Bayesian Semi-supervised Inference via a Debiased Modeling Approach

Gözde Sert[1], Abhishek Chakrabortty[1], Anirban Bhattacharya[1,*]

[1]Department of Statistics, Texas A&M University

Abstract

Inference in semi-supervised (SS) settings has received substantial attention in recent years due to increased relevance in modern big-data problems. In a typical SS setting, there is a much larger sized unlabeled data, containing observations only for a set of predictors, in addition to a moderately sized labeled data containing observations for both an outcome and the set of predictors. Such data arises naturally from settings where the outcome, unlike the predictors, is costly or difficult to obtain. One of the primary statistical objectives in SS settings is to explore whether parameter estimation can be improved by exploiting the unlabeled data. A novel Bayesian approach to SS inference for the population mean estimation problem is proposed. The proposed approach provides improved and optimal estimators both in terms of estimation efficiency as well as inference. The method itself has several interesting artifacts. The central idea behind the method is to model certain summary statistics of the data in a targeted manner, rather than the entire raw data itself, along with a novel Bayesian notion of debiasing. Specifying appropriate summary statistics crucially relies on a debiased representation of the population mean that incorporates unlabeled data through a flexible nuisance function while also learning its estimation bias. Combined with careful usage of sample splitting, this debiasing approach mitigates the effect of bias due to slow rates or misspecification of the nuisance parameter from the posterior of the final parameter of interest, ensuring its robustness and efficiency. Concrete theoretical results, via Bernstein–von Mises theorems, are established, validating all claims, and are further supported through extensive numerical studies. To our knowledge, this is possibly the first work on Bayesian inference in SS settings, and its central ideas also apply more broadly to other Bayesian semi-parametric inference problems.

Keywords: Robustness and efficiency; Bayesian semi-parametric inference; Debiasing; Sample splitting and cross-fitting; Bernstein–von Mises theorem.

1 Introduction and overview of contributions

Semi-supervised (SS) learning has emerged as an exciting and active research area in statistics and machine learning in recent years. A typical SS setting involves two types of data sets: (i) a small or

*Corresponding author.

Email addresses: gozdesert@stat.tamu.edu (Gözde Sert), abhishek@stat.tamu.edu (Abhishek Chakrabortty), anirbanb@stat.tamu.edu (Anirban Bhattacharya).

This work has been accepted for publication in Econometrics and Statistics (to appear). [Link]

moderate sized *labeled* (or supervised) data $\mathcal{L}$ with observations for both an outcome (or label) Y and a set of predictors $\mathbf{X}$, and (ii) a *much larger* sized *unlabeled* (or unsupervised) data $\mathcal{U}$ containing observations only for $\mathbf{X}$. SS settings arise naturally when the outcome is difficult or costly to obtain, but observations for the predictors are plenty and easy to access. Typically, this scenario occurs in many modern big-data problems involving large (electronic) databases, such as speech recognition, text mining, and more recently, biomedical applications like electronic health records (Chapelle et al., 2006; Zhu, 2008; Kohane, 2011; Chakrabortty and Cai, 2018). In a standard SS setup, one of the primary statistical goals is to investigate whether and how parameter estimation and accuracy of inference can be improved by making use of the unlabeled data $\mathcal{U}$, unlike supervised methods, which use only the labeled data $\mathcal{L}$ and completely ignore $\mathcal{U}$. SS inference in this spirit has been studied in the recent frequentist literature for various problems, including mean estimation (Zhang et al., 2019; Zhang and Bradic, 2022) and linear regression (Chakrabortty and Cai, 2018; Azriel et al., 2022), among others. However, Bayesian approaches for SS inference are largely lacking in the literature to the best of our knowledge.

We propose a *Bayesian debiased modeling and inference* (BDMI) procedure for estimating the *population mean* $\theta_0 := \mathbb{E}(Y)$ of Y under the SS setting, as a prototypical example. A fundamental idea behind BDMI is to carefully model certain *summary statistics* of the data in a *targeted* manner, rather than specifying a probability model for the raw data itself, along with developing and exploiting a novel *Bayesian notion of debiasing of nuisance parameters* (that are inherently involved in the procedure). Most existing SS approaches for estimating θ_0 (or similar parameters/functionals of the distribution of Y) naturally require estimation of the possibly high dimensional *regression function* $m_0(\mathbf{X}) := \mathbb{E}(Y|\mathbf{X})$ to exploit $\mathcal{U}$ (Chakrabortty and Cai, 2018; Zhang et al., 2019; Cai and Guo, 2020; Zhang and Bradic, 2022). $m_0(\cdot)$ therefore acts as a *nuisance function* here, that is needed (for exploiting $\mathcal{U}$) but is not of primary interest. In general, the presence of such a nuisance parameter and its own estimation bias can drastically affect the final estimator's asymptotic behavior in the first order. In recent years, a popular frequentist debiasing procedure called double machine learning (DML) based on Neyman orthogonalization has been developed to rectify the impact of bias in learning a nuisance parameter (Chernozhukov et al., 2018). A key contribution of this work is to develop a *Bayesian analogue* of such debiasing procedures, that ensures *robust, efficient* and *nuisance-insensitive* Bayesian inference for θ_0 (the *target*) while *allowing for slow/inefficient (or even inconsistent) learning of* m_0.

BDMI encapsulates a new principle of *disentangling* the nuisance parameter that is amenable to *Bayesian* modeling and inference. It crucially relies on a *debiased representation* (Section 3.1) of θ_0 in terms of m_0 (specifically, its estimator or a posterior sample) that simultaneously exploits $\mathcal{U}$ and also captures the nuisance bias incurred. Exploiting this representation, we then propose to model carefully chosen summary statistics of the data (see Section 3.2). Modeling summary statistics of the data has been sporadically considered in the Bayesian literature for estimation and hypothesis testing (Pratt, 1965; Savage, 1969; Doksum and Lo, 1990; Clarke and Ghosh, 1995; Johnson, 2005; Lewis et al., 2021) as well as in likelihood-free inference methods like Approximate Bayesian Computation (ABC) (Marjoram et al., 2003; Fearnhead and Prangle, 2012; Drovandi et al., 2015). In the present setting, the summary statistics are exploited to: (i) carefully pinpoint the target and the bias induced from the nuisance, and (ii) learn them jointly by constructing a robust working likelihood (that can be justified under mild assumptions on the data generating mechanism) which can then be combined with default prior distributions on the model parameters to arrive at a posterior distribution. Further, a key feature of our approach is the careful usage of

sample-splitting and *cross-fitting (CF)* (Chernozhukov et al., 2018; Newey and Robins, 2018) – *not* just as a technical artifact (as is common in the frequentist literature) but as an *integral component* of the debiasing process itself. It helps create *independent* sub-folds of the entire data that crucially enable the disentangling of the nuisance estimation process from the summary statistics modeling process. Further, to ensure usage of the full data overall, we use CF by rotating the roles of the splits and using each sub-fold in turn, and thereafter *aggregating* the posteriors from all sub-folds using a consensus Monte Carlo type approach (Scott et al., 2022). It is worth mentioning that, while commonplace in the modern frequentist literature on semi-parametric inference, handling sample splitting (and CF) under a Bayesian framework is more challenging since it requires combining *distributions* (posteriors) and not just point estimators. Our final CF-based version of BDMI is given in Section 3.3 and summarized in Algorithm 1.

