FAB-PPI: Frequentist, Assisted by Bayes, Prediction-Powered Inference

Stefano Cortinovis [1] **François Caron** [1]

Abstract

Prediction-powered inference (PPI) enables valid statistical inference by combining experimental data with machine learning predictions. When a sufficient number of high-quality predictions is available, PPI results in more accurate estimates and tighter confidence intervals than traditional methods. In this paper, we propose to inform the PPI framework with prior knowledge on the quality of the predictions. The resulting method, which we call frequentist, assisted by Bayes, PPI (FAB-PPI), improves over PPI when the observed prediction quality is likely under the prior, while maintaining its frequentist guarantees. Furthermore, when using heavy-tailed priors, FAB-PPI adaptively reverts to standard PPI in low prior probability regions. We demonstrate the benefits of FAB-PPI in real and synthetic examples.

1. Introduction

Statistical inference crucially relies on the availability of high-quality labelled data to draw actionable conclusions. As the scale of machine learning models keeps growing, their increasingly accurate predictions become a tempting alternative to labelled data in fields where the latter are traditionally scarce, such as proteomics (Jumper et al., 2021). However, blindly using potentially biased predictions as a surrogate for labelled data voids the statistical validity of the conclusions drawn. To address this, prediction-powered inference (Angelopoulos et al., 2023b) provides a general framework for statistical inference in the presence of a large number of black-box predictions by combining them with a smaller number of labelled observations, which are used to *correct* for the discrepancy between the predictions and the true labels. The estimators and confidence intervals (CIs) resulting from PPI are statistically valid regardless of the machine learning model used. Moreover, when the predictions are good, PPI results in more accurate estimates and shorter CIs than traditional methods that rely solely on labelled data.

More formally, for an input/output pair $(X, Y) \sim \mathbb{P} = \mathbb{P}_X \times \mathbb{P}_{Y|X}$ and a convex loss function $\mathcal{L}_\theta(x, y)$, where $\theta \in \mathbb{R}^d$, we wish to estimate

$$\theta^\star = \underset{\theta \in \mathbb{R}^d}{\arg\min}\ \mathbb{E}[\mathcal{L}_\theta(X, Y)]. \tag{1}$$

For instance, if $\mathcal{L}_\theta(x, y) = (\theta - y)^2/2$ is the squared loss, then $\theta^\star = \mathbb{E}[Y]$. We assume that we have n labelled observations $\{(X_i, Y_i)\}_{i=1}^n$ iid from $\mathbb{P}$ and N unlabelled observations $\{\widetilde{X}_i\}_{i=1}^N$ iid from $\mathbb{P}_X$, which are also independent of the labelled data. The number of unlabelled observations is typically much larger than the number of labelled ones, $N \gg n$. Additionally, we are provided with a machine learning prediction rule f, that can be used to predict an output $f(x)$ at any input x. PPI aims to obtain an estimator $\widehat{\theta}$ and a $(1-\alpha)$ confidence interval $\mathcal{C}_\alpha^{\text{pp}}$ for $\theta^\star$, which take advantage of f. Under mild assumptions, $\theta^\star$ can be expressed as the solution to

$$g_{\theta^\star} := \mathbb{E}[\mathcal{L}'_{\theta^\star}(X, Y)] = 0, \tag{2}$$

where $\mathcal{L}'_\theta$ is a subgradient of $\mathcal{L}_\theta$ with respect to θ. It is easy to see that the quantity above can be decomposed as $g_\theta = m_\theta + \Delta_\theta$, where

$$m_\theta := \mathbb{E}[\mathcal{L}'_\theta(X, f(X))], \tag{3}$$

$$\Delta_\theta := \mathbb{E}[\mathcal{L}'_\theta(X, Y) - \mathcal{L}'_\theta(X, f(X))]. \tag{4}$$

In this setting, m_θ represents a measure of fit of the predictor, whereas Δ_θ, called the *rectifier*, accounts for the discrepancy between the predicted outputs $f(X)$ and the true outputs Y, effectively quantifying prediction quality. For example, under the squared loss, $\Delta_\theta = \mathbb{E}[f(X) - Y]$ and a *good* predictor f is one such that Δ_θ is close to zero, i.e. $f(x) \simeq \mathbb{E}[Y|X = x]$. Note that, while in this case Δ_θ does not depend on θ, this is not true in general.

By estimating the two quantities m_θ and Δ_θ, Angelopoulos et al. (2023a) derive an estimator and a CI for $\theta^\star$, which use both labelled and unlabelled data. The resulting CI is shorter than the classical confidence interval based solely on the labelled data when $N \gg n$ and f is accurate because, in this case, m_θ can be estimated with low variance using the unlabelled data, while Δ_θ is close to zero.

[1]Department of Statistics, University of Oxford. Correspondence to: Stefano Cortinovis <cortinovis@stats.ox.ac.uk>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

Standard PPI employs off-the-shelf estimation and CI procedures for Δ_θ, which do not take advantage of any prior knowledge on the quality of the machine learning model f. However, in many applications, we expect the latter's predictions to be (i) usually very good, but (ii) sometimes prone to large errors and hallucinations. We propose to encode such an inductive bias with a horseshoe prior π_θ on Δ_θ (Carvalho et al., 2010), which accommodates the aforementioned properties by exhibiting (i) an infinitely tall spike at the origin, and (ii) Cauchy-like tails at infinity. In order to construct valid confidence regions for Δ_θ using the horseshoe prior, we resort to the frequentist-assisted by Bayes (FAB) framework (Pratt, 1961; 1963; Yu & Hoff, 2018). This approach provides confidence regions such that their expected length is lower for rectifiers Δ_θ that have high probability under π_θ, and larger otherwise. While the resulting confidence regions have exact coverage for any prior π_θ, the horseshoe prior is particularly well-suited for PPI. Being concentrated around the origin, it produces shorter confidence regions when the predictions are good, i.e. $||\Delta_\theta|| \simeq 0$. At the same time, its heavy tails ensure robustness when the predictions are poor. Indeed, as shown by Cortinovis & Caron (2024), if $||\Delta_\theta|| \gg 0$, the FAB procedure with the horseshoe prior reverts to the traditional CI based on the sample mean.

In this work, we introduce FAB-PPI, a Bayes-assisted approach for PPI that encodes prior information on the quality of the machine learning predictions by specifying a prior for the rectifier Δ_θ. FAB-PPI is:

- Statistically valid, as its confidence regions have correct coverage for any choice of prior;
- Efficient, as its confidence regions have smaller expected length when the predictions are good;
- Robust, as it reverts to standard PPI when the predictions are poor, if the horseshoe prior is used;
- Modular, as it can be used in conjunction with power tuning (Angelopoulos et al., 2023a).

The remainder of the paper is organised as follows. Section 2 reviews related work. Section 3 provides background on control variates, PPI, and FAB confidence regions. Section 4 describes our novel approach for PPI, called FAB-PPI. Section 5 demonstrates the benefits of FAB-PPI on synthetic and real data. Finally, Section 6 discusses limitations and further extensions of our approach.

2. Related Work

PPI (Angelopoulos et al., 2023b) was introduced to obtain shorter CIs for the parameters of interest by leveraging machine learning predictions in semi-supervised settings. PPI has since been extended in multiple directions. PPI++ (Angelopoulos et al., 2023a) proposes a different, loss-based formulation of PPI, leading to a more computationally efficient procedure, along with an additional power-tuning parameter to enhance PPI's performance. Stratified PPI (Fisch et al., 2024) improves upon PPI by employing a data stratification strategy. Cross PPI (Zrnic & Candès, 2024b) demonstrates how the training of f can be included in the PPI pipeline. Active statistical inference (Zrnic & Candès, 2024a) applies an active learning approach to select which inputs from the unlabelled set should be labelled. Closer to our work, Bayesian PPI (Hofer et al., 2024) considers an alternative PPI estimator motivated by Bayesian ideas. However, their approach provides Bayesian credible intervals, which do not offer frequentist guarantees. Additionally, their approach achieves similar experimental performance to PPI, while we demonstrate that FAB-PPI may significantly improve upon PPI.

As discussed in Angelopoulos et al. (2023b;a), PPI has close ties with control variates for variance reduction (Glasserman, 2003, §4.1). In the case of mean estimation, the form of the PPI estimator is similar to the one proposed by Zhang et al. (2019). PPI is also related to work in semiparametric inference with missing data (Robins & Rotnitzky, 1995).

The concept of Bayes-optimal confidence regions originates from the work of Pratt (1961; 1963). Pratt's approach, which has been given the name FAB by Yu & Hoff (2018), has since been extended in multiple directions (Brown et al., 1995; Farchione & Kabaila, 2008; Kabaila & Giri, 2013; Kabaila & Farchione, 2022; Yu & Hoff, 2018; Hoff & Yu, 2019; Hoff, 2023). In particular, Cortinovis & Caron (2024) show that, when combined with priors with power-law tails, FAB provides robust confidence regions that revert to classical ones in the presence of outliers. Hoff (2023) applied FAB in a predictive supervised context, showing that it can lead to more accurate predictions than standard methods.

3. Background

3.1. Control Variates

The method of control variates is a standard variance reduction technique in Monte Carlo approximation (Glasserman, 2003, §4.1). For simplicity, we present the method in the scalar case, but extensions to the multivariate setting are available. Let (Z, Y) be a pair of real-valued random variables, and assume we are interested in estimating $\mathbb{E}[Y]$ based on an iid sample $\{(Z_i, Y_i)\}_{i=1}^n$. Assuming $\mu = \mathbb{E}[Z]$ is known, one defines the control variate estimator (CVE)

$$\widehat{Y}_\lambda^{\text{cv}} = \overline{Y} - \lambda(\overline{Z} - \mu) = \frac{1}{n}\sum_{i=1}^{n} \left(Y_i - \lambda(Z_i - \mu)\right), \quad (5)$$

where $\lambda \in \mathbb{R}$ is a tuning coefficient and $\overline{Z}$ and $\overline{Y}$ are the sample means of (Z_i) and (Y_i), respectively. The centered random variable $Z_i - \mu$ serves as a control variate to estimate $\mathbb{E}[Y]$. The CVE is a consistent and unbiased esti-

mator of $\mathbb{E}[Y]$ with $\mathrm{var}(\widehat{Y}^{\mathrm{cv}}_{\lambda}) = (\mathrm{var}(Y) - 2\lambda\mathrm{cov}(Z,Y) + \lambda^2\mathrm{var}(Z))/n$, while $\mathrm{var}(\overline{Y}) = \mathrm{var}(Y)/n$. Therefore, CVE achieves smaller variance whenever $\lambda < \frac{2\mathrm{cov}(Z,Y)}{\mathrm{var}(Z)}$, and the optimal coefficient is given by $\lambda^* = \mathrm{cov}(Z,Y)/\mathrm{var}(Z)$. In this case, $\mathrm{var}(\widehat{Y}^{\mathrm{cv}}_{\lambda^*}) = (1-\rho^2_{Z,Y})\mathrm{var}(\overline{Y})$, where $\rho_{Z,Y}$ is the correlation between Z and Y. The more correlated Z and Y, the larger the variance reduction. By plugging the estimator

$$\widehat{\lambda} = \frac{\sum_{i=1}^n (Z_i - \overline{Z})(Y_i - \overline{Y})}{\sum_{i=1}^n (Z_i - \overline{Z})^2} \tag{6}$$

for λ in Equation (5), one has

$$\frac{\widehat{Y}^{\mathrm{cv}}_{\widehat{\lambda}} - \mathbb{E}[Y]}{s/\sqrt{n}} \to \mathcal{N}(0,1)$$

as $n \to \infty$, where s is the sample standard deviation of $\{(Y_i - \widehat{\lambda}Z_i)\}_{i=1,\dots,n}$. Hence, $\widehat{Y}^{\mathrm{cv}}_{\widehat{\lambda}} \pm z_{1-\alpha/2}s/\sqrt{n}$ is an asymptotically valid $(1-\alpha)$ CI for $\mathbb{E}[Y]$, whose asymptotic width is $2z_{1-\alpha/2}\sqrt{1-\rho^2_{ZY}}\sqrt{\mathrm{var}(Y)}/\sqrt{n}$.

3.2. Prediction-Powered Inference

PPI (Angelopoulos et al., 2023b) defines an estimator $\widehat{\theta}$ and a CI $\mathcal{C}^{\mathrm{pp}}_{\alpha}$ for a parameter of interest $\theta^\star$ satisfying Equation (2). In particular, let $\widehat{m}_\theta$ and $\widehat{\Delta}_\theta$ be some estimators of m_θ and Δ_θ. Using Equation (2), the estimator $\widehat{\theta}$ is defined as the solution, in θ, to the equation

$$\widehat{m}_\theta + \widehat{\Delta}_\theta = 0. \tag{7}$$

Similarly, let $\mathcal{R}_\delta$ and $\mathcal{T}_{\alpha-\delta}$ be $1-\delta$ and $1-(\alpha-\delta)$ CIs for Δ_θ and m_θ, respectively. Then, the PPI confidence interval $\mathcal{C}^{\mathrm{pp}}_{\alpha}$ is defined as

$$\mathcal{C}^{\mathrm{pp}}_{\alpha} = \{\theta \mid 0 \in \mathcal{R}_\delta + \mathcal{T}_{\alpha-\delta}\}, \tag{8}$$

where $+$ denotes the Minkowski sum. Typical choices for $\widehat{m}_\theta$ and $\mathcal{T}_{\alpha-\delta}$ are the sample mean of the unlabelled data,

$$\widehat{m}_\theta = \frac{1}{N}\sum_{i=1}^N \mathcal{L}'_\theta(\widetilde{X}_i, f(\widetilde{X}_i)), \tag{9}$$

and classical CIs for sample means, respectively. Different choices for $\widehat{\Delta}_\theta$ have been proposed in the literature, leading to different PPI estimators.

Standard PPI. Angelopoulos et al. (2023b) propose to use the sample mean

$$\widehat{\Delta}^{\mathrm{PP}}_\theta = \frac{1}{n}\sum_{i=1}^n (\mathcal{L}'_\theta(X_i, Y_i) - \mathcal{L}'_\theta(X_i, f(X_i))) \tag{10}$$

as an estimator for Δ_θ and the associated classical CIs to construct $\mathcal{R}_\delta$. For the squared loss, the estimator $\widehat{\theta}^{\mathrm{PP}}$ solving $\widehat{m}_\theta + \widehat{\Delta}_\theta = 0$ takes the control variate form

$$\widehat{\theta}^{\mathrm{PP}} = \overline{Y} - \left(\frac{1}{n}\sum_{i=1}^n f(X_i) - \frac{1}{N}\sum_{j=1}^N f(\widetilde{X}_j)\right) \tag{11}$$

with control variate $f(X_i) - \frac{1}{N}\sum_{j=1}^N f(\widetilde{X}_j)$ and $\lambda = 1$.

PPI++. Angelopoulos et al. (2023a) extend standard PPI by introducing an additional control variate parameter λ, which they call power tuning parameter. The chosen $\widehat{m}_\theta$ is still the sample mean (9), while $\widehat{\Delta}^{\mathrm{PP+}}_\theta$ now takes the control variate form

$$\begin{aligned}\widehat{\Delta}^{\mathrm{PP+}}_\theta = \frac{1}{n}\sum_{i=1}^n (\mathcal{L}'_\theta(X_i, Y_i) - \mathcal{L}'_\theta(X_i, f(X_i))) \\ - (\widehat{\lambda} - 1)\left(\frac{1}{n}\left[\sum_{i=1}^n \mathcal{L}'_\theta(X_i, f(X_i))\right] - \widehat{m}_\theta\right),\end{aligned} \tag{12}$$

where $\widehat{\lambda}$ is estimated from the data. In this case, the centered control variate is $\mathcal{L}'_\theta(X_i, f(X_i)) - \widehat{m}_\theta$, which depends only on the machine learning predictions. For the squared loss, we obtain

$$\widehat{\theta}^{\mathrm{PP+}} = \overline{Y} - \widehat{\lambda}\left(\frac{1}{n}\sum_{i=1}^n f(X_i) - \frac{1}{N}\sum_{j=1}^N f(\widetilde{X}_j)\right) \tag{13}$$

with plug-in estimator

$$\widehat{\lambda} = \frac{c_n}{(1+\frac{n}{N})v_{n+N}}, \tag{14}$$

where c_n is the sample covariance of $(Y_i, f(X_i))_{i=1}^n$ and v_{n+N} is the sample variance of $((f(X_i))_{i=1}^n, (f(\widetilde{X}_j))_{j=1}^N)$. The estimator (13) is closely related (though slightly different) to the one introduced by Zhang et al. (2019) for mean estimation in semi-supervised inference.

CLT-based CIs. While the definition of the PPI confidence interval (8) allows for merging any CIs $\mathcal{R}_\delta$ and $\mathcal{T}_{\alpha-\delta}$ for Δ_θ and m_θ, in practice these are often chosen to be CLT-based CIs that, once combined into $\mathcal{C}_\alpha$ give exact asymptotic coverage,

$$\liminf_{n,N\to\infty} \Pr(\theta^* \in \mathcal{C}_\alpha) \geq 1 - \alpha.$$

Such CLT-based CIs rely on the following standard assumption on the estimators $\widehat{m}_\theta$ and $\widehat{\Delta}_\theta$.

