcal{I}_{\mathcal{X},n}} \mathrm{I}\{\|\theta_i\| + \|X_i\| > C\}.$$

The first limit is zero by Assumption B.6. The last limit converges to zero as $C \to \infty$ by the second part of Assumption B.7 and Markov's inequality. This completes the verification of Assumption B.5.

We now verify Assumption B.2. Given $\mathcal{X} \in \mathcal{A}$ and given $\varepsilon > 0$, we can partition $\mathcal{X}$ into sets $\mathcal{X}_1, \ldots, \mathcal{X}_J$ such that, for some $c_1, \ldots, c_J$, we have $|c(\gamma, \sigma_i)^{\ell_k} - c_j^{\ell_k}| < \varepsilon$ for all $k = 1, \ldots, p$ whenever $i \in \mathcal{I}_{\mathcal{X}_j,n}$ for some $j$. Thus, for each $j$ and $k$,

$$\begin{aligned}
\frac{1}{N_{\mathcal{X}_j,n}} \sum_{i\in\mathcal{I}_{\mathcal{X}_j,n}} b_{i,n}^{\ell_k} - m_k(\tilde{X}_i) &= \frac{1}{N_{\mathcal{X}_j,n}} \sum_{i\in\mathcal{I}_{\mathcal{X}_j,n}} c(\gamma,\sigma_i)^{\ell_k} \left[(\theta_i - X_i'\delta)^{\ell_k} - \mu_{0,\ell_k}\right] \\
&= c_j^{\ell_k} \cdot \frac{1}{N_{\mathcal{X}_j,n}} \sum_{i\in\mathcal{I}_{\mathcal{X}_j,n}} \left[(\theta_i - X_i'\delta)^{\ell_k} - \mu_{0,\ell_k}\right] \\
&\quad + \frac{1}{N_{\mathcal{X}_j,n}} \sum_{i\in\mathcal{I}_{\mathcal{X}_j,n}} \left[c(\gamma,\sigma_i)^{\ell_k} - c_j^{\ell_k}\right] \left[(\theta_i - X_i'\delta)^{\ell_k} - \mu_{0,\ell_k}\right].
\end{aligned}$$

Under Assumption B.7, the first term converges to 0 and the second term is bounded up to an $o(1)$ term by $\varepsilon$ times a constant that depends only on $K$. Since the absolute value of $\frac{1}{N_{\mathcal{X},n}} \sum_{i\in\mathcal{I}_{\mathcal{X},n}} b_{i,n}^{\ell_k} - m_k(\tilde{X}_i)$ is bounded by the maximum over $j$ of the absolute value of the above display, and since $\varepsilon$ can be chosen arbitrarily small, the first part of Assumption B.2 follows.

For the second part of Assumption B.2, we have $\hat{m}_{i,k} - m_k(\tilde{X}_i) = c(\gamma,\sigma_i)\hat{\mu}_{\ell_j} - c(\gamma,\sigma_i)^{\ell_j}\mu_{0,\ell_j}$. By uniform continuity of $(\tilde{\gamma}', \sigma, \mu_{\ell_1}, \ldots, \mu_{\ell_p})' \mapsto (c(\gamma,\sigma_i)^{\ell_1}\mu_{\ell_1}, \ldots, c(\gamma,\sigma_i)^{\ell_p}\mu_{\ell_p})'$ in an open set containing $\{\gamma\} \times \mathcal{S}_1 \times \{(\mu_{0,\ell_1}, \ldots, \mu_{0,\ell_p})'\}$, for any $\varepsilon > 0$, there exists $\eta > 0$ such that $\|(\hat{\gamma}' - \gamma', \hat{\sigma}_i - \sigma, \hat{\mu}_{\ell_1} - \mu_{0,\ell_1}, \ldots, \hat{\mu}_{\ell_p} - \mu_{0,\ell_p})\| < \eta$ implies $\|\hat{m}_{i,k} - m_k(\tilde{X}_i)\| < \varepsilon$. Thus,

$$\max_{1\leq i\leq n} \tilde{P}(\|\hat{m}_i - m(\tilde{X}_i)\| \geq \varepsilon) \leq \max_{1\leq i\leq n} \tilde{P}(\|(\hat{\gamma}' - \gamma', \hat{\sigma}_i - \sigma, \hat{\mu}_{\ell_1} - \mu_{0,\ell_1}, \ldots, \hat{\mu}_{\ell_p} - \mu_{0,\ell_p})\| < \eta),$$

which converges to zero by Assumptions B.6 and B.7. This completes the verification of Assumption B.2.

Assumption B.3 follows immediately from compactness of the set $\mathcal{S}_1 \times \cdots \times \mathcal{S}_1$ and uniform continuity of $m()$ on this set. Assumption B.4 follows from Assumption B.7 and Lemma B.6. This completes the proof of Theorem B.2.

As a consequence of Theorem B.2, we obtain, under the exchangeability condition (36),

conditional empirical Bayes coverage, as defined in (37), for any distribution $P$ of the data and $\theta, \tilde{\nu}$ such that the conditions of Theorem B.2 hold with probability one with the sequence of probability measures $P(\cdot \mid \theta, \tilde{\nu}, X^{(n)}, \tilde{X}^{(n)})$ playing the role of $\tilde{P}$. This follows from the arguments in Appendix B.2.

**Corollary B.1.** *Let $\theta, \nu, X^{(n)}, \tilde{X}^{(n)}, Y_i$ follow a sequence of distributions $P$ such that the conditions of Theorem B.2 hold with $\tilde{X}_i$ taking on finitely many values, and $P(\cdot \mid \theta, \nu, X^{(n)}, \tilde{X}^{(n)})$ playing the role of $\tilde{P}$ with probability one, and such that the exchangeability condition (36) 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 empirical Bayes coverage condition (37).*

## B.5 $t$-statistic shrinkage

This appendix provides details on the $t$-statistic shrinkage approach discussed in Remark 3.8. Let $W_i = Y_i/\hat{\sigma}_i$ and let $\tau_i = \theta_i/\sigma_i$. Let $\hat{X}_i'\hat{\delta}$ be a regression estimate where $\hat{\delta}$ is an estimate of a regression parameter $\delta$ (typically the limit or probability limit of $(\sum_{i=1}^n X_i X_i')^{-1} \sum_{i=1}^n X_i \tau_i$, although we do not impose this). We apply the approach in Appendix B.4 with $W_i$ in place of $Y_i$ and $\tau_i$ in place of $\theta_i$, which leads to the estimate

$$\hat{\tau}_i = \hat{X}_i'\hat{\delta} + \hat{w} \cdot (W_i - \hat{X}_i'\hat{\delta})$$