We show through our theoretical results in Section 4 that the marginal posterior distribution $\Pi_{\boldsymbol{\theta}}$ for θ from BDMI inherits a *Bernstein–von Mises (BvM)-type limiting behavior* (van der Vaart, 2000, Chapter 10) with an asymptotically Gaussian shape, and contracts *always* around the true θ_0 at a parametric $n^{-1/2}$ rate (n being the size of $\mathcal{L}$) and with a spread *tighter* than the supervised counterpart – all holding *irrespective* of the choice/method used to obtain the nuisance posterior ($\Pi_{\mathbf{m}}$) for learning m_0. Further, $\Pi_{\boldsymbol{\theta}}$'s first order variability is *unaffected* by that of $\Pi_{\mathbf{m}}$ and is of the correct $n^{-1/2}$ rate *even if* the contraction rate of $\Pi_{\mathbf{m}}$ is arbitrarily *slow* or if it is even *misspecified* (i.e., does not contract around the true m_0). This makes BDMI *first-order insensitive* (Chernozhukov et al., 2018) to the nuisance estimation. Most importantly, from an SS inference perspective, $\Pi_{\boldsymbol{\theta}}$ (and its posterior mean) provably possess the desirable properties of *global robustness* and *efficiency improvement*: we show (i) the symmetric Bayesian credible intervals (CIs) from $\Pi_{\boldsymbol{\theta}}$ possess asymptotically correct frequentist coverage and sizes (of order $n^{-1/2}$) guaranteed to be tighter than their supervised counterpart; and (ii) the posterior mean is *always* $\sqrt{n}$-consistent, asymptotically Normal and *more efficient or at least as efficient* as the supervised estimator. Furthermore, when $\Pi_{\mathbf{m}}$ is correctly specified (with arbitrary contraction rate), $\Pi_{\boldsymbol{\theta}}$ and its posterior mean attain *optimal* efficiency, with variance matching the *semi-parametric efficiency bound*. All our claims above are validated through extensive simulations as well as a real data application in Section 5. It is also worth noting that BDMI is *computationally scalable,* with all ingredient posteriors (from each fold) in $\Pi_{\boldsymbol{\theta}}$ being convolutions of t-distributions (hence easy to sample from). To our knowledge, BDMI is the *first work* on Bayesian inference (with provable guarantees) in SS settings.

Aside from SS inference itself, this work also contributes more generally to the growing literature on *Bayesian semi-parametric inference* in modern big-data settings. The SS setting has a distinct semi-parametric flavor, with $m_0(\cdot)$ being the (potentially high dimensional) nuisance parameter and functionals like θ_0 being the target. There is a growing literature on frequentist properties of Bayesian semi-parametric inference procedures; see, e.g., Bickel and Kleijn (2012); Rivoirard and Rousseau (2012); Castillo and Rousseau (2015); Norets (2015); Ray and Szabo (2019); where the quantity of interest is the marginal posterior of the parameter of interest obtained upon marginalizing out the nuisance parameter. Under delicate conditions on the prior distribution of the nuisance parameter, BvM results have been established for the parameter of interest in some of these works. Moreover, there have been some recent developments in the Bayesian semi-parametric literature (primarily for missing data or causal inference problems) aimed at alleviating bias arising from the nuisance estimation with slow rates (Ray and van der Vaart, 2020; Luo et al., 2023; Breunig et al., 2025; Yiu et al., 2025). Most of these are based on careful prior selection/modification, or tailored posterior updating, to mimic the flavors of their frequentist counterparts. BDMI adds to

this literature by considering a different perspective and a principled approach to mitigate the bias of nuisance parameters. Another key feature of the approach is that it leaves the nuisance estimation method *entirely* to the user, and the nuisance posterior (or prior) does *not* require any form of adjustment or updating. While proposed albeit under the auspices of the SS inference problem, we believe the fundamental ideas of BDMI – Bayesian debiasing and targeted modeling via summary statistics – will also apply more generally to other Bayesian semi-parametric inference problems.

The rest of the article is organized as follows. We discuss the problem setup and some key preliminaries in Section 2. Our proposed methodology is presented in Section 3, with its various facets distributed across Sections 3.1–3.3. The theoretical properties of our method, including our main results (Theorems 4.1–4.2), are presented in Section 4, along with an alternative hierarchical version of our method and its theoretical properties discussed in Section 4.2. Finally, extensive simulation studies and real data analysis are presented in Section 5 to illustrate its empirical performance, followed by a concluding discussion in Section 6. All technical materials, including the proofs of all the main theoretical results, along with supporting lemmas and their proofs, as well as additional numerical results and methodological discussions that could not be accommodated in the main paper, are collected in the Supplementary Material (Sections S1–S5).

2 The problem setup and key preliminary ideas

Let $Y \in \mathbb{R}$ be the outcome variable, $\mathbf{X} \in \mathbb{R}^p$ be the covariate (or predictor) vector, and $\mathbb{P}_{\mathbf{Z}} \equiv \mathbb{P}_{Y|\mathbf{X}} \otimes \mathbb{P}_{\mathbf{X}}$ be the unknown joint distribution of $\mathbf{Z} := (Y, \mathbf{X}')'$, where $\mathbb{P}_{Y|\mathbf{X}}$ and $\mathbb{P}_{\mathbf{X}}$ denote the conditional distribution of $Y \mid \mathbf{X}$ and the marginal distribution of $\mathbf{X}$, respectively. The *available data* under the SS setting is denoted as: $\mathcal{D} := \mathcal{L} \cup \mathcal{U}$, with $\mathcal{L} := \{\mathbf{Z}_i \equiv (Y_i, \mathbf{X}_i')' : i = 1, \ldots, n\}$ being the labeled data containing n independent and identically distributed (i.i.d.) samples of $\mathbf{Z} \sim \mathbb{P}_{\mathbf{Z}}$, and $\mathcal{U} := \{\mathbf{X}_i : i = n+1, \ldots, n+N\}$ being the unlabeled data containing N i.i.d. samples of $\mathbf{X} \sim \mathbb{P}_{\mathbf{X}}$, and $\mathcal{L}$ and $\mathcal{U}$ are independent, denoted as $\mathcal{L} \perp\!\!\!\perp \mathcal{U}$.

Assumption 2.1 (Standard features of SS settings)**.** We assume throughout that: (i) the unlabeled data size N grows at least as fast as (and typically faster than) the labeled data size n, such that $n/N \to c$ as $n, N \to \infty$, where $0 \leq c < 1$ ($c = 0$ being a key focus); and (ii) the observations for $\mathbf{Z}$ in $\mathcal{L}$ and those for $\mathbf{Z}$ underlying the unlabeled $\mathbf{X}$ in $\mathcal{U}$ arise from the same distribution $\mathbb{P}_{\mathbf{Z}}$ above, and $\mathbf{Z}$ has finite second moments.

Remark 2.1. Assumption 2.1 is fairly standard in the SS inference literature (Chapelle et al., 2006; Kawakita and Kanamori, 2013). The condition (i) encodes a key (and *unique*) feature of SS settings, allowing for disproportionate sizes of $\mathcal{L}$ and $\mathcal{U}$. For example, while the size of $\mathcal{L}$ may be of the order of hundreds, the size of $\mathcal{U}$ could be of the order of tens of thousands. Further, since the outcome Y is missing in $\mathcal{U}$, one can view SS inference as a missing data problem by assuming Y is 'missing completely at random' (Tsiatis, 2006). However, since $\lim_{n,N\to\infty} n/N \to c = 0$ is allowed, it naturally violates the positivity assumption (on the proportion of Y observed) standard in the missing data literature (Tsiatis, 2006), and makes the SS setting fundamentally *different* and more challenging (due to *non-standard asymptotics*) from the missing data setup. The condition (ii) asserts that the underlying distributions of $\mathcal{L}$ and $\mathcal{U}$ are the same, which is standard and often implicit in the SS inference literature (Kawakita and Kanamori, 2013; Chakrabortty and Cai, 2018; Zhang et al., 2019; Zhang and Bradic, 2022), along with a mild moment assumption on $\mathbf{Z}$ to ensure

$\mathbb{E}(Y \mid \mathbf{X})$ and $\mathbb{E}(Y)$ exist. Finally, we clarify that we allow *high dimensional* settings throughout (p can diverge with n).