Assumption 3.1 (CLT assumption for PPI and PPI++)**.** Let $\widehat{m}_\theta$ be the sample mean (9) and consider some estimator $(\widehat{\sigma}^f_\theta)^2$ of $\mathrm{var}(\widehat{m}_\theta)$, with $(\widehat{\sigma}^f_\theta)^2/\mathrm{var}(\widehat{m}_\theta) \to 1$ almost surely. Let $\widehat{\Delta}_\theta$ be either the PPI estimator (10) or the PPI++ estimator (12) and consider some estimator $\widehat{\sigma}_\theta$ of $\mathrm{var}(\widehat{\Delta}_\theta)$ with $\widehat{\sigma}_\theta/\mathrm{var}(\widehat{\Delta}_\theta) \to 1$ a.s. Assume that, as $\min(n,N) \to \infty$,

$$(\widehat{m}_\theta - m_\theta)/\widehat{\sigma}^f_\theta \to \mathcal{N}(0,1) \tag{15}$$

$$(\widehat{\Delta}_\theta - \Delta_\theta)/\widehat{\sigma}_\theta \to \mathcal{N}(0,1). \tag{16}$$

3.3. Bayes-Optimal Confidence Regions

The FAB framework (Pratt, 1961; 1963; Yu & Hoff, 2018) aims to construct valid confidence regions with smaller expected volume. Let $W \mid \beta \sim \mathcal{N}(\beta, \sigma^2)$ with some prior $\pi_0(\beta)$ and denote by $\pi(w) = \int p(w \mid \beta)\pi_0(\beta)d\beta$ the corresponding marginal likelihood. For $\alpha \in (0,1)$, let $\mathcal{C}_\alpha(w)$ be an exact $(1-\alpha)$ confidence region for β based on the data w. That is, for any fixed β_0,

$$\Pr(\beta \in \mathcal{C}_\alpha(W) \mid \beta = \beta_0) = 1 - \alpha. \tag{17}$$

Let $\mathrm{vol}(\mathcal{C}_\alpha(w)) = \int_{\beta' \in \mathcal{C}_\alpha(w)} d\beta'$ be the volume of $\mathcal{C}_\alpha(w)$, and consider its expected value under the marginal likelihood π,

$$\mathbb{E}[\mathrm{vol}(\mathcal{C}_\alpha(W))] = \int \mathrm{vol}(\mathcal{C}_\alpha(w))\pi(w)dw. \tag{18}$$

Definition 3.2. For $\alpha \in (0,1)$, $\sigma > 0$ and a prior $\pi_0(\beta)$, the FAB confidence region $\mathcal{C}_\alpha$ for the mean parameter β of the normal model $Y \mid \beta \sim \mathcal{N}(\beta, \sigma^2)$, is the minimiser of the (Bayesian) expected volume

$$\mathcal{C}_\alpha = \arg\min_{\widetilde{\mathcal{C}}_\alpha} \mathbb{E}[\mathrm{vol}(\widetilde{\mathcal{C}}_\alpha(W))] \tag{19}$$

subject to the (frequentist) coverage constraint (17). We write $\mathcal{C}_\alpha(w) = \text{FAB-CR}(w; \pi_0, \sigma^2, \alpha)$.

The solution to Equation (19), which exists and is unique if $\pi_0(\beta)$ is not degenerate (Cortinovis & Caron, 2024, Theorem 2.1), may be found numerically as long as the marginal likelihood $\pi(w)$ can be evaluated pointwise. Additional details are provided in Appendix S1.2. Intuitively, the FAB confidence region $\mathcal{C}_\alpha(w)$ constructed through Equation (19) will be smaller for values of w that are likely under the marginal likelihood, and larger otherwise. As a result of this, while FAB guarantees the right coverage for any prior, one that assigns high probability to the value of β that generated the data is required to achieve smaller expected volume compared to the standard CI $(w \pm \sigma z_{1-\alpha/2})$, whose width does not depend on w.

Bayes-Assisted Estimator. A natural estimator to use alongside the FAB confidence region $\mathcal{C}_\alpha(w)$ is the posterior mean $\widehat{\beta}(W) = \mathbb{E}[\beta \mid W]$. As shown by (Cortinovis & Caron, 2024, Theorem 2.1), it is always contained within the confidence region: $\widehat{\beta}(w) \in \mathcal{C}_\alpha(w)$ for any $w \in \mathbb{R}$ and any $\alpha \in (0,1)$. We refer to $\widehat{\beta}(W)$ as the Bayes-assisted estimator.

4. FAB-PPI

Our approach, which we call FAB-PPI, combines the PPI framework with the FAB construction of confidence regions by specifying a prior on the rectifier Δ_θ. To ease the presentation, here we describe the method for $Y, \theta \in \mathbb{R}$. The general multivariate case is discussed in Appendix S4.

As in PPI, we use the sample mean (9) as the estimator of m_θ. For Δ_θ, we start by considering a consistent estimator $\widehat{\Delta}_\theta$, such as the sample mean (10) used in PPI, or the control variate estimator (12) used in PPI++. Throughout this section, we assume that Assumption 3.1 is satisfied. That is, a CLT holds for $\widehat{m}_\theta$ and $\widehat{\Delta}_\theta$ with respect to some estimators $(\widehat{\sigma}^f_\theta)^2$ and $\widehat{\sigma}^2_\theta$ of $\mathrm{var}(\widehat{m}_\theta)$ and $\mathrm{var}(\widehat{\Delta}_\theta)$, respectively. In this setting, let $\pi_0(\Delta_\theta; \tau_n)$ be a prior on Δ_θ with scale parameter τ_n, which may depend on the labelled data through $\widehat{\sigma}_\theta$. Denote by $\ell(w; \sigma, \tau)$ the log-marginal likelihood, evaluated at w, of a Gaussian likelihood model with mean Δ and variance σ^2 under the prior $\pi_0(\Delta; \tau)$,

$$\ell(w; \sigma, \tau) = \log \int_{\mathbb{R}} \mathcal{N}(w; \Delta, \sigma^2)\pi_0(\Delta; \tau)d\Delta.$$

4.1. Bayes-Assisted PPI Estimators

Consider the Bayes-assisted estimator

$$\widehat{\Delta}^{\texttt{FABPP}}_\theta = \widehat{\Delta}_\theta + \widehat{\sigma}^2_\theta \ell'\left(\widehat{\Delta}_\theta; \widehat{\sigma}_\theta, \tau_n\right) \tag{20}$$

for the rectifier Δ_θ. By Tweedie's formula (Efron, 2011), the above estimator is the posterior mean of the mean parameter of a Gaussian likelihood model under the prior π_0. Note however that we do not assume here that $\widehat{\Delta}_\theta$ is normally distributed for a fixed n.

The FAB-PPI estimator of $\theta^\star$, denoted by $\widehat{\theta}^{\texttt{FABPP}}$, is then obtained as the solution, in θ, to the equation

$$\widehat{m}_\theta + \widehat{\Delta}^{\texttt{FABPP}}_\theta = 0.$$

4.2. FAB-PPI Confidence Regions

As in PPI, let $\mathcal{T}_{\alpha-\delta}(\widehat{m}_\theta)$ denote a standard $1-(\alpha-\delta)$ confidence interval for m_θ. For Δ_θ, we apply the FAB framework with the prior π_0 to obtain a $1-\delta$ confidence region $\mathcal{R}^{\texttt{FABPP}}_\delta(\widehat{\Delta}_\theta) = \text{FAB-CR}(\widehat{\Delta}_\theta; \pi_0(\cdot\,; \tau_n), \widehat{\sigma}_\theta, \delta)$. Then, the FAB-PPI confidence region $\mathcal{C}^{\texttt{FABPP}}_\alpha$ is obtained as

$$\mathcal{C}^{\texttt{FABPP}}_\alpha = \left\{\theta \mid 0 \in \mathcal{R}^{\texttt{FABPP}}_\delta(\widehat{\Delta}_\theta) + \mathcal{T}_{\alpha-\delta}(\widehat{m}_\theta)\right\}. \tag{21}$$

Algorithm 1 summarises the steps of the FAB-PPI approach in a general convex estimation problem.

4.3. Choosing the Prior

FAB-PPI is motivated by applications in which the PPI predictor f is expected to be generally accurate, as measured by the rectifier Δ_θ. Such a property may be encoded in $\pi_0(\Delta_\theta; \tau_n)$ by choosing a prior that concentrates around

Algorithm 1 FAB-PPI for convex estimation

Input: labelled $\{(X_i, Y_i)\}_{i=1}^n$, unlabelled $\{\widetilde{X}_j\}_{j=1}^N$, predictor f, prior $\pi_0(\cdot\,;\tau_n)$, error levels α, δ

Set $\widehat{\lambda} = 1$ (FAB-PPI) or estimate $\widehat{\lambda}$ from data (FAB-PPI++) as in Angelopoulos et al. (2023a).

for $\theta \in \Theta_{\text{grid}}$ **do**

$\widehat{m}_\theta \leftarrow \frac{1}{N}\sum_{i=1}^N \mathcal{L}'_\theta(\widetilde{X}_i, f(\widetilde{X}_i))$

$\widehat{\xi} \leftarrow \frac{1}{n}\sum_{i=1}^n \Big(\mathcal{L}'_\theta(X_i, Y_i) - \widehat{\lambda}\mathcal{L}'_\theta(X_i, f(X_i))\Big)$

$\widehat{\Delta}_\theta \leftarrow \widehat{\xi} + (\widehat{\lambda}-1)\widehat{m}_\theta$

$(\widehat{\sigma}^f_\theta)^2 \leftarrow \frac{1}{N-1}\sum_{i=1}^N \Big(\mathcal{L}'_\theta(\widetilde{X}_i, f(\widetilde{X}_i)) - \widehat{m}_\theta\Big)^2$

$\widehat{\sigma}^2_\xi \leftarrow \frac{1}{n-1}\sum_{i=1}^n \Big(\mathcal{L}'_\theta(X_i, Y_i) - \widehat{\lambda}\mathcal{L}'_\theta(X_i, f(X_i)) - \widehat{\xi}\Big)^2$

$\widehat{\sigma}^2_\theta \leftarrow \frac{1}{n}\widehat{\sigma}^2_\xi + \frac{(\widehat{\lambda}-1)^2}{N}\widehat{\sigma}^2_m$

$\mathcal{T}_{\alpha-\delta}(\widehat{m}_\theta) \leftarrow \Big(\widehat{m}_\theta \pm \widehat{\sigma}^f_\theta z_{1-(\alpha-\delta)/2}\Big)$

$\mathcal{R}^{\text{FABPP}}_\delta(\widehat{\Delta}_\theta) \leftarrow \text{FAB-CR}(\widehat{\Delta}_\theta; \pi_0(\cdot\,;\tau_n), \widehat{\sigma}_\theta, \delta)$

$\widehat{\Delta}^{\text{FABPP}}_\theta \leftarrow \widehat{\Delta}_\theta + \widehat{\sigma}^2_\theta \ell'\Big(\widehat{\Delta}_\theta; \widehat{\sigma}_\theta, \tau_n\Big)$

end for

Outputs: estimator $\widehat{\theta}^{\text{FABPP}} = \arg\min_{\Theta_{\text{grid}}} \Big|\widehat{m}_\theta + \widehat{\Delta}^{\text{FABPP}}_\theta\Big|$ and CR $\mathcal{C}^{\text{FABPP}}_\alpha = \Big\{\theta \mid 0 \in \mathcal{R}^{\text{FABPP}}_\delta(\widehat{\Delta}_\theta) + \mathcal{T}_{\alpha-\delta}(\widehat{m}_\theta)\Big\}$

zero. As mentioned in Section 3.3, the FAB construction of $\mathcal{R}^{\text{FABPP}}_\delta(\widehat{\Delta}_\theta)$ will exhibit smaller volume compared to the classical CI, and hence result in downstream efficiency gains over standard PPI, if the true rectifier Δ_θ is likely under π_0. In particular, the prior scale τ_n controls the size of the potential efficiency gains and losses of FAB-PPI over PPI: the smaller τ_n, the more the resulting CR will shrink (resp. grow) when $\Delta_\theta \simeq 0$ (resp. $|\Delta_\theta| \gg 0$). Experimentally, we find that the choice $\tau_n = \widehat{\sigma}_\theta$ results in a parameter-free approach that strikes a good compromise. More general choices of τ_n are briefly mentioned in Section 6.

A seemingly natural proposal for π_0 that meets the requirements above is the Gaussian prior

$$\pi_{\text{N}}(\Delta_\theta; \widehat{\sigma}_\theta) = \mathcal{N}(\Delta_\theta; 0, \widehat{\sigma}_\theta). \tag{22}$$

However, as we will discuss in Section 4.4, π_{N} exhibits undesirable properties for FAB-PPI. Instead, we propose to use the horseshoe prior (Carvalho et al., 2010)

$$\pi_{\text{HS}}(\Delta_\theta; \widehat{\sigma}_\theta) = \int_0^\infty \mathcal{N}(\Delta_\theta; 0, \nu^2\widehat{\sigma}^2_\theta) C^+(\nu; 0, 1) d\nu, \tag{23}$$

where $C^+(\nu; 0, 1)$ denotes the pdf of the half-Cauchy distribution with location parameter 0 and scale parameter 1. In the case of π_{HS}, the choice of scaling $\tau_n = \widehat{\sigma}_\theta$ is further motivated by Piironen & Vehtari (2017, §3.3). Furthermore, the horseshoe prior has power-law tails, making it a particularly robust choice for FAB-PPI, as discussed in Section 4.4.

Crucially, for both priors π_{N} and π_{HS}, the marginal likelihood under a Gaussian model with standard deviation $\widehat{\sigma}_\theta$ can be expressed in terms of standard functions (see Appendix S1.1 for the horseshoe), enabling us to compute $\widehat{\Delta}^{\text{FABPP}}_\theta$ and $\mathcal{R}^{\text{FABPP}}_\delta(\widehat{\Delta}_\theta)$ in Algorithm 1.

4.4. Theoretical Properties

As shown by the following result, proved in Appendix S3.1, the FAB-PPI CR has exact asymptotic coverage.

Theorem 4.1 (Asymptotic coverage)**.** *For $\alpha \in (0, 1)$, let $\mathcal{C}^{FABPP}_\alpha$ be the FAB-PPI confidence region (21) under the Gaussian prior (22) or the horseshoe prior (23). Then, under Assumption 3.1,*

$$\liminf_{\min(n,N)\to\infty} \Pr(\theta^* \in \mathcal{C}^{FABPP}_\alpha) \geq 1 - \alpha.$$

The proof of Theorem 4.1 crucially relies on showing exact asymptotic coverage of the FAB CR $\mathcal{R}^{\text{FABPP}}_\delta(\widehat{\Delta}_\theta)$. While the latter holds for both priors introduced in the previous sections, the two limits are very different. In particular, as discussed in Remark S3.5, the volume of $\mathcal{R}^{\text{FABPP}}_\delta(\widehat{\Delta}_\theta)$ vanishes asymptotically under π_{HS}, while it does not under π_{N}.

The behaviour of $\mathcal{R}^{\text{FABPP}}_\delta(\widehat{\Delta}_\theta)$ under the two priors also differs for large values of observed $\widehat{\Delta}_\theta$. In case of increasing disagreement between the prior and the data, Gaussian FAB confidence regions are known to become arbitrarily large (Yu & Hoff, 2018). On the other hand, thanks to its power-law tails, the horseshoe results in confidence regions that revert to the corresponding standard CI (Cortinovis & Caron, 2024). Here, we state the implication of this property on FAB-PPI informally, and provide a formal proof in Appendix S3.2.

Proposition 4.2 (Robustness under the horseshoe, informal)**.** *For $\alpha \in (0, 1)$, let $\mathcal{C}^{FABPP}_\alpha$ and $\mathcal{C}^{PP}_\alpha$ denote, respectively, the FAB-PPI confidence region (21) under the horseshoe prior (23) and the standard CLT-based PPI CI for θ, both viewed as functions of $\widehat{\Delta}_\theta$. If $|\widehat{\Delta}_\theta| \gg 0$, then*

$$\mathcal{C}^{FABPP}_\alpha \simeq \mathcal{C}^{PP}_\alpha.$$

In practice, this means that, in the presence of heavily biased predictors, FAB-PPI with the horseshoe prior reverts to standard PPI. In a sense, this represents a form of robustness to prior misspecification of FAB-PPI under the horseshoe.

Overall, Remark S3.5 and Proposition 4.2 provide strong support for preferring π_{HS} over π_{N} within the FAB-PPI framework.

4.5. FAB-PPI for Mean Estimation

To provide a concrete example, a specialised version of Algorithm 1 under the squared loss is derived in Ap-

pendix S2.1. Here, we briefly discuss the differences between the FAB-PPI mean estimator and its standard PPI counterpart, as well as the asymptotic behaviour of the former. Under the squared loss, the rectifier $\Delta := \Delta_\theta$ does not depend on θ and the FAB-PPI estimator $\widehat{\theta}^{\text{FABPP}}$ corresponding to the chosen estimator $\widehat{\Delta}$ (PPI or PPI++) is given by

$$\begin{aligned}\widehat{\theta}^{\text{FABPP}} &= \widehat{\theta} - \widehat{\sigma}^2 \ell'(\widehat{\Delta}; \widehat{\sigma}, \tau_n) \\ &= \overline{Y} - \widehat{\lambda}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \frac{1}{N}\sum_{j=1}^{N} f(\widetilde{X}_j)\right) \\ &\quad - \widehat{\sigma}^2 \ell'\left(\widehat{\Delta}; \widehat{\sigma}, \tau_n\right),\end{aligned} \tag{24}$$

where $\widehat{\sigma}^2$ is an estimator of $\text{var}(\widehat{\Delta})$, $\widehat{\theta}$ is the PPI estimator corresponding to $\widehat{\Delta}$, and $\widehat{\lambda}$ is set either to one (PPI) or (14) (PPI++). In both cases, the estimator $\widehat{\theta}^{\text{FABPP}}$ takes the form

$$\text{Classic Estimator + PPI correction + } \textbf{Bayes correction},$$

where the last component depends on the chosen prior. The following proposition, proved in Appendix S3.3, further differentiates between the priors presented in Section 4.3 in favour of the horseshoe.

Proposition 4.3 (Consistency of FAB-PPI mean estimators)**.** *Let $\widehat{\theta}^{\text{FABPP}}_{HS}$ and $\widehat{\theta}^{\text{FABPP}}_{N}$ be the FAB-PPI estimators (24) under the horseshoe (23) and Gaussian (22) priors, respectively. If the PPI estimator $\widehat{\theta}$ is a consistent estimator of $\theta^\star$, then $\widehat{\theta}^{\text{FABPP}}_{HS}$ is a consistent estimator of $\theta^\star$, while $\widehat{\theta}^{\text{FABPP}}_{N}$ is not.*

Intuitively, this is due to the fact that the influence of π_{HS} vanishes asymptotically, while for π_{N} it does not.

5. Experiments

We compare FAB-PPI and power-tuned FAB-PPI (FAB-PPI++) to classical inference, PPI and power-tuned PPI (PPI++) on both synthetic and real estimation problems. For FAB-PPI, we use (HS) and (N) to indicate the use of the horseshoe and Gaussian priors defined in Section 4.3. As already mentioned, PPI is motivated by settings in which labelled data are scarce, while unlabelled data are abundant. Moreover, the application of FAB to PPI specifically targets the estimation of the rectifier Δ_θ. For these reasons, we choose to focus on cases where $N \gg n$ is large enough to rule out any uncertainty on the measure of fit m_θ, which we estimate using the sample mean $\widehat{m}_\theta$ (9). As a result of this, given a $1-\delta$ confidence interval (FAB or not) $\mathcal{R}_\delta$ for Δ_θ, the corresponding $1-\alpha$ CI for $\theta^\star$ is obtained simply by setting $\delta = \alpha$ and shifting $\mathcal{R}_\delta$ by $\widehat{m}_\theta$. This simplification allows us to evaluate the direct effect of FAB on the procedure, eliminating concerns about the loss of tightness in the CI on θ^* due to the Minkowski sum in Equation (21). In all experiments, we check empirically that N is large enough to make this assumption by monitoring the coverage of the resulting intervals against both the nominal level $1-\alpha$ and the coverage of PPI intervals that also consider the uncertainty on m_θ (denoted with PPI (full) and PPI++ (full) in Appendix S6).