for $\tau_i$, where $\hat{w}$ is a shrinkage coefficient, which is an estimate of some unknown constant $w > 0$ (for example, the choice $\hat{w} = \hat{\mu}_2/(\hat{\mu}_2 + 1)$, with $\hat{\mu}_2$ an estimate of the second moment of $\theta_i - X_i'\delta$, optimizes mean squared error for estimating $\theta_i/\tau_i$). The standard error of this estimate is $\hat{w}$, so that an interval for $\tau_i$ with critical value $\chi$ is given by $\{\hat{\tau}_i \pm \hat{w} \cdot \chi\}$. This leads to the interval $\{\hat{\theta}_i \pm \mathrm{se}_i \cdot \chi\}$ where $\hat{\theta}_i = \hat{\tau}_i\hat{\sigma}_i$ and $\mathrm{se}_i = \hat{w}\hat{\sigma}_i$. Then

$$Z_i = \frac{\hat{\theta}_i - \theta_i}{\mathrm{se}_i} = \frac{\hat{\sigma}_i \cdot \hat{X}_i'\hat{\delta} + \hat{\sigma}_i \cdot \hat{w} \cdot (W_i - \hat{X}_i'\hat{\delta}) - \theta_i}{\hat{w}\hat{\sigma}_i} = \frac{Y_i - \theta_i}{\hat{\sigma}_i} + \frac{\hat{w} - 1}{\hat{w}} \left( \frac{\theta_i}{\hat{\sigma}_i} - \hat{X}_i'\hat{\delta} \right).$$

Let $\tilde{b}_i = \frac{\hat{w}-1}{\hat{w}} \left( \frac{\theta_i}{\hat{\sigma}_i} - \hat{X}_i'\hat{\delta} \right)$ and let $b_{i,n} = \frac{w-1}{w} \left( \frac{\theta_i}{\sigma_i} - X_i'\delta \right)$. Let $g(b) = (b^{\ell_1}, \ldots, b^{\ell_p})'$ where $\ell_1, \ldots, \ell_p$ are as in Appendix B.4. Let $\hat{\mu}$ be an estimate of the (unconditional) moments of $\frac{\theta_i}{\sigma_i} - X_i'\delta$. This leads to the estimate $\hat{m}_j = [(\hat{w} - 1)/\hat{w}]^{\ell_j} \hat{\mu}_{\ell_j}$ of the $\ell_j$th moment of the $b_{i,n}$'s. Let $\hat{m} = (\hat{m}_1, \ldots, \hat{m}_p)'$ and let $\hat{\chi} = \mathrm{cva}_{\alpha,g}(\hat{m})$. We consider unconditional average coverage, and we verify the conditions of Theorem B.1 with $\mathcal{A}$ containing only one set, which contains all observations.

We use conditions similar to those in Appendix B.4, but we replace Assumption B.7 with the following assumption, which does not impose any independence between the conditional

moments of the $b_{i,n}$ and $\sigma_i$.

**Assumption B.8.** *For some $\mu_0$ 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 $K$, we have, for each $j = 1, \ldots, p$*

$$\frac{1}{n}\sum_{i=1}^{n}(\theta_i/\sigma_i - \delta_0' X_i)^{\ell_j} \to \mu_{0,\ell_j}, \quad \limsup_n \frac{1}{n}\sum_{i=1}^{n}|\theta_i|^{\ell_j} \le K, \quad \limsup_n \frac{1}{n}\sum_{i=1}^{n}\|X_i\|^{\ell_j} \le K.$$

*and $\hat{\mu}_{\ell_j}$ converges in probability to $\mu_{0,\ell_j}$ under $\tilde{P}$.*

**Theorem B.3.** *Let $\hat{\theta}_i$, $se_i$ and $\hat{\chi}_i$ be defined above, and suppose that Assumptions B.5, B.6 and B.8 hold. Suppose $\hat{w}$ converges in probability to $w > 0$ under $\tilde{P}$. Then $\frac{1}{n}\sum_{i=1}^{n}\tilde{P}(\theta_i \notin \{\hat{\theta}_i \pm se_i \cdot \hat{\chi}\}) \le \alpha + o(1)$. If, in addition, $(Y_i, \hat{\sigma}_i)$ is independent over $i$ under $\tilde{P}$, then $\frac{1}{n}\sum_{i=1}^{n}\mathrm{I}\{\theta_i \notin \{\hat{\theta}_i \pm se_i \cdot \hat{\chi}\}\} \le \alpha + o_{\tilde{P}}(1)$.*

To prove Theorem B.3, we verify the conditions of Theorem B.1. The first part of Assumption B.1 is immediate from Assumption B.5. For the second part, we have

$$\tilde{b}_i - b_{i,n} = \frac{\hat{w}-1}{\hat{w}}\left(\frac{\theta_i}{\hat{\sigma}_i} - \hat{X}_i'\hat{\delta}\right) - \frac{w-1}{w}\left(\frac{\theta_i}{\sigma_i} - X_i'\delta\right) = f(\hat{w}, \hat{\sigma}_i, \hat{X}_i, \hat{\delta}, \theta_i) - f(w, \sigma_i, X_i, \delta, \theta_i)$$

where $f(w, \sigma_i, X_i, \delta, \theta_i) = \frac{w-1}{w}\left(\frac{\theta_i}{\sigma_i} - X_i'\delta\right)$ is uniformly continuous on any compact set on which $\sigma_i$ is bounded away from zero. Let $C > 0$ be given. It follows that, for any $\varepsilon > 0$, there exists $\eta$ such that $\|(\hat{\sigma}_i - \sigma_i, \hat{w} - w, \hat{X}_i' - X_i', \hat{\delta}' - \delta')'\| \le \eta$ and $\|\theta_i\| + \|X_i\| \le C$ implies $|\tilde{b}_i - b_{i,n}| < \varepsilon$ (where we use the fact that $\hat{\sigma}_i$ and $\sigma_i$ are bounded away from zero once $\eta$ is small enough by Assumption B.6). Thus,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\tilde{P}(|\tilde{b}_i - b_{i,n}| \ge \varepsilon) \\
&\qquad \le \frac{1}{n}\sum_{i=1}^{n}\tilde{P}(\|(\hat{\sigma}_i - \sigma_i, \hat{w} - w, \hat{X}_i' - X_i', \hat{\delta}' - \delta')'\| > \eta)\,\mathrm{I}\{\|\theta_i\| + \|X_i\| \le C\|\} \\
&\qquad\qquad + \frac{1}{n}\sum_{i=1}^{n}\mathrm{I}\{\|\theta_i\| + \|X_i\| > C\|\}.
\end{aligned}$$

The first term converges to zero by Assumption B.6 and the assumption that $\hat{w}$ converges in probability to $w$. The last term can be made arbitrarily small by Assumption B.8. This completes the verification of Assumption B.1.

