Robust and Efficient Empirical Bayes Confidence Intervals using $\gamma$-Divergence

Daisuke Kurisu$^{1,4}$, Takuya Ishihara$^2$ and Shonosuke Sugawasa$^{3,4}$

$^1$Department of Industrial Engineering and Economics, Tokyo Institute of Technology
$^2$Graduate School of Economics and Management, Tohoku University
$^3$Center for Spatial Information Science, The University of Tokyo
$^4$Nospare Inc.

Abstract

Although parametric empirical Bayes confidence intervals of multiple normal means are fundamental tools for compound decision problems, their performance can be sensitive to the misspecification of the parametric prior distribution (typically normal distribution), especially when some strong signals are included. We suggest a simple modification of the standard confidence intervals such that the proposed interval is robust against misspecification of the prior distribution. Our main idea is using well-known Tweedie’s formula with robust likelihood based on $\gamma$-divergence. An advantage of the new interval is that the interval lengths are always smaller than or equal to those of the parametric empirical Bayes confidence interval so that the new interval is efficient and robust. We prove asymptotic validity that the coverage probability of the proposed confidence intervals attain a nominal level even when the true underlying distribution of signals is contaminated, and the coverage accuracy is less sensitive to the contamination ratio. The numerical performance of the proposed method is demonstrated through simulation experiments and a real data application.

Key words: hierarchical model; signal estimation; small area estimation; Tweedie’s formula; two-stage normal model
1 Introduction

Empirical Bayes techniques to increase estimation accuracy by “borrowing strength” of information each other have been widely adopted in many scientific fields such as small area estimation (Rao and Molina, 2015), meta-analysis (DerSimonian and Laird, 2015), disease risk estimate (Lawson, 2018) and gene expression analysis (Efron et al., 2001). Although there are now many available techniques of empirical Bayes estimation, effective ways for uncertainty quantification of the empirical Bayes estimates are not fully discussed, especially when the assumed parametric prior (latent) distribution is subject to misspecification. In particular, when representative outlying subjects (so-called “hotspot”) exist whose signals are extremely larger than the other subjects, the standard normality assumption for the prior distribution is severely misspecified. In this case, the standard parametric confidence interval is highly affected even when the proportion of such outliers is small and fails to provide an efficient inference result.

To overcome this issue, we employ Tweedie’s formula (e.g. Efron, 2011) and robust divergence techniques to produce a novel robust empirical Bayes confidence interval when the underlying distribution is contaminated. Our basic consideration is that the form of marginal likelihood fully characterizes the standard parametric empirical confidence intervals; both posterior mean and variance are obtained from the derivatives of the marginal likelihood through Tweedie’s formula, and the model parameters are estimated by maximizing the marginal likelihood. This observation gives us a natural way to propose a robust confidence interval by replacing the marginal likelihood with robust alternatives. Specifically, we here employ $\gamma$-divergence (Jones et al., 2001; Fujisawa and Eguchi, 2008) which is known to hold strong robustness against outliers and produces a closed-form expression of the objective function under normality. The $\gamma$-divergence has a tuning parameter that controls its robustness. We propose a simple selection method by minimizing the average of posterior variances, ensuring that the proposed interval produces shorter intervals than the standard parametric intervals. For the theoretical justification, we first prove that the selection method for $\gamma$ is consistent in the sense that the probability of $\gamma = 0$, corresponding to the standard parametric confidence interval, is asymptotically one when the underlying distribution is not contaminated. We also prove that the strictly positive
value of $\gamma$ is selected asymptotically when the underlying distribution is contaminated, indicating an adaptive property of the proposed interval. Further, we show that the coverage probability of the proposed interval is approximately equal to the nominal level even when the true underlying distribution is subject to contamination. The proposed method will also be numerically shown to provide reasonable confidence intervals whose empirical coverage is close to the nominal level under a wide range of scenarios for the underlying distribution of signals.

We here review some relevant works. Empirical Bayes confidence intervals that are robust against misspecification of the prior (random effect) distribution have been studied since Kiefer and Wolfowitz (1956). Koenker and Mizera (2014) and Efron et al. (2019) proposed semi-parametric techniques to estimate underlying prior distribution. More recently, Armstrong et al. (2020) developed a robust empirical Bayes confidence interval using only higher moment information. In the context of small area estimation, there exist many works concerned with accurate and efficient empirical Bayes confidence intervals and comprehensive asymptotic theory (e.g. Chatterjee et al., 2008; Hall and Maiti, 2006; Yoshimori and Lahiri, 2014). However, all the results are based on the assumption that the parametric prior distribution is correctly specified; thereby, our framework in which the true prior distribution is subject to contamination is completely different from the conventional one. Hence, we need to develop a new theory for the proposed method. Tweedie’s formula is known to be a useful tool to characterize posterior moments (e.g. De Luca et al., 2021), and robust empirical Bayes techniques based on Tweedie’s formula has been considered in Gu and Koenker (2017), Liu et al. (2020), and Sugasawa (2020). Nevertheless, there have been no attempt to construct a robust empirical Bayes confidence interval using Tweedie’s formula.

This paper is organized as follows. In Section 2 we introduce new confidence intervals through Tweedie’s formula and $\gamma$-divergence. In Section 3 we provide asymptotic theory regarding the selection performance of the tuning parameter and coverage probability when the underlying distribution is contaminated. The numerical performance of the proposed methods is demonstrated through simulation studies and data analysis in Section 4 and 5, respectively. Conclusions and discussions are given in Section 6. All the technical proofs are given in the Appendix. The R code implementing the proposed method is available at
GitHub repository (https://github.com/sshonosuke/GDCI).

2 Robust and efficient empirical Bayes confidence intervals

2.1 Two-stage normal models and the standard parametric confidence intervals

Suppose we observe \( n \) independent normally distributed estimates

\[
y_i|\theta_i, \sigma_i \sim N(\theta_i, \sigma_i^2), \quad i = 1, \ldots, n,
\]

of the parameter vector \( \theta = (\theta_1, \ldots, \theta_n)^\top \). Although \( \sigma_i^2 \) is usually estimated as the asymptotic variance of the estimate \( y_i \), we condition \( \sigma_i^2 \) in what follows, following the existing works (e.g. Armstrong et al., 2020; Efron, 2016). Further, we often suppress the explicit dependency on \( \sigma_i \) if there is no confusion. A standard approach to improve the raw estimator \( y_i \) is to assume that \( \theta_i \)'s are independent realizations from a normal population, namely,\[
\theta_i \sim N(\mu, \tau^2), \quad i = 1, \ldots, n,
\]

where \( \mu \) and \( \tau^2 \) are unknown location and variance parameters, respectively. This setting results in a well-known two-stage normal model for \( y_i \).

Under the two-stage model, the posterior (conditional) distribution of \( \theta_i \) given \( y_i \) and the true parameters is \( N(\tilde{\theta}_i, \tilde{s}_i^2) \), where

\[
\tilde{\theta}_i(\mu, \tau^2) = y_i - \frac{\sigma_i^2}{\tau^2 + \sigma_i^2}(y_i - \mu), \quad \tilde{s}_i^2(\mu, \tau^2) = \frac{\tau^2\sigma_i^2}{\tau^2 + \sigma_i^2}.
\]

Then, the empirical Bayes confidence interval for \( \theta_i \) with \( 1 - \alpha \) nominal level is given by

\[
\hat{C}_{i, 1-\alpha} = (\tilde{\theta}_i - z_{\alpha/2}\tilde{s}_i, \tilde{\theta}_i + z_{\alpha/2}\tilde{s}_i),
\]

where \( \tilde{\theta}_i = \tilde{\theta}_i(\hat{\mu}, \hat{\tau}^2) \) and \( \tilde{s}_i^2 = \tilde{s}_i^2(\hat{\mu}, \hat{\tau}^2) \), and \( z_{\alpha/2} \) is the upper \( 100(\alpha/2)\% \)-quantile of the standard normal distribution. Here \( \hat{\mu} \) and \( \hat{\tau}^2 \) are estimators of \( \mu \) and \( \tau^2 \) based on the observations \( y_i \)'s. The most standard approach to estimating these unknown parameters
is to use the marginal likelihood of $y_i \sim N(\mu, \sigma_i^2 + \tau^2)$, given by

$$L(\mu, \tau^2) = -\frac{1}{2} \sum_{i=1}^{n} \log\{2\pi(\tau^2 + \sigma_i^2)\} - \frac{1}{2} \sum_{i=1}^{n} \frac{(y_i - \mu)^2}{\tau^2 + \sigma_i^2}. \quad (5)$$

The maximum marginal likelihood estimators of $\mu$ and $\tau^2$ are defined as the maximizer of $(5)$. When the assumed normal prior $(2)$ is approximately correct, the standard interval $(4)$ provides an efficient interval estimation whose interval length is shorter than that of the crude confidence interval

$$\hat{C}^c_{i,1-\alpha} = (y_i - z_{\alpha/2}\sigma_i, y_i + z_{\alpha/2}\sigma_i) \quad (6)$$

based on the first stage model (1) (e.g. Morris, 1983).

However, when the true distribution for $\theta_i$ is contaminated, that is, there exist some representative outliers that cannot be captured by the assumed normal distribution $(2)$, the standard parametric interval $(4)$ is no more reasonable. We first demonstrate that the standard interval $(4)$ is approaching to the crude interval $(6)$ when the outlier location gets more extreme. To simplify the situation, we consider a situation that the absolute value of the first observation goes to infinity, namely, $|y_1| \to \infty$, while other observed values are fixed. Such framework is widely adopted in signal estimation (e.g. Carvalho et al., 2010) to investigate the robustness of shrinkage estimation. Under the framework, we prove the following property:

**Proposition 1.** Assume that $n$ is fixed and let $\hat{\mu}$ and $\hat{\tau}^2$ be the maximum likelihood estimator based on the marginal likelihood $(5)$. Then as $|y_1| \to \infty$, it holds that $\hat{\theta}_i - y_i \to 0$ and $\hat{s}_i^2 \to \sigma_i^2$ for all $i = 1, \ldots, n$, so the standard confidence interval $(4)$ approaches to the crude confidence interval $(6)$.

Proposition 1 implies that if there are some outliers in the observations, every standard empirical Bayes confidence intervals $\hat{C}^c_{i,1-\alpha}$ are affected by the outliers, and the intervals are close to the crude confidence intervals for all $i = 1, \ldots, n$. This means that the resulting confidence intervals are inefficient by failing “borrowing strength” from the information of other subjects. This leads to poor performance of the confidence intervals $\hat{C}^c_{i,1-\alpha}$ infinite sample, as we can see in our simulation study. Our aim is to establish a new confidence
interval that can successfully borrow information of other information among non-outlying subjects to produce an efficient interval.

2.2 New robust confidence intervals

We first note that the standard empirical Bayes confidence interval (4) is characterized by the posterior mean and variance of \( \theta_i \), and Tweedie’s formula (Efron, 2011) admit the following expressions based on the marginal distribution:

\[
\tilde{\theta}_i = y_i + \sigma_i^2 \frac{d}{dy_i} \log m(y_i), \quad \tilde{s}_i^2 = \sigma_i^2 + \sigma_i^4 \frac{d^2}{dy_i^2} \log m(y_i),
\]

where \( m(y_i) \) is the marginal density of \( y_i \sim N(\mu, \tau^2 + \sigma_i^2) \). Moreover, the estimator \( \hat{\mu} \) and \( \hat{\tau}^2 \) is obtained through the marginal distribution of \( y_i \), which indicates that the marginal likelihood of \( y_i \) completely characterizes the parametric interval (4).

When the distribution of \( \theta_i \) is misspecified, the marginal distribution of \( y_i \) is also misspecified, and this would be the main reason for the undesirable property of the standard interval (4) as demonstrated in Proposition 1. To overcome the problem, our proposal is very simple; we replace the marginal log-likelihood with a robust alternative against outliers. Specifically, we use an objective function based on \( \gamma \)-divergence (Jones et al., 2001; Fujisawa and Eguchi, 2008) given by

\[
D_\gamma(y_i; \psi) \equiv \frac{1}{\gamma} \phi(y_i; \mu, \tau^2 + \sigma_i^2)^\gamma \left\{ \int \phi(t; \mu, \tau^2 + \sigma_i^2)^{1+\gamma} dt \right\}^{-\gamma/(1+\gamma)}
\]

\[
= \frac{1}{\gamma} \phi(y_i; \mu, \tau^2 + \sigma_i^2)^\gamma c_\gamma(\tau^2),
\]

where \( c_\gamma(\tau^2) = \left\{ 2\pi(\tau^2 + \sigma_i^2) \right\}^{\gamma^2/2(1+\gamma)} \) and \( \phi(y; \mu, \tau^2) \) is the density function of the normal distribution \( N(\mu, \tau^2) \) and \( \gamma > 0 \) is a tuning parameter related to the robustness for which a selection procedure will be discussed later. Note that \( \lim_{\gamma \to 0} \{ D_\gamma(y_i; \psi) - \gamma^{-1} \} = \log m(y_i) \), that is, the objective function is very similar the marginal log-likelihood under small \( \gamma \). Hence, the objective function (8) is regarded as a natural generalization of the marginal log-likelihood.