5.1. Synthetic Data

The simulated experiments below have a common structure. We sample two datasets, n labelled observations $\{(X_i, Y_i)\}_{i=1}^n$ iid from $\mathbb{P}$ and N unlabelled observations $\{\widetilde{X}_i\}_{i=1}^N$ iid from $\mathbb{P}_X$. We use a prediction rule f to obtain predictions $\{f(X_i)\}_{i=1}^n$ and $\{f(\widetilde{X}_i)\}_{i=1}^N$. We apply the different procedures to obtain estimates and $1-\alpha$ confidence regions for the mean $\theta^* = \mathbb{E}[Y]$. For all experiments, we set $\alpha = 0.1$ and report the average mean squared error (MSE), interval volume, and coverage over 1000 repetitions.

Biased Predictions. We sample $X_i \overset{iid}{\sim} \mathcal{N}(0,1)$ and $Y_i = X_i + \epsilon_i$ with $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0,1)$, so that $\theta^* = \mathbb{E}[Y] = 0$. The prediction rule is defined as $f(X_i) = X_i + \gamma$, where $\gamma \in \mathbb{R}$. For this choice, the bias of f is controlled by γ, since $\text{MSE}(f) = \gamma^2 + 1$. For this experiment, we assume that N is infinite, set $n = 200$, and vary γ between -1.5 and 1.5. Figure 1 shows the average interval volume as a function of γ for classical inference, PPI++, and FAB-PPI++ with both a horseshoe and a Gaussian prior. Results for the non power-tuned methods, as well as MSE and coverage plots, are reported in Figure S6. Except for the version with the Gaussian prior, all the PPI procedures outperform classical inference for every bias level γ, but the behaviour exhibited by PPI++ deserves attention, as its CI volume is approximately constant across values of γ. This is due to the fact that, since N is taken to be infinite and n is fairly large, $\widehat{\lambda} \simeq \text{cov}(Y, f(X)) = 1$ and the rectifier is accurately estimated with similar variance across all values of γ. On the other hand, the CI volume for the FAB-PPI methods varies greatly with γ. When the bias is small ($\gamma \simeq 0$), the observed rectifier has a value close to

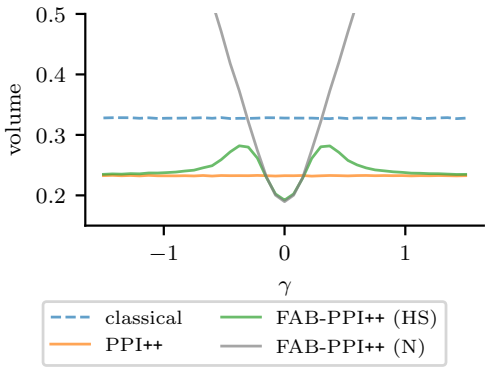

Figure 1. Biased predictions study. The panel shows the average CI volume as the bias level γ varies.

0, leading to smaller CIs. As the bias increases, the volume of the confidence intervals grows, until it surpasses that of the PPI intervals. At this point, the two FAB-PPI procedures behave differently: the volume of the Gaussian intervals grows unbounded, whereas the horseshoe intervals eventually revert to the PPI ones. This example clearly shows that FAB-PPI with a horseshoe prior allows to obtain smaller CIs when the predictions are good, while ensuring robustness as the quality of the predictions decreases (Proposition 4.2).

Noisy Predictions. We consider the mean estimation example of Angelopoulos et al. (2023a, §7.1.1), which does not involve any covariate X. We sample $Y_i \overset{iid}{\sim} \mathcal{N}(0,1)$, so that $\theta^* = \mathbb{E}[Y] = 0$. The prediction rule is defined as $f(X_i) = Y_i + \sigma_Y \epsilon_i$, where $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0,1)$ and σ_Y is successively set to 0.1, 1, and 2. For this experiment, we set $N = 10^6$ and vary n from 100 to 1000. Figure 2 shows the average interval volume as a function of n for the different methods as the noise level σ_Y varies, while similar plots for the MSE and coverage are reported in Figure S7. In this

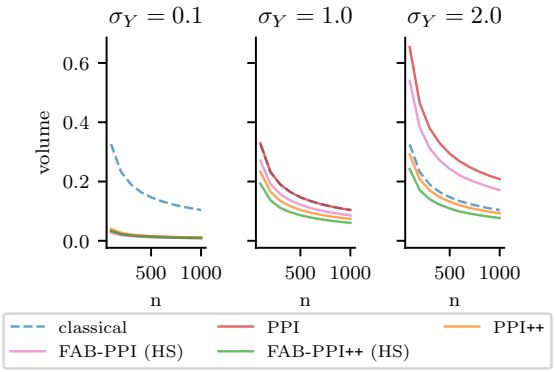

Figure 2. Noisy predictions study. The left, middle and right panels show the average CI volume for noise levels $\sigma_Y = 0.1, 1, 2$.

case, the effect of power tuning matches the observations of Angelopoulos et al. (2023a): as the noise level increases, $\hat{\lambda}$ decreases and less weight is given to the predicted labels. When the noise is small, all PPI procedures perform similarly, and much better than classical inference. When the noise is large, the power-tuned procedures perform similar to or better than classical inference, whereas the non-tuned alternatives lose ground. At the intermediate noise level, the power-tuned methods clearly outperform the other baselines. Crucially, FAB-PPI outperforms the PPI counterpart at all noise levels, with FAB-PPI++ being the best performer overall. This is because, in this setting, while predictions exhibit increasing variance with σ_Y, they remain unbiased. As a result of this, regardless of the value of λ used, any additional shrinkage performed on the rectifier by FAB-PPI is beneficial. This example shows that FAB-PPI++ retains the benefits of power tuning, while also taking advantage of the adaptive shrinkage provided by the FAB procedure.

5.2. Real Data

We consider several estimation experiments using the datasets presented in Angelopoulos et al. (2023b) and briefly described in Appendix S5.1. Each dataset comes with covariate/label/prediction triples $\{X_i, Y_i, f(X_i)\}_{i=1}^N$, which we randomly split into two subsets with n labelled and $N-n$ unlabelled observations, for varying values of n. For all experiments and methods, we report the average estimation MSE, CI volume and coverage across multiple repetitions.

We begin with four experiments, where the machine learning predictions provided are of high quality, and whose goals are as follows. Two of them are mean estimation tasks performed on the GALAXIES and FOREST datasets. The third one, performed on the ALPHAFOLD dataset, is an odds ratio estimation task, for which the construction of confidence intervals also indirectly involves mean estimation as detailed in Appendix S5.1. The fourth one, involving the HEALTHCARE dataset, is a logistic regression task. Figure 3 shows the results for classical inference, PPI++, and FAB-PPI++ applied to the datasets involving mean estimation.

The mean estimation results for the non power-tuned methods are reported in Figure S8, whereas the ones for the logistic regression experiment are reported in Figure S11. In all cases, FAB-PPI/FAB-PPI++ outperform classical inference and the corresponding PPI methods, both in terms of MSE and CI volume, while achieving comparable coverage. These examples suggest that the quality of the predictions of existing machine learning models on several real datasets may fall into the regime where the adaptive shrinkage provided by the FAB framework leads to a further improvement over standard PPI. In these settings, as the predictions are good, FAB-PPI under the horseshoe and Gaussian priors exhibit similar gains, as already seen in Figure 1.

However, the same is not true in the presence of bad predictions. For instance, Figure S12 shows the results of a quantile estimation experiment on the GENES dataset, where predictions are heavily biased. In this case, the behaviour of the FAB-PPI methods under the horseshoe and Gaussian priors differs significantly: the former matches the performance of the PPI methods, which outperform classical inference, whereas the latter leads to much larger MSE and CIs. As previously discussed, such desirable behaviour of FAB-PPI under the horseshoe prior is due to its robustness against large bias levels (Proposition 4.2). Similarly, Figure S13 reports the results of a linear regression experiment on the CENSUS dataset. For one of the two parameters considered (panel (a)), FAB-PPI underperforms the alternatives under both priors for small n. However, as n increases, the performance under the horseshoe prior improves and even-

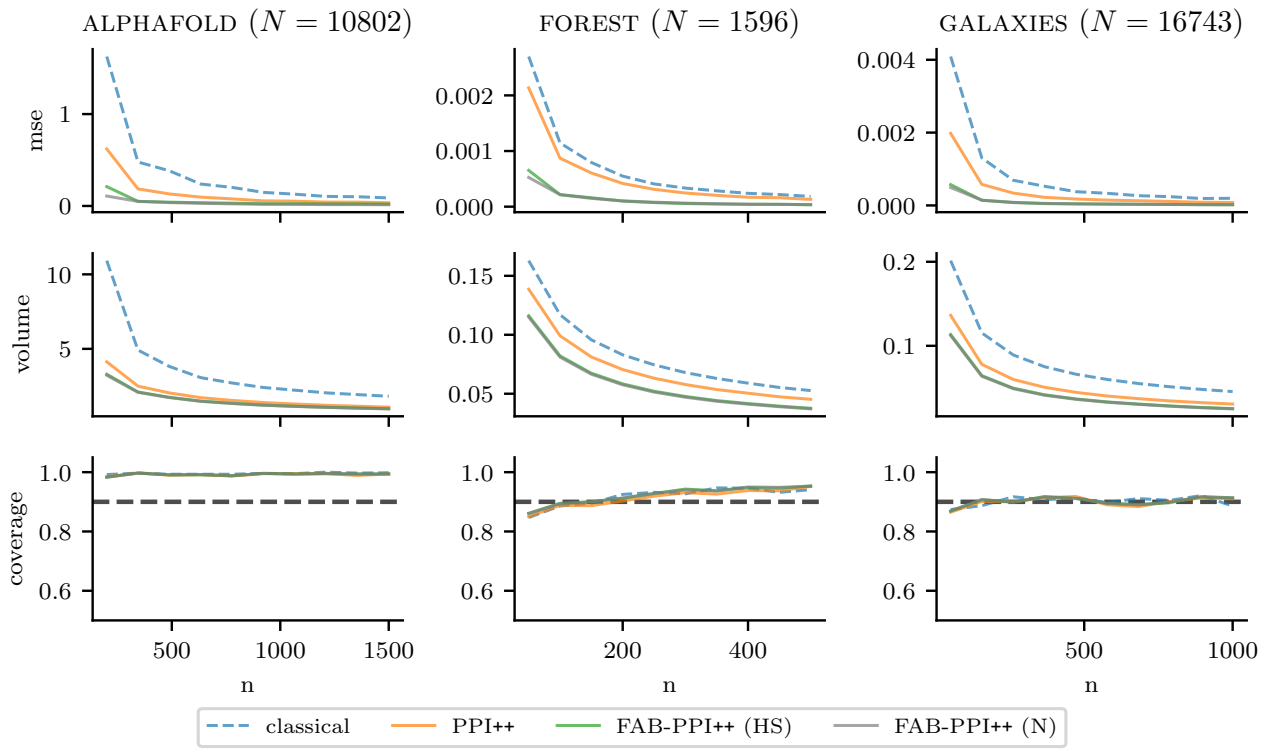

Figure 3. Real data mean estimation study. The left, middle, and right panels correspond to the ALPHAFOLD, GALAXIES, and FOREST datasets. The top, middle, and bottom rows show average MSE, CI volume, and CI coverage over 1000 repetitions for $\alpha = 0.1$.

tually matches that of the PPI methods, while the Gaussian prior does not. This example shows another facet of the horseshoe's robustness: even for moderate bias levels, as the available labelled sample size grows, disagreements between the prior and the data become apparent (i.e. $\mathrm{var}(\widehat{\Delta}_\theta)$ decreases), eventually leading Proposition 4.2 to kick in.

6. Discussion and Extensions

We proposed FAB-PPI as a Bayes-informed method to significantly improve the performance of PPI in the presence of high-quality predictions. In doing so, we showed that the horseshoe represents a sensible default prior for FAB-PPI, contrary to the seemingly natural choice of a Gaussian prior. However, several options may be worth exploring.

In particular, the horseshoe prior was chosen due to its popularity and key properties: (i) its spike at zero (ii) power-law tails and (iii) the closed-form expression for the marginal density $\pi(y)$. However, many other scale-mixture of Gaussians models share these properties. For example, the family of priors with a beta prime (aka inverted beta) prior over the variance (Polson & Scott, 2012), which includes the horseshoe, normal-exponential-gamma (Griffin & Brown, 2011) and other robust priors (Berger, 1980; Strawderman, 1971) as special cases, share the same three properties. On the other hand, some other standard priors such as the Laplace prior (Park & Casella, 2008) or normal-gamma prior (Caron & Doucet, 2008; Griffin & Brown, 2010) do not have power-law tails and therefore do not offer the same robustness guarantees. Other priors, such as the student-t, lack an analytical expression for $\pi(y)$, therefore requiring additional numerical approximation to be applied to FAB-PPI.

Furthermore, we used the scale σ of the noise in the generative model as the scale for both the horseshoe and Gaussian priors, as this allows us to obtain a simple, parameter-free approach, which generally performs well. Alternatively, one could consider a prior scale of $\eta\sigma$, where η is a hyperparameter to be tuned using a validation set. However, in the case of the horseshoe, this renders the marginal likelihood intractable. While using a rescaled horseshoe prior for FAB-PPI remains feasible through numerical integration, as shown in Figure S10, this increases the computational cost of the method. By contrast, a rescaled Gaussian prior would not encounter this issue. Furthermore, we conjecture that choosing a scale that does not depend on σ may resolve the inconsistency of the estimator based

on a Gaussian prior, which was discussed in Section 4.5.

As a potential drawback, FAB-PPI shares the computational limitations of the PPI approach (Angelopoulos et al., 2023b), which are discussed in Angelopoulos et al. (2023a). In particular, except for particular cases such as mean estimation and linear regression, the method requires evaluating $\widehat{m}_\theta + \widehat{\Delta}_\theta$ over a grid of values of θ. This can be computationally expensive, especially in high-dimensional settings.

Supplementary Material and Code. The supplementary material contains additional background, proofs, and experiments. All sections, figures, and equations in the supplementary material are prefixed with 'S' for clarity. Code for reproducing the experiments is available at `https://github.com/stefanocortinovis/fab-ppi`.

Acknowledgements

Stefano Cortinovis is supported by the EPSRC Centre for Doctoral Training in Modern Statistics and Statistical Machine Learning (EP/S023151/1).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Andrews, D. and Mallows, C. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(1):99–102, 1974.

Angelopoulos, A., Duchi, J., and Zrnic, T. PPI++: Efficient prediction-powered inference. *arXiv preprint arXiv:2311.01453*, 2023a.

Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., and Zrnic, T. Prediction-powered inference. *Science*, 382 (6671):669–674, 2023b.

Berger, J. A robust generalized Bayes estimator and confidence region for a multivariate normal mean. *The Annals of Statistics*, pp. 716–761, 1980.

Bludau, I., Willems, S., Zeng, W.-F., Strauss, M. T., Hansen, F. M., Tanzer, M. C., Karayel, O., Schulman, B. A., and Mann, M. The structural context of posttranslational modifications at a proteome-wide scale. *PLoS biology*, 20(5):e3001636, 2022.

Brown, L. D., Casella, G., and G. Hwang, J. Optimal confidence sets, bioequivalence, and the limacon of Pascal. *Journal of the American Statistical Association*, 90(431): 880–889, 1995.

Bullock, E. L., Woodcock, C. E., Souza Jr, C., and Olofsson, P. Satellite-based estimates reveal widespread forest degradation in the Amazon. *Global Change Biology*, 26 (5):2956–2969, 2020.

Caron, F. and Doucet, A. Sparse Bayesian nonparametric regression. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 88–95, 2008.

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

Cortinovis, S. and Caron, F. Robust Bayes-assisted confidence regions. *arXiv preprint arXiv:2410.20169*, 2024.

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

Farchione, D. and Kabaila, P. Confidence intervals for the normal mean utilizing prior information. *Statistics & Probability Letters*, 78(9):1094–1100, 2008.

Fisch, A., Maynez, J., Hofer, R. A., Dhingra, B., Globerson, A., and Cohen, W. W. Stratified prediction-powered inference for hybrid language model evaluation. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024.

Ghosh, J. K. On the relation among shortest confidence intervals of different types. *Calcutta Statistical Association Bulletin*, 10(4):147–152, 1961.

Glasserman, P. *Monte Carlo Methods in Financial Engineering*. Springer, 2003.

Griffin, J. E. and Brown, P. J. Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1):171–188, 2010.

Griffin, J. E. and Brown, P. J. Bayesian hyper-lassos with non-convex penalization. *Australian & New Zealand Journal of Statistics*, 53(4):423–442, 2011.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hofer, R., Maynez, J., Dhingra, B., Fisch, A., Globerson, A., and Cohen, W. Bayesian prediction-powered inference. *arXiv preprint arXiv:2405.06034*, 2024.

Hoff, P. Bayes-optimal prediction with frequentist coverage control. *Bernoulli*, 29(2):901–928, 2023.

Hoff, P. and Yu, C. Exact adaptive confidence intervals for linear regression coefficients. *Electronic Journal of Statistics*, 13:94–119, 2019.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Kabaila, P. and Farchione, D. Confidence intervals that utilize sparsity. *Stat*, 11(1):e434, 2022.

Kabaila, P. and Giri, K. Further properties of frequentist confidence intervals in regression that utilize uncertain prior information. *Australian & New Zealand Journal of Statistics*, 55(3):259–270, 2013.

Park, T. and Casella, G. The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

Piironen, J. and Vehtari, A. On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pp. 905–913, 2017.

Polson, N. G. and Scott, J. G. On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis*, 7(4):887–902, 2012.

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

Pratt, J. W. Shorter confidence intervals for the mean of a normal distribution with known variance. *The Annals of Mathematical Statistics*, pp. 574–586, 1963.

Puza, B. and O'Neill, T. Interval estimation via tail functions. *Canadian Journal of Statistics*, 34(2):299–310, 2006.

Robins, J. M. and Rotnitzky, A. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

Slater, L. J. *Confluent hypergeometric functions*. Cambridge University Press, 1960.

Strawderman, W. E. Proper Bayes minimax estimators of the multivariate normal mean. *The Annals of Mathematical Statistics*, 42(1):385–388, 1971.

Vaishnav, E. D., de Boer, C. G., Molinet, J., Yassour, M., Fan, L., Adiconis, X., Thompson, D. A., Levin, J. Z., Cubillos, F. A., and Regev, A. The evolution, evolvability and engineering of gene regulatory DNA. *Nature*, 603 (7901):455–463, 2022.