For Assumption B.2, letting $m_j = [(w-1)/w]^{\ell_j}\mu_{0,\ell_j}$, it is immediate from Assumption B.8 and the assumption that $\hat{w}$ converges in probability to $\hat{w}$ that $\hat{m}$ converges to $m$ under $\tilde{P}$,

which gives the second part of Assumption B.2. Furthermore, the first part of Assumption B.2 holds since

$$\frac{1}{n}\sum_{i=1}^{n} b_{n,i}^{\ell_j} - m_j = \left(\frac{w-1}{w}\right)^{\ell_j} \frac{1}{n}\sum_{i=1}^{n}\left[(\theta_i/\sigma_i - X_i'\delta)^{\ell_j} - \mu_{0,\ell_j}\right] \to 0$$

by Assumption B.8.

Assumption B.3 is vacuous since there is no covariate $\tilde{X}_i$ and $m = m(\tilde{X}_i)$ takes on only one value. Assumption B.4 holds by Lemma B.6. This completes the proof of Theorem B.3.

## B.6 Primitive Conditions for Assumption B.5

To verify Assumption B.5, we will typically have to define $\theta_i$ to be scaled by a rate of convergence. Let $\tilde{Y}_i$ be an estimator of a parameter $\beta_{i,n}$ with rate of convergence $\kappa_n$ and asymptotic variance estimate $\hat{\sigma}_i^2$. Suppose that

$$\lim_{n\to\infty} \max_{1\le i\le n} \sup_{t\in\mathbb{R}} \left| P\left(\frac{\kappa_n(\tilde{Y}_i - \beta_{i,n})}{\hat{\sigma}_i} \le t\right) - \Phi(t)\right| = 0. \tag{38}$$

Then Assumption B.5 holds with $\theta_i = \kappa_n\beta_{i,n}$ and $Y_i = \kappa_n\tilde{Y}_i$. Consider an affine estimator $\hat{\beta}_i = a_i/\kappa_n + w_i\tilde{Y}_i = (a_i + w_iY_i)/\kappa_n$ with standard error $\widetilde{\text{se}}_i = w_i\hat{\sigma}_i/\kappa_n$. The corresponding affine estimator of $\theta_i$ is $\hat{\theta}_i = \kappa_n\hat{\beta}_i = a_i + w_iY_i$ with standard error $\text{se}_i = \kappa_n \cdot \widetilde{\text{se}}_i = w_i\hat{\sigma}_i$. Then $\beta_{i,n} \in \{\hat{\beta}_i \pm \widetilde{\text{se}}_i \cdot \hat{\chi}_i\}$ iff. $\theta_i \in \{\hat{\theta}_i \pm \text{se}_i \cdot \hat{\chi}_i\}$. Thus, Theorem B.2 guarantees average coverage of the intervals $\{\hat{\beta}_i \pm \widetilde{\text{se}}_i \cdot \hat{\chi}_i\}$ for $\beta_{i,n}$. Note that, in order for the moments of $\theta_i$ to converge to a non-degenerate constant, we will need to consider triangular arrays $\beta_{i,n}$ that converge to zero at a $\kappa_n$ rate.

As an example, consider the case where the estimate is a sample mean: $\tilde{Y}_i = \bar{X}_i = \frac{1}{T_{i,n}}\sum_{i=1}^{T_{i,n}} X_{i,t}$, where $X_{i,t}$ is a sequence of random variables that is independent across both $i$ and $t$ and identically distributed across $i$ with the same $t$, with mean $\beta_{i,n} = EX_{i,t}$. Letting $s_i^2$ denote the variance of $X_{i,t}$ and $\hat{s}_i^2$ the sample variance, we can then define $\kappa_n^2 = \bar{T}_n = \frac{1}{n}\sum_{i=1}^{n} T_{i,n}$ and $\hat{\sigma}_i^2 = \hat{s}_i^2\bar{T}_n/T_{i,n}$ so that

$$\frac{\kappa_n(\tilde{Y}_i - \beta_{i,n})}{\hat{\sigma}_i} = \frac{\sqrt{T_{i,n}}(\tilde{Y}_i - \beta_{i,n})}{\hat{s}_i}.$$

If $\min_{1\le i\le n} T_{i,n} \to \infty$ and the family of distributions of $X_{i,t} - \beta_{i,n}$ satisfy the uniform integrability condition (11.77) in Lehmann and Romano (2005), then (38) holds by applying Theorem 11.4.4 in Lehmann and Romano (2005) along any sequence $i_n$.

## B.7 Technical lemmas

**Lemma B.5.** *Suppose that $\mu$ is in the interior of the set of values of $\int g(b)\,dF(b)$ as $F$ ranges over all probability measures with respect to the Borel sigma algebra, where $g : \mathbb{R} \to \mathbb{R}^p$. Then $(1, \mu')'$ is in the interior of the set of values of $\int (1, g(b)')'\,dF(b)$ as $F$ ranges over all measures with respect to the Borel sigma algebra.*

*Proof.* Let $\mu$ be on the interior of the set of values of $\int g(b)\,dF(b)$ as $F$ ranges over all probability measures with respect to the Borel sigma algebra. We need to show that, for any $a, \tilde{\mu}$ with $(a, \tilde{\mu}')'$ close enough to $(1, \mu')$, there exists a measure $F$ such that $\int (1, g(b)')dF(b) = (a, \tilde{\mu}')'$. To this end, note that, $\tilde{\mu}/a$ can be made arbitrarily close to $\mu$ by making $(a, \tilde{\mu}')'$ close to $(1, \mu')$. Thus, for $(a, \tilde{\mu}')'$ close enough to $(1, \mu')$, there exists a probability measure $\tilde{F}$ with $\int g(b)\,d\tilde{F}(b) = \tilde{\mu}/a$. Let $F$ be the measure defined by $F(A) = a\tilde{F}(A)$ for any measurable set $A$. Then $\int (1, g(b)')'dF(b) = a \int (1, g(b)')'d\tilde{F}(b) = (a, \tilde{\mu})$. This completes the proof. □

**Lemma B.6.** *Suppose that, as $F$ ranges over all probability measures with respect to the Borel sigma algebra, $(\mu_{\ell_1}, \ldots, \mu_{\ell_p})'$ is in the interior of the set of values of $\int (b^{\ell_1}, \ldots, b^{\ell_p})'\,dF(b)$ Let $c \in \mathbb{R}$. Then, as $F$ ranges over all probability measures with respect to the Borel sigma algebra, $(c^{\ell_1}\mu_{\ell_1}, \ldots, c^{\ell_p}\mu_{\ell_p})'$ is also in the interior of the set of values of $\int (b^{\ell_1}, \ldots, b^{\ell_p})'\,dF(b)$.*

*Proof.* We need to show that, for any vector $r$ with $\|r\|$ small enough, there exists a probability measure $F$ such that $\int (b^{\ell_1}, \ldots, b^{\ell_p})'\,dF(b) = (c^{\ell_1}\mu_{\ell_1} + r_1, \ldots, c^{\ell_p}\mu_{\ell_p} + r_p)'$. Let $\tilde{\mu}_{\ell_k} = \mu_{\ell_k} + r_k/c^{\ell_k}$. For $\|r\|$ small enough, there exists a probability measure $\tilde{F}$ with $\int b^{\ell_k}\,dF(b) = \tilde{\mu}_{\ell_k}$ for each $k$. Let $F$ denote the probability measure of $cB$ when $B$ is a random variable distributed according to $\tilde{F}$. Then $\int b^{\ell_k}\,dF(b) = c^{\ell_k} \int b^{\ell_k}\,d\tilde{F} = c^{\ell_k}\tilde{\mu}_{\ell_k} = c^{\ell_k}\mu_{\ell_k} + r_k$ as required. □

# References


Angrist, J. D., Hull, P. D., Pathak, P. A., and Walters, C. R. (2017). Leveraging Lotteries for School Value-Added: Testing and Estimation. *The Quarterly Journal of Economics*, 132(2):871–919.