Our main idea is to replace \( \log m(y_i) \) with \( D_\gamma(y_i) \) in the Tweedie’s formula (7) to
obtain robust posterior mean and variance. A straightforward calculation shows that

\[
\tilde{\theta}_i^{(\gamma)} \equiv y_i + \sigma_i^2 \frac{d}{dy_i} D_\gamma(y_i; \psi) = y_i - w_\gamma(y_i) \frac{\sigma_i^2}{\tau^2 + \sigma_i^2} (y_i - \mu),
\]

\[
\tilde{s}_i^{2(\gamma)} \equiv \sigma_i^2 + \sigma_i^4 \frac{d^2}{dy_i^2} D_\gamma(y_i; \psi)
\]

\[
= \sigma_i^2 + w_\gamma(y_i) \frac{\sigma_i^4}{(\tau^2 + \sigma_i^2)^2} \{ \gamma(y_i - \mu)^2 - (\tau^2 + \sigma_i^2) \},
\]

where \( w_\gamma(y_i) = \phi(y_i; \mu, \tau^2 + \sigma_i^2) \gamma c_\gamma(\tau^2) \). It can be easily checked that \( \tilde{\theta}_i^{(\gamma)} \to \hat{\theta}_i \) and \( \tilde{s}_i^{2(\gamma)} \to \hat{s}_i^2 \) under \( \gamma \to 0 \), so that the expression in (7) can be seen as robust generalization of (3). For the unknown parameters \( \psi = (\mu, \tau^2) \), we define the robust estimator as

\[
\hat{\psi}^{(\gamma)} = \arg\max_\psi \sum_{i=1}^n D_\gamma(y_i; \psi),
\]

where \( D_\gamma(y; \psi) \) is given in (8). Note that \( \hat{\psi}^{(\gamma)} = (\hat{\mu}_\gamma, \hat{\tau}_\gamma^2) \) reduces to the maximum likelihood estimator as the maximizer of (3) as \( \gamma \to 0 \). Since the above objective function is an analytic form of \( \psi \), it can be easily optimized, for example, by the Newton-Raphson algorithm. Then, we proposed a robust empirical Bayes confidence interval as

\[
\hat{C}_{i, 1-\alpha}^{(\gamma)} = (\hat{\theta}_i^{(\gamma)} - z_{\alpha/2} \hat{s}_i^{(\gamma)}, \hat{\theta}_i^{(\gamma)} + z_{\alpha/2} \hat{s}_i^{(\gamma)}),
\]

here \( \hat{\theta}_i^{(\gamma)} \) and \( \hat{s}_i^{(\gamma)} \) are obtained by replacing \( \psi \) with \( \hat{\psi}^{(\gamma)} \) in \( \tilde{\theta}_i^{(\gamma)} \) and \( \tilde{s}_i^{2(\gamma)} \), respectively. A notable property of the new interval (11) is that not only \( \tilde{\theta}_i^{(\gamma)} \) but \( \tilde{s}_i^{2(\gamma)} \) depends on the observed value \( y_i \), which results in the following quite different property from one given in Proposition 1.

**Proposition 2.** Let \( \gamma \in (0, 1] \). Assume that \( n \) is fixed. As \( |y_1| \to \infty \), it holds that \( \tilde{\theta}_1^{(\gamma)} - y_1 \to 0 \) and \( \tilde{s}_1^{2(\gamma)} \to \sigma_1 \) while \( \tilde{\theta}_i^{(\gamma)} \to \hat{\theta}_i^{(\gamma)} \) and \( \tilde{s}_i^{2(\gamma)} \to \hat{s}_i^{2(\gamma)} \) for \( i \neq 1 \), where \( \tilde{\theta}_i^{(\gamma)} \) and \( \tilde{s}_i^{2(\gamma)} \) are defined by replacing \( \hat{\psi}^{(\gamma)} \) with \( \hat{\psi}_i^{(\gamma)} \), which is the maximizer of \( n^{-1} \sum_{i=2}^n D_\gamma(y_i; \psi) \).

Proposition 2 implies two adaptive properties of the proposed confidence interval. For non-outlying signals, the proposed intervals produce efficient intervals by borrowing the strength of information from other non-outlying signals. In other word, the proposed interval can automatically identify the outlying subjects and successfully eliminate such
data to provide a stable interval for non-outlying observations. The proposed interval reduces to the crude interval for outlying signals, which makes sense because there would be no information to borrow strength from non-outlying observations. Such property can be recognized as an extension of “tail robustness” (Carvalho et al., 2010) for the shrinkage estimation.

2.3 Tuning parameter selection

The choice of $\gamma$ in (11) controls the trade-off between statistical efficiency without outliers and robustness against outliers. We here propose selecting a suitable value of $\gamma$ to obtain intervals with minimum posterior variances. Let $\Gamma \equiv \{\gamma_1, \ldots, \gamma_M\}$ be a set of candidate values for $\gamma$, where $0 = \gamma_1 < \gamma_2 < \cdots < \gamma_M = 1$. Then, we propose selecting $\gamma$ minimizing the total robust variances, that is, we define the optimal $\gamma$ as

$$\gamma_{op} = \arg\min_{\gamma \in \Gamma} \sum_{i=1}^{n} a_i s_i^2(\gamma),$$

(12)

where $a_i$ is a fixed weight, for example, $a_i = 1/\sigma_i^2$. We set $a_i = 1$ in our theoretical argument for simplicity, but the extension is quite straightforward.

2.4 Some extensions

We here discuss two important extensions of the proposed method. The first issue is that we have some auxiliary vector $x_i$ associated with $y_i$. In this case, the prior distribution will be changed to $\theta_i \sim N(x_i^\top \beta, \tau^2)$, and all the results can be rewritten by replacing $\mu$ with $x_i^\top \beta$. The asymptotic properties that will be developed in the subsequent section still hold as long as standard conditions for $x_i$ (e.g. Chatterjee et al., 2008; Yoshimori and Lahiri, 2014) are satisfied, and the dimension of $x_i$ is fixed. An example with auxiliary information is considered in Section 5.

The second issue is related to the zero-estimate of $\tau^2$ under which empirical Bayes confidence intervals cannot be obtained because of the zero-estimate of the posterior variance. To avoid such undesirable estimates, we can employ the adjustment maximum likelihood (Li and Lahiri, 2010). This can be easily implemented by adding $\log \tau^2$ to the objective function given in (10). Since the contribution of the adjustment term is negligible under
$n \to \infty$, this modification does not affect the asymptotic properties given in the subsequent section.

3 Theoretical properties

In this section, we provide the asymptotic properties of the proposed method. In particular, we reveal selection performance for $\gamma$ based on (12) and asymptotic coverage probability of the robust confidence intervals. For theoretical development, we consider two scenarios where the underlying distribution of $\theta_i$ is correctly specified or contaminated. We first discuss the former case (correctly specified case) in Section 3.1 corresponding to the justification under the conventional settings (e.g. Morris, 1983), and consider the latter case (contamination case) in Section 3.2, which is the first attempt to discuss coverage probability under contamination. Throughout this section, we assume that the first stage model (1) is correctly specified, that is, contamination does not occur in the first stage. This situation is very likely in practice because $y_i$ is an estimator or summary statistics of some quantities, and the normality assumption is approximately true. Throughout this section, we assume the following regularity conditions:

**Assumption 1. (Regularity conditions)**

(C1) $\{\sigma_i\}$ is a sequence of independent and identically distributed random variables and there exist positive constants $\sigma_L$ and $\sigma_U$ such that $0 < \sigma_L \leq \sigma_i \leq \sigma_U < \infty$.

(C2) $\Psi \subset \mathbb{R} \times (0, \infty)$ is a compact parameter space of $\psi_* := (\mu_*, \tau_2^2)$, where $\psi_*$ is the true parameter, and $\tau_2^2 + \sigma_L^2 > (2\pi)^{-1}e^{-3}$.

The uniform boundedness of $\sigma_i$ in (C1) is widely adopted in this context (e.g. Armstrong et al. 2020; Yoshimori and Lahiri 2014). The assumption that $\sigma_i$ is a random variable in (C1) reflects the practical situation where $\sigma_i$’s are estimated variance of $y_i$’s. The latter part in (C2) is satisfied when the lower bound of $\sigma_i^2$ is not small. For example, if $y_i$ can be regarded $z$-score of some estimand of interest, then $\sigma_i = 1$ for all $i$ and the condition is clearly satisfied.
3.1 Properties under correctly specified case

**Theorem 1.** Suppose that $\theta_i \sim N(\mu_*^*, \tau_*^2)$, that is, the model (4) is correctly specified. Then, under Assumption 1, it holds that $P(\gamma_{op} = 0) \rightarrow 1$ as $n \rightarrow \infty$.

Theorem 1 indicates that if the distribution of $\theta_i$ is correctly specified, namely, $\theta_i \sim N(\mu_*^*, \tau_*^2)$, then $\gamma_{op}$ converge in probability to 0 as $n \rightarrow \infty$. This implies that our method is asymptotically consistent with the standard parametric empirical Bayes confidence intervals when the observations are not contaminated with some (non-Gaussian) distribution or do not include outliers.

We next consider asymptotic coverage probabilities of the proposed interval when the assumed model is correctly specified.

**Theorem 2.** Suppose that $\theta_i \sim N(\mu_*^*, \tau_*^2)$. Then, under Assumption 1, we have

$$P(\theta_i \in \hat{C}_{i,1-\alpha}^{(\gamma_{op})}) = 1 - \alpha + o(1) \text{ as } n \rightarrow \infty.$$  

Theorem 2 implies that if the data generating process is correctly specified, then our robust empirical Bayes confidence intervals using $\gamma_{op}$ selected via a data-dependent manner using (12) have asymptotically correct coverage probabilities.

**Remark 1.** Let $\xi_1, \ldots, \xi_n$ be i.i.d. standard normal random variables, and for $\alpha \in (0, 1)$, let $q_\alpha$ satisfy $P(\max_{1 \leq i \leq n} |\xi_i| > q_\alpha) = \alpha$. Define $\hat{C}_{i,1-\alpha}^{(\gamma)}$ by replacing $z_{\alpha/2}$ with $q_\alpha$. If we can take $\epsilon_n = O(n^{-\zeta})$ for some $\zeta > 0$, then applying almost the same argument in the proof of Theorem 2, we can also show $P(\theta_i \in \hat{C}_{i,1-\alpha}^{(\gamma_{op})}, i = 1, \ldots, n) = 1 - \alpha + o(1)$ as $n \rightarrow \infty$.

**Remark 2.** Although we have not shown in detail, the coverage accuracy of the proposed interval is roughly $O(n^{-1})$, which can be improved by using, for instance, parametric bootstrap (e.g. Chatterjee et al., 2008; Hall and Maiti, 2006) to make the coverage error $O(n^{-3/2})$ when $n$ is not very large.
3.2 Properties under contaminated case

We next consider a contaminated case. Suppose that $\theta_i$ are drawn from the density given by

$$f_{\psi*}(\theta_i) = (1 - \omega)\phi(\theta_i; \mu*, \tau_*^2) + \omega\delta(\theta_i).$$

where $\omega \in [0, 1/2)$ is a contamination ratio and $\delta$ is a contamination distribution. We define the following quantity:

$$\rho \equiv \max_{k \in \{0, 1, 2\}} \sup_{\sigma_i \in [\sigma_L, \sigma_U]} \int f_{\delta}(y_i)\phi(y_i; \mu, \tau_*^2 + \sigma_i^2)\gamma|y_i - \mu|^k dy_i \equiv \rho(\psi*, \gamma, \delta),$$  \hspace{1cm} (13)

where

$$f_{\delta}(y_i) = \int \phi(y_i; \theta_i, \sigma_i^2)\delta(\theta_i)d\theta_i$$

is the marginal distribution of $y_i$ when $\theta_i \sim \delta$. Note that (13) measures separability between the genuine distribution $N(\mu*, \tau_*^2)$ and contamination distribution $\delta$; $\rho$ is almost 0 when the two distributions are well separated, that is, the density $f_{\delta}$ mostly lies on the tail of the density $\phi(y_i; \mu, \tau_*^2 + \sigma_i^2)$. Let $E_{f_{\psi}}[\cdot]$ be the expectation with respect to the model $y_i|\theta_i, \sigma_i \sim N(\theta_i, \sigma_i)$ that $\theta_i$ are drawn from the density $f_{\psi}(\theta_i)$, and let $E_{\psi}[\cdot]$ and $E_{\delta}[\cdot]$ denote the expectation with respect to $\phi(\theta_i; \mu, \tau_*^2 + \sigma_i^2)$ and $f_{\delta}$.

Define $\partial_{\psi}D_\gamma(y_i, \sigma_i^2; \psi) = \partial D_\gamma(y_i, \sigma_i^2; \psi)/\partial \psi$. Assume that $\gamma > 0$ and let $\psi^{(\gamma)}_1 = (\mu_{\gamma,1}, \tau_{\gamma,1}^2)$ be a root of $E_{f_{\psi}}[\partial_{\psi}D_\gamma(y_1, \sigma_1^2; \psi)] = 0$, and let $\tilde{\psi}^{(\gamma)}_n = (\tilde{\mu}_\gamma, \tilde{\tau}_\gamma^2)$ be a root of the estimating equation $n^{-1}\sum_{i=1}^n \partial_{\psi}D_\gamma(y_i, \sigma_i^2; \psi) = 0$. As a result of Lemmas A4 and A5 in Appendix A, we can show

$$\tilde{\psi}^{(\gamma)}_n = \psi^{(\gamma)}_1 + O_p(n^{-1/2}) = \psi* + O(\omega \rho) + O_p(n^{-1/2}).$$

This enables us to establish the results on the asymptotic properties of $\gamma_{op}$ (Theorem 3) and the proposed confidence intervals (Theorem 4).

We first show the selection performance of $\gamma$. 

11
Theorem 3. Define

\[ Q(\gamma; \psi) := E_{\psi_i} \left[ w_{\gamma}(y_i) \frac{\sigma_i^4}{(\tau^2 + \sigma_i^2)^2} \{\gamma(y_i - \mu)^2 - (\tau^2 + \sigma_i^2)^2\} \right]. \]

Assume that \( \theta_i \sim f(\theta_i) \) with \( \omega \in (0, 1/2) \), \( \mu_\delta := \int \theta_i \delta(\theta_i) d\theta_i < \infty \), \( \tau_\delta^2 := \int (\theta_i - \mu_\delta)^2 \delta(\theta_i) d\theta_i < \infty \), and \((\mu_* - \mu_\delta)^2 \geq \tau_*^2 + \sigma_\delta^2\). Further, suppose that Conditions (a)-(e) in Lemma A4 hold and Assumption 1 is satisfied. If there exists \( \gamma_* \in \Gamma \setminus \{0\} \) such that \( Q(0; \psi) > (1 - \omega)Q(\gamma_*; \psi) \) and \( \rho(\psi, \gamma_* , \delta) \) is sufficiently small, then we have \( P(\gamma_{\text{op}} > 0) \to 1 \) as \( n \to \infty \).