Willett, K. W., Lintott, C. J., Bamford, S. P., Masters, K. L., Simmons, B. D., Casteels, K. R., Edmondson, E. M., Fortson, L. F., Kaviraj, S., Keel, W. C., et al. Galaxy Zoo 2: detailed morphological classifications for 304 122 galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, 435(4):2835–2860, 2013.

Yu, C. and Hoff, P. D. Adaptive multigroup confidence intervals with constant coverage. *Biometrika*, 105(2): 319–335, 2018.

Zhang, A., Brown, L. D., and Cai, T. T. Semi-supervised inference: general theory and estimation of means. *The Annals of Statistics*, 47(5):2538–2566, 2019.

Zrnic, T. and Candès, E. J. Active statistical inference. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 62993–63010, 2024a.

Zrnic, T. and Candès, E. J. Cross-prediction-powered inference. *Proceedings of the National Academy of Sciences*, 121(15):e2322083121, 2024b.

S1. Additional Background Material

S1.1. Horseshoe Prior

Consider the Gaussian likelihood model

$$Y \mid \beta \sim \mathcal{N}(\beta, \sigma^2)$$

with standard deviation $\sigma > 0$ and mean parameter $\beta \in \mathbb{R}$. The horseshoe prior (Carvalho et al., 2010) with density π_{HS} can be represented as a scale mixture of normals (Andrews & Mallows, 1974)

$$\beta \mid \nu^2 \sim \mathcal{N}(0, \eta^2\sigma^2\nu^2) \tag{S25}$$

$$\nu \sim C^+(0, 1), \tag{S26}$$

where $\eta > 0$ and $C^+(0, 1)$ is the half-Cauchy distribution with location parameter 0 and scale parameter 1. Throughout this section and the main text, we assume $\eta = 1$. The rationale for this choice, along with a discussion of the general case $\eta \neq 1$, is provided at the end of this section.

The marginal likelihood is given by

$$\begin{aligned}
\pi(y) &= \int_{-\infty}^{\infty} \mathcal{N}(y \mid \beta, \sigma^2)\pi_{\text{HS}}(\beta)d\beta \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_0^{\infty} \frac{1}{\sqrt{1+\nu^2}} e^{-\frac{y^2}{2\sigma^2(1+\nu^2)}} p(\nu)d\nu \\
&= \frac{2}{\pi\sqrt{2\pi\sigma^2}} \int_0^{\infty} e^{-\frac{y^2}{2\sigma^2(1+\nu^2)}} \frac{1}{(1+\nu^2)^{3/2}} d\nu.
\end{aligned}$$

Using the change of variable $u = \frac{1}{1+\nu^2}$, we obtain

$$\begin{aligned}
\pi(y) &= \frac{1}{\pi\sqrt{2\pi\sigma^2}} \int_0^1 e^{-\frac{uy^2}{2\sigma^2}} (1-u)^{-1/2} du \\
&= \frac{2}{\pi\sqrt{2\pi\sigma^2}} \, {}_1F_1\left(1, \frac{3}{2}, -\frac{y^2}{2\sigma^2}\right),
\end{aligned}$$

where ${}_1F_1$ is (Kummer's) confluent hypergeometric function of the first kind, with integral representation

$${}_1F_1(a, b, z) = \frac{\Gamma(b)}{\Gamma(a)\Gamma(b-a)} \int_0^1 e^{zt} t^{a-1}(1-t)^{b-a-1} dt.$$

Alternatively, the marginal can be expressed in function of the imaginary error function (erfi) or Dawson function (aka Dawson integral) as

$$\begin{aligned}
\pi(y) &= \frac{1}{\pi\sqrt{2\sigma^2}} e^{-y^2/(2\sigma^2)} \frac{\operatorname{erfi}(|y|/\sqrt{2\sigma^2})}{|y|/(\sqrt{2\sigma^2})} \\
&= \frac{2}{\pi^{3/2}} \frac{1}{|y|} D\left(\frac{|y|}{\sqrt{2\sigma^2}}\right)
\end{aligned}$$

where Dawson's function is defined as

$$D(z) = e^{-z^2} \int_0^z e^{t^2} dt.$$

The marginal likelihood exhibits power-law tails

$$\pi(y) \sim C \frac{1}{|y|^2} \quad \text{as } |y| \to \infty$$

for some constant $C > 0$. Let $\ell(y) = \log \pi(y)$ denote the log-marginal likelihood. Kummer's function has the derivative

$$\frac{d}{dz} {}_1F_1(a, b, z) = \frac{a}{b} {}_1F_1(a+1, b+1, z).$$

It follows that

$$\ell'(y) = \frac{\pi'(y)}{\pi(y)} = -\frac{2}{3}\frac{y}{\sigma^2}\frac{{}_1F_1\left(2, \frac{5}{2}, -\frac{y^2}{2\sigma^2}\right)}{{}_1F_1\left(1, \frac{3}{2}, -\frac{y^2}{2\sigma^2}\right)}.$$

Applying Tweedie's formula (Efron, 2011), we obtain the posterior mean

$$\mathbb{E}[\beta \mid y] = y + \sigma^2 \ell'(y; \sigma) = (1 - \kappa(y))y, \tag{S27}$$

where the shrinkage function $\kappa(y) \in (0, 1)$ is given by

$$\kappa(y) = \frac{2}{3}\frac{{}_1F_1\left(2, \frac{5}{2}, -\frac{y^2}{2\sigma^2}\right)}{{}_1F_1\left(1, \frac{3}{2}, -\frac{y^2}{2\sigma^2}\right)}.$$

Using the asymptotic expansion (Slater, 1960, Chapter 4, Eq. (4.I.3))

$${}_1F_1(a, b, -z) \sim z^{-a}\frac{\Gamma(b)}{\Gamma(b-a)}$$

as $z \to \infty$, we find

$$\kappa(y) \sim \frac{2\sigma^2}{y^2}$$
$$|\mathbb{E}[\beta \mid y] - y| = \sigma^2|\ell'(y)| \sim \frac{2\sigma^2}{|y|}$$

as $|y| \to \infty$.

The horseshoe prior π_{HS} has two key properties: an infinite spike at zero, inducing strong shrinkage near $y = 0$, and Cauchy-like tails, ensuring that strong signals remain largely unshrunk ($\kappa(y) \to 0$ and $|\mathbb{E}[\beta \mid y] - y| \to 0$ as $|y| \to \infty$). This is illustrated in Figure S4.

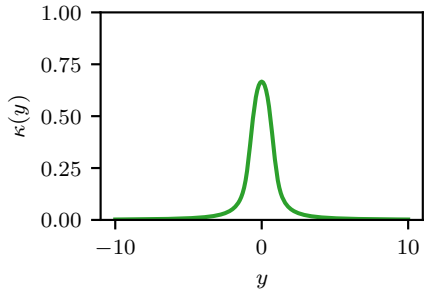

Figure S4. Shrinkage function $\kappa(y)$ for the horseshoe prior when $\sigma^2 = 0.1$.

Remark S1.1 (Parameterisation). In this section and in the main text, we focused on the specific parametrisation $\eta = 1$. For a general η, similar expressions can be derived for the marginal likelihood and posterior mean, replacing Kummer's ${}_1F_1$ function with the more general degenerate hypergeometric function of two variables, Φ_1 (see (Carvalho et al., 2010, Equations (4) in the main text and (A1) in the appendix)). While Kummer's ${}_1F_1$ function is implemented in many standard scientific libraries, such as SciPy, Φ_1 is not. Consequently, computing the marginal likelihood when $\eta \neq 1$ requires numerical integration. Since the evaluation of the marginal likelihood is crucial to our approach, it is therefore reasonable to set $\eta = 1$ here.

S1.2. FAB Framework

In this section, we provide additional background on the FAB framework (Pratt, 1961; 1963; Yu & Hoff, 2018).

Let $Y \mid \beta \sim \mathcal{N}(\beta, \sigma^2)$ with some prior $\pi_0(\beta)$. Denote by $\pi(y) = \int_{\mathbb{R}} p(y \mid \beta)\pi_0(\beta)d\beta$ the corresponding marginal likelihood. For $\alpha \in (0,1)$, let $\mathcal{C}_\alpha$ be the confidence procedure that solves the constrained optimisation problem

$$\mathcal{C}_\alpha = \arg\min_{\widetilde{\mathcal{C}}_\alpha} \mathbb{E}[\mathrm{vol}(\widetilde{\mathcal{C}}_\alpha(Y))]$$
$$\text{under the constraints } \Pr(\beta \in \mathcal{C}_\alpha(Y) \mid \beta = \beta') = 1-\alpha \text{ for all fixed } \beta',$$

where $\mathrm{vol}(\mathcal{C}_\alpha(y)) = \int_{\beta' \in \mathcal{C}_\alpha(y)} d\beta'$ is the volume of $\mathcal{C}_\alpha(y)$ and

$$\mathbb{E}[\mathrm{vol}(\mathcal{C}_\alpha(Y))] = \int_{\mathbb{R}} \mathrm{vol}(\mathcal{C}_\alpha(y))\pi(y)dy \tag{S28}$$

is the expected volume under the marginal distribution $\pi(y)$. By the Ghosh-Pratt identity (Ghosh, 1961; Pratt, 1961),

$$\begin{aligned}\mathbb{E}[\mathrm{vol}(\mathcal{C}_\alpha(Y))] &= \int_{\mathbb{R}} \mathrm{vol}(\mathcal{C}_\alpha(y))\pi(y)dy \\ &= \int_{\mathbb{R}}\int_{\mathbb{R}} 1_{\beta' \in \mathcal{C}_\alpha(y)} d\beta' \pi(y)dy \\ &= \int_{\mathbb{R}} \Pr(\beta' \in \mathcal{C}_\alpha(Y))d\beta'.\end{aligned}$$

That is, minimising $\mathbb{E}[\mathrm{vol}(\mathcal{C}_\alpha(Y))]$ is equivalent to minimising $\Pr(\beta' \in \mathcal{C}_\alpha(Y))$ for each $\beta' \in \mathbb{R}$. Define the acceptance region

$$A_\alpha(\beta') = \{y \mid \beta' \in \mathcal{C}_\alpha(y)\}.$$

The constrained optimisation problem above then reduces to solving, for each β',

$$A_\alpha(\beta') = \arg\max_{\widetilde{A}_\alpha} \Pr(Y \notin \widetilde{A}_\alpha(\beta'))$$
$$\text{such that } \Pr(Y \notin A_\alpha(\beta) \mid \beta = \beta') = \alpha.$$

The term $\Pr(Y \notin A_\alpha(\beta'))$ may be interpreted as the power of a size-α test

$$H_0 : \beta = \beta' \text{ vs } H_1 : \beta \sim \pi_0$$

where $Y \mid \beta \sim \mathcal{N}(\beta, \sigma^2)$. By the Neyman-Pearson lemma, the most powerful test is of the form

$$A_\alpha(\beta') = \left\{ y \mid \frac{\pi(y)}{p(y \mid \beta')} \le k_\alpha(\beta') \right\}$$

where $k_\alpha(\beta')$ is such that $\Pr(Y \in A_\alpha(\beta) \mid \beta = \beta') = 1-\alpha$. The acceptance region is an interval $[\underline{A}_\alpha(\beta'), \overline{A}_\alpha(\beta')]$ (Cortinovis & Caron, 2024, Theorem 2.1). Defining

$$w_\alpha(\beta') = \frac{1}{\alpha}\Phi\left(\frac{\underline{A}_\alpha(\beta') - \beta'}{\sigma}\right),$$

the confidence region is given by

$$\mathcal{C}_\alpha(y) = \{\beta' \mid \underline{A}_\alpha(\beta') = \beta' - \sigma z_{1-\alpha w_\alpha(\beta')} \le y \le \beta' + \sigma z_{1-\alpha(1-w_\alpha(\beta'))} = \overline{A}_\alpha(\beta')\}.$$

The function $w_\alpha(\beta') \in [0,1]$ is called the spending function or tail function (Puza & O'Neill, 2006; Yu & Hoff, 2018), and represents the proportion of the α rejection budget allocated to the left tail of the acceptance interval $[\underline{A}_\alpha(\beta'), \overline{A}_\alpha(\beta')]$.

The spending function w_α satisfies several key properties, which will be useful for our asymptotic analysis. Most of these originate from Cortinovis & Caron (2024). Under mild assumptions on the prior, satisfied for the models considered in this paper, $w_\alpha(\beta)$ is continuous in β. If the prior π_0 is symmetric around zero, we have

$$w_\alpha(-\beta') = 1 - w_\alpha(\beta'). \tag{S29}$$

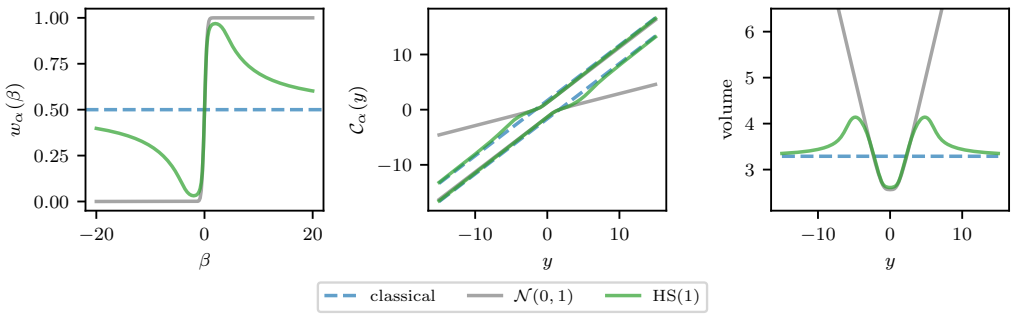

Figure S5. Comparison of the FAB procedures under a Gaussian ($\tau^2 = 1$) and a horseshoe ($\eta = 1$) priors when $\sigma^2 = 1$ and $\alpha = 0.1$.

Additionally, if the prior $\pi_0(\beta) := \pi_0(\beta; \sigma)$ admits σ as a scale parameter, writing $w_\alpha(\beta; \sigma)$ for the corresponding tail function, we have

$$w_\alpha(\beta; \sigma) = w_\alpha\left(\frac{\beta}{\sigma}; 1\right). \tag{S30}$$

We now describe other properties of the spending function in the case of a Gaussian prior and of a prior with power-law tails, such as the horseshoe.

Proposition S1.2 (FAB with a Gaussian prior (Pratt, 1963; Yu & Hoff, 2018))**.** *If the prior $\pi_0(\beta) = \mathcal{N}(\beta; 0, \tau^2\sigma^2)$ is Gaussian, the spending function is given by $w_\alpha(\beta) = g_\alpha^{-1}\left(\frac{2\beta}{\sigma\tau^2}\right)$ where $g_\alpha : (0,1) \to \mathbb{R}$ is the one-to-one function*

$$g_\alpha(\omega) = \Phi^{-1}(\alpha\omega) - \Phi^{-1}(\alpha(1-\omega)). \tag{S31}$$

w_α is strictly increasing and

$$\lim_{\beta\to\infty} w_\alpha(\beta) = 1.$$

Proposition S1.3 (FAB with a prior with power-law tails (Cortinovis & Caron, 2024))**.** *Let $\pi_0(\beta; \sigma)$ be a symmetric prior such that the marginal density $\pi(y)$ has power-law tails:*

$$\pi(y) \sim \frac{C_\sigma}{|y|^{2\beta+1}} \text{ as } |y| \to \infty$$

for some constant C_σ and some $\beta > 0$. Then,

$$\lim_{\beta\to\infty} w_\alpha(\beta) = \lim_{\beta\to-\infty} w_\alpha(\beta) = \frac{1}{2}.$$

The difference between the spending functions of the two priors greatly affects the resulting FAB confidence regions. In particular, while both priors lead to confidence regions that are shorter than the classical one when the observed y is close to zero, their behaviour differs as the disagreement between the prior and the data increases. In particular, the FAB confidence regions under the Gaussian prior become unbounded as $|y|$ grows, while the horseshoe prior leads to confidence regions that eventually revert to the classical confidence interval. This is illustrated in Figure S5.

S2. Derivations

S2.1. FAB-PPI for Mean Estimation

Here, we outline the steps to derive the FAB-PPI mean estimator presented in Equation (24), as well as the corresponding FAB-PPI confidence region.

The convex loss function that corresponds to estimating $\theta^* = \mathbb{E}[Y]$ is the squared loss $\mathcal{L}_\theta(x, y) = \frac{1}{2}(\theta - y)^2$. In this case, the subgradient of $\mathcal{L}_\theta$ with respect to θ is given by $\mathcal{L}'_\theta(x, y) = \theta - y$. As a result of this, the measure of fit m_θ and the

rectifier Δ_θ take the form

$$m_\theta = \mathbb{E}[\mathcal{L}'_\theta(X, f(X))] = \theta - \mathbb{E}[f(X)],$$
$$\Delta_\theta = \mathbb{E}[\mathcal{L}_\theta(X, Y) - \mathcal{L}'_\theta(X, f(X))] = \mathbb{E}[f(X) - Y].$$

In particular, under the squared loss, the rectifier Δ_θ does not depend on θ, and we indicate this by dropping the subscript θ and writing $\Delta := \Delta_\theta$.

In order to apply FAB-PPI to this setting, we follow the steps outlined in Section 4. In particular, we use the sample mean of the unlabelled data (9) as the estimator $\widehat{m}_\theta$ of m_θ,

$$\widehat{m}_\theta = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}'_\theta(X_i, f(X_i)) = \theta - \frac{1}{N}\sum_{i=1}^{N} f(\widetilde{X}_i),$$

and either the sample mean (10) or the control variate estimator (12) as the estimator $\widehat{\Delta}$ of Δ, as in PPI and PPI++, respectively. To avoid repetitions, in this section we write $\widehat{\Delta}$ as the following general control variate estimator with tuning parameter $\lambda \in \mathbb{R}$,

$$\begin{aligned}
\widehat{\Delta} &= \frac{1}{n}\sum_{i=1}^{n} (\mathcal{L}'_\theta(X_i, Y_i) - \mathcal{L}'_\theta(X_i, f(X_i))) - (\lambda - 1)\left(\frac{1}{n}\left[\sum_{i=1}^{n} \mathcal{L}'_\theta(X_i, f(X_i))\right] - \widehat{m}_\theta\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} (\mathcal{L}'_\theta(X_i, Y_i) - \lambda\mathcal{L}'_\theta(X_i, f(X_i))) + (\lambda - 1)\widehat{m}_\theta \\
&= -\overline{Y} + \lambda\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \frac{1}{N}\sum_{j=1}^{N} f(\widetilde{X}_j)\right) + \frac{1}{N}\sum_{j=1}^{N} f(\widetilde{X}_j).
\end{aligned}$$

The sample mean estimator (10) and the control variate estimator (12) under the squared loss are recovered by setting λ to 1 and $\widehat{\lambda}$ (14), respectively. From this, the standard PPI mean estimators (11) and (13) are obtained by solving the equation $\widehat{m}_\theta + \widehat{\Delta} = 0$ for θ.