Armstrong, T. B. and Kolesár, M. (2018). Optimal Inference in a Class of Regression Models. *Econometrica*, 86(2):655–683.

Bai, J. and Ng, S. (2002). Determining the Number of Factors in Approximate Factor Models. *Econometrica*, 70(1):191–221.

Bai, J. and Ng, S. (2008). Large Dimensional Factor Analysis. *Foundations and Trends in Econometrics*, 3(2):89–163.

Bai, J. and Wang, P. (2015). Identification and Bayesian Estimation of Dynamic Factor Models. *Journal of Business & Economic Statistics*, 33(2):221–240.

Bonhomme, S. and Weidner, M. (2020). Posterior Average Effects. arXiv: 1906.06360.

Boyd, S. P. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Cai, T. T., Low, M., and Ma, Z. (2014). Adaptive Confidence Bands for Nonparametric Regression Functions. *Journal of the American Statistical Association*, 109(507):1054–1070.

Carlin, B. P. and Louis, T. A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall/CRC, New York, NY, 2 edition.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014). Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. *American Economic Review*, 104(9):2593–2632.

Chetty, R. and Hendren, N. (2018). The Impacts of Neighborhoods on Intergenerational Mobility II: County-Level Estimates. *The Quarterly Journal of Economics*, 133(3):1163–1228.

Fessler, P. and Kasy, M. (2019). How to Use Economic Theory to Improve Estimators: Shrinking Toward Theoretical Restrictions. *Review of Economics and Statistics*, 101(4):681–698.

Finkelstein, A., Gentzkow, M., Hull, P., and Williams, H. (2017). Adjusting Risk Adjustment—Accounting for Variation in Diagnostic Intensity. *New England Journal of Medicine*, 376(7):608–610.

Hansen, B. E. (2016). Efficient shrinkage in parametric models. *Journal of Econometrics*, 190(1):115–132.

Hull, P. (2020). Estimating Hospital Quality with Quasi-Experimental Data. Unpublished manuscript, University of Chicago.

Ignatiadis, N. and Wager, S. (2019). Bias-Aware Confidence Intervals for Empirical Bayes Analysis. arXiv: 1902.02774.

Jacob, B. A. and Lefgren, L. (2008). Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education. *Journal of Labor Economics*, 26(1):101–136.

James, W. and Stein, C. M. (1961). Estimation with Quadratic Loss. In Neyman, J., editor, *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 361–379, Berkeley, CA. University of California Press.

Jiang, W. and Zhang, C.-H. (2009). General Maximum Likelihood Empirical Bayes Estimation of Normal Means. *The Annals of Statistics*, 37(4):1647–1684.

Kane, T. and Staiger, D. (2008). Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation. Technical Report 14607, National Bureau of Economic Research, Cambridge, MA.

Kitagawa, T., Giacomini, R., and Uhlig, H. (2019). Estimation under Ambiguity. Cemmap Working Paper 24/19.

Koopman, S. J. and Mesters, G. (2017). Empirical Bayes Methods for Dynamic Factor Models. *Review of Economics and Statistics*, 99(3):486–498.

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. Springer.

Liu, L., Moon, H. R., and Schorfheide, F. (2019). Forecasting with a Panel Tobit Model. Unpublished manuscript, University of Pennsylvania.

Morris, C. N. (1983). Parametric Empirical Bayes Inference: Theory and Applications. *Journal of the American Statistical Association*, 78(381):47–55.

Nychka, D. (1988). Bayesian Confidence Intervals for Smoothing Splines. *Journal of the American Statistical Association*, 83(404):1134–1143.

Pratt, J. W. (1961). Length of Confidence Intervals. *Journal of the American Statistical Association*, 56(295):549–567.

Robbins, H. (1951). Asymptotically Subminimax Solutions of Compound Statistical Decision Problems. In Neyman, J., editor, *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 131–149. University of California Press, Berkeley, California.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton University Press, Princeton, NJ.

Smith, J. E. (1995). Generalized Chebychev Inequalities: Theory and Applications in Decision Analysis. *Operations Research*, 43(5):807–825.

Stock, J. H. and Watson, M. W. (2016). Factor Models and Structural Vector Autoregressions in Macroeconomics. In Taylor, J. B. and Uhlig, H., editors, *Handbook of Macroeconomics*, volume 2, pages 415–525. Elsevier.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge, UK; New York, NY.

Wahba, G. (1983). Bayesian "Confidence Intervals" for the Cross-Validated Smoothing Spline. *Journal of the Royal Statistical Society: Series B (Methodological)*, 45(1):133–150.

Wasserman, L. (2006). *All of Nonparametric Statistics*. Springer, New York, NY.

Xie, X., Kou, S. C., and Brown, L. D. (2012). SURE Estimates for a Heteroscedastic Hierarchical Model. *Journal of the American Statistical Association*, 107(500):1465–1479.

Table 1: Statistics for 90% EBCIs for neighborhood effects.

| | Baseline | | $t$-stat shrinkage | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Percentile | 25th | 75th | 25th | 75th |
| Panel A: Summary statistics | | | | |
| $\sqrt{\mu_2}$ | 0.079 | 0.044 | 0.377 | 0.395 |
| $\kappa$ | 345.3 | 5024.9 | 27.2 | 71.4 |
| $E[\mu_2/\sigma_i^2]$ | 0.142 | 0.040 | | |
| $\delta_{\text{intercept}}$ | −1.441 | −2.162 | −4.060 | −4.584 |
| $\delta_{\text{perm. resident}}$ | 0.032 | 0.038 | 0.092 | 0.079 |
| $E[w_{EB,i}]$ | 0.093 | 0.033 | 0.124 | 0.135 |
| $E[w_{opt,i}]$ | 0.191 | 0.100 | 0.259 | 0.269 |
| $E$[non-cov of parametric EBCI$_i$] | 0.227 | 0.278 | 0.186 | 0.181 |
| Panel B: $E$[half-length$_i$] | | | | |
| Robust EBCI | 0.195 | 0.122 | 0.398 | 0.517 |
| Optimal robust EBCI | 0.149 | 0.090 | 0.313 | 0.410 |
| Parametric EBCI | 0.123 | 0.070 | 0.277 | 0.365 |
| Unshrunk CI | 0.786 | 0.993 | 0.786 | 0.993 |
| Panel C: Efficiency relative to robust EBCI | | | | |
| Optimal robust EBCI | 1.312 | 1.352 | 1.271 | 1.261 |
| Parametric EBCI | 1.582 | 1.731 | 1.437 | 1.417 |
| Unshrunk CI | 0.248 | 0.123 | 0.507 | 0.521 |