From Theorem 3, the probability that a positive value \( \gamma_{\text{op}} \) is selected approaches 1 as \( n \to \infty \) if the true distribution of \( \theta_i \) is misspecified. The condition \( Q(0; \psi) > (1 - \omega)Q(\gamma_*; \psi) \) implies that if \( \omega \) is small, then \( \gamma_* \) should be close to 0. Practically, it is sufficient to take the number of candidates \( M \) in \( \Gamma \) sufficiently large. In order to make \( \rho(\psi, \gamma_* , \delta) \) sufficiently small when \( \gamma_* \) is close to 0, the distribution \( \delta \) should be well-separated from \( \mathcal{N}(\mu_* , \tau_*^2) \). Note that the condition \((\mu_* - \mu_\delta)^2 \geq \tau_*^2 + \sigma_\delta^2\) is also related to the separability.

Next, we provide the asymptotic property of the proposed confidence intervals. In the following argument, we set \( \rho_{\psi_*} = \rho \) to indicate that \( \rho \) depends on \( \psi_* = (\mu_* , \tau_*^2) \).

Theorem 4. Suppose that Conditions (a)-(e) in Lemma A4 hold and Assumption 1 is satisfied. Let \( 0 < \gamma < \gamma_* \). Assume that \( \theta_i \sim f(\theta_i) \), \( \omega \rho_{\psi_*} = O(n^{-\zeta}) \) as \( n \to \infty \) for some \( \zeta \in (0, 1/2) \) and \( E_{\delta}[y_i^2] < \infty \). Then for \( \gamma \in \left[ \frac{1}{2} , \gamma_* \right] \), we have

\[ P \left( \theta_i \in \hat{C}_{i,1-\alpha}^{(\gamma)} \right) = 1 - \alpha + O \left( (\omega \rho_{\psi_*})^{1/4} + \gamma \right) \quad \text{as} \quad n \to \infty. \]

The term \( O((\omega \rho_{\psi_*})^{1/4}) \) in the approximation error of the confidence interval \( \hat{C}_{i,1-\alpha}^{(\gamma)} \) comes from the bias for estimating \( \psi_* \) under the contaminated model \( \theta \sim f(\theta_i) \). If \( \rho_{\psi_*} \) is small, that is, the contamination density \( f_\delta(y_i) \) mostly lies on the tail of the density \( \phi(y_i; \mu_* , \tau_*^2 + \sigma_\delta^2) \), then the term \( O((\omega \rho_{\psi_*})^{1/4}) \) can be small even if \( \omega \) is not small. In this case, the observations could be heavily contaminated. When \( \rho_{\psi_*} \) is not small, \( O((\omega \rho_{\psi_*})^{1/4}) \) can be small if the contamination ratio \( \omega \) is small. The term \( O(\gamma) \) reflects the effect of robustification using the \( \gamma \)-divergence for constructing \( \hat{C}_{i,1-\alpha}^{(\gamma)} \) instead of \( \hat{C}_{i,1-\alpha} \), the
standard parametric empirical Bayes confidence interval. Hence, the term $O(\gamma)$ is likely to be positive since the posterior variance $\tilde{s}_i(\gamma)$ increases with $\gamma$ (when $\gamma^2 + \sigma_L^2 > (2\pi)^{-1}$) and the resulting interval has wider length than the parametric interval. This means that the term $O(\gamma)$ is thought to work in the direction of increasing the coverage probability. Moreover, we will demonstrate that the effect of such coverage error is very limited in our simulation studies as long as the genuine distribution and contamination distribution are well separated.

4 Simulation study

We compare the performance of the proposed interval estimation with some existing methods through simulation studies. Throughout this study, we generate observations $y_i$ from $y_i \sim N(\theta_i, \sigma_i^2)$ with $\sigma_i = 1$, where $\theta_i$ is the true signal. For the underlying generating process for $\theta_i$, we adopted the following four scenarios:

(i) $\theta_i \sim N(0, V)$,
(ii) $\theta_i \sim 0.95N(0, V) + 0.05N(7, V)$,
(iii) $\theta_i \sim LN(0, V)$,
(iv) $\theta_i \sim 0.9N(0, (0.2)^2) + 0.1N(10V, (0.2)^2),$

where LN($a, b$) denotes a log-normal distribution with log-mean $a$ and log-variance $b$. In each scenario, four values of $V$ are considered, $V \in \{0.1, 0.5, 1, 2\}$, following Koenker (2020). Remark that scenario (i) corresponds to the situation where there are no outliers, and the assumed model is correctly specified to check the efficiency loss of the proposed method compared with the standard empirical Bayes confidence intervals. On the other hand, the underlying distribution of $\theta_i$ is quite different from a normal distribution in the other scenarios. In particular, signals in scenarios (iii) and (iv) produce some outlying values, which would make the standard empirical Bayes approach inefficient as the crude interval.

Based on the simulated dataset, we computed 95% confidence intervals of $\theta_i$ using the proposed robust empirical Bayes confidence interval using $\gamma$-divergence (GD) and the standard parametric empirical Bayes confidence interval (EB). We also employed confidence intervals via empirical Bayes deconvolution by Efron (2016) and recently proposed robust confidence intervals by Armstrong et al. (2020), which are denoted by EF and AKM, re-
spectively. We implemented AKM by using R packages “ebci” (Kolesar 2021). Using 1000 simulated datasets, we computed coverage probability (CP) and average length (AL) of 95% confidence intervals, averaged over all the subjects. The results under $n \in \{200, 1000\}$ are shown in Table 1, and additional simulation results are provided in the Appendix. The averaged value of the selected $\gamma$ in the proposed method is also reported in Table 1.

It is observed that the proposed GD interval provides CP around the nominal level (95%) and smaller AL than EB and AKM in almost all the scenarios, clearly showing the efficiency and robustness of GD. In scenario (i), the average value of $\gamma$ is 0, so that the proposed GD method does not lose any efficiency when the assumed distribution is correct, which is consistent with our theoretical result given in Theorem 1. On the other hand, once the underlying distribution of $\theta_i$ is misspecified or $\theta_i$ contains outlying values, the optimal value of $\gamma$ tends to be strictly positive, as proved in Theorem 3. In most scenarios, it is also observed that AKM tends to be inefficient, even more than EB. On the other hand, EF produces a very efficient confidence interval under a relatively simple underlying structure like scenario (ii), but its CP can be small under some scenarios such as (i) and (iv).

5 Example: crime count in Tokyo

We demonstrate the proposed methods using a dataset of the number of police-recorded crimes in the Tokyo metropolitan area, obtained from the University of Tsukuba and publicly available online (“GIS database of the number of police-recorded crime at O-aza, chome in Tokyo, 2009-2017”, available at https://commons.sk.tsukuba.ac.jp/data-en). This study focuses on the number of violent crimes in 2,855 local towns in the Tokyo metropolitan area. The number of violent crimes per year is observed for nine years (from 2009 to 2017) in each town. We first excluded areas where no violent crime has occurred in the nine years, resulting in $n = 2826$ areas used in the analysis. We first computed the average number of crimes per unit km$^2$ (denoted by $Y_i$) and its sampling variances (denoted by $\sigma_i^2$) for $i = 1, \ldots, n$. The dataset contains five area-level auxiliary information, entire population density (PD), day-time population density (DPD), the density of foreign people (FD), percentage of single-person households (SH), and average
Table 1: Average values of optimal $\gamma$, empirical coverage probability (CP) and average length (AL) of the four 95% confidence intervals under $n \in \{200, 1000\}$.

| scenario | $V$ | $\gamma$ | $m = 200$ | $m = 1000$ |
|----------|-----|----------|-----------|-----------|
| (i)      |     |          |           |           |
|          | 0.1 | 0.00     | 94.6      | 94.6      |
|          | 0.5 | 0.00     | 94.1      | 94.1      |
|          | 1   | 0.00     | 94.5      | 94.5      |
|          | 2   | 0.00     | 94.8      | 94.8      |
| (ii)     |     |          |           |           |
|          | 0.1 | 0.17     | 94.4      | 94.4      |
|          | 0.5 | 0.19     | 95.1      | 95.0      |
|          | 1   | 0.09     | 95.5      | 94.7      |
|          | 2   | 0.02     | 94.9      | 94.8      |
| (iii)    |     |          |           |           |
|          | 0.1 | 0.00     | 94.4      | 94.4      |
|          | 0.5 | 0.01     | 95.1      | 95.0      |
|          | 1   | 0.08     | 95.8      | 94.9      |
|          | 2   | 0.18     | 96.0      | 94.9      |
| (iv)     |     |          |           |           |
|          | 0.1 | 0.00     | 93.0      | 93.0      |
|          | 0.5 | 0.21     | 96.9      | 94.3      |
|          | 1   | 0.14     | 99.4      | 95.0      |
|          | 2   | 0.11     | 99.5      | 94.9      |
| (i)      |     |          |           |           |
|          | 0.1 | 0.00     | 94.9      | 94.9      |
|          | 0.5 | 0.00     | 95.0      | 95.0      |
|          | 1   | 0.00     | 94.9      | 94.9      |
|          | 2   | 0.00     | 94.9      | 94.9      |
| (ii)     |     |          |           |           |
|          | 0.1 | 0.18     | 99.5      | 94.6      |
|          | 0.5 | 0.20     | 97.9      | 94.7      |
|          | 1   | 0.12     | 95.9      | 94.8      |
|          | 2   | 0.00     | 94.9      | 94.9      |
| (iii)    |     |          |           |           |
|          | 0.1 | 0.00     | 94.5      | 94.5      |
|          | 0.5 | 0.00     | 95.4      | 95.4      |
|          | 1   | 0.10     | 96.3      | 95.2      |
|          | 2   | 0.19     | 96.2      | 95.0      |
| (iv)     |     |          |           |           |
|          | 0.1 | 0.00     | 92.3      | 92.3      |
|          | 0.5 | 0.32     | 98.7      | 94.4      |
|          | 1   | 0.13     | 99.4      | 95.0      |
|          | 2   | 0.12     | 99.5      | 95.0      |
Letting $x_i$ be a six-dimensional vector of intercept and the five covariates, we employ the following model:

$$y_i | \theta_i \sim N(\theta_i, \sigma_i^2), \quad \theta_i \sim N(x_i^\top \beta, \tau^2), \quad i = 1, \ldots, n,$$

where $\theta_i$ is the true average number of crimes in the $i$th area.

We first applied the proposed robust interval with $\gamma$-divergence (GD), where we searched the optimal value of $\gamma$ from

$$\gamma \in \{0, 0.005, 0.01, \ldots, 0.195, 0.200\}.$$

The selection criterion (12) with weight $\omega_i = 1/\sigma_i^2$ for each value of $\gamma$ is shown in the left panel in Figure 1. We can observe that the criterion has a unique minimizer and found that $\gamma_{opt} = 0.095$ (bounded away from 0), indicating that the dataset contains some outliers and the normality assumption for the distribution of $\theta_i$ seems to be violated. The estimated values of the regression coefficients and the variance parameter ($\tau^2$) based on the GD method with the optimal value of $\gamma$ are reported in Table 2, where the estimates from the empirical Bayes (EB) method are also reported for comparison. From Table 2, it can be seen that the estimates of $\beta$ and $\tau^2$ in EB tend to be larger than those in GD, because of the existence of outliers. In particular, the estimate of $\tau^2$ in EB is 20 times larger than that in GD, and such a large value of $\tau^2$ leads to inefficient interval estimation. For comparison, we also considered advanced approaches to robust confidence intervals. Since most existing robust methods such as Efron et al. (2019) are not directly applicable under the existence of covariates, we here employed the AMK method by Armstrong et al. (2020) using a weight $1/\sigma_i^2$ in the parameter estimation. The estimated parameters of AMK are relatively similar to those of GD, except for $\tau^2$, as shown in Table 2.

We computed 95% confidence intervals of $\theta_i$ based on the three methods (GD, EB, and AKM) and computed its length $L_i$. To see the efficiency of the interval estimation, we calculated average length, $n^{-1} \sum_{i=1}^n L_i / \sigma_i$, which are reported in Table 2. The result shows that GD provides the most efficient interval among the three methods, and the order of interval length is consistent with that of the estimates of $\tau^2$. To see more details of the
obtained confidence intervals, we computed ratios of the interval of the three methods to that of the crude maximum likelihood (ML) confidence interval, \( (y_i + z_{0.025}\sigma_i, y_i + z_{0.975}\sigma_i) \), which are shown in the right panel in Figure 2. The figure shows that the three methods tend to produce more efficient confidence intervals than ML. Especially, the proposed GD method produces the most efficient intervals when the observed values are moderate, i.e., \( 1 \leq y_i/\sigma_i \leq 5 \). On the other hand, the interval lengths of EB and AKM in areas with large observed values (i.e., \( y_i/\sigma_i \geq 10 \)) are much smaller than those of ML and GD, while the lengths of GD are almost the same as those of ML. However, this phenomenon of excessively short lengths of EB and AKM is caused by so-called “over-shrinkage” rather than the efficiency of empirical Bayes methods. To see this, we show 95% confidence intervals in areas having the top 30 largest \( y_i \) (i.e., outlying areas), in the left panel of Figure 2 which reveals the undesirable properties of both EB and AKM; not only the interval length but also the interval location of EB and AKM are much smaller than the observation \( y_i \). On the other hand, the confidence intervals of GD are almost identical to those of ML, owing to the robustness property given in Proposition 2 while there still exist efficiency gain through “borrowing strength” from other areas in non-outlying areas, as shown in the right panel in Figure 2. Finally, we remark that the computation time of GD, including the selection step for \( \gamma \) was about 8 seconds while AKM using the R package “ebci” takes about 50 seconds on a standard laptop computer. This shows the computational efficiency of GD under large-scale situations.

Table 2: Estimates of the regression coefficients \( \beta \) and variance parameter \( \tau^2 \), and average interval lengths which are non-scaled and scaled by \( \sigma_i \), of the proposed method using \( \gamma \)-divergence (GD), the standard empirical Bayes (EB) and the AMK method by Armstrong et al. (2020).

| method | Int | PD | DPD | FD | SH | AYL | \( \tau^2 \) | Average length |
|--------|-----|----|-----|----|----|-----|----------|----------------|
| GD     | 7.81| 1.42| 4.49| 1.28| 0.79| 0.35| 11.65    | 3.57           |
| EB     | 12.20| 0.60| 8.64| 2.96| 2.64| 1.64| 231.54   | 3.81           |
| AMK    | 9.13| 1.06| 7.30| 1.69| 1.09| 0.86| 121.43   | 3.79           |
**Figure 1:** Selection criterion as a function of $\gamma$ in the proposed GD method (left) and ratios of 95% confidence interval lengths of the three methods to those of the crude ML confidence interval ($y_i + z_{0.025}\sigma_i$, $y_i + z_{0.975}\sigma_i$) against the scaled observed value $y_i/\sigma_i$ (right), obtained in the analysis of crime data.