Instead, we first define the Bayes assisted estimator (20) under the chosen prior $\pi_0(\Delta; \tau_n)$,

$$\widehat{\Delta}^{\texttt{FABPP}} = \widehat{\Delta} + \widehat{\sigma}^2 \ell'\left(\widehat{\Delta}; \widehat{\sigma}, \tau_n\right),$$

where $\widehat{\sigma}^2$ is an estimator of $\mathrm{var}(\widehat{\Delta})$ and $\ell'_\theta(z; \sigma, \tau)$ is the derivative of the log-marginal likelihood of a Gaussian likelihood model with mean Δ and variance σ^2 under the prior $\pi_0(\Delta, \tau)$. Then, the FAB-PPI mean estimator $\widehat{\theta}^{\texttt{FABPP}}$ under π_0 is given by the solution to the equation

$$\widehat{m}_\theta + \widehat{\Delta}^{\texttt{FABPP}} = 0$$

in θ, that takes the form

$$\widehat{\theta}^{\texttt{FABPP}} = \overline{Y} - \lambda\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \frac{1}{N}\sum_{j=1}^{N} f(\widetilde{X}_j)\right) - \widehat{\sigma}^2 \ell'\left(\widehat{\Delta}; \widehat{\sigma}, \tau_n\right),$$

which matches the expression in Equation (24). Furthermore, by recognising that the first two terms in the above expression match (13), we can alternatively write the FAB-PPI mean estimator as

$$\widehat{\theta}^{\texttt{FABPP}} = \widehat{\theta} - \widehat{\sigma}^2 \ell'\left(\widehat{\Delta}; \widehat{\sigma}, \tau_n\right),$$

where $\widehat{\theta}$ is the corresponding standard PPI mean estimator.

Given $\alpha \in (0, 1)$, we construct the FAB-PPI confidence region $\mathcal{C}_\alpha^{\texttt{FABPP}}$ for the mean as described in Section 4.2. In particular, let $\mathcal{T}_{\alpha-\delta}(\widehat{m}_\theta)$ denote a standard $1 - (\alpha - \delta)$ confidence interval for m_θ,

$$\mathcal{T}_{\alpha-\delta}(\widehat{m}_\theta) = (\widehat{m}_\theta \pm \widehat{\sigma}^f z_{1-(\alpha-\delta)/2}) = \left(\theta - \widehat{\theta}^f \pm \widehat{\sigma}^f z_{1-(\alpha-\delta)/2}\right),$$

where $\widehat{\theta}^f := \frac{1}{N}\sum_{i=1}^N f(\widetilde{X}_i)$ for conciseness, and $(\widehat{\sigma}^f)^2$ is an estimator of $\mathrm{var}(\widehat{m}_\theta)$. Then, we apply the FAB framework under the prior $\pi_0(\Delta;\tau_n)$ to obtain a $1-\delta$ confidence region for Δ,

$$\mathcal{R}_\delta^{\texttt{FABPP}}(\widehat{\Delta}) = \text{FAB-CR}(\widehat{\Delta};\pi_0(\cdot\,;\tau_n),\widehat{\sigma},\delta),$$

where, again, $\widehat{\sigma}^2$ is an estimator of $\mathrm{var}(\widehat{\Delta})$. Finally, to avoid making assumptions on the specific form of $\mathcal{R}_\delta^{\texttt{FABPP}}(\widehat{\Delta})$, we use $(\inf\mathcal{R}_\delta^{\texttt{FABPP}}), \sup(\mathcal{R}_\delta^{\texttt{FABPP}})) \supseteq \mathcal{R}_\delta^{\texttt{FABPP}}(\widehat{\Delta})$ in the definition of $\mathcal{C}_\alpha^{\texttt{FABPP}}$ (21) to obtain the FAB-PPI interval

$$\begin{aligned}
\mathcal{C}_\alpha^{\texttt{FABPP}} &= \left\{\theta \mid 0 \in \left(\theta - \widehat{\theta}^f \pm \widehat{\sigma}^f z_{1-(\alpha-\delta)/2}\right) + (\inf(\mathcal{R}_\delta^{\texttt{FABPP}}), \sup(\mathcal{R}_\delta^{\texttt{FABPP}}))\right\} \\
&= \left\{\theta \mid 0 \in \left(\theta - \widehat{\theta}^f - \widehat{\sigma}^f z_{1-(\alpha-\delta)/2} + \inf(\mathcal{R}_\delta^{\texttt{FABPP}}), \theta - \widehat{\theta}^f + \widehat{\sigma}^f z_{1-(\alpha-\delta)/2} + \sup(\mathcal{R}_\delta^{\texttt{FABPP}})\right)\right\} \\
&= \left(\widehat{\theta}^f - \widehat{\sigma}^f z_{1-(\alpha-\delta)/2} - \sup(\mathcal{R}_\delta^{\texttt{FABPP}}), \widehat{\theta}^f + \widehat{\sigma}^f z_{1-(\alpha-\delta)/2} - \inf(\mathcal{R}_\delta^{\texttt{FABPP}})\right).
\end{aligned}$$

Algorithm 2 summarises the FAB-PPI approach under the squared loss, where $\widehat{\xi}$ is defined for notational convenience and the corresponding sample variances are used as $(\widehat{\sigma}^f)^2$ and $\widehat{\sigma}^2$.

Algorithm 2 FAB-PPI for mean estimation

Input: labelled $\{(X_i,Y_i)\}_{i=1}^n$, unlabelled $\{\widetilde{X}_j\}_{j=1}^N$, predictor f, prior $\pi_0(\cdot\,;\tau_n)$, error levels α, δ

Set $\widehat{\lambda}=1$ (FAB-PPI) or estimate $\widehat{\lambda}$ from data (FAB-PPI++) using Equation (14)
$\widehat{\theta}^f \leftarrow \frac{1}{N}\sum_{j=1}^N f(\widetilde{X}_j)$
$\widehat{\xi} \leftarrow \frac{1}{n}\sum_{i=1}^n(\widehat{\lambda}f(X_i)-Y_i)$
$\widehat{\Delta} \leftarrow \widehat{\xi} - (\widehat{\lambda}-1)\widehat{\theta}^f$
$(\widehat{\sigma}^f)^2 \leftarrow \frac{1}{N-1}\sum_{j=1}^N(f(\widetilde{X}_j)-\widehat{\theta}^f)^2$
$\widehat{\sigma}_\xi^2 = \frac{1}{n-1}\sum_{i=1}^n(\widehat{\lambda}f(X_i)-Y_i-\widehat{\xi})^2$
$\widehat{\sigma}^2 \leftarrow \frac{1}{n}\widehat{\sigma}_\xi^2 + \frac{(\widehat{\lambda}-1)^2}{N}(\widehat{\sigma}^f)^2$
$\mathcal{R}_\delta^{\texttt{FABPP}} \leftarrow \text{FAB-CR}(\widehat{\Delta};\pi_0(\cdot\,;\tau_n),\widehat{\sigma},\delta)$

Outputs: estimator $\widehat{\theta}^{\texttt{FABPP}} = \widehat{\theta}^f - \widehat{\Delta} - \widehat{\sigma}^2\ell'(\widehat{\Delta};\widehat{\sigma},\tau_n)$ and CR $\mathcal{C}_\alpha^{\texttt{FABPP}} = (\widehat{\theta}^f - \sigma^f z_{1-(\alpha-\delta)/2} - \sup(\mathcal{R}_\delta^{\texttt{FABPP}}), \widehat{\theta}^f + \sigma^f z_{1-(\alpha-\delta)/2} - \inf(\mathcal{R}_\delta^{\texttt{FABPP}}))$

S3. Proofs

S3.1. Theorem 4.1 - Asymptotic Coverage of FAB-PPI under the Gaussian and Horseshoe Priors

Under the prior $\pi_0(\cdot;\widehat{\sigma}_\theta)$, the FAB confidence region for the rectifier Δ_θ is

$$\mathcal{R}_\delta^{\texttt{FABPP}}(\widehat{\Delta}_\theta;\widehat{\sigma}_\theta) = \{\Delta_\theta \mid \Delta_\theta - \widehat{\sigma}_\theta z_{1-\delta w_\delta(\Delta_\theta;\widehat{\sigma}_\theta)} < \widehat{\Delta}_\theta < \Delta_\theta + \widehat{\sigma}_\theta z_{1-\delta(1-w_\delta(\Delta_\theta;\widehat{\sigma}_\theta))}\}, \tag{S32}$$

where $w_\delta(\cdot;\widehat{\sigma}_\theta)$ is the FAB spending function. The proof of asymptotic coverage is organised as follows. First, we show that $\mathcal{R}_\delta^{\texttt{FABPP}}(\widehat{\Delta}_\theta;\widehat{\sigma}_\theta)$ is asymptotically a $1-\delta$ confidence interval. This is established via Lemma S3.1 for the Gaussian prior, and via Lemmas S3.2 and S3.3 for the horseshoe prior. This result is then combined with the asymptotic coverage of the standard sample mean estimator for m_θ to conclude the asymptotic coverage of the FAB-PPI estimator of $\theta^\star$.

We first prove the following lemma for the Gaussian prior, demonstrating that the rectifier has the correct asymptotic coverage.

Lemma S3.1. *Let $\widehat{\Delta}_\theta$ be a consistent estimator of Δ_θ such that a CLT holds for $\widehat{\Delta}_\theta$, i.e.*

$$\frac{\widehat{\Delta}_\theta - \Delta_\theta}{\widehat{\sigma}_\theta} \to \mathcal{N}(0,1)$$

as $\min(n,N)\to\infty$, where $\frac{\widehat{\sigma}_\theta^2}{\mathrm{var}(\widehat{\Delta}_\theta)} \to 1$ almost surely. Let $\pi_0(\cdot;\widehat{\sigma}_\theta)$ be the Gaussian prior (22) for Δ_θ and consider the corresponding $1-\delta$ FAB confidence region

$$\mathcal{R}_\delta^{FABPP}(\widehat{\Delta}_\theta;\widehat{\sigma}_\theta) = \textit{FAB-CR}\left(\widehat{\Delta}_\theta;\pi_0\left(\cdot\,;\widehat{\sigma}_\theta\right),\widehat{\sigma}_\theta,\delta\right).$$

Then

$$\lim_{\min(n,N)\to\infty} \Pr(\Delta_\theta \in \mathcal{R}_\delta^{FABPP}(\widehat{\Delta}_\theta; \widehat{\sigma}_\theta) \mid \Delta_\theta) \geq 1-\delta. \tag{S33}$$

Proof. Using Proposition S1.2, for any $x > 0$, $\sigma > 0$,

$$\begin{aligned}\frac{2x}{\sigma} &= g_\delta(g_\delta^{-1}(2x/\sigma)) \\ &= g_\delta(w_\delta(x;\sigma)) \\ &= z_{1-\delta(1-w_\delta(x;\sigma))} - z_{1-\delta w_\delta(x;\sigma)}\end{aligned}$$

and

$$\begin{aligned}\sigma z_{1-\delta(1-w_\delta(x;\sigma))} &= \sigma z_{1-\delta(1-w_\delta(x/\sigma;1))} \\ &= 2x + \sigma z_{1-\delta w_\delta(x/\sigma;1)}.\end{aligned}$$

The FAB confidence region (S32) can therefore be written as

$$\begin{aligned}\mathcal{R}_\delta^{\text{FABPP}}(\widehat{\Delta}_\theta; \widehat{\sigma}_\theta) = \{\Delta_\theta > 0 \mid \Delta_\theta - \widehat{\sigma}_\theta z_{1-\delta w_\delta(\Delta_\theta;\widehat{\sigma}_\theta)} < \widehat{\Delta}_\theta < 3\Delta_\theta + \widehat{\sigma}_\theta z_{1-\delta w_\delta(\Delta_\theta;\widehat{\sigma}_\theta)}\} \\ \cup \{\Delta_\theta < 0 \mid 3\Delta_\theta - \widehat{\sigma}_\theta z_{1-\delta(1-w_\delta(\Delta_\theta;\widehat{\sigma}_\theta))} < \widehat{\Delta}_\theta < \Delta_\theta + \widehat{\sigma}_\theta z_{1-\delta(1-w_\delta(\Delta_\theta;\widehat{\sigma}_\theta))}\}.\end{aligned}$$

Additionally, for any $x > 0$,

$$\begin{aligned}z_{1-\delta w_\delta(x;\widehat{\sigma}_\theta)} &\to z_{1-\delta} \\ z_{1-\delta(1-w_\delta(-x;\widehat{\sigma}_\theta))} &\to z_{1-\delta}\end{aligned}$$

almost surely as $\min(n, N) \to \infty$.

It follows from Equation (S32) that, for any $z_{1-\delta} > \epsilon > 0$, there exist N_0 such that for all n, N with $\min(n, N) \geq N_0$, the FAB confidence region $\mathcal{R}_\delta^{\text{FABPP}}(\widehat{\Delta}_\theta; \widehat{\sigma}_\theta)$ contains the set

$$\begin{aligned}\mathcal{S}_\delta(\widehat{\Delta}_\theta; \widehat{\sigma}_\theta) = \left\{\Delta_\theta > 0 \mid \Delta_\theta - \widehat{\sigma}_\theta(z_{1-\delta} - \epsilon) < \widehat{\Delta}_\theta < 3\Delta_\theta + \widehat{\sigma}_\theta(z_{1-\delta} - \epsilon)\right\} \\ \cup \left\{\Delta_\theta < 0 \mid 3\Delta_\theta - \widehat{\sigma}_\theta(z_{1-\delta} - \epsilon) < \widehat{\Delta}_\theta < \Delta_\theta - \widehat{\sigma}_\theta(z_{1-\delta} - \epsilon)\right\}.\end{aligned}$$

For any fixed $\Delta_\theta > 0$,

$$\Pr(\Delta_\theta \in \mathcal{S}_\delta(\widehat{\Delta}_\theta; \widehat{\sigma}_\theta)) = \Pr\left(-(z_{1-\delta} - \epsilon) < \frac{\widehat{\Delta}_\theta - \Delta_\theta}{\widehat{\sigma}_\theta} < \frac{2\Delta_\theta}{\widehat{\sigma}_\theta} + z_{1-\delta} - \epsilon\right). \tag{S34}$$

Noting that $\frac{2\Delta_\theta}{\widehat{\sigma}_\theta} + z_{1-\delta} - \epsilon \to \infty$ a.s. as $\min(n, N) \to \infty$, we obtain that

$$\lim_{\min(n,N)\to\infty} \Pr(\Delta_\theta \in \mathcal{R}_\delta^{\text{FABPP}}(\widehat{\Delta}_\theta; \widehat{\sigma}_\theta) \mid \Delta_\theta) \geq \lim_{\min(n,N)\to\infty} \Pr(\Delta_\theta \in \mathcal{S}_\delta(\widehat{\Delta}_\theta; \widehat{\sigma}_\theta) \mid \Delta_\theta) \geq 1-\delta. \tag{S35}$$

□

Lemma S3.2. *For $\delta \in (0, 1)$, let $w_\delta(\cdot; \widehat{\sigma}_\theta)$ be the FAB spending function associated to the FAB confidence region*

$$\mathcal{R}_\delta^{FABPP}(\bullet) = \textit{FAB-CR}(\bullet; \pi_0(\cdot\,; \widehat{\sigma}_\theta), \widehat{\sigma}_\theta, \delta).$$

Assume $\widehat{\sigma}_\theta \to 0$ almost surely as $\min(n, N) \to \infty$. If π_0 is the horseshoe prior (23), then, for any $z \in \mathbb{R}$,

$$w_\delta(z; \widehat{\sigma}_\theta) \to \frac{1}{2} \textit{ a.s. as } \min(n, N) \to \infty.$$

Proof. As described in Appendix S1.2, the spending function w_δ, which is continuous, satisfies, for any $z \in \mathbb{R}$ and $\sigma > 0$,

$$w_\delta\left(z;\sigma\right) = w_\delta\left(\frac{z}{\sigma};1\right).$$

By Proposition S1.3, we have

$$\lim_{z\to\infty} w_\delta(z;1) = \lim_{z\to-\infty} w_\delta(z;1) = w_\delta(0;1) = \frac{1}{2}.$$

Since $\widehat{\sigma}_\theta \to 0$ almost surely and w_δ is continuous, for any $z \in \mathbb{R}$, we conclude that

$$w_\delta\left(z;\widehat{\sigma}_\theta\right) \to \frac{1}{2} \text{ a.s. as } \min(n,N) \to \infty.$$

□

Lemma S3.3. *Let $\widehat{\Delta}_\theta$ be a consistent estimator of Δ_θ such that a CLT holds for $\widehat{\Delta}_\theta$, i.e.*

$$\frac{\widehat{\Delta}_\theta - \Delta_\theta}{\widehat{\sigma}_\theta} \to \mathcal{N}(0,1)$$

as $\min(n,N) \to \infty$, where $\frac{\widehat{\sigma}_\theta^2}{\mathrm{var}(\widehat{\Delta}_\theta)} \to 1$ almost surely. Specify a symmetric prior $\pi_0(\cdot;\widehat{\sigma}_\theta)$ for Δ_θ and consider the corresponding $1-\delta$ FAB confidence region

$$\mathcal{R}_\delta^{FABPP}(\widehat{\Delta}_\theta;\widehat{\sigma}_\theta) = \textit{FAB-CR}\left(\widehat{\Delta}_\theta;\pi_0\left(\cdot\,;\widehat{\sigma}_\theta\right),\widehat{\sigma}_\theta,\delta\right),$$

with associated weight function $w_\delta(\Delta_\theta;\widehat{\sigma}_\theta)$. If, for any $x \in \mathbb{R}$,

$$w_\delta\left(x;\widehat{\sigma}_\theta\right) \to \frac{1}{2} \textit{ a.s. as } \min(n,N) \to \infty, \tag{S36}$$

then the confidence region $\mathcal{R}_\delta^{FABPP}(\widehat{\Delta}_\theta;\widehat{\sigma}_\theta)$ reverts to the classical $1-\delta$ z-interval for Δ_θ, i.e., almost surely,

$$\lim_{\min(n,N)\to\infty} \frac{\sup(\mathcal{R}_\delta^{FABPP}(\widehat{\Delta}_\theta;\widehat{\sigma}_\theta)) - \widehat{\Delta}_\theta}{\widehat{\sigma}_\theta} = \lim_{\min(n,N)\to\infty} \frac{\widehat{\Delta}_\theta - \inf(\mathcal{R}_\delta^{FABPP}(\widehat{\Delta}_\theta;\widehat{\sigma}_\theta))}{\widehat{\sigma}_\theta} = z_{1-\delta/2},$$

with

$$\lim_{\min(n,N)\to\infty} \Pr(\Delta_\theta \in \mathcal{R}_\delta^{FABPP}(\widehat{\Delta}_\theta;\widehat{\sigma}_\theta) \mid \Delta_\theta) = 1-\delta. \tag{S37}$$

Proof. As Φ^{-1} is continuous, Equation (S36) implies that, for any $x \in \mathbb{R}$,

$$\lim_{\min(n,N)\to\infty} z_{1-\delta w_\delta(x;\widehat{\sigma}_\theta)} = \lim_{\min(n,N)\to\infty} z_{1-\delta(1-w_\delta(x;\widehat{\sigma}_\theta))} = z_{1-\delta/2}$$

almost surely. From Equation (S32), it follows that, for any $\epsilon > 0$ and for $\min(n,N)$ large enough,

$$z_{1-\delta/2} - \epsilon \le \frac{\sup(\mathcal{R}_\delta^{\mathrm{FABPP}}(\widehat{\Delta}_\theta;\widehat{\sigma}_\theta)) - \widehat{\Delta}_\theta}{\widehat{\sigma}_\theta} \le z_{1-\delta/2} + \epsilon$$
$$z_{1-\delta/2} - \epsilon \le \frac{\widehat{\Delta}_\theta - \inf(\mathcal{R}_\delta^{\mathrm{FABPP}}(\widehat{\Delta}_\theta;\widehat{\sigma}_\theta))}{\widehat{\sigma}_\theta} \le z_{1-\delta/2} + \epsilon.$$

(S37) then follows directly, which completes the proof.