*Notes:* Columns (1) and (2) correspond to shrinking $Y_i$ as in the baseline implementation. Columns (3) and (4) shrink the $t$-statistic $Y_i/\hat{\sigma}_i$, as in Remark 3.8. "$E$[non-cov of parametric EBCI$_i$]": average of maximal non-coverage probability of parametric EBCI, given the estimated moments. In the "baseline" case, $\hat{\delta}$ is computed by regressing $Y_i$ onto a constant and outcomes for permanent residents, while in the "$t$-stat" case, the outcome in this regression is given by $Y_i/\sigma$. $\mu_2$ and $\kappa$ refer to moments of $\theta_i - X_i'\delta$ ("baseline") or of $\theta_i/\sigma_i - X_i'\delta$ ("$t$-stat").

Table 2: Statistics for 95% EBCIs for structural breaks in the Eurozone DFM.

| | Baseline | | $t$-stat shrinkage | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Moments used | 2nd | 2nd+4th | 2nd | 2nd+4th |
| Panel A: Summary statistics | | | | |
| $\sqrt{\mu_2}$ | 0.291 | | 1.640 | |
| $\kappa$ | | 2.994 | | 3.479 |
| $E[\mu_2/\sigma_i^2]$ | 2.727 | | | |
| $E[w_{EB,i}]$ | 0.647 | | 0.729 | |
| $E[w_{opt,i}]$ | 0.721 | 0.664 | 0.776 | 0.743 |
| $E$[non-cov of parametric EBCI$_i$] | 0.062 | 0.052 | 0.056 | 0.051 |
| Panel B: $E$[half-length$_i$] | | | | |
| Robust EBCI | 0.370 | 0.333 | 0.381 | 0.372 |
| Optimal robust EBCI | 0.344 | 0.333 | 0.377 | 0.371 |
| Parametric EBCI | 0.330 | | 0.370 | |
| Unshrunk CI | 0.433 | | 0.433 | |
| Panel C: Efficiency relative to robust EBCI | | | | |
| Optimal robust EBCI | 1.075 | 1.001 | 1.011 | 1.001 |
| Parametric EBCI | 1.122 | 1.009 | 1.031 | 1.004 |
| Unshrunk CI | 0.855 | 0.768 | 0.880 | 0.858 |

*Notes:* Columns (1) and (2) correspond to shrinking $Y_i$ as in the baseline implementation. Columns (3) and (4) shrink the $t$-statistic $Y_i/\hat{\sigma}_i$, as in Remark 3.8. Columns (1) and (3) impose only a constraint on the second moment of $\theta_i$, while columns (2) and (4) also impose the fourth moment. "$E$[non-cov of parametric EBCI$_i$]": average of maximal non-coverage probability of parametric EBCI, given the estimated moments. $\mu_2$ and $\kappa$ refer to moments of $\theta_i$ ("baseline") or of $\theta_i/\sigma_i$ ("$t$-stat").

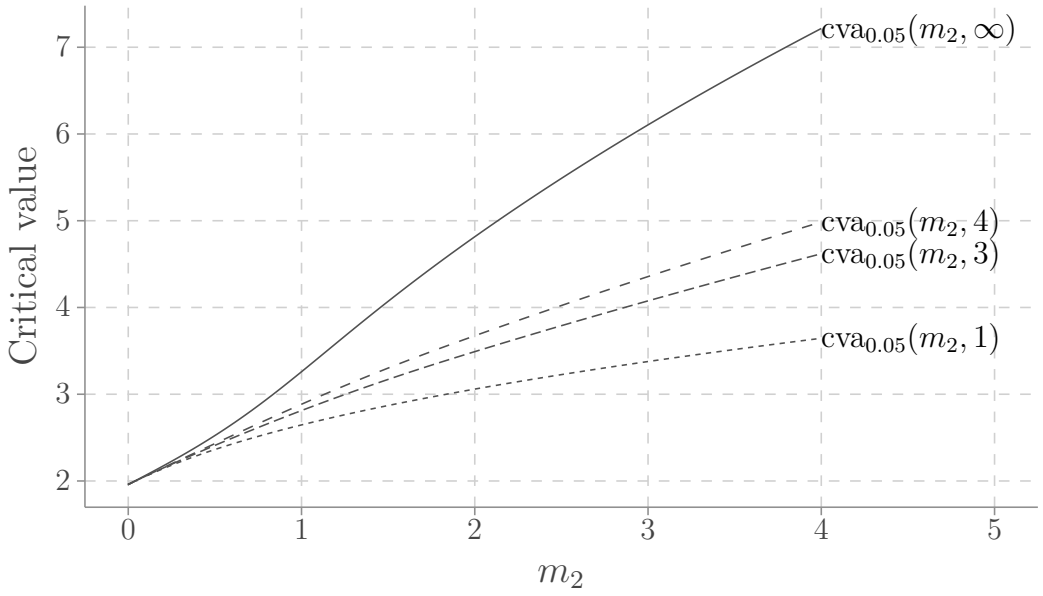


Figure 1: Function $\text{cva}_\alpha(m_2, \kappa)$ for $\alpha = 0.05$ and selected values of $\kappa$. The function $\text{cva}_\alpha(m_2)$, defined in Section 2, that only imposes a constraint on the second moment, corresponds to $\text{cva}_\alpha(m_2, \infty)$.