**Figure 2:** 95% confidence intervals in outlying areas having top 30 largest $y_i$ (left) and in 30 randomly selected non-outlying areas (right).

## 6 Conclusions and Discussion

In this work, we developed a novel robust empirical Bayes confidence interval when the assumed prior distribution is subject to contamination. The proposed method uses Tweedie’s
formula to compute posterior mean and variance where the marginal likelihood is replaced with a robust alternative based on $\gamma$-divergence. Under some regularity conditions, we gave an asymptotic evaluation of the coverage probability of the proposed interval when the assumed prior is correctly specified or contaminated.

Although we employed the $\gamma$-divergence as a robust alternative of the standard likelihood, it would be possible to adopt other robust divergences such as density power divergence [Basu et al. (1998)] and $\alpha\beta$-divergence [Cichocki et al. (2011)]. The main reason to use $\gamma$-divergence is its strong robustness against contamination; that is, the robust estimator converges to a quantity with a small (latent) bias against the true value, which was the basis of our theoretical development. However, it would be interesting to compare several robust divergences in interval estimation, which leaves an interesting future study.

Acknowledgement

This work is partially supported by the Japan Society for Promotion of Science (KAKENHI) grant numbers 20K13468 and 21H00699.

Appendix

A1 Proofs for Section 2

A1.1 Proof of Proposition 1

Note that the maximum likelihood estimators $\hat{\mu}$ and $\hat{\tau}^2$ based on the marginal likelihood ([5]) solve the equations

\[
\sum_{i=1}^{n} \frac{y_i}{\tau^2 + \sigma_i^2} = \mu \sum_{i=1}^{n} \frac{1}{\tau^2 + \sigma_i^2}, \quad (A1)
\]

\[
\sum_{i=1}^{n} \frac{(y_i - \mu)^2}{(\tau^2 + \sigma_i^2)^2} = \sum_{i=1}^{n} \frac{1}{\tau^2 + \sigma_i^2}. \quad (A2)
\]
From \([A1]\), we have

\[
\left| \left( \sum_{i=1}^{n} \frac{1}{\tau_i^2 + \sigma_i^2} \right) \left( \frac{y_1}{\tau^2 + \sigma_1^2} \right)^{-1} \tilde{\mu} - 1 \right| = O \left( \frac{1}{|y_1|} \right) \to 0 \text{ as } |y_1| \to \infty.
\]

Replacing \(\mu\) with \(\tilde{\mu} = \{\sum_{i=1}^{n} (\tau_i^2 + \sigma_i^2)^{-1}\}^{-1} y_1 / (\tau^2 + \sigma_1^2)\) in \([A2]\), we have

\[
\sum_{i=1}^{n} \frac{(y_i - \tilde{\mu})^2}{(\tau^2 + \sigma_1^2)^2} = \frac{y_1^2}{(\tau^2 + \sigma_1^2)^2} \left[ 1 - \left( \sum_{i=1}^{n} \frac{1}{\tau_i^2 + \sigma_i^2} \right)^{-1} \left( \frac{1}{\tau^2 + \sigma_1^2} \right)^2 \right] + \frac{y_1^2}{(\tau^2 + \sigma_1^2)^2} \left[ \sum_{i=1}^{n} \frac{1}{\tau_i^2 + \sigma_i^2} \sum_{i=2}^{n} \frac{1}{(\tau_i^2 + \sigma_i^2)^2} + \sum_{i=1}^{n} \frac{y_i^2}{(\tau_i^2 + \sigma_i^2)^2} \right] - \frac{2y_1}{\tau^2 + \sigma_1^2} \left( \sum_{i=1}^{n} \frac{1}{\tau_i^2 + \sigma_i^2} \right)^{-1} \sum_{i=2}^{n} \frac{y_i}{\tau_i^2 + \sigma_i^2} + \sum_{i=1}^{n} \frac{y_i^2}{(\tau_i^2 + \sigma_i^2)^2} \right] \right) + O \left( \frac{|y_1|}{\tau^4} \right)
\]

as \(|y_1| \to \infty\). In this case, \([A2]\) implies

\[
\left( \sum_{i=1}^{n} \frac{1}{\tau_i^2 + \sigma_i^2} \right)^3 = \frac{y_1^2}{(\tau^2 + \sigma_1^2)^2} \left[ \left( \sum_{i=1}^{n} \frac{1}{\tau_i^2 + \sigma_i^2} \right)^2 - \sum_{i=1}^{n} \frac{1}{(\tau_i^2 + \sigma_i^2)^2} \right] - 2 \left( \sum_{i=1}^{n} \frac{1}{\tau_i^2 + \sigma_i^2} \right)^{-1} \left( \frac{1}{\tau^2 + \sigma_1^2} \right) \right) + O \left( \frac{|y_1|}{\tau^4} \right) \left( \sum_{i=1}^{n} \frac{1}{\tau_i^2 + \sigma_i^2} \right)^2.
\]

Since the left hand side is \(O(\tau^{-6})\) and the right hand side is \(O(|y_1|^2 \tau^{-8})\), we have \(\tilde{\tau}^2 = O(|y_1|^2)\). Therefore,

\[
\tilde{\hat{\theta}}_i - y_i = - \frac{\sigma_i^2}{\tau_i^2 + \sigma_i^2} (y_i - \tilde{\mu})
\]

\[
= - \frac{\sigma_i^2}{\tau_i^2 + \sigma_i^2} \left\{ y_i - \left( \sum_{i=1}^{n} \frac{1}{\tau_i^2 + \sigma_i^2} \right)^{-1} \left( \frac{y_1}{\tau^2 + \sigma_1^2} \right) \left( 1 + o(1) \right) \right\} \to 0
\]
and $s^2_i = \frac{\tilde{\tau}^2 \sigma^2_i}{\gamma_i + \sigma^2_i}$ → $\sigma^2_i$ as $|y_1| \to \infty$.

A1.2 Proof of Proposition 2

Define $D_{\gamma, y_1}(\psi) := n^{-1} \sum_{i=1}^n D_{\gamma}(y_i; \psi)$. Let $\hat{\psi}_1^{(\gamma)} = (\hat{\mu}_1^{(\gamma)}, \hat{\tau}_1^{(\gamma)})$ be the maximizer of

$$D_{\gamma, \infty}(\psi) := \frac{1}{n} \sum_{i=2}^n D_{\gamma}(y_i; \psi)$$

and let $\Psi_{\infty} \subset \mathbb{R} \times (0, \infty)$ be a compact set such that $\hat{\psi}_1^{(\gamma)} \in \Psi_{\infty}$.

First we prove $\hat{\psi}_1^{(\gamma)} \to \hat{\psi}_1^{(\gamma)*}$ as $|y_1| \to \infty$. Note that $\hat{\psi}_1^{(\gamma)} \in \Psi_{\infty}$ for sufficiently large $|y_1|$, and as $|y_1| \to \infty$, $\sup_{\psi \in \Psi_{\infty}} |D_{\gamma, y_1}(\psi) - D_{\gamma, \infty}(\psi)| \to 0$. Then we have

$$|D_{\gamma, \infty}(\hat{\psi}_1^{(\gamma)}) - D_{\gamma, \infty}(\hat{\psi}_1^{(\gamma)*})| \leq |D_{\gamma, \infty}(\hat{\psi}_1^{(\gamma)}) - D_{\gamma, y_1}(\hat{\psi}_1^{(\gamma)})| + |D_{\gamma, y_1}(\hat{\psi}_1^{(\gamma)}) - D_{\gamma, y_1}(\hat{\psi}_1^{(\gamma)*})|
+ |D_{\gamma, y_1}(\hat{\psi}_1^{(\gamma)}) - D_{\gamma, \infty}(\hat{\psi}_1^{(\gamma)})|$$
$$\leq 2 \sup_{\psi \in \Psi_{\infty}} |D_{\gamma, y_1}(\psi) - D_{\gamma, \infty}(\psi)| + |D_{\gamma, y_1}(\hat{\psi}_1^{(\gamma)}) - D_{\gamma, y_1}(\hat{\psi}_1^{(\gamma)*})|
\to 0 \text{ as } |y_1| \to \infty. \quad (A3)$$

Since $D_{\gamma, \infty}(\psi)$ is continuous over $\psi \in \Psi_{\infty}$, (A3) implies that

$$\hat{\psi}_1^{(\gamma)} \to \hat{\psi}_1^{(\gamma)*} \text{ as } |y_1| \to \infty. \quad (A4)$$

Next we prove (i) $\hat{\theta}_1^{(\gamma)} - y_1 \to 0$ and $\hat{s}_1^{(\gamma)} \to \sigma_1^2$, and (ii) $\hat{\theta}_i^{(\gamma)} \to \bar{\theta}_i^{(\gamma)}$ and $\hat{s}_i^{(\gamma)} \to s_i^{(\gamma)}$ for $i = 2, \ldots, n$ as $|y_1| \to \infty$. It is straightforward to show (ii) by applying (A4). Since $|y_1|^2 \varphi(y_1 : \hat{\mu}_1^{(\gamma)}, (\hat{\tau}_1^{(\gamma)})^2 + \sigma_1^2) \gamma \to 0$ as $|y_1| \to \infty$, we have $\hat{\theta}_1^{(\gamma)} - y_1 \to 0$ and $\hat{s}_1^{(\gamma)} \to \sigma_1^2$.

A2 Proofs for Section 3

We first prove the results on $\gamma_{\text{op}}$ (Theorems 1 and 2) and then prove the results on confidence intervals (Theorems 2 and 4).
A2.1 Proof of Theorem 1

Before we see the asymptotic property of $\gamma_{op}$, we prepare some auxiliary results. Our approach is based on general theory for $M$-estimators (e.g. van der Vaart (1998)). Lemma [A1] is a uniform law of large numbers for a class of functions. Lemma [A2] provides the consistency of $\hat{\psi}(\gamma)$.

Lemma A1. Let $\{(Z_i, \sigma_i^2)\}_{i \geq 1}$ be a sequence of i.i.d. random vectors such that $0 < \sigma_L \leq \sigma_1 \leq \sigma_U < \infty$ and $E[Z_i^2] < \infty$. Define $\mathcal{F} = \{ (y_1, y_2) \mapsto f_{\psi, \gamma}(y_1, y_2) : \psi \in \Psi, \gamma \in [0, 1] \}$ and

$$f_{\psi, \gamma}(y_1, y_2) = \phi(y_1 : \mu, \tau^2 + y_2)^\gamma \frac{y_2^2}{(\tau^2 + y_2)^2} \times \{ \gamma(y_1 - \mu)^2 - (\tau^2 + y_2) \} \{ 2\pi(\tau^2 + y_2) \}^{\gamma/2(1+\gamma)}$$

where $\psi = (\mu, \tau^2)$ and $\Psi$ is a compact subset of $\mathbb{R} \times (0, \infty)$. Then we have

$$\sup_{f \in \mathcal{F}} |P_n f - Pf| \overset{P}{\to} 0 \text{ as } n \to \infty,$$

where $P_n f = \frac{1}{n} \sum_{i=1}^n f(Z_i, \sigma_i^2)$ and $P f = E[f(Z_i, \sigma_i^2)]$.

Proof. Because $0 < \sigma_L \leq \sigma_1 \leq \sigma_U < \infty$, for $\psi \in \Psi$ and $\gamma \in [0, 1]$, there exists a positive constant $C$ such that

$$|f_{\psi, \gamma}(y_1, y_2)| \leq C(y_1 - \mu)^2.$$  

Because $\Psi$ is compact and $E[Z_i^2] < \infty$, there exists a function $F$ such that $E[F(Z_i, \sigma_i^2)] < \infty$ and $\sup_{\psi \in \Psi, \gamma \in [0, 1]} |f_{\psi, \gamma}(y_1, y_2)| \leq F(y_1, y_2)$. In addition, the map $(\psi, \gamma) \mapsto f_{\psi, \gamma}(y_1, y_2)$ is continuous for every $(y_1, y_2)$ and $\Psi \times [0, 1]$ is compact. Hence, from Example 19.8 in van der Vaart (1998), $\mathcal{F}$ is Glivenko-Cantelli.