2.1 Preliminaries: Notational conventions and the supervised approach

We use the following *notational conventions* throughout the paper. Let $\mathbb{E}(\cdot) \equiv \mathbb{E}_{\mathbf{Z}}(\cdot)$, $\mathbb{E}_{Y|\mathbf{X}}(\cdot)$ and $\mathbb{E}_{\mathbf{X}}(\cdot)$ denote expectations under the distributions $\mathbb{P} \equiv \mathbb{P}_{\mathbf{Z}}$, $\mathbb{P}_{Y|\mathbf{X}}$ and $\mathbb{P}_{\mathbf{X}}$, respectively. For any dataset/collection (or its subset/functions) $\mathcal{C}$ on $\mathbf{Z}$, let $\mathbb{E}_{\mathcal{C}}(\cdot)$ and $\mathbb{P}_{\mathcal{C}}(\cdot)$ denote expectations and probability under the joint distribution of $\mathcal{C}$. Let W be a generic random variable (or vector) with an underlying probability distribution $\mathbb{P}_W$, and let f be any measurable $\mathbb{R}$-valued deterministic function of W. Then, the expectation of $f(W)$ is defined as: $\mathbb{E}_W\{f(W)\} \equiv \mathbb{E}_{W\sim\mathbb{P}_W}\{f(W)\} := \int f(w)\mathrm{d}\mathbb{P}_W(w)$, whenever the Lebesgue integral exists. Further, for any $d \geq 1$, let $\mathbb{L}_d(\mathbb{P}_{\mathbf{Z}})$ and $\mathbb{L}_d(\mathbb{P}_{\mathbf{X}})$ denote the spaces of all $\mathbb{R}$-valued measurable functions g of $\mathbf{Z}$, and h of $\mathbf{X}$, such that $\|g(\mathbf{Z})\|^d_{\mathbb{L}_d(\mathbb{P}_{\mathbf{Z}})} := \mathbb{E}_{\mathbf{Z}}\{|g(\mathbf{Z})|^d\} < \infty$ and $\|h(\mathbf{X})\|^d_{\mathbb{L}_d(\mathbb{P}_{\mathbf{X}})} := \mathbb{E}_{\mathbf{X}}\{|h(\mathbf{X})|^d\} < \infty$, respectively. Let $\mathcal{N}(\mu, \sigma^2)$ denote the Normal (Gaussian) distribution with mean μ and variance σ^2, and $t_\nu(\mu, c^2)$ denote the t-distribution with degrees of freedom $\nu > 0$, center μ and scale c. We also use $\mathcal{N}(x; \mu, \sigma^2)$ and $t_\nu(x; \mu, c^2)$ to denote their respective probability density functions (pdfs) evaluated at $x \in \mathbb{R}$. For given probability measures P and Q on a measurable space $(\Omega, \mathcal{F})$, the total variation (TV) distance between P and Q is $\|P - Q\|_{\mathrm{TV}} := \sup_{B\in\mathcal{F}} |P(B) - Q(B)|$. For a sequence $b_n > 0$ and a sequence of random variables X_n, we say $X_n = o_{\mathbb{P}}(b_n)$ if and only if (iff) $|X_n|/b_n \xrightarrow{\mathbb{P}} 0$ as $n \to \infty$. If $X_n \xrightarrow{\mathbb{P}} 0$, we write $X_n = o_{\mathbb{P}}(1)$. Similarly, a sequence of random variables $W_n = O_{\mathbb{P}}(b_n)$ iff for any $\varepsilon > 0$, there exist $B_\varepsilon > 0$ and n_ε such that $\mathbb{P}(|W_n| \leq B_\varepsilon b_n) > 1 - \varepsilon$ for all $n \geq n_\varepsilon$. Furthermore, $W_n = o_{\mathbb{P}}(1)$ iff for some sequence $b_n \to 0$, $W_n = O_{\mathbb{P}}(b_n)$. Lastly, for $\psi_0 \equiv \psi_0(\mathbb{P})$ denoting any functional of interest for any distribution $\mathbb{P}$, we let ψ represent the corresponding *random variable* (or vector, function, etc., as applicable) in a Bayesian framework, and denote its posterior distribution by Π_ψ. *This convention is used consistently, without mention, throughout the paper.*

Before discussing any SS approaches, we first introduce the standard *supervised* Bayesian approach for estimating θ_0 using $\mathcal{L}$ only, to set a benchmark. In the supervised setting, one can adopt a Bayesian framework by modeling $\mathcal{L}$ (i.e., the Y_i's $\in \mathcal{L}$) with a working Gaussian likelihood with mean θ and variance σ^2, combined with a joint prior on (θ, σ^2). This yields a marginal *posterior* Π_{sup} for θ which, under mild regularity conditions on the prior, satisfies a BvM result (van der Vaart, 2000, Chapter 10.2): $\Pi_{\mathrm{sup}} \approx \mathcal{N}(\widehat{\theta}_{\mathrm{sup}}, \sigma^2_Y/n)$ as $n \to \infty$, where $\widehat{\theta}_{\mathrm{sup}} := \overline{Y} \equiv n^{-1}\sum_{i=1}^n Y_i$ and $\sigma^2_Y := \mathrm{Var}(Y)$. Thus, Π_{sup} yields $\widehat{\theta}_{\mathrm{sup}} \equiv \overline{Y}$ as a natural (supervised) *point estimator* of θ_0, as well as CIs of sizes $\propto \sigma_Y/\sqrt{n}$. Further, σ^2_Y is the *best* achievable variance in the *supervised* setting and attains the semi-parametric efficiency bound under a fully non-parametric model (van der Vaart, 2000, Chapter 25.3) for estimating θ_0. We will therefore use the limiting supervised posterior $\mathcal{N}(\widehat{\theta}_{\mathrm{sup}}, \sigma^2_Y/n)$ as a *benchmark* for asymptotic estimation/inference efficiency comparisons with BDMI later.

2.2 A motivating imputation-type Bayesian SS approach

The construction of the supervised posterior Π_{sup} (and $\widehat{\theta}_{\mathrm{sup}}$) naturally does not utilize the large unlabeled data $\mathcal{U}$ on $\mathbf{X}$ available in the SS setting. By virtue of its large size, $\mathcal{U}$ essentially informs us on the distribution, $\mathbb{P}_{\mathbf{X}}$, of $\mathbf{X}$. Thus, whenever $\mathbb{P}_{\mathbf{X}}$ is informative about the parameter of interest (Zhang and Oles, 2000; Seeger, 2002), one may hope to utilize $\mathcal{U}$ and come up with an improved SS

Bayesian estimation procedure with a more efficient, i.e., tighter posterior contracting around θ_0 (albeit at a $\sqrt{n}$-rate, since information on Y is still limited to n observations), and accordingly a $\sqrt{n}$-consistent point estimator of θ_0 that is more efficient than $\widehat{\theta}_{\text{sup}}$. We now discuss such an intuitive *imputation-based* approach with a natural Bayesian flavor, along with its potential drawbacks, which form a crucial basis for our final formulation of the BDMI method in Section 3.

Recalling $m_0(\mathbf{X}) \equiv \mathbb{E}(Y \mid \mathbf{X})$, the functional $\theta_0 \equiv \theta_0(\mathbb{P}_{\mathbf{Z}}) = \mathbb{E}(Y)$ can be written via iterated expectations as: $\theta_0 \equiv \theta_0(\mathbb{P}_{\mathbf{X}}; m_0) = \mathbb{E}_{\mathbf{X}}\{\mathbb{E}_{Y|\mathbf{X}}(Y \mid \mathbf{X})\} = \mathbb{E}_{\mathbf{X}}\{m_0(\mathbf{X})\}$. This representation clearly explains the *connection* between $\mathbb{P}_{\mathbf{X}}$ and θ_0, and the potential for $\mathcal{U}$ to be exploited through bringing in the *nuisance* function m_0 (unknown but *estimable* via $\mathcal{L}$). One can then construct an imputation-based Bayesian SS approach as follows.