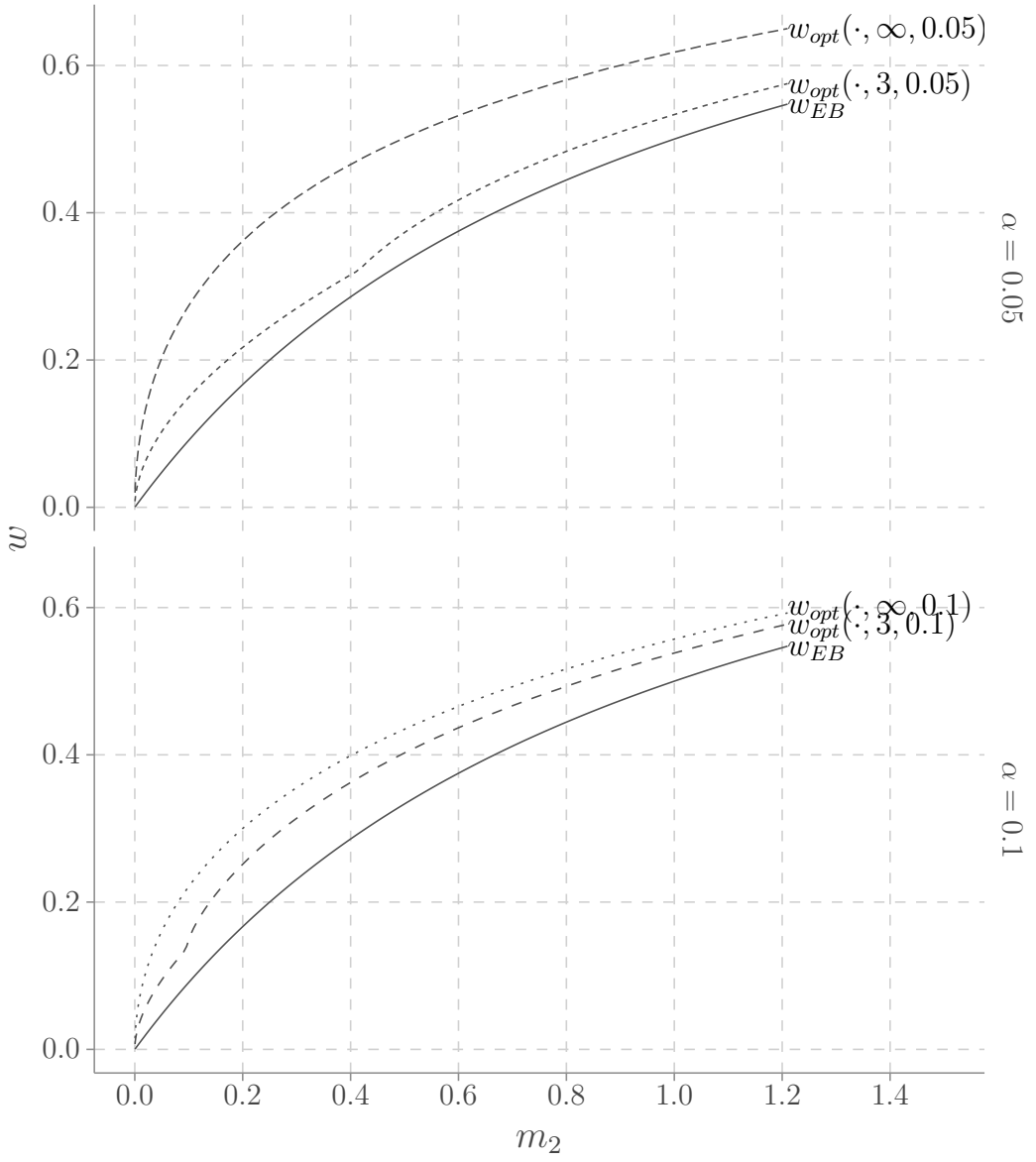


Figure 2: Optimal linear shrinkage $w_{opt}(m_2, \kappa, \alpha)$, and EB shrinkage $w_{EB} = m_2/(m_2 + 1)$ plotted as a function of the signal-to-noise ratio $m_2 = \mu_2/\sigma^2$ for $\alpha \in \{0.05, 0.1\}$ and $\kappa \in \{3, \infty\}$.