Lemma A2. Let $\{\sigma_i\}$ be a sequence of i.i.d. random variables and there exist positive constants $\sigma_L$ and $\sigma_U$ such that $0 < \sigma_L \leq \sigma_1 \leq \sigma_U < \infty$. Assume that $\Psi \subset \mathbb{R} \times (0, \infty)$ is a compact parameter space of $\psi_* := (\mu_*, \tau_*^2)$. When $y_i | \theta_i, \sigma_i \sim N(\theta_i, \sigma_i^2)$ and $\theta_i \sim N(\mu_*, \tau_*^2)$, for any $\gamma \in [0, 1]$ we have

$$\hat{\mu}_{\gamma} \overset{P}{\to} \mu_* \text{ and } \hat{\tau}_{\gamma}^2 \overset{P}{\to} \tau_*^2.$$
Proof. For any $\gamma \in (0, 1]$, we define

$$D_\gamma(y_i, \sigma_i^2; \psi) := \frac{1}{\gamma} \phi(y_i; \mu, \tau^2 + \sigma_i^2) \gamma c_\gamma(\tau^2) - \frac{1}{\gamma},$$

In addition, we define $D_0(y_i, \sigma_i^2; \psi) := \log \{ \phi(y_i; \mu, \tau^2 + \sigma_i^2) \}$. Then, we have $\hat{\psi}(\gamma) = (\hat{\mu}_\gamma, \hat{\tau}_\gamma^2) = \text{argmax}_{\psi \in \Psi} n^{-1} \sum_{i=1}^n D_\gamma(y_i, \sigma_i^2; \psi)$. Similar to Lemma A1 from Example 19.8 in van der Vaart (1998), for any $\gamma \in [0, 1]$ we have

$$\sup_{\psi \in \Psi} \left| \frac{1}{n} \sum_{i=1}^n D_\gamma(y_i, \sigma_i^2; \psi) - E[D_\gamma(y_i, \sigma_i^2; \psi)] \right| = o_p(1). \quad (A5)$$

We observe that

$$\frac{\partial}{\partial \mu} E[D_\gamma(y_i, \sigma_i^2; \psi)] = E \left[ E \left[ \phi(y_i; \mu, \tau^2 + \sigma_i^2) \gamma \left( \frac{y_i - \mu}{\tau^2 + \sigma_i^2} \right) c_\gamma(\tau^2) \mid \sigma_i \right] \right],$$

$$\frac{\partial}{\partial \tau^2} E[D_\gamma(y_i, \sigma_i^2; \psi)] = \frac{1}{2} E \left[ E \left[ \phi(y_i; \mu, \tau^2 + \sigma_i^2) \gamma \left\{ (y_i - \mu)^2 - \frac{\tau^2 + \sigma_i^2}{1 + \gamma} \right\} c_\gamma(\tau^2) \mid \sigma_i \right] \right].$$

Since $\phi(y_i; \mu, \tau^2 + \sigma_i^2)$ is symmetric around $y_i = \mu$, the first derivative is 0 when $\mu = \mu_*$ and $\tau^2 = \tau_*^2$. Regarding the second derivative, note that when $\mu = \mu_*$ and $\tau^2 = \tau_*^2$,

$$E[\phi(y_i; \mu, \tau^2 + \sigma_i^2) \gamma | \sigma_i] = V_i(1 + \gamma)^{-1/2},$$

$$E[\phi(y_i; \mu, \tau^2 + \sigma_i^2) \gamma(y_i - \mu)^2 | \sigma_i] = V_i(1 + \gamma)^{-3/2}(\tau^2 + \sigma_i^2), \quad (A6)$$

where $V_i = \{2\pi(\tau^2 + \sigma_i^2)\}^{-\gamma/2}$. This shows that the second derivative is also 0. This implies that $E[D_\gamma(y_i, \sigma_i^2; \psi)]$ is maximized at $\psi = (\mu_*, \tau_*^2)$. Hence, from Theorem 5.7 in van der Vaart (1998), we obtain consistency of $\hat{\mu}_\gamma$ and $\hat{\tau}_\gamma^2$ for all $\gamma \in [0, 1]$. \hfill \Box

Now we move on to the proof of Theorem [1]. We define

$$\hat{Q}_n(\gamma; \mu, \tau^2) := \frac{1}{n} \sum_{i=1}^n \phi(y_i; \mu, \tau^2 + \sigma_i^2) \gamma \left( \frac{\sigma_i^4}{\tau^2 + \sigma_i^2} \right) \left\{ \gamma(y_i - \mu)^2 - (\tau^2 + \sigma_i^2) \right\} c_\gamma(\tau^2),$$

$$Q(\gamma; \mu, \tau^2) := E \left[ \phi(y_i; \mu, \tau^2 + \sigma_i^2) \gamma \left( \frac{\sigma_i^4}{\tau^2 + \sigma_i^2} \right) \left\{ \gamma(y_i - \mu)^2 - (\tau^2 + \sigma_i^2) \right\} c_\gamma(\tau^2) \right],$$

then $\gamma_{op} = \text{argmin}_{\gamma \in \Gamma} Q_n(\gamma; \hat{\mu}_\gamma, \hat{\tau}_\gamma^2)$. Because $\hat{\psi}_\gamma := (\hat{\mu}_\gamma, \hat{\tau}_\gamma^2) = \text{argmax}_{\psi \in \Psi} n^{-1} \sum_{i=1}^n D_\gamma(y_i, \sigma_i^2; \psi),$
we have \((\hat{\mu}_n, \hat{\tau}_n^2) \in \Psi\) with probability 1. From the triangular inequality, we obtain

\[
\left| \hat{Q}_n(\gamma; \hat{\mu}_n, \hat{\tau}_n^2) - Q_n(\gamma; \mu^*, \tau^2_n) \right|
\leq \left| \hat{Q}_n(\gamma; \hat{\mu}_n, \hat{\tau}_n^2) - Q(\gamma; \hat{\mu}_n, \hat{\tau}_n^2) \right| + \left| Q(\gamma; \hat{\mu}_n, \hat{\tau}_n^2) - Q(\gamma; \mu^* \tau^2_n) \right|
\leq 2 \sup_{(\mu, \tau^2) \in \Psi} \left| \hat{Q}_n(\gamma; \mu^* \tau^2_n) - Q(\gamma; \mu^* \tau^2_n) \right| + \left| Q(\gamma; \hat{\mu}_n, \hat{\tau}_n^2) - Q(\gamma; \mu^* \tau^2_n) \right|.
\]

Because \(\hat{\mu}_n \xrightarrow{p} \mu^* \) and \(\hat{\tau}_n^2 \xrightarrow{p} \tau^2_n\) from Lemma A2 and \(Q(\gamma; \mu, \tau^2)\) is continuous in \(\mu\) and \(\tau^2\), it follows from Lemma A1 that \(\hat{Q}_n(\gamma; \hat{\mu}_n, \hat{\tau}_n^2) = \hat{Q}_n(\gamma; \mu^*, \tau^2_n) + o_p(1) = Q(\gamma; \mu^*, \tau^2_n) + o_p(1)\).

From (A6), we have

\[
Q(\gamma; \mu^*, \tau^2_n) = -E \left[ \frac{\sigma_1^4}{\tau_n^2 + \sigma_1^2} \left\{ 2\pi(\tau_n^2 + \sigma_1^2) \right\}^{-\gamma/2(1+\gamma)}(1+\gamma)^{-3/2} \right].
\]

Then, it follows that

\[
\frac{\partial}{\partial \gamma} Q(\gamma; \mu^*, \tau^2_n) = \frac{1}{2} E \left[ \frac{\sigma_1^4}{\tau_n^2 + \sigma_1^2} \left\{ 2\pi(\tau_n^2 + \sigma_1^2) \right\}^{-\gamma/2(1+\gamma)}(1+\gamma)^{-5/2} \left\{ \frac{\log(2\pi(\tau_n^2 + \sigma_1^2))}{1+\gamma} + 3 \right\} \right] 
\geq \frac{\sigma_1^4}{2(\tau_n^2 + \sigma_1^2)} \left\{ 2\pi(\tau_n^2 + \sigma_1^2) \right\}^{-\gamma/2(1+\gamma)}(1+\gamma)^{-5/2} \left\{ \frac{\log(2\pi(\tau_n^2 + \sigma_1^2))}{1+\gamma} + 3 \right\}.
\]

Then, under the condition \(\tau_n^2 + \sigma_1^2 > (2\pi)^{-1}e^{-3}\), we have

\[
\frac{\partial}{\partial \gamma} Q(\gamma; \mu^*, \tau_n^2) > 0 \text{ for } \gamma \in [0, 1].
\]

This implies that \(Q(\gamma; \mu^*, \tau^2_n)\) is strictly increasing on \(\gamma \in [0, 1]\). Hence, \(\hat{Q}_n(0; \hat{\mu}_0, \hat{\tau}_0^2) < \min_{\gamma \in \Gamma \setminus \{0\}} Q_n(\gamma; \hat{\mu}_n, \hat{\tau}_n^2)\) holds with probability approaching 1. This implies that \(P(\gamma_{op} = 0) \to 1\).

A2.2 Proof of Theorem \(\beta\)

Fix \(\omega \in (0, 1)\) and \(\delta\). From Lemmas A4 and A5, we obtain \(\hat{\psi}_n^{(\gamma)} = \psi_1^{(\gamma)} + O_p(n^{-1/2})\) and \(\psi_1^{(\gamma)} = \psi_\ast + O(\omega p(\psi_\ast, \gamma, \delta))\) for any \(\gamma \in \Gamma \setminus \{0\}\). Similar to Lemma A2, for any \(\gamma \in \Gamma\) we
have

\[
\sup_{\psi \in \Psi} \left| \hat{Q}_n (\gamma; \psi) - \hat{Q} (\gamma; \psi) \right| = o_p(1),
\]

where \( \hat{Q}(\gamma; \psi) := E_{f,\psi} \left[ \phi(y_i; \mu, \tau, \sigma_i^2) \right] \). For any \( \gamma \in \Gamma \), we obtain

\[
\hat{Q}_n (\gamma; \hat{\psi}_n^{(n)}) \to_p \hat{Q} (\gamma; \hat{\psi}_n^{(n)}).
\]

Because \( \hat{Q}(\gamma; \psi) \) is differentiable in \( \psi \) and \( \hat{\psi}_n^{(n)} = \psi_\ast + O (\omega \rho (\psi_\ast, \gamma, \delta)) \), \( \hat{Q} (\gamma; \hat{\psi}_n^{(n)}) = \hat{Q} (\gamma; \psi_\ast) + O (\omega \rho (\psi_\ast, \gamma, \delta)) \) holds for any \( \gamma \in \Gamma \setminus \{0\} \).

Next, we consider \( \gamma = 0 \). Let \( \psi_0^{(n)} := (\mu_0, \tau_0^2) \) be the probability limit of \( \hat{\psi}_n^{(0)} := (\mu_0, \tau_0^2) \) under \( \theta_i \sim f(\theta_i) \). Similar to the above argument, we have

\[
\hat{Q}_n (0; \hat{\psi}_n^{(0)}) = -\frac{1}{n} \sum_{i=1}^{n} \frac{\sigma_i^4}{\tau_i^2 + \sigma_i^4} E \left[ \frac{\sigma_i^4}{\tau_i^2 + \sigma_i^4} \right].
\]

Then, \( (\mu_0, \tau_0^2) \) satisfies the following equations:

\[
E_{f,\psi_0} \left[ \frac{y_i - \mu_0}{\tau_0^2 + \sigma_i^2} \right] = 0 \quad \text{and} \quad E_{f,\psi_0} \left[ \frac{(y_i - \mu_0)^2 - (\tau_0^2 + \sigma_i^2)}{(\tau_0^2 + \sigma_i^2)^2} \right] = 0.
\]

Because \( E[y_i | \sigma_i] = (1 - \omega) \mu_i + \omega \mu_\delta \), the first equality implies \( \mu_0 = (1 - \omega) \mu_\ast + \omega \mu_\delta \). Hence, we obtain

\[
E_{f,\psi_0} \left[ \frac{(y_i - \mu_0)^2}{(\tau_0^2 + \sigma_i^2)^2} \right] = \frac{(1 - \omega) \left\{ (\tau_0^2 + \sigma_i^2) + (\mu_0 - \mu_\delta)^2 \right\} + \omega \left\{ (\tau_0^2 + \sigma_i^2) + (\mu_0 - \mu_\delta)^2 \right\}}{(\tau_0^2 + \sigma_i^2)^2} = \frac{(1 - \omega)(\tau_0^2 + \sigma_i^2) + \omega(\tau_0^2 + \sigma_i^2) + \{\omega^2 + (1 - \omega)^2\} (\mu_\ast - \mu_\delta)^2}{(\tau_0^2 + \sigma_i^2)^2}.
\]

Combined with the second equation, \( \tau_0^2 \) must satisfy the following equation:

\[
E_{f,\psi_0} \left[ \frac{(1 - \omega)(\tau_0^2 + \sigma_i^2) + \omega(\tau_0^2 + \sigma_i^2) + \{\omega^2 + (1 - \omega)^2\} (\mu_\ast - \mu_\delta)^2 - (\tau_0^2 + \sigma_i^2)}{(\tau_0^2 + \sigma_i^2)^2} \right] = 0.
\]

Because \( \omega \in (0, 1/2) \) and \( \{\omega^2 + (1 - \omega)^2\} (\mu_\ast - \mu_\delta)^2 \geq \frac{1}{2} (\tau_0^2 + \sigma_i^2) \), we have

\[
(1 - \omega)(\tau_0^2 + \sigma_i^2) + \omega(\tau_0^2 + \sigma_i^2) + \{\omega^2 + (1 - \omega)^2\} (\mu_\ast - \mu_\delta)^2 - (\tau_0^2 + \sigma_i^2) \\
\geq \omega(\tau_0^2 + \sigma_i^2) + (\tau_0^2 - \tau_i^2) \geq \omega \sigma_i^2 + (\tau_0^2 - \tau_i^2).
\]
Hence, we have $\tau_1^2 - \tau_2^* > \omega \sigma_L^2$. As a result, we obtain

$$Q(0; \psi_1^{(0)}) := -E\left[\frac{\sigma_1^4}{\tau_1^2 + \sigma_1^2}\right] > Q(0; \psi_1) := -E\left[\frac{\sigma_1^4}{\tau_2^* + \sigma_1^2}\right].$$

We note that the difference is strictly positive uniformly with respect to $\delta$ that satisfies $(\mu_1 - \mu_2)^2 > (\tau_2^* + \sigma_1^2)$. 

From the definition of $\rho(\psi, \gamma, \delta)$, for any $\gamma \in \Gamma \setminus \{0\}$ we observe that $\tilde{Q}(\gamma; \psi_1^{(0)}) = E_{\psi_1^{(0)}}\left[w_\gamma(y_i) \frac{\sigma_1^4}{(\tau_2^* + \sigma_1^2)^2} \left\{\gamma(y_i - \mu_2)^2 - (\tau_2^* + \sigma_1^2)\right\} c_\gamma(\tau_2^*)\right]$

$$= (1 - \omega)E_{\psi_1^{(0)}}\left[w_\gamma(y_i) \frac{\sigma_1^4}{(\tau_2^* + \sigma_1^2)^2} \left\{\gamma(y_i - \mu_2)^2 - (\tau_2^* + \sigma_1^2)\right\} c_\gamma(\tau_2^*)\right] + O(\omega \rho(\psi_1, \gamma, \delta))$$

$$= (1 - \omega)Q(\gamma; \psi_1^{(0)}) + O(\omega \rho(\psi_1, \gamma, \delta)),$$

where $E_{\psi_1^{(0)}}$ denotes the expectation under $y_i|\sigma_i \sim N(\mu_2, \tau_2^* + \sigma_1^2)$. As discussed in the proof of Theorem 1, $Q(\gamma; \psi_1^{(0)})$ is continuous in $\gamma$ and $Q(\gamma; \psi_1^{(0)}) > Q(0; \psi_1^{(0)}) = Q(0; \psi_1)$. As a result, we obtain

$$Q(\gamma_1; \psi_1^{(0)}) = Q(\gamma_1; \psi_1) + O(\omega \rho(\psi_1, \gamma_1, \delta)) = (1 - \omega)Q(\gamma_1; \psi_1) + O(\omega \rho(\psi_1, \gamma_1, \delta))$$

$$< Q(0; \psi_1) + O(\omega \rho(\psi_1, \gamma_1, \delta)) < Q(0; \psi_1^{(0)}) + O(\omega \rho(\psi_1, \gamma_1, \delta)).$$

Hence, if $\rho(\psi_1, \gamma_1, \delta)$ is sufficiently small, then we have $P(\gamma > 0) \to 1$ as $n \to \infty$.