Suppose one learns $m_0(\cdot)$ from $\mathcal{L}$ via *any* reasonable Bayesian regression method (see Remark 3.4 for some examples) that provides a *nuisance posterior* $\Pi_{\mathbf{m}} \equiv \Pi_{\mathbf{m}}(\cdot\,; \mathcal{L})$ for m. Then, using the identity $\theta_0 = \mathbb{E}_{\mathbf{X}}\{m_0(\mathbf{X})\}$, and replacing $\mathbb{E}_{\mathbf{X}}$ therein with an empirical average over $\mathcal{U}$, one may obtain an *induced posterior* Π_{imp} for θ via a natural *imputation* approach, i.e., for samples $\widetilde{m} \sim \Pi_{\mathbf{m}}$, we let $\theta_{\text{imp}} \equiv \theta_{\text{imp}}(\widetilde{m}) := N^{-1} \sum_{\mathbf{X}_i \in \mathcal{U}} \widetilde{m}(\mathbf{X}_i) \sim \Pi_{\text{imp}}$. Further, by linearity of expectation, it is easy to show that, $\widehat{\theta}_{\text{imp}} := N^{-1} \sum_{\mathbf{X}_i \in \mathcal{U}} \widehat{m}(\mathbf{X}_i)$ is the posterior mean of Π_{imp} (and hence, a point estimate of θ_0), where $\widehat{m}(\cdot) := \mathbb{E}_{\widetilde{m} \sim \Pi_{\mathbf{m}}}\{\widetilde{m}(\cdot) \mid \mathcal{L}\}$ is the posterior mean of $\Pi_{\mathbf{m}}$.

There are two major issues with this approach: (i) potential *misspecification* of $\Pi_{\mathbf{m}}$ in learning the true m_0; and (ii) more importantly, effect of the *nuisance* $\Pi_{\mathbf{m}}$'s *first-order properties (its rate/bias and variability) directly impacting* the *target* Π_{imp}'s *first-order behavior*. To illustrate, consider the *ideal* case: $N = \infty$. Then, the posterior sample θ_{imp} equals $\mathbb{E}_{\mathbf{X}}\{\widetilde{m}(\mathbf{X}) | \widetilde{m}\} \equiv \theta_0 + \mathbb{E}_{\mathbf{X}}\{\widetilde{m}(\mathbf{X}) - m_0(\mathbf{X}) | \widetilde{m}\}$. Thus, when misspecification is allowed, i.e., $\mathbb{E}_{\widetilde{m} \sim \Pi_{\mathbf{m}}}\{\|\widetilde{m}(\mathbf{X}) - m^*(\mathbf{X})\|_{\mathbb{L}_2(\mathbb{P}_{\mathbf{X}})} \mid \mathcal{L}\} \xrightarrow{\mathbb{P}} 0$ (under $\mathbb{P}_{\mathcal{L}}$) for *some* function $m^*(\cdot) \in \mathbb{L}_2(\mathbb{P}_{\mathbf{X}})$ possibly $\neq m_0(\cdot)$, then Π_{imp} may become *inconsistent* (i.e., not contracting around the true θ_0). More fundamentally, *even if* $m^*(\cdot) = m_0(\cdot)$, the *entire* first-order behavior (rate, shape, and variability) of Π_{imp} depends *directly* on the corresponding behavior of (posterior of): $\widetilde{m}(\cdot) - m_0(\cdot)$, the 'bias term', making Π_{imp} *sensitive*, in the *first order*, to $\Pi_{\mathbf{m}}$'s first order properties, and accordingly, the choice of the *method* used therein. In particular, if $\Pi_{\mathbf{m}}$ has a contraction rate, a_n, *slower* than $n^{-1/2}$, then so will Π_{imp}. More importantly, the *variability* of Π_{imp} itself (after scaling by its rate) will be directly impacted by that of $\Pi_{\mathbf{m}}$. Overall, this indicates that to obtain a BvM-type result on Π_{imp} – necessary to ensure provably valid estimation and inference on θ_0 – one *requires* the availability of a corresponding semi-parametric BvM-type result under the nuisance $\Pi_{\mathbf{m}}$, which may necessitate delicate conditions/control on specifics of $\Pi_{\mathbf{m}}$'s construction. This becomes especially challenging when using non-smooth or complex methods, e.g., sparse regression in high dimensions or non-parametric machine learning methods, as nuisance estimators. These methods, while highly relevant and popular, have rates slower than $n^{-1/2}$, as well as unclear first-order properties with often intractable posteriors and limited availability (or feasibility) of corresponding BvM results. In general, this first-order sensitivity of Π_{imp} and its reliance on such intricate aspects of $\Pi_{\mathbf{m}}$, therefore, jeopardizes rate-optimal and provably valid inference on θ_0 with the correct variance. In Section S2 of the Supplementary Material, we present a detailed case study on Π_{imp} (and also compare it to BDMI) showcasing its sensitivity and failure to provide a valid inference on θ_0.

3 Bayesian debiased modeling and inference: BDMI

This section introduces the BDMI approach, which addresses the limitations of the imputation approach discussed in Section 2.2, by appropriately accounting for nuisance estimation bias within a Bayesian likelihood framework. BDMI is based on the principle of disentangling the nuisance parameter, and jointly *learning* its bias with the parameter of interest via targeted summary statistics *amenable* to Bayesian modeling. Incorporating this debiasing idea and the targeted modeling approach are our key methodological contributions towards *Bayesian semi-parametric inference*, in general, for robust and efficient inference in the presence of high dimensional nuisances, drawing parallels to the recent frequentist DML literature (Chernozhukov et al., 2018).

3.1 Bayesian debiasing: Overcoming the bias from nuisance estimation within the Bayesian framework

For exposition of the BDMI approach and its salient features, we assume for the time being that there exists a dataset $\mathcal{S}$ which is an *independent copy* of the labeled data $\mathcal{L}$. The sample size s_n of $\mathcal{S}$ is assumed to be of the same order as n; see Section 3.3 for more details. Suppose the nuisance estimation is performed on this $\mathcal{S}$, using *any* reasonable Bayesian (or frequentist) method by constructing a likelihood for the nuisance parameter m on $\mathcal{S}$, combining with a suitable prior on m, to obtain a posterior $\Pi_{\mathbf{m}}$ for m. For our primary goal of inference on θ_0, the specific construction of $\Pi_{\mathbf{m}}$ is not crucial, provided it satisfies some basic regularity conditions (see Section 4 for details). Henceforth, we assume access to a *generic* posterior $\Pi_{\mathbf{m}}$ for m, noting that $\Pi_{\mathbf{m}}(\cdot) \equiv \Pi_{\mathbf{m}}(\cdot; \mathcal{S})$ is itself a *random* distribution dependent on $\mathcal{S}$. For simplicity, this dependence is suppressed in the notations whenever clear from context. The dataset $\mathcal{S}$ can be viewed as *training data*, used *solely* to obtain the nuisance posterior $\Pi_{\mathbf{m}}$ for m. In contrast, $\mathcal{D} = \mathcal{L} \cup \mathcal{U}$ serves as *test data*, used to obtain the posterior for the parameter of interest θ via the BDMI procedure. In practice, we construct such pairs of independent training and test datasets from the original data $\mathcal{D}$ itself via *sample splitting*; see Section 3.3.

Let $\widetilde{m} : \mathbb{R}^p \to \mathbb{R}$ be any *random function* (van der Vaart, 2000, Ch. 19.4) output from $\mathcal{S}$ (e.g., a posterior sample from a Bayesian regression model fitted to $\mathcal{S}$). More formally, $\widetilde{m} : (\Omega_{\mathcal{S}}, \mathbb{P}_{\mathcal{S}}) \times \mathbb{R}^p \to \mathbb{R}$ is a measurable map, i.e., $\{\widetilde{m}(x)\}_{x \in \mathbb{R}^p} \equiv \{\widetilde{m}(\omega; x)\}_{x \in \mathbb{R}^p}$ is a *stochastic process*, with sample paths $\widetilde{m}(\omega; \cdot)$ for $\omega \in \Omega_{\mathcal{S}}$, where $(\Omega_{\mathcal{S}}, \mathbb{P}_{\mathcal{S}})$ denotes the probability space underlying the randomness of $\mathcal{S}$ and any derived measures (e.g., posteriors) from it. Suppose now the argument $\mathbf{x}$ (or domain) of $\widetilde{m}(\mathbf{x}) \equiv \widetilde{m}(\omega; \mathbf{x})$ is *measurized* (randomized) *independently* as: $\mathbf{X} \sim \mathbb{P}_{\mathbf{X}} \perp\!\!\!\perp \mathbb{P}_{\mathcal{S}}$, e.g., $\mathbf{X} \in \mathcal{D}$ meets this requirement, since $\mathcal{D} \perp\!\!\!\perp \mathcal{S}$ by construction. Consider the *doubly* random variable $\widetilde{m}(\mathbf{X})$ – having two sources of randomness that are independent – (i) the process $\widetilde{m}(\cdot)$ *itself* from $\mathcal{S}$, and (ii) its random argument $\mathbf{X} \sim \mathbb{P}_{\mathbf{X}}$ from $\mathcal{D}$ ($\perp\!\!\!\perp \mathcal{S}$). We can then write $\theta_0 \equiv \mathbb{E}(Y)$ as:

$$\theta_0 \;=\; \mathbb{E}_{\mathbf{X}}\{\mathbb{E}(Y \mid \mathbf{X})\} \;\equiv\; \mathbb{E}_{\mathbf{X}}\{m_0(\mathbf{X})\} \;=\; \underbrace{\mathbb{E}_{\mathbf{X} \in \mathcal{D}}[\{m_0(\mathbf{X}) - \widetilde{m}(\mathbf{X})\} \mid \widetilde{m}]}_{:= \, b(\widetilde{m}) \rightsquigarrow \text{Bias induced from } \widetilde{m}(\cdot)} + \underbrace{\mathbb{E}_{\mathbf{X} \in \mathcal{D}}\{\widetilde{m}(\mathbf{X}) \mid \widetilde{m}\}}_{\text{Imputation via } \widetilde{m}(\cdot)}; \quad (1)$$

$$\equiv\; b(\widetilde{m}) + \mathbb{E}_{\mathbf{X} \in \mathcal{D}}\{\widetilde{m}(\mathbf{X}) | \widetilde{m}\} \;=\; \mathbb{E}_{\mathbf{Z} \in \mathcal{L}}\{Y - \widetilde{m}(\mathbf{X}) | \widetilde{m}\} + \mathbb{E}_{\mathbf{X} \in \mathcal{U}}\{\widetilde{m}(\mathbf{X}) | \widetilde{m}\} \quad [\mathcal{D} \equiv \mathcal{L} \cup \mathcal{U} \perp\!\!\!\perp \widetilde{m}(\cdot)]. \quad (2)$$

The steps in both (1)–(2) use $\widetilde{m}(\cdot)$ from $\mathcal{S}$ is $\perp\!\!\!\perp$ of $\mathbf{X}$ (and $\mathbf{Z}$) $\in \mathcal{D}$. This independence is *crucial* and necessary to derive (1), which we refer to as the *debiased representation* of θ_0. For notational clarity, we emphasize that for a given $\widetilde{m}$, $b(\widetilde{m})$ should be interpreted as a parameter dependent on

$\widetilde{m}$, i.e., a *function* of $\widetilde{m}$. Finally, we reiterate that the above representations (1)–(2) remain valid if $\widetilde{m}(\cdot)$ is a random draw from the posterior $\Pi_{\mathbf{m}}$, and $\mathbf{X} = \mathbf{X}_i \in \mathcal{D}$ $(i = 1, \ldots, n+N)$, and $\mathbf{Z} = \mathbf{Z}_i \in \mathcal{L}$ $(i = 1, \ldots, n)$, since $\Pi_{\mathbf{m}}$ is constructed from $\mathcal{S}$ which is independent of $\mathcal{D}$. Subsequent references to (1)–(2) are with respect to (w.r.t.) these particular choices.

Note that the first term $b(\widetilde{m})$ in (1) is essentially the expected *bias*, which is the price of replacing $m_0(\cdot)$ with a random sample $\widetilde{m}(\cdot)$. As noted in Section 2.2, this is precisely the primary cause of the issues with the imputation approach. Modeling this $b(\widetilde{m})$ itself, along with θ_0, is the central idea of BDMI. Note further that:

$$b(\widetilde{m}) \;\equiv\; \mathbb{E}_{\mathbf{X}\in\mathcal{D}}\big[\{m_0(\mathbf{X}) - \widetilde{m}(\mathbf{X})\}\,|\,\widetilde{m}\big] \;=\; \mathbb{E}_{\mathbf{X}}\{m_0(\mathbf{X}) - m^*(\mathbf{X})\} + \mathbb{E}_{\mathbf{X}\in\mathcal{D}}\big[\{m^*(\mathbf{X}) - \widetilde{m}(\mathbf{X})\}\,|\,\widetilde{m}\big].$$

This shows $b(\widetilde{m})$ captures two pivotal aspects: (i) when $m^*(\cdot) \neq m_0(\cdot)$, the first term measures its average deviation from $m_0(\cdot)$, and (ii) the second term importantly reflects the *variability* of $\widetilde{m}(\cdot)$ itself as a sample from $\Pi_{\mathbf{m}}$ (which is further random through $\mathcal{S}$). From the perspective of statistical learning theory (Vapnik, 1998), one could think of the first term as *approximation error* and the second term as *estimation error*.

Most importantly, observe that (2) implies we also have *i.i.d. replicates* $\{Y_i - \widetilde{m}(\mathbf{X}_i)\}_{i\in\mathcal{L}}$ and $\{\widetilde{m}(\mathbf{X}_i)\}_{i\in\mathcal{U}}$ from *conditionally (given $\widetilde{m}$) independent sources* that *target* $b(\widetilde{m})$ and $\theta_0 - b(\widetilde{m})$, respectively, through their expectations. Thus, $b(\widetilde{m})$ and $\theta_0 - b(\widetilde{m})$ can be seen as *functionals* of the underlying distribution of $\mathcal{L}$ and $\mathcal{U}$, specifically depending on the *summary statistics* (means) of $Y - \widetilde{m}(\mathbf{X})$ in $\mathcal{L}$ and $\widetilde{m}(\mathbf{X})$ in $\mathcal{U}$ (given $\widetilde{m}$ from an *independent* source), respectively. The *basic premise* of BDMI is: to model the data for these *target-specific parameters* – $b(\widetilde{m})$ and $\theta_0 - b(\widetilde{m})$ – via summary statistics, since they *directly inform us on θ_0, while also learning the bias* induced by $\widetilde{m}$. This *targeted modeling* of summary statistics (instead of the entire data as in traditional Bayesian approaches) is a salient feature of BDMI. Further, its modeling of the bias $b(\widetilde{m})$ *encodes a Bayesian form of debiasing* which plays a crucial role in ensuring nuisance-insensitive inference for θ_0.

3.2 Targeted modeling of summary statistics: Likelihood construction and final posterior

We are now ready to introduce the target-specific model construction discussed in the previous section. Given $\widetilde{m} \sim \Pi_{\mathbf{m}}$ (from $\mathcal{S}$), the i.i.d. replicates $\{Y_i - \widetilde{m}(\mathbf{X}_i)\}_{i=1}^{n}$ and $\{\widetilde{m}(\mathbf{X}_i)\}_{i=n+1}^{n+N}$ from $\mathcal{D}$ $(\perp\!\!\!\perp \mathcal{S})$ target $b(\widetilde{m})$ and $\theta_0 - b(\widetilde{m})$, respectively, in terms of their means. These variables are now treated as our 'observables' on the data $\mathcal{D} \mid \widetilde{m}$, and we now present a working likelihood construction for these observables on this data. To proceed, let us first define $\sigma_1^2(\widetilde{m}) := \mathrm{Var}_{\mathbf{Z}}\{Y - \widetilde{m}(\mathbf{X})\}$ and $\sigma_2^2(\widetilde{m}) := \mathrm{Var}_{\mathbf{X}}\{\widetilde{m}(\mathbf{X})\}$. Then, *given* $\widetilde{m}$, $Y_i - \widetilde{m}(\mathbf{X}_i)$ are i.i.d. with mean $b(\widetilde{m})$ and variance $\sigma_1^2(\widetilde{m})$ for $i \in \{1, \ldots, n\}$, and $\widetilde{m}(\mathbf{X}_i)$ are i.i.d. with mean $\theta_0 - b(\widetilde{m})$ and variance $\sigma_2^2(\widetilde{m})$ for $i \in \{n+1, \ldots, n+N\}$. Since these observables are i.i.d., a natural choice of a working model for such data could be based on Normal distributions with unknown variances, as follows:

$$\begin{aligned} Y_i - \widetilde{m}(\mathbf{X}_i) \mid \widetilde{m}, b(\widetilde{m}), \sigma_1^2(\widetilde{m}) &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(b(\widetilde{m}), \sigma_1^2(\widetilde{m})), \quad i \in \{1, \ldots, n\}; \;\text{ and} \\ \widetilde{m}(\mathbf{X}_i) \mid \widetilde{m}, b(\widetilde{m}), \theta_0, \sigma_2^2(\widetilde{m}) &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta_0 - b(\widetilde{m}), \sigma_2^2(\widetilde{m})), \quad i \in \{n+1, \ldots, n+N\}. \end{aligned} \tag{3}$$