□

With the above two lemma, we can now prove Theorem 4.1, which we restate here in extended form.

Theorem S3.4. *Consider a convex estimation problem whose solution can be expressed as in Equation (2). For all $\theta \in \mathbb{R}$, define $\widehat{\Delta}_\theta$ and $\widehat{m}_\theta$ as in Section 4 and let*

$$\mathcal{R}_\delta^{FABPP}(\widehat{\Delta}_\theta; \widehat{\sigma}_\theta) = \textit{FAB-CR}\left(\widehat{\Delta}_\theta; \pi_0\left(\cdot\,; \widehat{\sigma}_\theta\right), \widehat{\sigma}_\theta, \delta\right),$$
$$\mathcal{T}_{\alpha-\delta}(\widehat{m}_\theta; \widehat{\sigma}_\theta^f) = \left(\widehat{m}_\theta - \widehat{\sigma}_\theta^f z_{1-(\alpha-\delta)/2}, \widehat{m}_\theta + \widehat{\sigma}_\theta^f z_{1-(\alpha-\delta)/2}\right),$$

where $\frac{\widehat{\sigma}_\theta^2}{\mathrm{var}(\widehat{\Delta}_\theta)} \to 1$ and $\frac{(\widehat{\sigma}_\theta^f)^2}{\mathrm{var}(\widehat{m}_\theta)} \to 1$ almost surely as $\min(n, N) \to \infty$. Then, the FAB-PPI confidence region $\mathcal{C}_\alpha^{FABPP}$, defined as

$$\mathcal{C}_\alpha^{FABPP} = \left\{\theta \mid 0 \in \mathcal{R}_\delta^{FABPP}(\widehat{\Delta}_\theta; \widehat{\sigma}_\theta) + \mathcal{T}_{\alpha-\delta}(\widehat{m}_\theta; \widehat{\sigma}_\theta^f)\right\},$$

has correct asymptotic coverage, i.e. it satifies

$$\liminf_{\min(n,N)\to\infty} \Pr(\theta^* \in \mathcal{C}_\alpha^{FABPP}) = \liminf_{\min(n,N)\to\infty} \Pr(0 \in \mathcal{R}_\delta^{FABPP}(\widehat{\Delta}_{\theta^*}; \widehat{\sigma}_{\theta^*}) + \mathcal{T}_{\alpha-\delta}(\widehat{m}_{\theta^*}; \widehat{\sigma}_{\theta^*}^f)) \geq 1 - \alpha.$$

Proof. By Lemma S3.1 (Gaussian prior) and Lemmas S3.2 and S3.3 (horseshoe prior), $\mathcal{R}_\delta^{\mathrm{FABPP}}(\widehat{\Delta}_{\theta^\star}; \widehat{\sigma}_{\theta^\star})$ is an asymptotically valid $1-\delta$ confidence region for $\Delta_{\theta^\star}$, that is

$$\liminf_{n,N\to\infty} \Pr(\Delta_{\theta^\star} \in \mathcal{R}_\delta^{\mathrm{FABPP}}(\widehat{\Delta}_{\theta^\star}; \widehat{\sigma}_{\theta^\star})) \geq 1 - \delta.$$

Similarly, the CLT for $\widehat{m}_\theta$ implies that

$$\liminf_{n,N\to\infty} \Pr(m_{\theta^*} \in \mathcal{T}_{\alpha-\delta}(\widehat{m}_{\theta^*}; \widehat{\sigma}_{\theta^*}^f)) \geq 1 - (\alpha - \delta).$$

Consider the event

$$E = \{\Delta_{\theta^*} \in \mathcal{R}_\delta^{\mathrm{FABPP}}(\widehat{\Delta}_{\theta^*}; \widehat{\sigma}_{\theta^*})\} \cap \{m_{\theta^*} \in \mathcal{T}_{\alpha-\delta}(\widehat{m}_{\theta^*}; \widehat{\sigma}_{\theta^*}^f)\}.$$

By Boole's inequality,

$$\begin{aligned}
\liminf_{n,N\to\infty} \Pr(E) &\geq 1 - \limsup_{n,N\to\infty} \Pr(\{\Delta_{\theta^*} \notin \mathcal{R}_\delta^{\mathrm{FABPP}}(\widehat{\Delta}_{\theta^*}; \widehat{\sigma}_{\theta^*})\} \cup \{m_{\theta^*} \notin \mathcal{T}_{\alpha-\delta}(\widehat{m}_{\theta^*}; \widehat{\sigma}_{\theta^*}^f)\}) \\
&\geq 1 - \limsup_{n\to\infty} \Pr(\{\Delta_{\theta^*} \notin \mathcal{R}_\delta^{\mathrm{FABPP}}(\widehat{\Delta}_{\theta^*}; \widehat{\sigma}_{\theta^*})\}) - \limsup_{N\to\infty} \Pr(\{m_{\theta^*} \notin \mathcal{T}_{\alpha-\delta}(\widehat{m}_{\theta^*}; \widehat{\sigma}_{\theta^*}^f)\}) \\
&\geq 1 - \delta - (\alpha - \delta) \\
&= 1 - \alpha.
\end{aligned}$$

Furthermore, on the event E, we have that

$$0 = \Delta_{\theta^*} + m_{\theta^*} \in \mathcal{R}_\delta^{\mathrm{FABPP}}(\widehat{\Delta}_{\theta^*}; \widehat{\sigma}_{\theta^*}) + \mathcal{T}_{\alpha-\delta}(\widehat{m}_{\theta^*}; \widehat{\sigma}_{\theta^*}^f),$$

where the first equality follows from Equation (2). As a result of this,

$$\liminf_{n,N\to\infty} \Pr(0 \in \mathcal{R}_\delta^{\mathrm{FABPP}}(\widehat{\Delta}_{\theta^*}; \widehat{\sigma}_{\theta^*}) + \mathcal{T}_{\alpha-\delta}(\widehat{m}_{\theta^*}; \widehat{\sigma}_{\theta^*}^f)) \geq 1 - \alpha,$$

as desired. □

Remark S3.5. With both the Gaussian and horseshoe priors, we obtain asymptotic coverage. However, the asymptotic confidence regions differ significantly. In the Gaussian case, the volume of the confidence region does not vanish asymptotically. Instead, the confidence region converges to $(\frac{\Delta_\theta}{3}, \Delta_\theta)$, with volume of $\frac{2}{3}|\Delta_\theta|$. In contrast, when using the horseshoe prior (23), we revert to the usual CLT-based confidence intervals, and the volume of the confidence region converges to zero almost surely.

S3.2. Proposition 4.2 - Robustness of FAB-PPI under the Horseshoe Prior

Let π_0 be the horseshoe prior (23), and consider the FAB confidence region $\mathcal{R}_\delta^{\texttt{FABPP}}(\widehat{\Delta}_\theta; \widehat{\sigma}_\theta)$ for Δ_θ, as defined in Equation (S32).

We first state a corollary of Cortinovis & Caron (2024, Theorem 2.2), which follows from the power-law tails of the marginal likelihood under the horseshoe prior (see Appendix S1.1). The corollary states that, if $|\widehat{\Delta}_\theta|$ is very large, then the standard CLT-based confidence interval, $(\widehat{\Delta}_\theta \pm \widehat{\sigma}_\theta z_{1-\delta/2})$, is recovered.

Corollary S3.6. *(Cortinovis & Caron (2024, Theorem 2.2)) For any $\sigma > 0$,*

$$\lim_{|\Delta|\to\infty} \sup(\mathcal{R}_\delta^{FABPP}(\Delta;\sigma)) - \Delta = \lim_{|\Delta|\to\infty} \Delta - \inf(\mathcal{R}_\delta^{FABPP}(\Delta;\sigma)) = \sigma z_{1-\delta/2}.$$

Define

$$\mathcal{S}_{\alpha,\delta}(\widehat{\Delta}_\theta, \widehat{\sigma}_\theta, \widehat{m}_\theta, \widehat{\sigma}_\theta^f) = \mathcal{R}_\delta^{\texttt{FABPP}}(\widehat{\Delta}_\theta; \widehat{\sigma}_\theta) + \mathcal{T}_{\alpha-\delta}(\widehat{m}_\theta; \widehat{\sigma}_\theta^f), \tag{S38}$$

where

$$\mathcal{T}_{\alpha-\delta}(\widehat{m}_\theta; \widehat{\sigma}_\theta^f) = \left(\widehat{m}_\theta - \widehat{\sigma}_\theta^f z_{1-(\alpha-\delta)/2}, \widehat{m}_\theta + \widehat{\sigma}_\theta^f z_{1-(\alpha-\delta)/2}\right)$$

is the standard CLT-based confidence interval for m_θ. From Corollary S3.6, for any fixed $\sigma > 0$, $m \in \mathbb{R}$, $\sigma^f > 0$,

$$\lim_{|\Delta|\to\infty} \sup(\mathcal{S}_{\alpha,\delta}(\Delta,\sigma,m,\sigma^f)) - (\Delta+m) = \lim_{|\Delta|\to\infty} (\Delta+m) - \inf(\mathcal{S}_{\alpha,\delta}(\Delta,\sigma,m,\sigma^f)) = \sigma z_{1-\delta/2} + \sigma^f z_{1-(\alpha-\delta)/2}.$$

Therefore, if $|\widehat{\Delta}_\theta| \gg 0$, the confidence region $\mathcal{S}_{\alpha,\delta}(\widehat{\Delta}_\theta, \widehat{\sigma}_\theta, \widehat{m}_\theta, \widehat{\sigma}_\theta^f)$ reverts to the standard interval

$$(\widehat{\Delta}_\theta + \widehat{m}_\theta \pm (\widehat{\sigma}_\theta z_{1-\delta/2} + \widehat{\sigma}_\theta^f z_{1-(\alpha-\delta)/2})).$$

It follows that, if $\inf_{\theta'\in\mathbb{R}} |\widehat{\Delta}_{\theta'}| \gg 0$ the confidence region for $\theta^\star$, defined as

$$\mathcal{C}_\alpha^{\texttt{FABPP}} = \left\{\theta \mid 0 \in \mathcal{R}_\delta^{\texttt{FABPP}}(\widehat{\Delta}_\theta; \widehat{\sigma}_\theta) + \mathcal{T}_{\alpha-\delta}(\widehat{m}_\theta; \widehat{\sigma}_\theta^f)\right\},$$

reverts to the standard, CLT-based PPI confidence region

$$\mathcal{C}_\alpha^{\texttt{PP}} = \left\{\theta \in \mathbb{R} \mid -\widehat{\sigma}_\theta z_{1-\delta/2} - \widehat{\sigma}_\theta^f z_{1-(\alpha-\delta)/2} \le \widehat{\Delta}_\theta + \widehat{m}_\theta \le \widehat{\sigma}_\theta z_{1-\delta/2} + \widehat{\sigma}_\theta^f z_{1-(\alpha-\delta)/2}\right\}.$$

S3.3. Proposition 4.3 - Consistency of FAB-PPI Mean Estimators

Below, we use `BPP` and `BPP+` to distinguish between the estimators $\widehat{\theta}^{\texttt{FABPP}}$, $\widehat{\theta}$, $\widehat{\Delta}$ and $\widehat{\sigma}$ in the two cases of FAB-PPI and FAB-PPI++. Then, the FAB-PPI and FAB-PPI++ mean estimators are given by

$$\widehat{\theta}^{\texttt{BPP}} = \widehat{\theta}^{\texttt{PP}} - (\widehat{\sigma}^{\texttt{PP}})^2 \ell'\left(\widehat{\Delta}^{\texttt{PP}}; \widehat{\sigma}^{\texttt{PP}}, \widehat{\sigma}^{\texttt{PP}}\right),$$
$$\widehat{\theta}^{\texttt{BPP+}} = \widehat{\theta}^{\texttt{PP+}} - (\widehat{\sigma}^{\texttt{PP+}})^2 \ell'\left(\widehat{\Delta}^{\texttt{PP+}}; \widehat{\sigma}^{\texttt{PP+}}, \widehat{\sigma}^{\texttt{PP}}\right),$$

where we recall that $\ell(y;\sigma,\tau) = \log \int_{\mathbb{R}} \mathcal{N}(y;\Delta,\sigma^2)\pi_0(\Delta;\tau)d\Delta$ where τ is a scale parameter of the prior π_0. By assumption, both PPI estimators, $\widehat{\theta}^{\texttt{PP}}$ and $\widehat{\theta}^{\texttt{PP+}}$, are strongly consistent estimators of $\theta^\star$. It remains to prove that

$$(\widehat{\sigma}^{\texttt{PP}})^2 \ell'\left(\widehat{\Delta}^{\texttt{PP}}; \widehat{\sigma}^{\texttt{PP}}, \widehat{\sigma}^{\texttt{PP}}\right) \to 0 \tag{S39}$$
$$(\widehat{\sigma}^{\texttt{PP+}})^2 \ell'\left(\widehat{\Delta}^{\texttt{PP+}}; \widehat{\sigma}^{\texttt{PP+}}, \widehat{\sigma}^{\texttt{PP}}\right) \to 0 \tag{S40}$$

almost surely, as $\min(n, N) \to \infty$. For any $\sigma > 0$, we have $\ell'(y;\sigma,\sigma) = \frac{1}{\sigma}\ell'(y/\sigma; 1, 1)$.

Under the horseshoe prior (23), $\ell'_{\text{HS}}(y; 1, 1)$ is bounded. Therefore (S39) and (S40) hold almost surely by sandwiching.

Under the Gaussian prior (22),

$$\ell'_{\mathrm{N}}(y;\sigma,\sigma) = -\frac{y}{2\sigma^2}.$$

Hence, since $\widehat{\Delta}^{\mathtt{PP}} \to \Delta$ and $\widehat{\Delta}^{\mathtt{PP+}} \to \Delta$ almost surely, where we recall that $\Delta = \mathbb{E}[f(X) - Y]$, we obtain

$$(\widehat{\sigma}^{\mathtt{PP}})^2 \ell'_{\mathrm{N}}\left(\widehat{\Delta}^{\mathtt{PP}}; \widehat{\sigma}^{\mathtt{PP}}, \widehat{\sigma}^{\mathtt{PP}}\right) \to -\frac{\Delta}{2}$$
$$(\widehat{\sigma}^{\mathtt{PP+}})^2 \ell'_{\mathrm{N}}\left(\widehat{\Delta}^{\mathtt{PP+}}; \widehat{\sigma}^{\mathtt{PP+}}, \widehat{\sigma}^{\mathtt{PP}}\right) \to -\frac{\Delta}{2}$$

almost surely as $\min(n, N) \to \infty$, which implies that the FAB-PPI mean estimators under the Gaussian prior (22) are not consistent.

S4. Multivariate FAB-PPI

Here we extend FAB-PPI to the multivariate case, where $\theta, m_\theta, \Delta_\theta \in \mathbb{R}^d$. While most of the methodology remains the same as in the univariate case, we now need to specify a multivariate prior for Δ_θ, for which we consider independent horseshoe priors on each dimension.