### A2.3 Proof of Theorem 2

The following lemma (Nazarov’s inequality) plays an important role to show the asymptotic validity of the proposed confidence intervals.

**Lemma A3** (Nazarov (2003); Lemma A.1 in Chernozhukov et al. (2017)). Let $Y = (Y_1, \ldots, Y_p)^T$ be a centered Gaussian random vector in $\mathbb{R}^p$ such that $E[Y_i^2] \geq \sigma_0^2$ for all $i = 1, \ldots, p$ and some constant $\sigma_0 > 0$. Then for every $y \in \mathbb{R}^p$ and $\delta > 0$,

$$P(Y \leq y + \delta) - P(Y \leq y) \leq \frac{\delta}{\sigma_0}(\sqrt{2\log p} + 2).$$
Now we move on to the proof of Theorem 2. Recall that \( \hat{C}_{i,1-\alpha}^{(\gamma)} = (\hat{\theta}_1^{(\gamma)} - z_{\alpha/2} \tilde{s}_i^{(\gamma)}, \hat{\theta}_1^{(\gamma)} + z_{\alpha/2} \tilde{s}_i^{(\gamma)}) \). Then \( \hat{C}_{i,1-\alpha}^{(0)} = \hat{C}_{i,1-\alpha} \) (see (4) for the definition of \( \hat{C}_{i,1-\alpha} \)). Since \( P(\gamma_{op} = 0) \to 1 \) as \( n \to \infty \), we have

\[
P \left( \theta_i \in \hat{C}_{i,1-\alpha}^{(\gamma_{op})} \right) = P \left( \left\{ \theta_i \in \hat{C}_{i,1-\alpha}^{(\gamma_{op})} \right\} \cap \{ \gamma_{op} = 0 \} \right) + P \left( \left\{ \theta_i \in \hat{C}_{i,1-\alpha}^{(\gamma_{op})} \right\} \cap \{ \gamma_{op} > 0 \} \right)
\leq P \left( \theta_i \in \hat{C}_{i,1-\alpha} \right) + P(\gamma_{op} > 0) = P \left( \theta_i \in \hat{C}_{i,1-\alpha} \right) + o(1).
\]

Likewise, we have

\[
P \left( \theta_i \in \tilde{C}_{i,1-\alpha} \right) = P \left( \left\{ \tilde{s}_i^{-1} (\theta_i - \tilde{\theta}_i) \right\} \leq z_{\alpha/2} \right) = E \left[ P \left( \left\{ \tilde{s}_i^{-1} (\theta_i - \tilde{\theta}_i) \right\} \leq z_{\alpha/2} \right| y_i, \sigma_i \right]
\]

where \( \Phi(\cdot) \) is the cumulative distribution function of the standard normal distribution.

Now we will show \( P \left( \theta_i \in \tilde{C}_{i,1-\alpha} \right) = P \left( \theta_i \in \hat{C}_{i,1-\alpha} \right) + o(1) \). Since \( \tilde{\tau}^2 \overset{P}{\to} \tau^*_s \) and \( \tilde{\mu} \overset{P}{\to} \mu_s \) as \( n \to \infty \) from Lemma A2 there exists a sequence of constants \( \epsilon_n \to 0 \) such that \( P(\Omega_n) \to 1 \) where \( \Omega_n = \{ \max \{|\tilde{\tau}^2 - \tau^*_s|, |\tilde{\mu} - \mu_s|\} \leq \epsilon_n \} \). Observe that

\[
P \left( \theta_i \leq \tilde{\theta}_i + \tilde{s}_i z_{\alpha/2} \right)
= P \left( \theta_i \leq \tilde{\theta}_i + \tilde{s}_i z_{\alpha/2} + (\tilde{\theta}_i - \tilde{\theta}_i) + (\tilde{s}_i - \tilde{s}_i) z_{\alpha/2} \right)
= P \left( \left\{ \theta_i \leq \tilde{\theta}_i + \tilde{s}_i z_{\alpha/2} + (\tilde{\theta}_i - \tilde{\theta}_i) + (\tilde{s}_i - \tilde{s}_i) z_{\alpha/2} \right\} \cap \Omega_n \cap \{ |y_i - \mu_s| \leq \epsilon_n^{-1/2} \} \right)
+ P \left( \left\{ \theta_i \leq \tilde{\theta}_i + \tilde{s}_i z_{\alpha/2} + (\tilde{\theta}_i - \tilde{\theta}_i) + (\tilde{s}_i - \tilde{s}_i) z_{\alpha/2} \right\} \cap \Omega_n \cap \{ |y_i - \mu_s| > \epsilon_n^{-1/2} \} \right)
+ P \left( \left\{ \theta_i \leq \tilde{\theta}_i + \tilde{s}_i z_{\alpha/2} + (\tilde{\theta}_i - \tilde{\theta}_i) + (\tilde{s}_i - \tilde{s}_i) z_{\alpha/2} \right\} \cap \Omega_n^c \cap \{ |y_i - \mu_s| \leq \epsilon_n^{-1/2} \} \right)
+ P \left( |y_i - \mu_s| > \epsilon_n^{-1/2} \right) + P(\Omega_n^c)
= P \left( \left\{ \theta_i \leq \tilde{\theta}_i + \tilde{s}_i z_{\alpha/2} + (\tilde{\theta}_i - \tilde{\theta}_i) + (\tilde{s}_i - \tilde{s}_i) z_{\alpha/2} \right\} \cap \Omega_n \cap \{ |y_i - \mu_s| \leq \epsilon_n^{-1/2} \} \right) + o(1).
\]

(A7)
On $\Omega_n \cap \{|y_i - \mu_s| \leq \epsilon_{n}^{-1/2}\}$, we have

\[
|\hat{\theta}_i - \tilde{\theta}_i| \leq \left| \frac{\sigma^2}{\tau^2 + \sigma^2} - \frac{\sigma^2}{\tau^2 + \sigma^2} \right| |y_i - \mu_s| + \frac{\sigma^2}{\tau^2 + \sigma^2} |\hat{\mu} - \mu_s|
\]
\[
+ \left| \frac{\sigma^2}{\tau^2 + \sigma^2} - \frac{\sigma^2}{\tau^2 + \sigma^2} \right| |\hat{\mu} - \mu_s|
\]
\[
\leq \frac{\sigma^2}{\left(\tau^2 + \sigma^2 - \epsilon_n^2\right)\epsilon_n^{-1/2}} + \frac{\sigma^2}{\tau^2 + \sigma^2} \epsilon_n + \frac{\sigma^2}{\left(\tau^2 + \sigma^2 - \epsilon_n^2\right)\epsilon_n^{-1/2}}
\]
\[
\leq C_1(\sigma^2_L, \sigma^2_U, \tau_s)\epsilon_n^{1/2}, \tag{A8}
\]

and

\[
|\tilde{s}_i - \hat{s}_i| \leq \frac{\sigma^2}{\tilde{s}_i^2 - \hat{s}_i^2} \frac{1}{|\tilde{s}_i^2 - \hat{s}_i^2|}
\]
\[
\leq \left( \frac{\sigma^2_{L}^2 - \tau^2_s}{\tau^2_s + \sigma^2_s} + \frac{\sigma^2_{L}^2 - \tau^2_s}{\tau^2_s + \sigma^2_s} \right) \frac{1}{|\tilde{s}_i^2 - \hat{s}_i^2|}
\]
\[
\leq \left( \frac{\sigma^2_{L}^2 - \tau^2_s}{\tau^2_s + \sigma^2_s} + \frac{\sigma^2_{L}^2 - \tau^2_s}{\tau^2_s + \sigma^2_s} \right) \frac{1}{|\tilde{s}_i^2 - \hat{s}_i^2|}
\]
\[
\leq C_2(\sigma^2_L, \sigma^2_U, \tau_s)\epsilon_n \tag{A9}
\]

for some positive constants $C_1(\sigma^2_L, \sigma^2_U, \tau_s)$ and $C_2(\sigma^2_L, \sigma^2_U, \tau_s)$. Combining (A8), (A9), and Lemma A3 we have

\[
P\left( \left\{ \theta_i \leq \hat{\theta}_i + \tilde{s}_i z_{\alpha/2} + \left(\tilde{s}_i - \hat{s}_i\right) z_{\alpha/2} / 2 \right\} \cap \Omega_n \cap \{|y_i - \mu_s| \leq \epsilon_{n}^{-1/2}\} \right| y_i, \sigma_i
\]
\[
\leq P\left( \theta_i \leq \hat{\theta}_i + \tilde{s}_i z_{\alpha/2} / 2 + C_1(\sigma^2_L, \sigma^2_U, \tau_s)\epsilon_n^{1/2} + C_2(\sigma^2_L, \sigma^2_U, \tau_s)\epsilon_n z_{\alpha/2} \right| y_i, \sigma_i
\]
\[
\leq P\left( \theta_i \leq \hat{\theta}_i + \tilde{s}_i z_{\alpha/2} / 2 \right| y_i, \sigma_i + C_3(\sigma^2_L, \sigma^2_U, \tau_s)\epsilon_n^{1/2}
\]
\[
= 1 - \alpha/2 + C_3(\sigma^2_L, \sigma^2_U, \tau_s)\epsilon_n^{1/2} \tag{A10}
\]

for some positive constant $C_3(\sigma^2_L, \sigma^2_U, \tau_s)$. Combining (A7) and (A10), we have $P\left( \theta_i \leq \hat{\theta}_i + \tilde{s}_i z_{\alpha/2} / 2 \right) \leq 1 - \alpha/2 + o(1)$. Likewise, we have $P\left( \theta_i \leq \hat{\theta}_i + \tilde{s}_i z_{\alpha/2} / 2 \right) \geq 1 - \alpha/2 - o(1)$. Almost the same argument yields $P\left( \theta_i \leq \hat{\theta}_i - \tilde{s}_i z_{\alpha/2} / 2 \right) = \alpha/2 + o(1)$. Therefore, we have

\[
P\left( \theta_i \in \tilde{\mathcal{C}}_{\text{top}}(\gamma_{\text{top}}) \right) = 1 - \alpha + o(1).
\]
A2.4 Proof of Theorem 4

To investigate the asymptotic properties of the confidence interval \( \hat{C}_{i,1-\alpha} \), we first prepare two auxiliary results. From Lemma A4, \( \hat{\psi}^{(\gamma)} \) converge to a ”pseudo-true” parameter \( \psi_1^{(\gamma)} \), and from Lemma A5, \( \psi_1^{(\gamma)} \) has a bias \( O(\omega \rho_\psi) \) with respect to the true parameter \( \psi_* \).

Lemma A4 (Theorem 5.41 and 5.42 of van der Vaart [1998]: modified version of Theorem 4.1 of Fujisawa [2013]). Suppose that \( \gamma > 0 \) and \( x_1 = (y_1, \sigma_1^2) \), \( \ldots \), \( x_n = (y_n, \sigma_n^2) \) are i.i.d samples from the model \( y_i | \theta_i, \sigma_i \sim N(\theta_i, \sigma_i^2) \), \( \theta_i \sim f(\theta_i) \) and that there exist positive constants \( \sigma_L \) and \( \sigma_U \) such that \( 0 < \sigma_L \leq \sigma_i \leq \sigma_U < \infty \). We assume: (a)

The function \( \partial_\psi D_\gamma(x; \psi) \) is twice continuously differentiable with respect to \( \psi \) for any \( x = (y, \sigma^2) \in \mathbb{R} \times [\sigma_L^2, \sigma_U^2] \). (b) There exists a root \( \psi_1^{(\gamma)} \) such that \( E_{f_*} [\partial_\psi D_\gamma(x; \psi)] = 0 \).

(c) \( E_{f_*} [||\partial_\psi D_\gamma(x; \psi)||^2] < \infty \). (d) \( E_{f_*} [\partial(\partial_\psi D_\gamma(x; \psi))]/\partial\psi^\top \) exists and is nonsingular. (e) The second-order differentials of \( \partial_\psi D_\gamma(x; \psi) \) with respect to \( \psi \) are dominated by a fixed integrable function \( D(x) \) in a neighborhood of \( \psi = \psi_1^{(\gamma)} \).

Then there exists a sequence of roots \( \{\hat{\psi}_n^{(\gamma)}\}_{n=1}^\infty \), such that

(i) \( \hat{\psi}_n^{(\gamma)} \xrightarrow{p} \psi_1^{(\gamma)} \),

(ii) \( \sqrt{n}(\hat{\psi}_n^{(\gamma)} - \psi_1^{(\gamma)}) \xrightarrow{d} N(0, \tau_1^2(\psi_1^{(\gamma)})) \),

where \( \tau_1^2(\psi_N) = J_f(\psi) K_f(\psi) \{J_f(\psi)^\top\}^{-1}, J_f(\psi) = E_{f_*} [\partial(\partial_\psi D_\gamma(x; \psi))]/\partial\psi^\top \), \( K_f(\psi) = E_{f_*} [\partial_\psi D_\gamma(x; \psi)\partial_\psi D_\gamma(x; \psi)^\top] \).

Assumptions (a)-(e) in Lemma A4 are easy to verify for the estimators based on \( \gamma \)-divergence (see Fujisawa [2013] for example).