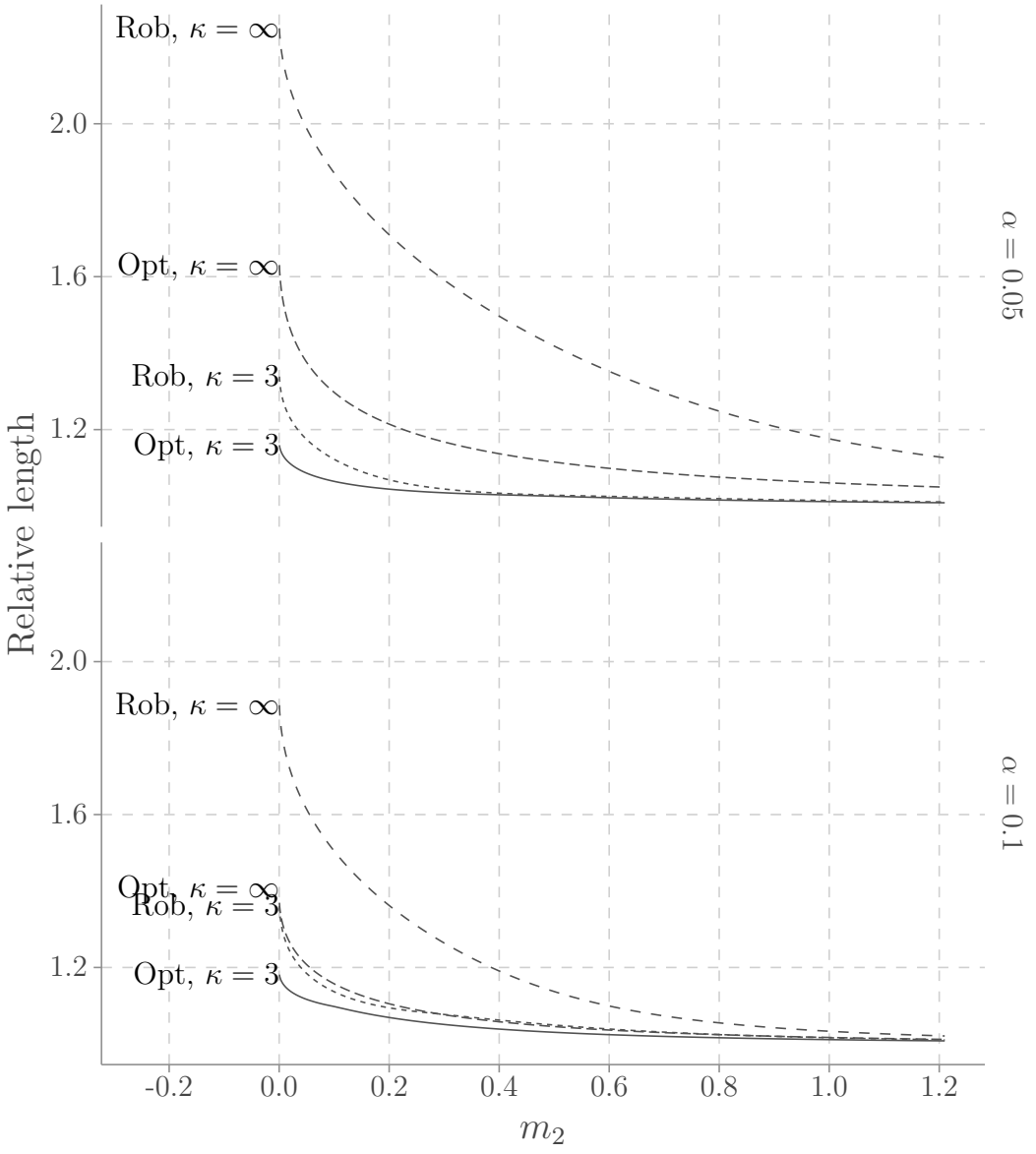


Figure 3: Relative efficiency of robust EBCI (Rob) and optimal robust EBCI (Opt) relative to the normal benchmark. The figures plot ratios of the length of the robust EBCI, $2\,\mathrm{cva}_\alpha(1/m_2, \kappa) \cdot \sigma m_2/(m_2+1)$, and the length of the optimal robust EBCI $2\,\mathrm{cva}_\alpha((1-1/w_{opt}(m_2,\kappa,\alpha))^2 m_2, \kappa) \cdot \sigma w_{opt}(m_2,\kappa,\alpha)$, relative to the parametric EBCI length $2z_{1-\alpha/2}\sqrt{m_2/(m_2+1)}\sigma$ as a function of the signal-to-noise ratio $m_2 = \mu_2/\sigma$.

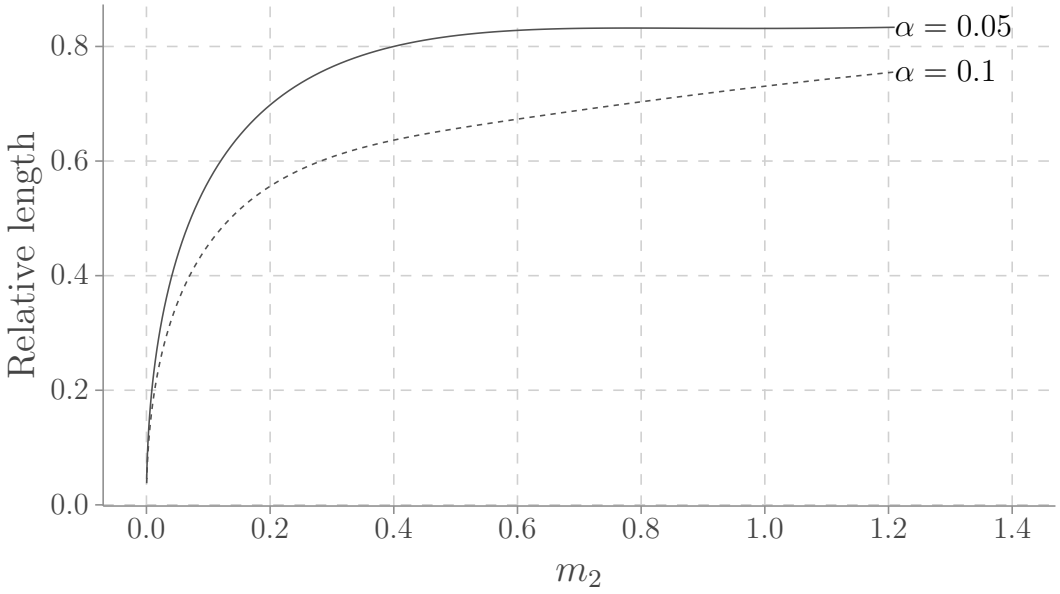


Figure 4: Relative efficiency of robust EBCI $\hat{\theta}_i \pm \text{cva}_\alpha(1/m_2, \kappa = \infty) \cdot \sigma m_2/(m_2 + 1)$ relative to the unshrunk CI $Y_i \pm z_{1-\alpha/2}\sigma$. The figure plots the ratio of the length of the robust EBCI relative to the unshrunk CI as a function of the signal-to-noise ratio $m_2 = \mu_2/\sigma$.

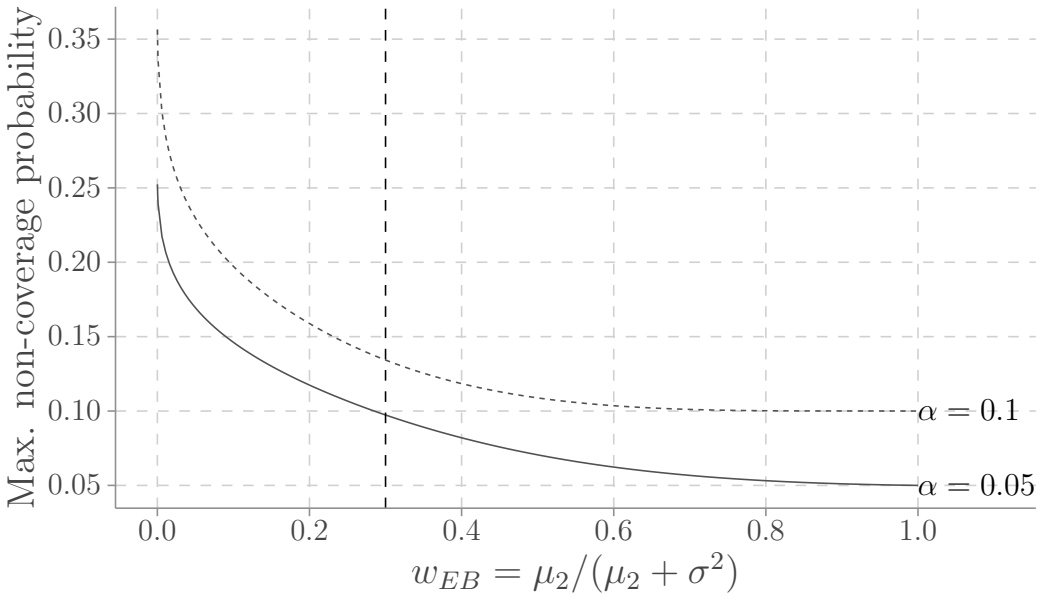


Figure 5: Maximal non-coverage probability of parametric EBCI, $\alpha \in \{0.05, 0.10\}$. The vertical line marks the "rule of thumb" value $w_{EB} = 0.3$, above which the maximal coverage distortion is less than 5 percentage points for these two values of $\alpha$.