S4.1. Multivariate Bayesian PPI Estimators

As in the univariate case, we use the sample mean $\widehat{m}_\theta$ as the estimator of m_θ. Similarly, we consider some consistent estimator $\widehat{\Delta}_\theta$ of Δ_θ, such as the sample mean (10), as in PPI, or the control variate estimator (12), as in PPI++. Crucially, we assume that a multivariate CLT holds for this estimator, that is

$$\widehat{\Sigma}_\theta^{-1/2}\left(\widehat{\Delta}_\theta - \Delta_\theta\right) \to \mathcal{N}(0, \mathrm{I})$$

as $\min(n, N) \to \infty$, where $\widehat{\Sigma}_\theta$ is an estimator of $\mathrm{cov}(\widehat{\Delta}_\theta)$. Again, this holds for both the PPI and PPI++ estimators (Angelopoulos et al., 2023b;a). We consider d independent priors, $\pi_0(\Delta_{\theta,k}; \widehat{\sigma}_{\theta,k})$ for $k = 1, \ldots, d$, on the components of Δ_θ, where $\widehat{\sigma}^2_{\theta,k}$ is the k-th diagonal element of $\widehat{\Sigma}_\theta$. The multivariate FAB-PPI estimator $\widehat{\Delta}_\theta^{\mathtt{FABPP}}$ is formed by stacking the individual estimators

$$\widehat{\Delta}_{\theta,k}^{\mathtt{FABPP}} = \widehat{\Delta}_{\theta,k} + \widehat{\sigma}^2_{\theta,k}\ell'\left(\widehat{\Delta}_{\theta,k}; \widehat{\sigma}_{\theta,k}\right)$$

for each dimension $k = 1, \ldots, d$. Importantly, note that the k-th dimension of $\widehat{\Delta}_\theta^{\mathtt{FABPP}}$ only depends on the k-th dimension of the observed $(\mathcal{L}'_\theta(X_i, Y_i) - \mathcal{L}'_\theta(X_i, f(X_i)))$ that are used to estimate $\Delta_{\theta,k}$. The FAB-PPI estimator of $\theta^\star$ then becomes the solution, in θ, to the equation

$$\widehat{m}_\theta + \widehat{\Delta}_\theta^{\mathtt{FABPP}} = \mathbf{0} \in \mathbb{R}^d.$$

S4.2. Multivariate FAB-PPI Confidence Regions

As in the univariate case, let $\mathcal{T}_{\alpha-\delta}(\widehat{m}_\theta)$ denote a standard $1-(\alpha-\delta)$ confidence interval for m_θ. For Δ_θ, we apply the FAB framework with independent horseshoe priors to each dimension $\Delta_{\theta,k}$ and use a union bound to obtain a $1-\delta$ confidence region for Δ_θ. In particular, let $\mathcal{R}^{\mathtt{FABPP}}_{\delta/d}(\widehat{\Delta}_{\theta,k}, \widehat{\sigma}_{\theta,k}) = \text{FAB-CR}(\widehat{\Delta}_{\theta,k}; \pi_0(\cdot\,; \widehat{\sigma}_{\theta,k}), \widehat{\sigma}_{\theta,k}, \delta/d)$ be a $1-\delta/d$ FAB confidence region for $\Delta_{\theta,k}$ under the horseshoe prior $\pi_0(\cdot\,; \widehat{\sigma}_{\theta,k})$. Then,

$$\mathcal{R}^{\mathtt{FABPP}}_{\delta}(\widehat{\Delta}_\theta, \widehat{\sigma}_\theta) = \left\{\Delta_\theta \mid \Delta_{\theta,k} \in \mathcal{R}^{\mathtt{FABPP}}_{\delta/d}(\widehat{\Delta}_{\theta,k}, \widehat{\sigma}_{\theta,k}),\ k = 1, \ldots, d\right\}$$

where $\widehat{\Delta}_\theta = (\widehat{\Delta}_{\theta,1}, \ldots, \widehat{\Delta}_{\theta,d})$, $\widehat{\sigma}_\theta = (\widehat{\sigma}_{\theta,1}, \ldots, \widehat{\sigma}_{\theta,d})$, is a $1-\delta$ multivariate FAB confidence region for Δ_θ by a union bound. With this, the multivariate FAB-PPI confidence region $\mathcal{C}^{\mathtt{FABPP}}_\alpha$ is given by

$$\mathcal{C}^{\mathtt{FABPP}}_\alpha = \left\{\theta \mid \mathbf{0} \in \mathcal{R}^{\mathtt{FABPP}}_{\delta}(\widehat{\Delta}_\theta, \widehat{\sigma}_\theta) + \mathcal{T}_{\alpha-\delta}(\widehat{m}_\theta, \widehat{\sigma}^f_\theta)\right\},$$

exactly as in the univariate case. Moreover, also multivariate FAB-PPI enjoys exact asymptotic coverage as $\min(n, N) \to \infty$. In particular, Theorem 4.1 can be easily extended to the multivariate case by applying a union bound over the dimensions of Δ_θ.

S5. Experimental Details

S5.1. Datasets

Here we provide a brief description of each dataset used for the real data experiments in Section 5.2. For additional details, the reader may refer to Angelopoulos et al. (2023b). All of the datasets were downloaded from the examples provided as part of the `ppi-py` package (Angelopoulos et al., 2023a).

Alphafold. The ALPHAFOLD dataset contains the following features for $N = 10802$ protein residues analysed by Bludau et al. (2022): whether the residue is phosphorylated ($Z_i \in \{0, 1\}$), whether the residue is part of an intrinsically disordered region (IDR, $Y_i \in \{0, 1\}$), and the prediction of the AlphaFold model (Jumper et al., 2021) for the probability of Y_i being equal to one ($f(X_i) \in [0, 1]$). The goal is to estimate the odds ratio of a protein being phosphorylated and being part of an IDR region, i.e.

$$\theta^\star = \frac{\mu_1/(1-\mu_1)}{\mu_0/(1-\mu_0)},$$

where $\mu_1 = \Pr(Y = 1 \mid Z = 1)$ and $\mu_0 = \Pr(Y = 1 \mid Z = 0)$. Following Angelopoulos et al. (2023b), given $\alpha \in (0, 1)$, we construct $1 - \alpha/2$ confidence intervals $\mathcal{C}_0 = [l_0, u_0]$ and $\mathcal{C}_1 = [l_1, u_1]$ for μ_0 and μ_1, respectively. Then, by a union bound, the interval

$$\mathcal{C} = \left\{ \frac{c_1}{1-c_1} \cdot \frac{1-c_0}{c_0} : c_0 \in \mathcal{C}_0, c_1 \in \mathcal{C}_1 \right\} = \left(\frac{l_1}{1-l_1} \cdot \frac{1-u_0}{u_0}, \frac{u_1}{1-u_1} \cdot \frac{1-l_0}{l_0} \right)$$

has coverage at least $1 - \alpha$. Note that the union bound above may result in a conservative confidence interval, leading to coverage significantly larger than $1 - \alpha$ in practice, as in the left panel of Figure 3.

Forest. The FOREST dataset contains the following features for $N = 1596$ parcels of land in the Amazon rainforest examined during field visits (Bullock et al., 2020): whether the parcel has been subject to deforestation ($Y_i \in \{0, 1\}$) and the prediction of a gradient-boosted tree model for the probability of Y_i being equal to one ($f(X_i) \in [0, 1]$). The goal is to estimate the fraction of Amazon rainforest lost to deforestation, i.e. $\theta^* = \mathbb{E}[Y]$.

Galaxies. The GALAXIES dataset contains the following features for $N = 16743$ images from the Galaxy Zoo 2 initiative (Willett et al., 2013): whether the galaxy has spiral arms ($Y_i \in \{0, 1\}$) and the prediction of a ResNet50 model (He et al., 2016) for the probability of Y_i being equal to one ($f(X_i) \in [0, 1]$). The goal is to estimate the fraction of galaxies with spiral arms, i.e. $\theta^* = \mathbb{E}[Y]$.

Genes. The GENES dataset contains the following features for $N = 61150$ gene promoter sequences: the expression level of the gene induced by the promoter and the prediction of a transformer model for the same quantity (Vaishnav et al., 2022). The goal is to estimate the median expression level across genes.

Census. The CENSUS dataset contains the following features for $N = 380091$ individuals from the 2019 California census: the individual's age, sex, and yearly income, as well as the prediction of a gradient-boosted tree model trained on the previous year's raw data for the individual's income. The goal is to estimate the ordinary least squares (OLS) regression coefficients when regressing income on age and sex.

Healthcare. The HEALTHCARE dataset contains the following features for $N = 318215$ individuals from the 2019 California census: the individual's yearly income and whether they have health insurance ($Y_i \in \{0, 1\}$), as well as the prediction of a gradient-boosted tree model trained on the previous year's raw data for the probability of Y_i being equal to one ($f(X_i) \in [0, 1]$). The goal is to estimate the logistic regression coefficient when regressing health insurance status on income.

S5.2. Implementation

Code implementing the FAB-PPI method is written in Python and made available at `https://github.com/stefanocortinovis/fab-ppi`. Comparisons with standard PPI are performed using the `ppi-py` package (Angelopoulos et al., 2023a). All of the experiments presented here were run locally on an Intel Core i7-11850H CPU.

S6. Additional Results

S6.1. Experiments with Synthetic Data

The complete results for the experiments discussed in Section 5 are presented here. The legend names for the figures are as in Section 5.

S6.1.1. Biased Predictions Simulation Study

Figure S6 shows the average MSE, CI volume, and CI coverage as a function of the bias level γ for the biased predictions study in Section 5.1. Compared to Figure 1, we include results for the non power-tuned methods, as well as for the ones that

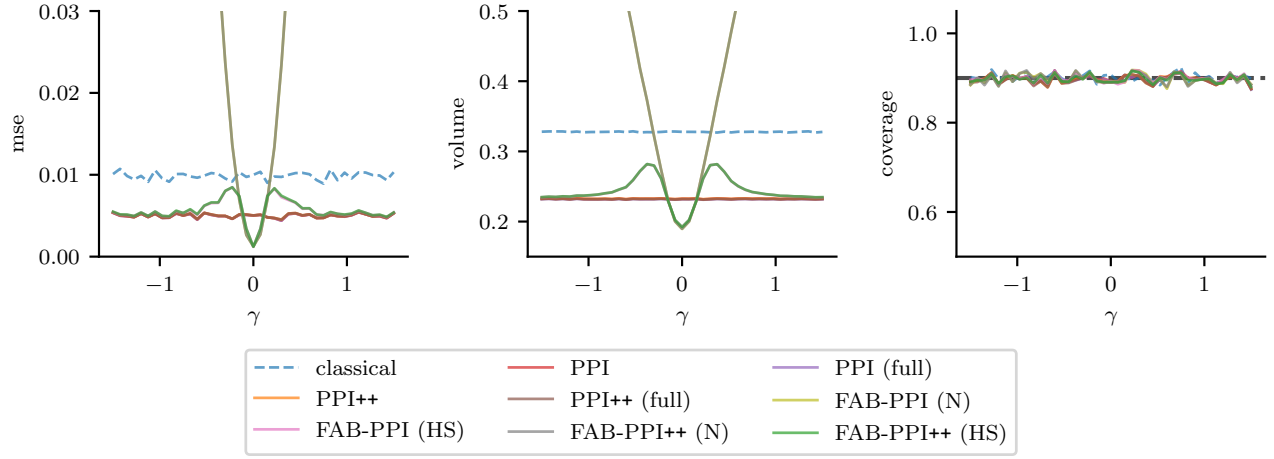

Figure S6. Full results for the biased predictions study. The left, middle, and right panels show the average MSE, CI volume, and CI coverage as the bias level γ varies.

take into account the uncertainty on the measure of fit m_θ (i.e. PPI (full) and PPI++ (full)). In this example, power-tuning does not play a significant role and the same conclusions as in Section 5.1 hold. In particular, standard PPI induces shorter CIs than classical inference with constant volume across bias levels. On the other hand, FAB methods induce shorter CIs when the predictions are good. As the prediction bias increases, the volume of the FAB CIs with Gaussian prior grows unbounded, while the horseshoe prior eventually reverts to the PPI intervals. Furthermore, the coverage plot shows that the methods tested achieve similar coverage to the nominal level and to PPI (full) and PPI++ (full).

S6.1.2. Noisy Predictions Simulation Study

Figure S7 shows the average MSE, CI volume, and CI coverage as a function of n for the values of σ_Y considered in the noisy predictions study of Section 5.1. Compared to Figure 2, we include results for the methods that use the Gaussian prior (FAB-PPI (N) and FAB-PPI++ (N)) and those that take into account the uncertainty on the measure of fit m_θ (i.e. PPI (full) and PPI++ (full)). Like the CI volume plots in the main text, the MSE plots clearly show the benefits of both power-tuning and adaptive shrinkage through the horseshoe prior: as σ_Y increases, the power-tuned methods clearly outperform the standard alternatives, while shrinkage always helps compared to standard PPI because the predictions remain unbiased. In this case, the Gaussian prior performs similarly to the horseshoe as the prediction rule f is unbiased. The coverage plots confirm that all methods achieve comparable coverage across noise levels.

S6.2. Experiments with Real Data

S6.2.1. Mean Estimation

Full Comparison. Figure S8 shows the average MSE, CI volume, and CI coverage as a function of n for the three datasets considered in Section 5.2. Compared to Figure 3, we include results for the non power-tuned methods, as well as for the ones that take into account the uncertainty on the measure of fit m_θ (i.e. PPI (full) and PPI++ (full)). The results are consistent

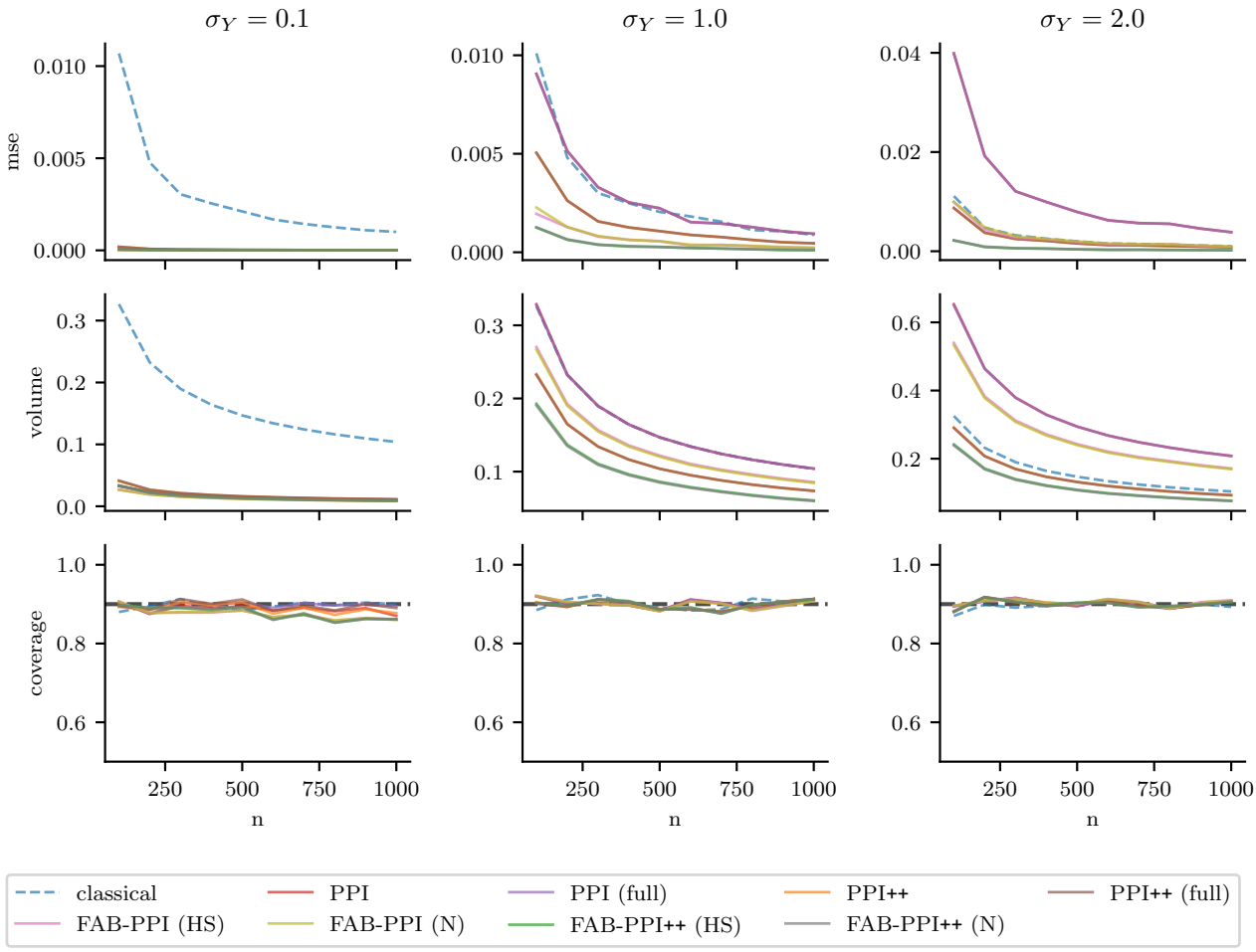

Figure S7. Full results for the noisy predictions study. The left, middle, and right panels correspond to noise levels $\sigma_Y = 0.1, 1, 2$, respectively. The top, middle, and bottom rows show average MSE, CI volume, and CI coverage, respectively.

with those presented in Section 5.2. In particular, FAB methods outperform the standard PPI alternatives and classical inference, while achieving comparable coverage. For the datasets and the values of n considered, power-tuned methods perform similarly to the non-tuned ones. Among the FAB methods, the horseshoe and Gaussian priors achieve similar performance.

Example Intervals. Figure S9 shows 10 randomly chosen intervals for the classical, PPI++, and FAB-PPI++ methods for the three datasets considered in Section 5.2 and different choices of the number of labelled observations n.

Varying the Prior Scale. We repeat the mean estimation experiment on the FOREST dataset while varying the scale of the horseshoe prior used for FAB-PPI++ in Appendix S6.2.1. In addition to the scale $\widehat{\sigma}$ used in the main text, we consider the sample independent scale $1/\sqrt{n}$ and the data independent scale 1. As already mentioned, the computation of the FAB-PPI confidence regions under a horseshoe prior with scale other than $\widehat{\sigma}$ involves numerical integration to compute the corresponding marginal likelihood. Figure S10 shows the average MSE, CI volume, and CI coverage for each of these choices, as well as for classical inference and PPI++. While the scale $\widehat{\sigma}$ achieves the best performance, the other scales also provide shorter CIs than classical inference and PPI++. In particular, the sample independent scale $1/\sqrt{n}$ results in good performance across all metrics without requiring the estimation of $\widehat{\sigma}$.