Then, the likelihood as a function of the parameters $\{\theta, b(\widetilde{m}), \sigma_1^2(\widetilde{m}), \sigma_2^2(\widetilde{m})\}$ is given by:

$$L\{\theta, b(\widetilde{m}), \sigma_1^2(\widetilde{m}), \sigma_2^2(\widetilde{m})\} \propto \prod_{i=1}^{n} \mathcal{N}(Y_i - \widetilde{m}(\mathbf{X}_i); b(\widetilde{m}), \sigma_1^2(\widetilde{m})) \prod_{i=n+1}^{n+N} \mathcal{N}(\widetilde{m}(\mathbf{X}_i); \theta - b(\widetilde{m}), \sigma_2^2(\widetilde{m})). \tag{4}$$

The (pseudo-) likelihood constructed above can be combined with a prior distribution on the model parameters $\{\theta, b(\widetilde{m}), \sigma_1^2(\widetilde{m}), \sigma_2^2(\widetilde{m})\}$ using Bayes' formula to yield a posterior, and thereafter a *marginal posterior* $\Pi_{\boldsymbol{\theta}}$ of θ.

We note that the Normal distributions in (3) above are only chosen as *working*, i.e., not necessarily correctly specified, distributions. Since a posterior depends on the data only through sufficient statistics, one could directly model the sample averages of $Y - \widetilde{m}(\mathbf{X})$ and $\widetilde{m}(\mathbf{X})$ as Normally distributed with appropriate parameters under modeling assumptions similar in spirit to (3), operationally leading to the same posterior. In that case, one could simply treat the sample means as the 'derived' observations, and since, given a sufficiently large number of observations, the sample averages are approximately Normal following the Central Limit Theorem (CLT), the Normality assumption on the sample averages would therefore be quite reasonable.

As a concrete *prior choice*, for the sake of theoretical and computational simplicity, we recommend using an improper prior on the model parameters $\{\theta, b(\widetilde{m}), \sigma_1^2(\widetilde{m}), \sigma_2^2(\widetilde{m})\}$ in (3), given by:

$$\pi\{\theta, b(\widetilde{m}) \mid \sigma_1^2(\widetilde{m}), \sigma_2^2(\widetilde{m})\} \propto 1, \quad \pi\{\sigma_1^2(\widetilde{m})\} \propto \{\sigma_1^2(\widetilde{m})\}^{-1} \quad \text{and} \quad \pi\{\sigma_2^2(\widetilde{m})\} \propto \{\sigma_2^2(\widetilde{m})\}^{-1}, \tag{5}$$

with $\sigma_1^2(\widetilde{m})$ and $\sigma_2^2(\widetilde{m})$ being independent. We note that more general prior choices could also be employed here (see Remark 3.2 for a discussion) without altering the asymptotic conclusions, such as the limiting posterior and related properties of the procedure, established in Section 4. For instance, by defining $\delta := \theta - b(\widetilde{m})$, one could place independent conjugate Normal-Inverse Gamma priors on $\{b(\widetilde{m}), \sigma_1^2(\widetilde{m})\}$ and $\{\delta, \sigma_2^2(\widetilde{m})\}$. The proposed improper prior in (5) can then be viewed as a limiting (diffused) version of such a proper prior.

We now explicitly compute the marginal posterior $\Pi_{\boldsymbol{\theta}}$ of θ under (3) and the prior choice (5), as follows.

Proposition 3.1. *Given the likelihood function $L\{\theta, b(\widetilde{m}), \sigma_1^2(\widetilde{m}), \sigma_2^2(\widetilde{m})\}$ in (4) and the improper prior in (5), the marginal posterior distribution $\Pi_{\boldsymbol{\theta}}$ of θ is the convolution of two t-distributions with the pdf $\pi_{\boldsymbol{\theta}}(\theta) = (f * g)(\theta) := \int f(\theta - w) g(w) \mathrm{d}w$, where $\pi_{\boldsymbol{\theta}}(\cdot)$, $f(\cdot)$ and $g(\cdot)$ are the pdfs of $\Pi_{\boldsymbol{\theta}}$, $t_{\nu_n}(\mu_n(\widetilde{m}), \widehat{\sigma}_{1,n}^2(\widetilde{m})/n)$ and $t_{\nu_N}(\mu_N(\widetilde{m}), \widehat{\sigma}_{2,N}^2(\widetilde{m})/N)$, respectively, where the parameters are given by: $\nu_n := n - 1$, $\nu_N := N - 1$,*

$$\begin{aligned} \mu_n(\widetilde{m}) &:= \frac{1}{n} \sum_{i=1}^{n} \{Y_i - \widetilde{m}(\mathbf{X}_i)\} \quad \text{and} \quad \frac{\widehat{\sigma}_{1,n}^2(\widetilde{m})}{n} := \frac{\sum_{i=1}^{n} \left[\{Y_i - \widetilde{m}(\mathbf{X}_i)\} - \mu_n(\widetilde{m})\right]^2}{n(n-1)}; \\ \mu_N(\widetilde{m}) &:= \frac{1}{N} \sum_{i=n+1}^{n+N} \widetilde{m}(\mathbf{X}_i) \quad \text{and} \quad \frac{\widehat{\sigma}_{2,N}^2(\widetilde{m})}{N} := \frac{\sum_{i=n+1}^{n+N} \{\widetilde{m}(\mathbf{X}_i) - \mu_N(\widetilde{m})\}^2}{N(N-1)}. \end{aligned} \tag{6}$$

Note that $\Pi_{\boldsymbol{\theta}}$, being a convolution of two t-distributions, is *easy to sample* from (e.g., for constructing CIs). Further, the *posterior mean:* $\widehat{\theta}_{\mathrm{BDM}}(\widetilde{m})$ of $\Pi_{\boldsymbol{\theta}}$ can be considered as a natural point estimator of θ_0. Note that the $\widetilde{m}$ in $\widehat{\theta}_{\mathrm{BDM}}(\widetilde{m})$ reflects that the estimator (and the posterior

$\Pi_{\boldsymbol{\theta}} \equiv \Pi_{\boldsymbol{\theta}}(\widetilde{m})$ itself) fundamentally depends on the nuisance posterior sample $\widetilde{m} \sim \Pi_{\mathbf{m}}$ used. From Proposition 3.1, it follows that $\widehat{\theta}_{\text{BDM}}(\widetilde{m}) = \mu_n(\widetilde{m}) + \mu_N(\widetilde{m})$. Note that $\widehat{\theta}_{\text{BDM}}(\widetilde{m})$ (and $\Pi_{\boldsymbol{\theta}}$, in general) utilize *both* $\mathcal{L}$ and $\mathcal{U}$, thereby justifying its billing as an SS approach. Also, as $n, N \to \infty$, it converges to θ_0 *even if* $\Pi_{\mathbf{m}}$ is misspecified. This is because the first term in $\widehat{\theta}_{\text{BDM}}(\widetilde{m})$ targets $\mathbb{E}_{\mathbf{Z}}[\{Y - \widetilde{m}(\mathbf{X})\} \mid \widetilde{m}]$, while the second term targets $\mathbb{E}_{\mathbf{X}}\{\widetilde{m}(\mathbf{X}) \mid \widetilde{m}\}$, hence canceling out $\widetilde{m}$'s effect. Thus, BDMI gives a posterior mean that is *always* a consistent point estimator. Moreover, one would expect the spread of the posterior $\Pi_{\boldsymbol{\theta}}$ to be of the correct rate $n^{-1/2}$, and also tighter than the supervised counterpart. These claims, along with other desirable properties of BDMI, are formally established later in Section 4.