In the following results, we suppose that the maximization of the objective function in (10) is performed on the set \( \Psi_n = \{\psi : \rho_\psi^{1/\gamma} \leq K \rho_\psi^{1/\gamma} \} \) such that \( \psi_* \), \( \psi_1^{(\gamma)} \in \Psi_n \), where \( K \) is a positive constant and \( \psi_* = (\mu_*, \tau_*^2) \) and \( \rho_\psi \) is defined in the same way as \( \rho_\psi \) by replacing \( \psi_* \) with \( \psi = (\mu, \tau^2) \).

Lemma A5 (Theorem 3.3 of Fujisawa and Eguchi [2008]). Suppose that \( \psi_* \), \( \psi_1^{(\gamma)} \in \Psi_n \). Then it holds that \( \psi_1^{(\gamma)} = \psi_* + O(\omega \rho_\psi) \).

From Lemmas A4 and A5 we have

\[ \hat{\psi}_n^{(\gamma)} = \psi_* + O(\omega \rho_\psi) + O_p(n^{-1/2}) \]
and this enables us to evaluate coverage probabilities of the proposed confidence intervals $\hat{C}_{i,1-\alpha}$.

Now we move on to the proof of Theorem 4. Since

$$\hat{\psi}^{(\gamma)} = (\hat{\mu}_i, \hat{\tau}_i^2) = \psi + O(\omega\rho_{\psi_*}) + O_p(1/\sqrt{n})$$

as $n \to \infty$ from Lemmas[A4] and [A5], there exists a sequence of constants $\hat{c}_n \to 0$ such that $\hat{c}_n = n^{-\zeta/2} + (\omega\rho_{\psi_*})^{1/2} = O(n^{-\zeta/4})$ and $P(\Omega_n) \to 1$ as $n \to \infty$ where $\Omega_n = \{\max\{\hat{\tau}_i^2 - \tau_i^2, |\hat{\mu}_i - \mu_*|\} \leq \hat{c}_n\}$. Observe that $\theta_i$ can be represented as $\theta_i = (1 - B_i)G_i + B_i\Delta_i$, where $B_i$, $G_i$, and $\Delta_i$ are i.i.d. random variables such that $B_i$ and $\{G_i, \Delta_i\}$ are independent, $P(B_i = 1) = \omega$ and $P(B_i = 0) = 1 - \omega$, $G_i \sim N(\mu_*, \tau_i^2)$, and $\Delta_i$ is a random variable with the density function $\delta$. Likewise, $y_i$ can be represented as $y_i = \theta_i + \sigma_iZ_i$, where $Z_i \sim N(0, 1)$. From the same argument to show [A7], we have

$$P\left(\theta_i \leq \hat{\theta}_i^{(\gamma)} + \hat{s}_i^{(\gamma)} z_{\alpha/2}\right)$$

$$= P\left(\left \{\theta_i \leq \hat{\theta}_i^{(\gamma)} + \hat{s}_i^{(\gamma)} z_{\alpha/2} + (\hat{\theta}_i^{(\gamma)} - \hat{\theta}_i^{(\gamma)}) + (\hat{s}_i^{(\gamma)} - \hat{s}_i^{(\gamma)}) z_{\alpha/2}\right \} \cap \Omega_n \cap \{|y_i - \mu_*| \leq \hat{c}_n^{-1/2}\}\right)$$

$$+ P(|y_i - \mu_*| > \hat{c}_n^{-1/2}) + P(\Omega_n').$$

From Lemmas[A4] and [A5], we have

$$P(\Omega_n') \leq P(||\hat{\psi}^{(\gamma)} - \psi_*|| > \hat{c}_n)$$

$$\leq P(||\hat{\psi}^{(\gamma)} - \psi^{(\gamma)}|| > \hat{c}_n/2) + P(||\psi^{(\gamma)} - \psi_*|| > \hat{c}_n/2)$$

$$\leq 4\hat{c}_n^{-2} \left(E[||\hat{\psi}^{(\gamma)} - \psi^{(\gamma)}||^2] + ||\psi^{(\gamma)} - \psi_*||^2\right)$$

$$= O\left(\frac{1}{n(n^{-\zeta/2} + \omega\rho_{\psi_*})} + \frac{(\omega\rho_{\psi_*})^2}{n^{-\zeta/2} + \omega\rho_{\psi_*}}\right)$$

$$= O\left(\frac{1}{n\omega\rho_{\psi_*} + \omega\rho_{\psi_*}}\right) = O(\omega\rho_{\psi_*}).$$
Further, we have

\[ P(|\gamma_i - \mu_*| > \epsilon_n^{-1/2}) = P \left( \{ |y_i - \mu_*| > \epsilon_n^{-1/2} \} \mid B_i = 0 \right) P(B_i = 0) \]
\[ + P \left( \{ |y_i - \mu_*| > \epsilon_n^{-1/2} \} \mid B_i = 1 \right) P(B_i = 1) \]
\[ \leq P \left( |G_i - \mu_* + \sigma_i Z_i| > \epsilon_n^{-1/2} \right) (1 - \omega) + \epsilon_n E_\delta (|y_i - \mu_*|^2) \omega \]
\[ \leq \epsilon_n^{1/2} (E [ |G_i - \mu_*| + \sigma_i E [ |Z_i| ] ) (1 - \omega) + O(\omega \epsilon_n) \]
\[ = O(\epsilon_n^{1/2}). \]

Then we have

\[ P \left( \theta_i \leq \hat{\theta}_i^{(\gamma)} + \hat{s}_i^{(\gamma)} z_{\alpha/2} \right) \]
\[ = P \left( \{ \theta_i \leq \hat{\theta}_i^{(\gamma)} + \hat{s}_i^{(\gamma)} z_{\alpha/2} + (\hat{\theta}_i^{(\gamma)} - \hat{\theta}_i^{(\gamma)}) + (\hat{s}_i^{(\gamma)} - \hat{s}_i^{(\gamma)}) z_{\alpha/2} \} \cap \Omega_n \cap \{ |y_i - \mu_*| \leq \epsilon_n^{-1/2} \} \right) \]
\[ + O(\epsilon_n^{1/2}) + O(\omega \psi_*). \]  \hspace{1cm} (A11)

On \( \Omega_n \cap \{ |y_i - \mu_*| \leq \epsilon_n^{-1/2} \} \), we can show

\[ |\hat{\theta}_i^{(\gamma)} - \hat{\theta}_i^{(\gamma)}| \leq C'_1(\sigma^2_L, \sigma^2_U, \psi_*) \epsilon_n^{1/2}, \]
\[ |\hat{s}_i^{(\gamma)} - \hat{s}_i^{(\gamma)}| \leq C'_1(\sigma^2_L, \sigma^2_U, \psi_*) \epsilon_n^{1/2} \]

for \( \gamma \in [\gamma, \bar{\gamma}] \) and some positive constant \( C'_1(\sigma^2_L, \sigma^2_U, \psi_*) \). Then we have

\[ P \left( \{ \theta_i \leq \hat{\theta}_i^{(\gamma)} + \hat{s}_i^{(\gamma)} z_{\alpha/2} + (\hat{\theta}_i^{(\gamma)} - \hat{\theta}_i^{(\gamma)}) + (\hat{s}_i^{(\gamma)} - \hat{s}_i^{(\gamma)}) z_{\alpha/2} \} \cap \Omega_n \cap \{ |y_i - \mu_*| \leq \epsilon_n^{-1/2} \} \right) \]
\[ \leq P \left( \theta_i \leq \hat{\theta}_i^{(\gamma)} + \hat{s}_i^{(\gamma)} z_{\alpha/2} + C'_1(\sigma^2_L, \sigma^2_U, \psi_*) \epsilon_n^{1/2} \right) \]
\[ = P \left( \theta_i \leq \hat{\theta}_i^{(\gamma)} + \hat{s}_i^{(\gamma)} z_{\alpha/2} + C'_1(\sigma^2_L, \sigma^2_U, \psi_*) \epsilon_n^{1/2} \mid B_i = 0 \right) P(B_i = 0) \]
\[ + P \left( \theta_i \leq \hat{\theta}_i^{(\gamma)} + \hat{s}_i^{(\gamma)} z_{\alpha/2} + C'_1(\sigma^2_L, \sigma^2_U, \psi_*) \epsilon_n^{1/2} \mid B_i = 1 \right) P(B_i = 1) \]
\[ = P_1(1 - \omega) + P_2 \omega \]  \hspace{1cm} (A12)
For $P_1$, we have

$$P_1 = P\left( \theta_i \leq \tilde{\theta}_i + \tilde{s}_i z_{\alpha/2} + (\tilde{\theta}_i^{(\gamma)} - \tilde{\theta}_i) + (\tilde{s}_i^{(\gamma)} - \tilde{s}_i) z_{\alpha/2} + C_1'(\sigma^2_L, \sigma^2_U, \psi_*) \epsilon^{1/2}_n \mid B_i = 0 \right)$$

$$= E\left[ P(\theta_i \leq \tilde{\theta}_i + \tilde{s}_i z_{\alpha/2} + (\tilde{\theta}_i^{(\gamma)} - \tilde{\theta}_i) + (\tilde{s}_i^{(\gamma)} - \tilde{s}_i) z_{\alpha/2} + C_1'(\sigma^2_L, \sigma^2_U, \psi_*) \epsilon^{1/2}_n \mid y_i, \sigma_i, B_i = 0 \mid B_i = 0 \right]$$

$$\leq E\left[ P(\theta_i \leq \tilde{\theta}_i + \tilde{s}_i z_{\alpha/2} \mid y_i, \sigma_i, B_i = 0) + 2|\tilde{\theta}_i^{(\gamma)} - \tilde{\theta}_i| + |\tilde{s}_i^{(\gamma)} - \tilde{s}_i| z_{\alpha/2} + 2C_1'(\sigma^2_L, \sigma^2_U, \psi_*) \epsilon^{1/2}_n \mid B_i = 0 \right]$$

$$= 1 - \alpha/2 + 2E_{\psi_*}\left[ |\tilde{\theta}_i^{(\gamma)} - \tilde{\theta}_i| + |\tilde{s}_i^{(\gamma)} - \tilde{s}_i| z_{\alpha/2} \right] + 2C_1'(\sigma^2_L, \sigma^2_U, \psi_*) \epsilon^{1/2}_n. \quad (A13)$$

For the second inequality, we used Lemma A3. Note that

$$\tilde{\theta}_i^{(\gamma)} - \tilde{\theta}_i = \frac{\sigma^2_1}{\tau^2 + \sigma^2_1} \left( \phi(y_i; \mu_*, \tau^2 + \sigma^2_1) \gamma c_1(\tau^2_1) - 1 \right) (y_i - \mu_*),$$

$$\tilde{s}_i^{(2)(\gamma)} - \tilde{s}_i^2 = \frac{\sigma^2_1}{\tau^2 + \sigma^2_1} \left\{ (1 - \phi(y_i; \mu_*, \tau^2 + \sigma^2_1) \gamma c_1(\tau^2_1)) + \frac{\gamma(y_i - \mu_*)^2}{\tau^2 + \sigma^2_1} \phi(y_i; \mu_*, \tau^2 + \sigma^2_1) \gamma c_1(\tau^2_1) \right\},$$

and

$$|\tilde{s}_i^{(\gamma)} - \tilde{s}_i| \leq \tilde{s}_i \left| \frac{\tilde{s}_i^{(\gamma)}}{\tilde{s}_i} - 1 \right| \frac{\tilde{s}_i^{(\gamma)}}{\tilde{s}_i} + 1 = \frac{1}{\tilde{s}_i} |\tilde{s}_i^{(\gamma)} - \tilde{s}_i^2| \leq \frac{\tau^2 + \sigma^2_1}{\sigma^2_L} |\tilde{s}_i^{(\gamma)} - \tilde{s}_i^2|.$$
\[ P_1 \leq 1 - \alpha/2 + O(\gamma) + 2C'_1(\sigma^2_L, \sigma^2_U, \psi_*) \bar{c}_n^{1/2}. \]  

(A16)

Define \( \Omega_{2,n} = \{ \phi(y_i; \mu_2, \tau^2_\xi | y_i - \mu_2 |^k \leq \rho_{\psi_2}, k = 0, 1, 2 \} \). On \( \Omega_{2,n} \), we have

\[
\tilde{\theta}_i^{(\gamma)} \leq \theta_i + \sigma_i Z_i + \frac{\sigma^2}{\tau^2_\xi} C_\gamma(\tau^2_\xi) \rho_{\psi_2}^{1/2}, \quad \tilde{s}_i^{(\gamma)} \leq \sigma_i^2 \left( 1 + \frac{\sigma^2}{\tau^2_\xi} (\gamma + 1) C_\gamma(\tau^2_\xi) \rho_{\psi_2}^{1/2} \right).
\]

Then we have

\[
P_2 = P \left( \left\{ \theta_i \leq \tilde{\theta}_i^{(\gamma)} + \tilde{s}_i^{(\gamma)} z_{\alpha/2} + C'_1(\sigma^2_L, \sigma^2_U, \psi_*) \bar{c}_n^{1/2} \right\} \cap \Omega_{2,n} \mid B_i = 1 \right) \\
+ P \left( \left\{ \theta_i \leq \tilde{\theta}_i^{(\gamma)} + \tilde{s}_i^{(\gamma)} z_{\alpha/2} + C'_1(\sigma^2_L, \sigma^2_U, \psi_*) \bar{c}_n^{1/2} \right\} \cap \Omega_{2,n} \mid B_i = 1 \right) \\
= : P_{2,1} + P_{2,2}.
\]