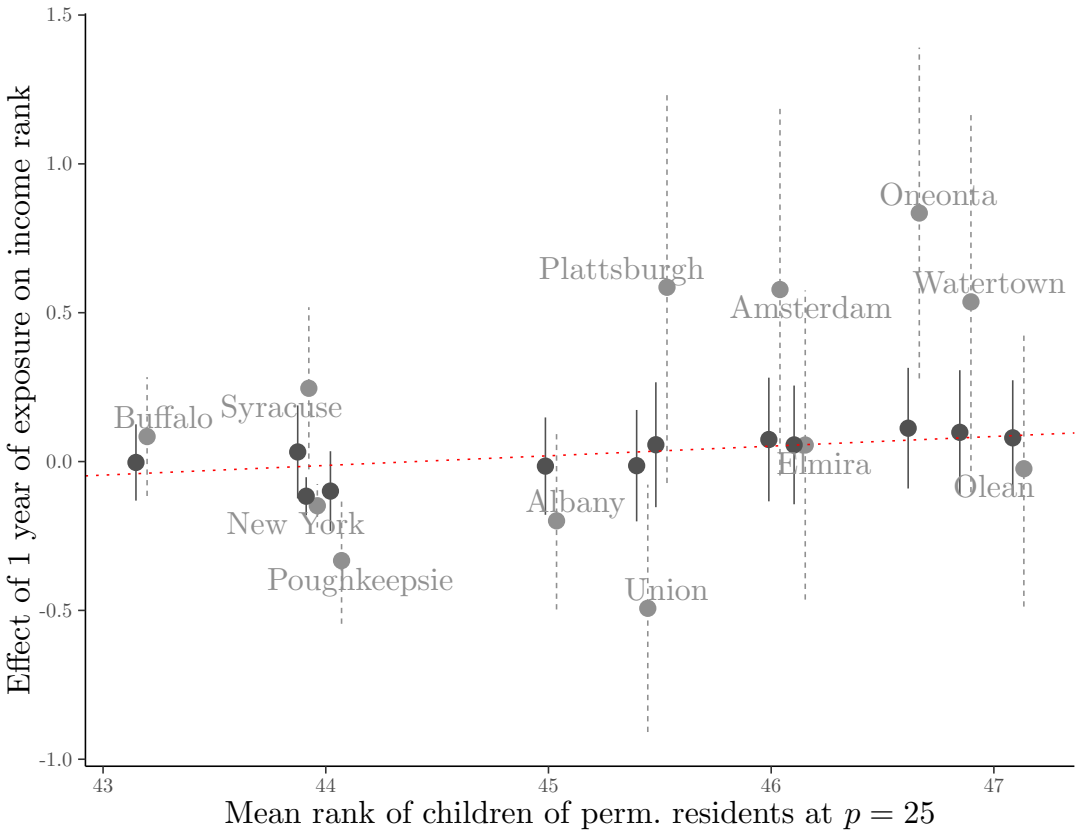


Figure 6: Neighborhood effects for New York and 90% robust EBCIs for children with parents at the $p = 25$ percentile of national income distribution, plotted against mean outcomes of permanent residents. Dashed gray lines correspond to CIs based on unshrunk estimates, and solid black lines correspond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

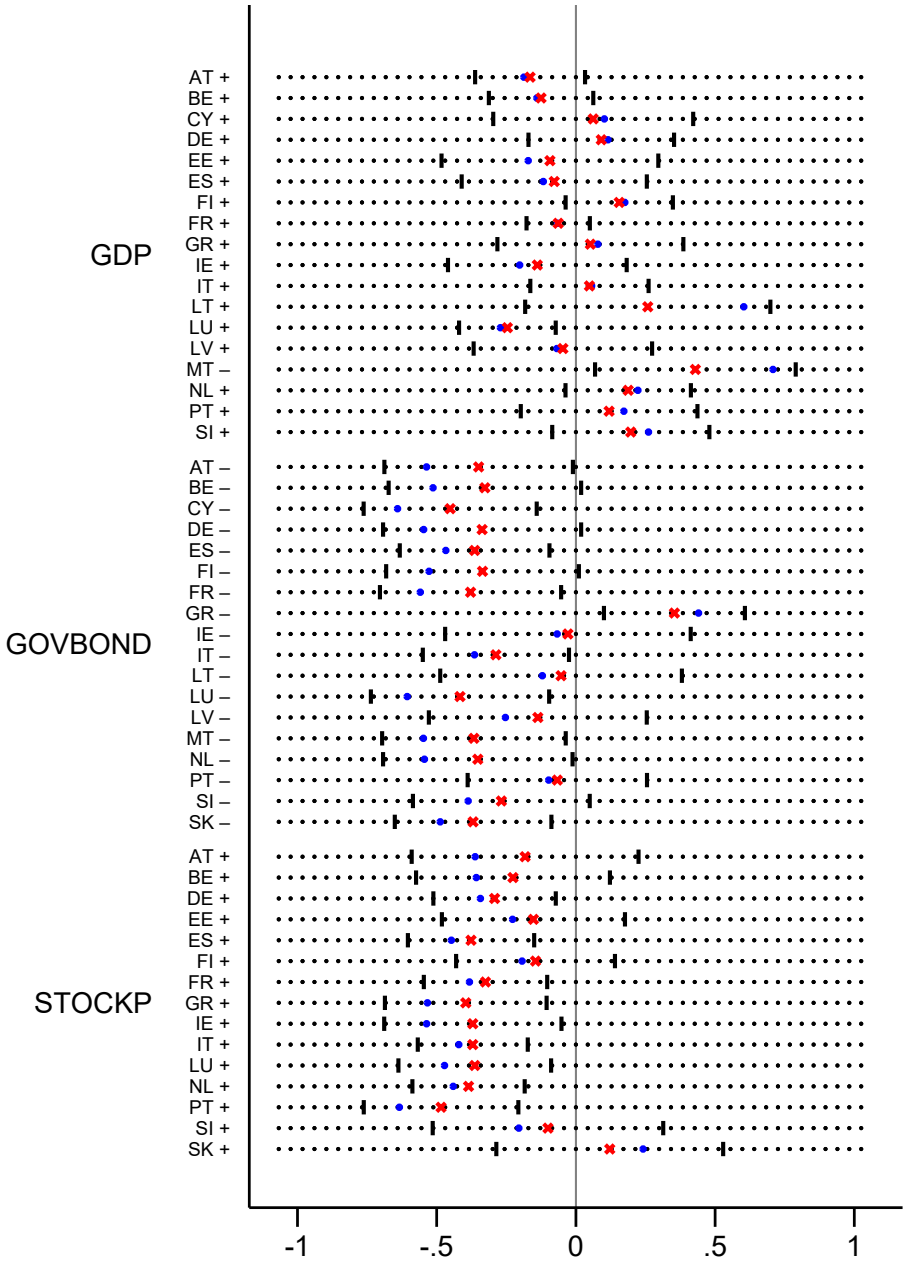


Figure 7: Shrinkage-estimated loading breaks (red crosses), corresponding 95% robust EBCIs (thick black vertical lines), and unshrunk loading break estimates (blue circles). Large text labels indicate the series type, while small text labels indicate the country. The sign (+/–) next to the country indicates the sign of the estimated pre-break loading. Series: real GDP growth (GDP), changes in 10-year government bond spread vs. Eurozone 3-month rate (GOVBOND), stock price growth (STOCKP). Robust EBCI type: $w_{EB}$, imposing both 2nd and 4th moments. Shrinkage type: baseline.