Remark 3.1. A notable feature of BDMI is that it needs *only one* sample $\widetilde{m}$ from the nuisance posterior $\Pi_{\mathbf{m}}$. However, one could also consider a more conventional version of BDMI based on a *hierarchical* construction, requiring use of *multiple* samples of $\widetilde{m}$. Section 4.2 rigorously discusses this alternative version, which we call *hierarchical-BDMI* (h-BDMI), and shows that it inherits the same BvM result as BDMI, but under a stronger assumption; see Theorem 4.3. Even empirically, based on extensive simulation studies, we observed that the two versions have mostly similar performances, both in estimation and inference; see Section 5 for details. Therefore, given that it is computationally simpler, we recommend the original BDMI as the final approach.

3.3 Sample splitting based version: BDMI with cross-fitting (BDMI-CF)

To practically implement the ideas introduced in Sections 3.1 and 3.2, we need to construct independent training and test dataset pairs $(\mathcal{S}, \mathcal{D})$ such that $\mathcal{S} \perp\!\!\!\perp \mathcal{D}$. To achieve this from the original data $\mathcal{D} = \mathcal{L} \cup \mathcal{U}$, we employ a K-fold sample splitting (with cross-fitting) procedure, where $K \geq 2$ is *fixed* (relative to n, N) and we assume without loss of generality (w.l.o.g.), that $|\mathcal{L}| = n$ and $|\mathcal{U}| = N$ are divisible by K. To construct independent training and test datasets required for the debiasing representation in (1), we perform K-fold sample splitting by randomly partitioning the indices $\{1, \ldots, n\}$ (for $\mathcal{L}$) and $\{n+1, \ldots n+N\}$ (for $\mathcal{U}$) into K disjoint folds $\{\mathcal{I}_k\}_{i=1}^K$ and $\{\mathcal{J}_k\}_{i=1}^K$, respectively, with each fold $\mathcal{I}_k$ of size $n_K := n/K$ and $\mathcal{J}_k$ of size $N_K := N/K$, for each $k \in \{1, \ldots, K\}$, define $\mathcal{I}_k^- := \{1, \ldots, n\} \backslash I_k$. Then, using these partitions, we construct pairs of training and test data folds $\{(\mathcal{S}_k, \mathcal{D}_k)\}_{k=1}^K$, where $\mathcal{S}_k := \{\mathbf{Z}_i : i \in \mathcal{I}_k^-\} \perp\!\!\!\perp \mathcal{D}_k := \mathcal{L}_k \cup \mathcal{U}_k$, with $\mathcal{L}_k := \{\mathbf{Z}_i : i \in \mathcal{I}_k\}$ and $\mathcal{U}_k := \{\mathbf{X}_i : i \in \mathcal{J}_k\}$. This provides K such (training, test) data pairs for constructing the BDMI approach on each pair. Importantly, the test datasets $\mathcal{D}_1, \ldots, \mathcal{D}_K$ are all disjoint and *independent*.

Adopting the BDMI construction from Section 3.2, we now detail the BDMI procedure for one pair $(\mathcal{S}_k, \mathcal{D}_k)$. Since $\mathcal{S}_k \perp\!\!\!\perp \mathcal{D}_k$, we use the training subfold $\mathcal{S}_k$ to obtain the nuisance posterior $\Pi_{\mathbf{m}}^{(k)}$ for m, as detailed in Section 3.1. Let $\widetilde{m}_k$ be *one* random sample from $\Pi_{\mathbf{m}}^{(k)}$. Following the same model construction in Section 3.2, we use the same likelihood formulation for the test subfold $\mathcal{D}_k$ as given in equations (3)–(4):

$$\begin{aligned} Y_i - \widetilde{m}_k(\mathbf{X}_i) \mid \widetilde{m}_k &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(b(\widetilde{m}_k), \sigma_1^2(\widetilde{m}_k)), \ i \in \mathcal{I}_k; \quad \text{and} \\ \widetilde{m}_k(\mathbf{X}_i) \mid \widetilde{m}_k &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta - b(\widetilde{m}_k), \sigma_2^2(\widetilde{m}_k)), \ i \in \mathcal{J}_k. \end{aligned} \tag{7}$$

Using the same improper prior on the model parameters $\{\theta, b(\widetilde{m}_k), \sigma_1^2(\widetilde{m}_k), \sigma_2^2(\widetilde{m}_k)\}$ from (5), and applying Proposition 3.1 with $(\mathcal{S}, \mathcal{D})$ therein set as $(\mathcal{S}_k, \mathcal{D}_k)$, we derive the marginal posterior $\Pi_{\boldsymbol{\theta}}^{(k)}$ for θ as follows:

Proposition 3.2. *Given the model construction in (7) and the improper prior in (5), the marginal posterior distribution* $\Pi_{\boldsymbol{\theta}}^{(k)}$ *of* θ *given* $\{\mathcal{D}_k, \widetilde{m}_k\}$ *is a convolution of the t-distributions:* $t_{\nu_{n_K}}(\mu_{n_K}(\widetilde{m}_k), \widehat{\sigma}^2_{1,n_K}(\widetilde{m}_k)/n_K)$ *and* $t_{\nu_{N_K}}(\mu_{N_K}(\widetilde{m}_k), \widehat{\sigma}^2_{2,N_K}(\widetilde{m}_k)/N_K)$*, where the parameters are given by:* $\nu_{n_K} := n_K - 1$, $\nu_{N_K} := N_K - 1$,

$$\begin{aligned}
\mu_{n_K}(\widetilde{m}_k) &:= \frac{1}{n_K}\sum_{i\in\mathcal{I}_k}\{Y_i - \widetilde{m}_k(\mathbf{X}_i)\} \quad \textit{and} \quad \frac{\widehat{\sigma}^2_{1,n_K}(\widetilde{m}_k)}{n_K} := \frac{\sum_{i\in\mathcal{I}_k}[\{Y_i - \widetilde{m}_k(\mathbf{X}_i)\} - \mu_{n_K}(\widetilde{m}_k)]^2}{n_K(n_K-1)}; \\
\mu_{N_K}(\widetilde{m}_k) &:= \frac{1}{N_K}\sum_{i\in\mathcal{J}_k}\widetilde{m}_k(\mathbf{X}_i) \quad \textit{and} \quad \frac{\widehat{\sigma}^2_{2,N_K}(\widetilde{m}_k)}{N_K} := \frac{\sum_{i\in\mathcal{J}_k}\left\{\widetilde{m}_k(\mathbf{X}_i) - \mu_{N_K}(\widetilde{m}_k)\right\}^2}{N_K(N_K-1)}.
\end{aligned} \tag{8}$$

Consistent with our earlier notation, let $\widehat{\theta}^{(k)}_{\mathrm{BDM}}(\widetilde{m}_k)$ denote the posterior mean of $\Pi_{\boldsymbol{\theta}}^{(k)}$. From Proposition 3.2, we have $\widehat{\theta}^{(k)}_{\mathrm{BDM}}(\widetilde{m}_k) = \mu_{n_K}(\widetilde{m}_k) + \mu_{N_K}(\widetilde{m}_k)$ and it retains the same properties as $\widehat{\theta}_{\mathrm{BDM}}(\widetilde{m})$ from Section 3.2.

While sample splitting enables us to obtain the debiased representation in (1), which is crucial for the BDMI approach, it uses only a subset $\mathcal{D}_k$ of the full dataset $\mathcal{D}$ to obtain a posterior for θ. This causes a notable lack of efficiency. Since sample splitting produces K splits, each data fold pair $(\mathcal{S}_k, \mathcal{D}_k)$ can be utilized to obtain a posterior $\Pi_{\boldsymbol{\theta}}^{(k)}$ of θ for $k = 1, \ldots, K$. We now introduce a method for combining these posteriors of θ, referred to as *BDMI with cross-fitting* (BDMI-CF), to construct an *aggregated* full-data posterior for θ. This approach addresses the efficiency loss discussed earlier by fully utilizing the available data and ensuring that the variance and contraction rates of the final procedure depend directly on n, as shown in Theorem 4.2.