For \( P_{2,1} \), we have

\[
P_{2,1} \leq P \left( \theta_i \leq \theta_i + \sigma_i Z_i + \frac{\sigma^2}{\tau^2_\xi} C_\gamma(\tau^2_\xi) \rho_{\psi_2}^{1/2} + \sigma_i \left( 1 + \frac{\sigma^2}{\tau^2_\xi} (\gamma + 1) C_\gamma(\tau^2_\xi) \rho_{\psi_2}^{1/2} \right)^{1/2} z_{\alpha/2} \right) \\
+ C'_1(\sigma^2_L, \sigma^2_U, \psi_*) \bar{c}_n^{1/2} \mid B_i = 1 \right) \\
= P \left( -\sigma_i Z_i \leq \frac{\sigma^2}{\tau^2_\xi} C_\gamma(\tau^2_\xi) \rho_{\psi_2}^{1/2} + \sigma_i \left( 1 + \frac{\sigma^2}{\tau^2_\xi} (\gamma + 1) C_\gamma(\tau^2_\xi) \rho_{\psi_2}^{1/2} \right)^{1/2} z_{\alpha/2} \right) \\
+ C'_1(\sigma^2_L, \sigma^2_U, \psi_*) \bar{c}_n^{1/2} \mid B_i = 1 \right) \\
= P \left( -Z_i \leq \frac{\sigma_i}{\tau^2_\xi} C_\gamma(\tau^2_\xi) \rho_{\psi_2}^{1/2} + \left( 1 + \frac{\sigma^2}{\tau^2_\xi} (\gamma + 1) C_\gamma(\tau^2_\xi) \rho_{\psi_2}^{1/2} \right)^{1/2} z_{\alpha/2} \right) \\
+ C'_1(\sigma^2_L, \sigma^2_U, \psi_*) \bar{c}_n^{1/2} \mid B_i = 1 \right) \\
\leq P \left( Z_i \leq \left( 1 + \frac{\sigma^2}{\tau^2_\xi} (\gamma + 1) C_\gamma(\tau^2_\xi) \rho_{\psi_2}^{1/2} \right)^{1/2} z_{\alpha/2} \right) + \frac{2\sigma_U}{\tau^2_\xi} C_\gamma(\tau^2_\xi) \rho_{\psi_2}^{1/2} + 2C'_1(\sigma^2_L, \sigma^2_U, \psi_*) \bar{c}_n^{1/2} \\
\leq P \left( Z_i \leq \left( 1 + \frac{\sigma^2}{\tau^2_\xi} (\gamma + 1) C_\gamma(\tau^2_\xi) \rho_{\psi_2}^{1/2} \right) z_{\alpha/2} \right) + \frac{2\sigma_U}{\tau^2_\xi} C_\gamma(\tau^2_\xi) \rho_{\psi_2}^{1/2} + 2C'_1(\sigma^2_L, \sigma^2_U, \psi_*) \bar{c}_n^{1/2} \\
\leq P \left( Z_i \leq z_{\alpha/2} + \frac{2\sigma_U}{\tau^2_\xi} C_\gamma(\tau^2_\xi) ((\gamma + 1) z_{\alpha/2} + 1) \rho_{\psi_2}^{1/2} + 2C'_1(\sigma^2_L, \sigma^2_U, \psi_*) \bar{c}_n^{1/2} \right) \\
\leq 1 - \alpha/2 + O(\rho_{\psi_2}^{1/2} + \bar{c}_n^{1/2}).
\]  

(A17)
For the second and fourth inequalities, we used Lemma \[\text{A3}\]

For \(P_{2,2}\), from the definition of \(\rho_{\psi^*}\), we have

\[
P_{2,2} \leq P(\Omega_{i,n}^C | B_i = 1)
= E[P(\Omega_{i,n}^C | \sigma_i, B_i = 1) | B_i = 1]
\leq \rho_{\psi^*}^{-1/2} \max_{k \in \{0,1,2\}} \sup_{\sigma_i \in [\sigma_L, \sigma_U]} E_d[\phi(y_i; \mu_*, \tau_i^2 + \sigma_i^2)^2 | y_i - \mu_* | \sigma_i]
\leq \rho_{\psi^*}^{1/2}.
\] (A18)

Combining \(\text{A17}\) and \(\text{A18}\), we have

\[
P_2 \leq 1 - \alpha/2 + O(\rho_{\psi^*}^{1/2} + \varepsilon_n^{1/2}).
\] (A19)

Combining \(\text{A11}, \text{A12}, \text{A16}, \text{A19}\), we have

\[
P\left(\theta_i \leq \hat{\theta}_i^{(\gamma)} + \hat{z}_{\alpha/2}^{(\gamma)} \varepsilon_i\right) \leq 1 - \alpha/2 + O(\omega \rho_{\psi^*})^{1/4}.
\]

Likewise, we have

\[
P\left(\theta_i \leq \hat{\theta}_i^{(\gamma)} + \hat{z}_{\alpha/2}^{(\gamma)} \varepsilon_i\right) \geq 1 - \alpha/2 - O\left((\omega \rho_{\psi^*})^{1/4}\right) - O(\gamma).\]

Almost the same argument yields

\[
P\left(\theta_i \leq \hat{\theta}_i - \hat{z}_{\alpha/2}^{(\gamma)} \varepsilon_i\right) = \alpha/2 + O\left((\omega \rho_{\psi^*})^{1/4}\right) + O(\gamma).\]

Therefore, we have

\[
P\left(\theta_i \in \hat{C}_{i,1-\alpha}^{(\gamma)}\right) = 1 - \alpha + O\left((\omega \rho_{\psi^*})^{1/4} + \gamma\right).
\]

\textbf{A3 Additional simulation results}

The simulation results under \(n \in \{100, 500\}\) are given in Table \[\text{A1}\]. The relative performance does not change very much compared with \(n \in \{200, 1000\}\).

\textbf{References}

Armstrong, T. B., M. Kolesár, and M. Plagborg-Møller (2020). Robust empirical bayes confidence intervals. \textit{arXiv preprint arXiv:2004.03448}.
Table A1: Average values of optimal $\gamma$, empirical coverage probability (CP) and average length (AL) of the four 95% confidence intervals under $n \in \{100, 500\}$.

| scenario | $V$ | $\gamma$ | CP | AL |
|----------|-----|----------|-----|----|
|          |     |          | GD  | EB | EF | AKM | GD  | EB | EF | AKM |
| (i)      | 0.1 | 0.00     | 95.5| 95.5| 95.3| 65.1| 1.49| 1.49| 1.46| 1.04 |
|          | 0.5 | 0.00     | 93.3| 93.3| 91.1| 94.3| 2.25| 2.25| 2.17| 2.50 |
| $m = 100$| 1   | 0.00     | 94.3| 94.3| 92.6| 97.3| 2.76| 2.76| 2.67| 3.22 |
|          | 2   | 0.00     | 94.6| 94.6| 92.8| 97.6| 3.02| 3.02| 2.92| 3.53 |
| (ii)     | 0.1 | 0.17     | 99.1| 94.5| 98.4| 96.6| 2.85| 3.30| 1.91| 3.86 |
|          | 0.5 | 0.16     | 97.0| 94.6| 94.1| 96.9| 3.20| 3.37| 2.38| 3.93 |
| $m = 100$| 1   | 0.08     | 95.3| 94.7| 93.5| 97.1| 3.40| 3.43| 2.79| 3.98 |
|          | 2   | 0.02     | 94.9| 94.8| 93.5| 97.3| 3.48| 3.48| 3.04| 4.02 |
| (iii)    | 0.1 | 0.00     | 95.3| 95.3| 94.9| 98.2| 1.51| 1.51| 1.48| 3.38 |
|          | 0.5 | 0.01     | 94.7| 94.6| 93.3| 97.5| 2.73| 2.74| 2.39| 3.87 |
| $m = 100$| 1   | 0.07     | 95.5| 94.8| 94.5| 97.2| 3.34| 3.45| 2.92| 4.12 |
|          | 2   | 0.15     | 95.7| 94.8| 94.1| 96.8| 3.53| 3.71| 3.23| 4.17 |
| (iv)     | 0.1 | 0.00     | 94.0| 94.0| 93.7| 73.0| 1.55| 1.55| 1.51| 1.22 |
|          | 0.5 | 0.14     | 95.8| 94.2| 99.0| 96.8| 3.19| 3.27| 1.90| 3.86 |
| $m = 500$| 1   | 0.14     | 99.3| 94.8| 99.6| 97.0| 2.68| 3.72| 1.88| 4.18 |
|          | 2   | 0.11     | 99.4| 95.0| 99.6| 96.7| 2.48| 3.83| 2.13| 4.18 |
| (i)      | 0.1 | 0.00     | 93.9| 93.9| 95.3| 81.5| 1.27| 1.27| 1.32| 1.09 |
|          | 0.5 | 0.00     | 94.6| 94.6| 90.4| 97.3| 2.26| 2.26| 2.13| 2.60 |
| $m = 500$| 1   | 0.00     | 94.9| 94.9| 92.2| 97.8| 2.77| 2.77| 2.67| 3.25 |
|          | 2   | 0.00     | 94.9| 94.9| 93.9| 97.8| 3.03| 3.03| 2.99| 3.56 |
| (ii)     | 0.1 | 0.18     | 99.4| 94.6| 98.6| 96.7| 2.69| 3.30| 1.65| 3.86 |
|          | 0.5 | 0.20     | 97.8| 94.7| 94.6| 97.0| 3.16| 3.37| 2.32| 3.93 |
| $m = 500$| 1   | 0.10     | 95.7| 94.7| 94.6| 97.2| 3.41| 3.44| 2.82| 3.99 |
|          | 2   | 0.01     | 94.8| 94.8| 94.7| 97.3| 3.49| 3.49| 3.10| 4.03 |
| (iii)    | 0.1 | 0.00     | 93.6| 93.6| 94.8| 98.4| 1.30| 1.30| 1.34| 3.40 |
|          | 0.5 | 0.00     | 95.3| 95.3| 94.9| 97.6| 2.79| 2.80| 2.47| 3.89 |
| $m = 500$| 1   | 0.09     | 96.1| 95.1| 94.8| 97.3| 3.40| 3.52| 3.13| 4.15 |
|          | 2   | 0.19     | 96.2| 95.0| 93.7| 96.8| 3.56| 3.77| 3.49| 4.18 |
| (iv)     | 0.1 | 0.00     | 92.7| 92.7| 93.3| 88.5| 1.38| 1.38| 1.39| 1.34 |
|          | 0.5 | 0.30     | 98.3| 94.4| 99.5| 96.9| 3.12| 3.27| 1.62| 3.86 |
| $m = 500$| 1   | 0.13     | 99.4| 94.9| 99.8| 97.1| 2.47| 3.72| 1.55| 4.18 |
|          | 2   | 0.12     | 99.5| 95.0| 99.8| 96.7| 2.30| 3.83| 1.67| 4.18 |
Basu, A., I. R. Harris, N. L. Hjort, and M. Jones (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika* 85(3), 549–559.

Carvalho, C. M., N. G. Polson, and J. G. Scott (2010). The horseshoe estimator for sparse signals. *Biometrika* 97(2), 465–480.

Chatterjee, S., P. Lahiri, H. Li, et al. (2008). Parametric bootstrap approximation to the distribution of eblup and related prediction intervals in linear mixed models. *The Annals of Statistics* 36(3), 1221–1245.

Chernozhukov, V., D. Chetverikov, and K. Kato (2017). Central limit theorems and bootstrap in high dimensions. *The Annals of Probability* 45, 2309–2352.

Cichocki, A., S. Cruces, and S.-i. Amari (2011). Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization. *Entropy* 13(1), 134–170.

De Luca, G., J. R. Magnus, and F. Peracchi (2021). Sampling properties of the bayesian posterior mean with an application to wals estimation. *Journal of Econometrics*.

DerSimonian, R. and N. Laird (2015). Meta-analysis in clinical trials revisited. *Contemporary Clinical Trials* 45, 139–145.

Efron, B. (2011). Tweedie’s formula and selection bias. *Journal of the American Statistical Association* 106(496), 1602–1614.

Efron, B. (2016). Empirical bayes deconvolution estimates. *Biometrika* 103(1), 1–20.

Efron, B. et al. (2019). Bayes, oracle bayes and empirical bayes. *Statistical Science* 34(2), 177–201.

Efron, B., R. Tibshirani, J. D. Storey, and V. Tusher (2001). Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association* 96(456), 1151–1160.

Fujisawa, H. (2013). Normalized estimating equation for robust parameter estimation. *Electronic Journal of Statistics* 7, 1587–1606.

Fujisawa, H. and S. Eguchi (2008). Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis* 99(9), 2053–2081.
Gu, J. and R. Koenker (2017). Unobserved heterogeneity in income dynamics: An empirical bayes perspective. *Journal of Business & Economic Statistics* 35(1), 1–16.

Hall, P. and T. Maiti (2006). On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68(2), 221–238.

Jones, M., N. L. Hjort, I. R. Harris, and A. Basu (2001). A comparison of related density-based minimum divergence estimators. *Biometrika* 88(3), 865–873.

Kiefer, J. and J. Wolfowitz (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, 887–906.

Koenker, R. (2020). Empirical bayes confidence intervals: an R vinaigrette. *available at http://www.econ.uiuc.edu/roger/research/ebayes/cieb.pdf*.

Koenker, R. and I. Mizera (2014). Convex optimization, shape constraints, compound decisions, and empirical bayes rules. *Journal of the American Statistical Association* 109(506), 674–685.

Kolesar, M. (2021). *ebci: Robust Empirical Bayes Confidence Intervals*. R package version 0.0.0.9002.

Lawson, A. B. (2018). *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology*. Chapman and Hall/CRC.

Li, H. and P. Lahiri (2010). An adjusted maximum likelihood method for solving small area estimation problems. *Journal of Multivariate Analysis* 101(4), 882–892.

Liu, L., H. R. Moon, and F. Schorfheide (2020). Forecasting with dynamic panel data models. *Econometrica* 88(1), 171–201.

Morris, C. N. (1983). Parametric empirical bayes inference: theory and applications. *Journal of the American Statistical Association* 78(381), 47–55.
Nazarov, F. (2003). On the maximal perimeter of a convex set in $\mathbb{R}^n$ with respect to a gaussian measure. *Geometric Aspects of Functional Analysis (2001-2002)* 1807, 169–187.

Rao, J. N. and I. Molina (2015). *Small Area Estimation*. John Wiley & Sons.

Sugasawa, S. (2020). Robust empirical bayes small area estimation with density power divergence. *Biometrika* 107(2), 467–480.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

Yoshimori, M. and P. Lahiri (2014). A second-order efficient empirical bayes confidence interval. *The Annals of Statistics* 42(4), 1233–1261.