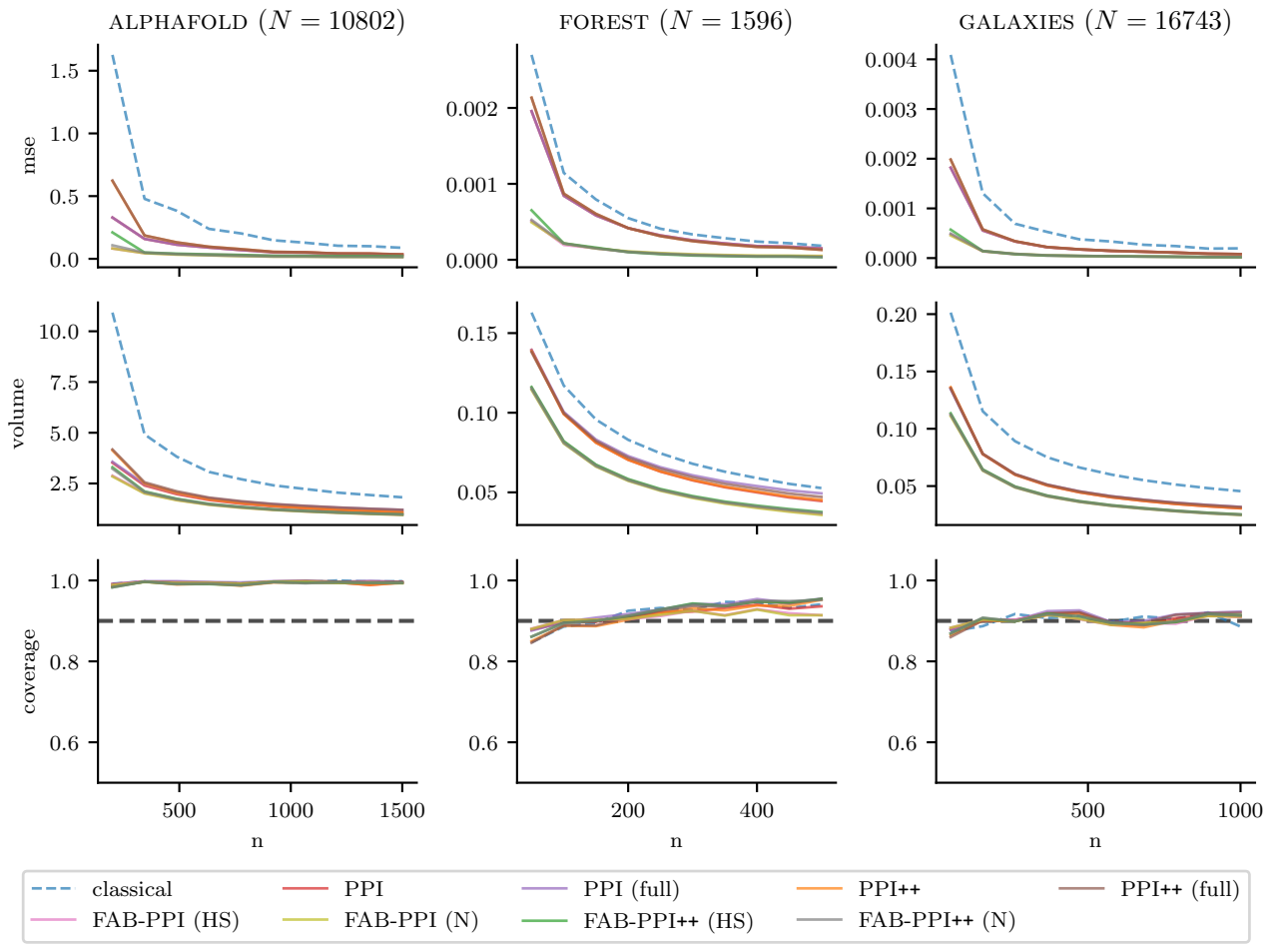

Figure S8. Full results for mean estimation experiment on real data. The left, middle, and right panels correspond to the ALPHAFOLD, GALAXIES, and FOREST datasets, respectively. The top, middle, and bottom rows show average MSE, CI volume, and CI coverage, respectively over 1000 repetitions for $\alpha = 0.1$.

S6.2.2. LOGISTIC REGRESSION

Figure S11 shows the average MSE, CI volume, and CI coverage as a function of n for the logistic regression experiment on the HEALTHCARE dataset mentioned in Section 5.2. As mentioned in the main text, FAB methods outperform the standard PPI alternatives and classical inference, while achieving comparable coverage. Among the FAB methods, the horseshoe and Gaussian priors achieve similar performance.

S6.2.3. QUANTILE ESTIMATION

Figure S12 shows the average MSE, CI volume, and CI coverage as a function of n for the quantile estimation experiment on the GENES dataset mentioned in Section 5.2. The predictions contained in this dataset are highly biased and this is reflected in the performance of the FAB-PPI methods. In particular, the Gaussian prior underperforms both classical inference and standard PPI, while the horseshoe prior achieves similar performance to standard PPI thanks to its robustness against large bias levels.

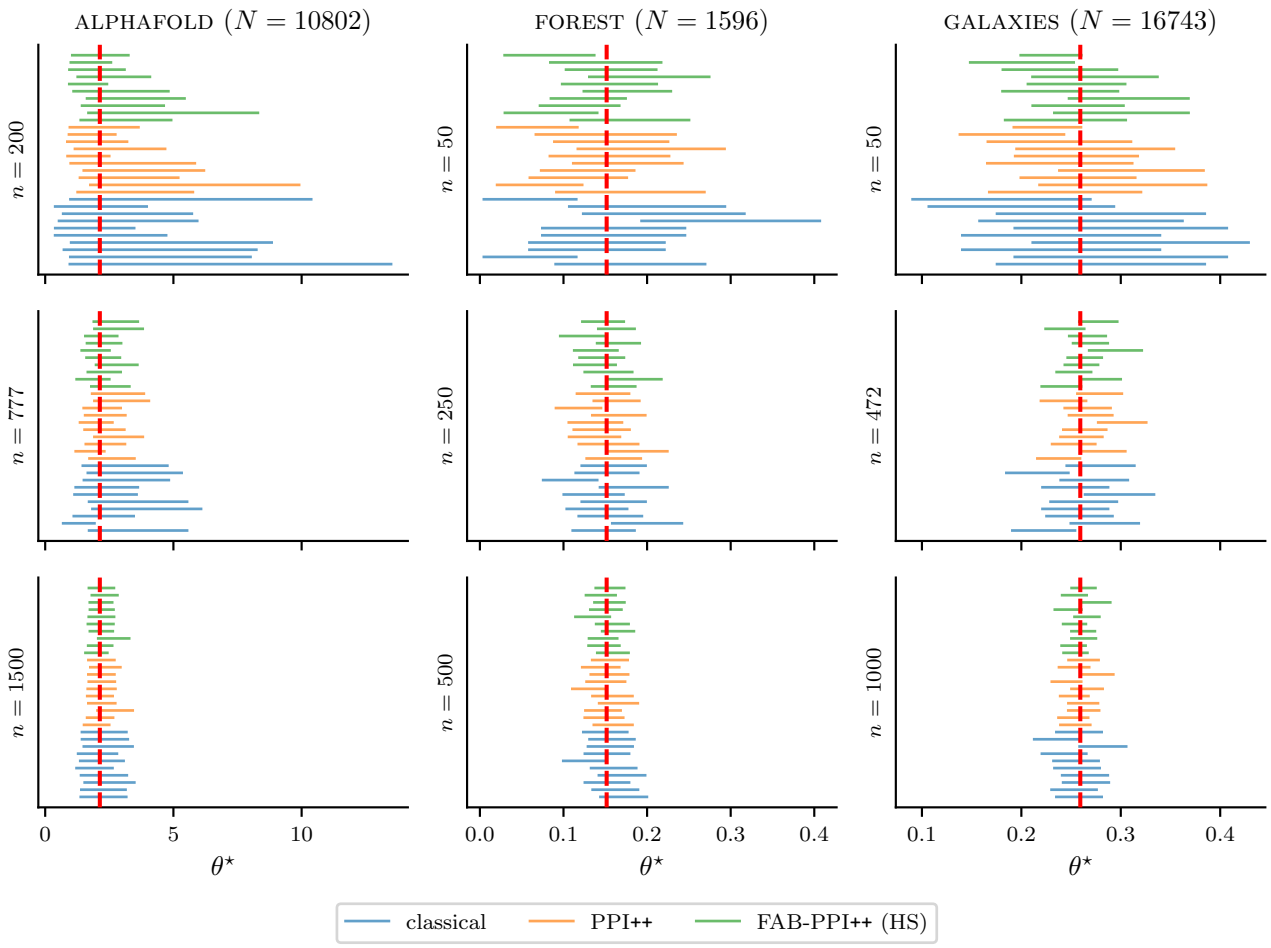

Figure S9. Each subfigure includes 10 randomly chosen intervals for the classical, PPI++ and FAB-PPI++ methods. The left, middle, and right panels refer to the ALPHAFOLD, GALAXIES, and FOREST datasets, respectively. The top, middle, and bottom rows refer correspond to different values of n.

S6.2.4. Linear Regression

Figure S12 shows the average MSE, CI volume, and CI coverage as a function of n for the linear regression experiment on the CENSUS dataset mentioned in Section 5.2. More specifically, panel (a) and (b) corresponds to the OLS parameters associated with the *age* and *sex* covariates, respectively. On the one hand, FAB-PPI seems to perform well for the *sex* covariate, with similar performance between the Gaussian and horseshoe priors, and slightly improved MSE and CI volume compared to classical inference and standard PPI. On the other hand, the performance of FAB-PPI for the *age* covariate seems to be affected by bias in the dataset predictions. In particular, FAB-PPI under the Gaussian prior underperforms the alternatives for all n. On the other hand, while the horseshoe prior achieves worse performance than the other methods for small n, its performance improve as n grows, and it eventually matches standard PPI. This suggests that, as n increases and $\text{var}(\widehat{\Delta}_\theta)$ decreases, the observed value of the rectifier is increasingly considered as extreme, causing the influence from the horseshoe prior to eventually vanish thanks to its robustness to extreme bias levels.

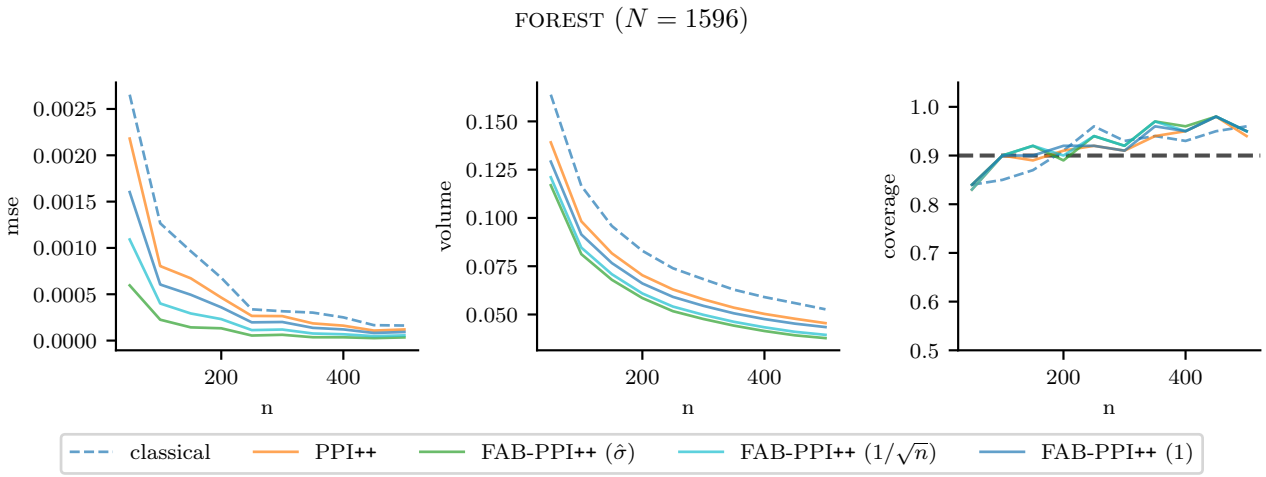

Figure S10. Mean estimation experiment on the FOREST dataset with varying horseshoe prior scale. The left, middle, and right panels show average MSE, CI volume, and CI coverage over 100 repetitions for $\alpha = 0.1$.

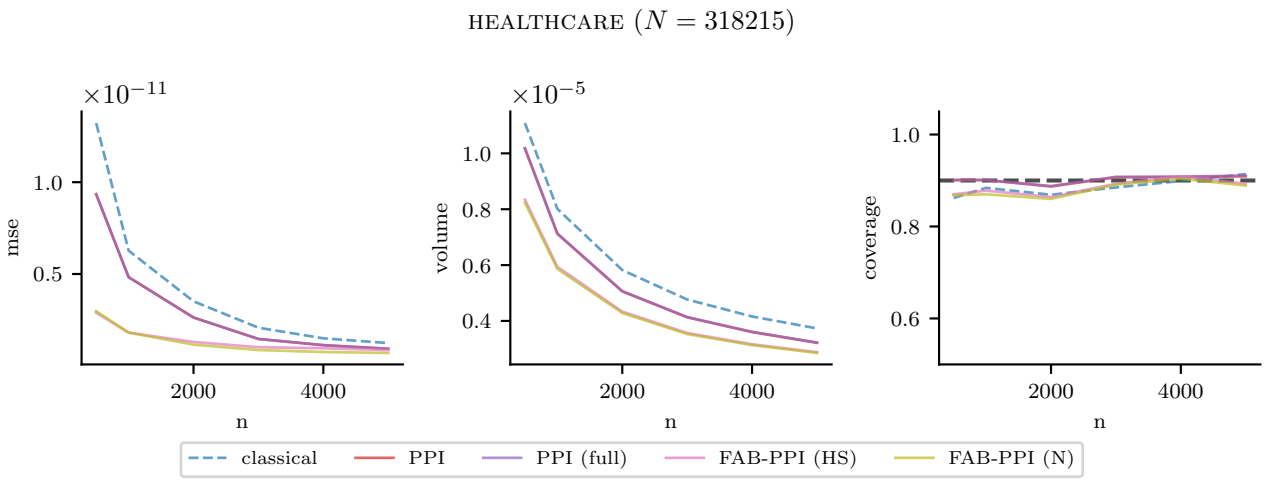

Figure S11. Logistic regression experiment on the HEALTHCARE dataset. The left, middle, and right panels show average MSE, CI volume, and CI coverage over 1000 repetitions for $\alpha = 0.1$.

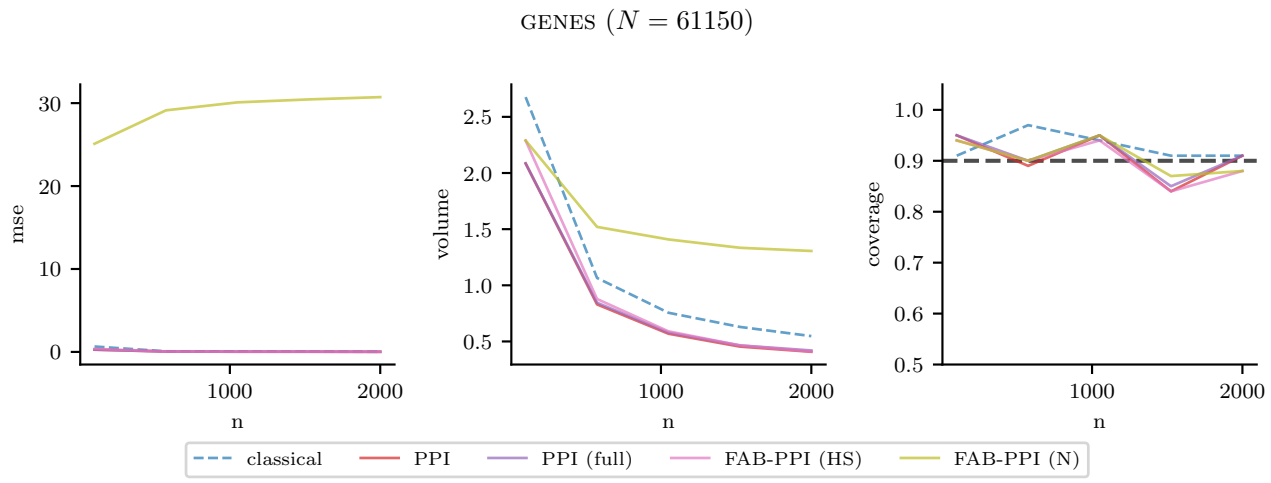

Figure S12. Quantile estimation experiment on the GENES dataset. The left, middle, and right panels show average MSE, CI volume, and CI coverage over 100 repetitions for $\alpha = 0.1$.

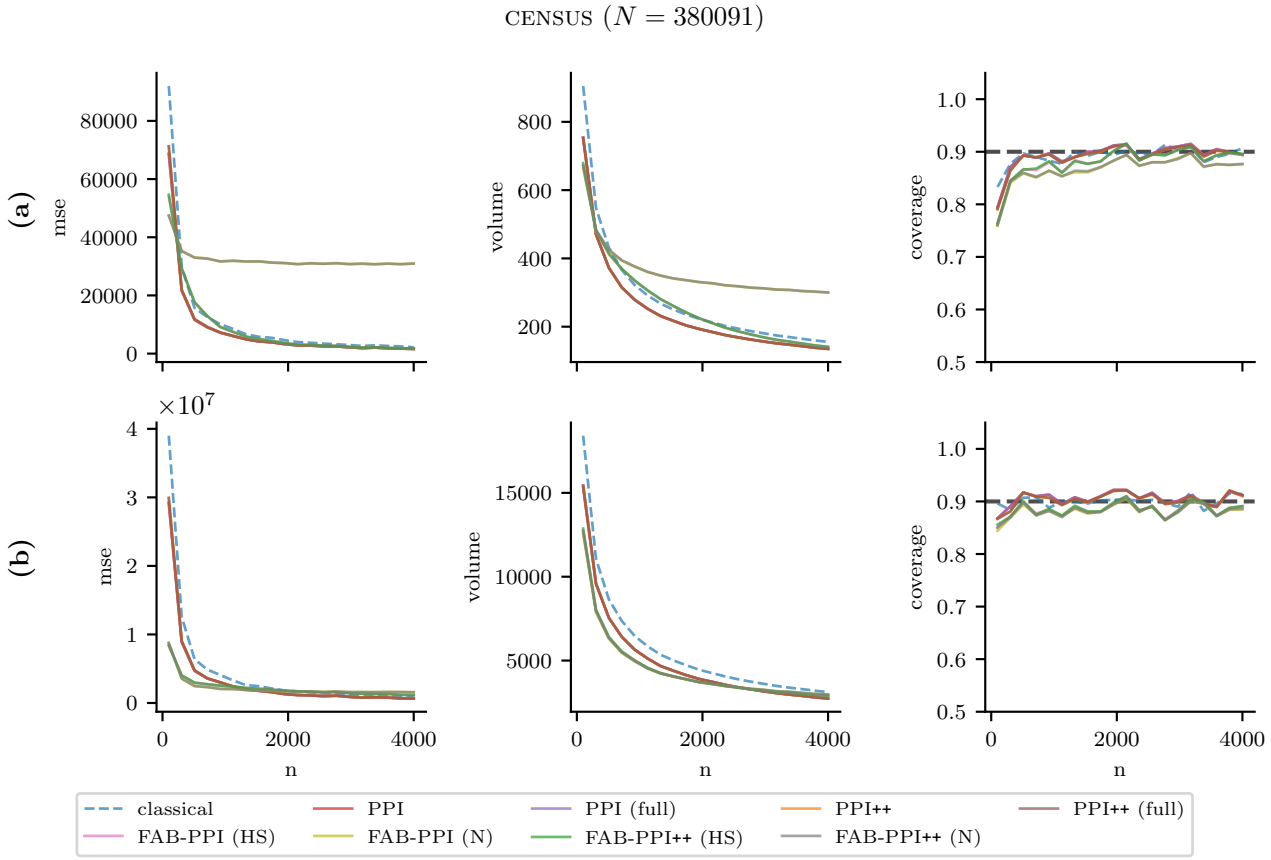

Figure S13. Linear regression experiment on the CENSUS dataset. The (a) and (b) panels correspond to the two covariates in the dataset. The left, middle, and right panels show average MSE, CI volume, and CI coverage over 1000 repetitions for $\alpha = 0.1$.