BDMI-CF is inspired by the frequentist cross-fitting (CF) idea (Chernozhukov et al., 2018), addressing challenges in high dimensional nuisance parameter estimation. The conventional CF approach has been used to (i) relax strong assumptions, e.g., Donsker class conditions (van der Vaart, 2000, Chapter 19), and (ii) make the sample splitting process efficient utilizing the full data in a 'cross-fitted' manner (Chernozhukov et al., 2018). CF techniques are widely used in the modern semi-parametric inference literature, where a combined estimator is obtained by averaging the estimators obtained from each split to regain full efficiency (Chernozhukov et al., 2018; Newey and Robins, 2018). In a *Bayesian* framework, however, additional care is required during the combination step, since entire *distributions (posteriors)* must be aggregated rather than point estimates. BDMI-CF addresses this issue by employing a consensus Monte Carlo-type approach (Scott et al., 2016) to suitably aggregate the posteriors from the sub-folds. This type of usage of cross-fitting (CF) for combining posteriors in Bayesian semi-parametric inference problems is not common. In the existing Bayesian literature, sample splitting has primarily been used to improve computational efficiency when handling large datasets (Scott et al., 2022). However, BDMI leverages sample splitting in a novel way: to ensure independence between the estimation of the nuisance parameter and the parameter of interest, and further via CF based aggregation, ensures efficient usage of the *entire* data. We now discuss the CF procedure.

Let $\theta_1, \ldots, \theta_K$ be independent random variables drawn from the corresponding posteriors $\Pi_{\boldsymbol{\theta}}^{(1)}, \ldots, \Pi_{\boldsymbol{\theta}}^{(K)}$ which are obtained from $(\mathcal{S}_1, \mathcal{D}_1), \ldots, (\mathcal{S}_K, \mathcal{D}_K)$, respectively. We then define a new

random variable:

$$\theta_{\text{BDM}} \;:=\; \frac{1}{K}\sum_{k=1}^{K}\theta_k, \quad \text{and let } \Pi_{\boldsymbol{\theta}} \text{ be the corresponding distribution of } \theta_{\text{BDM}}. \tag{9}$$

The distribution $\Pi_{\boldsymbol{\theta}}$ in (9) is referred to as the *final (aggregated)* posterior of θ from BDMI, specifically BDMI-CF. This final posterior $\Pi_{\boldsymbol{\theta}}$ is a (scaled) *convolution* of the posteriors $\Pi_{\boldsymbol{\theta}}^{(1)}, \ldots, \Pi_{\boldsymbol{\theta}}^{(K)}$ obtained from each data fold pair $(\mathcal{S}_1, \mathcal{D}_1), \ldots, (\mathcal{S}_K, \mathcal{D}_K)$. Hence, samples from $\Pi_{\boldsymbol{\theta}}$ can be easily generated by construction.

Further, by linearity of expectation, the posterior mean $\widehat{\theta}_{\text{BDM}}(\widetilde{m}_{\text{CF}})$ of $\Pi_{\boldsymbol{\theta}}$ is the average of the posterior means $\widehat{\theta}_{\text{BDM}}^{(1)}(\widetilde{m}_1), \ldots, \widehat{\theta}_{\text{BDM}}^{(K)}(\widetilde{m}_K)$ from the corresponding posteriors $\Pi_{\boldsymbol{\theta}}^{(1)}, \ldots, \Pi_{\boldsymbol{\theta}}^{(K)}$. More explicitly,

$$\widehat{\theta}_{\text{BDM}}(\widetilde{m}_{\text{CF}}) \;=\; \mu_n(\widetilde{m}_{\text{CF}}) + \mu_N(\widetilde{m}_{\text{CF}}) \;:=\; \frac{1}{n}\sum_{i=1}^{n}\left\{Y_i - \widetilde{m}_{\text{CF}}(\mathbf{X}_i)\right\} \;+\; \frac{1}{N}\sum_{i=n+1}^{n+N}\widetilde{m}_{\text{CF}}(\mathbf{X}_i), \tag{10}$$

where $\widetilde{m}_{\text{CF}}(\mathbf{X}_i) := \widetilde{m}_k(\mathbf{X}_i)$ for $i \in \mathcal{I}_k$ or $i \in \mathcal{J}_k$ where $\widetilde{m}_k$ is a random sample from the respective posterior $\Pi_{\mathbf{m}}^{(k)}$ of m for $k = 1, \ldots, K$. Naturally, we consider the posterior mean $\widehat{\theta}_{\text{BDM}}(\widetilde{m}_{\text{CF}})$ as a point estimator of θ_0. Furthermore, Theorem 4.2 guarantees the $\sqrt{n}$-consistency of $\widehat{\theta}_{\text{BDM}}(\widetilde{m}_{\text{CF}})$ as an estimator of θ_0. Detailed properties of $\widehat{\theta}_{\text{BDM}}(\widetilde{m}_{\text{CF}})$, and more generally the posterior $\Pi_{\boldsymbol{\theta}}$ in (9), are further examined in Section 4. We now present the final algorithm for our BDMI (specifically, BDMI-CF) approach in Algorithm 1.

Remark 3.2 (Discussion on Algorithm 1)**.** We first clarify that MC approximations are employed in Algorithm 1, particularly in the last step, to calculate posterior quantiles of θ. This involves using a sufficiently large number M of θ-samples to ensure that the statistical error margin dominates the MC error. Also, as detailed in Proposition 3.2, we calculated the posteriors $\{\Pi_{\boldsymbol{\theta}}^{(k)}\}_{k=1}^{K}$ for θ under the improper prior given in (5). Alternatively, users may pick a *different prior* (possibly non-conjugate) for $(\theta, b(\widetilde{m}_k), \sigma_1^2(\widetilde{m}_k), \sigma_2^2(\widetilde{m}_k))$. Using the same likelihood construction in (7), one can compute the posteriors $\{\Pi_{\boldsymbol{\theta}}^{(k)}\}_{k=1}^{K}$ of θ under the chosen prior. It is important to note that these posteriors $\{\Pi_{\boldsymbol{\theta}}^{(k)}\}_{k=1}^{K}$ would differ (possibly, not having a closed form) from those in Proposition 3.2. Despite such differences, one can *still* define a corresponding posterior mean $\widehat{\theta}_{\text{BDM}}(\widetilde{m}_{\text{CF}})$ (the average of the posterior means of the corresponding posteriors $\{\Pi_{\boldsymbol{\theta}}^{(k)}\}_{k=1}^{K}$) and use it as a valid point estimator for θ_0. When an exact expression for $\widehat{\theta}_{\text{BDM}}(\widetilde{m}_{\text{CF}})$ is unavailable (so (10) no longer holds), an MC average $M^{-1}\sum_{j=1}^{M}\theta_j$ of the M θ-samples (as obtained in Step 7 of Algorithm 1) can approximate $\widehat{\theta}_{\text{BDM}}(\widetilde{m}_{\text{CF}})$. To construct a $100 \times (1-\alpha)\%$ CI for θ_0, we still use MC approximations to calculate posterior quantiles of θ. Lastly, we highlight that BDMI provides a computationally efficient procedure for obtaining samples for θ. The primary computational cost lies in sampling from the nuisance posterior for m, as the remaining step of sampling θ from a convolution of two t-distributions is negligible. Moreover, by leveraging parallel computing, Steps 3–5 in Algorithm 1 can be executed in parallel to accelerate computation further.

Remark 3.3 (Recommendation for the choice of K)**.** As established in Section 4, the choice of K does *not* impact *asymptotic* properties or performance of BDMI-CF, provided that K is *fixed* (relative to n). However, in finite samples, K may influence performance and should be chosen carefully. The

Algorithm 1: The BDMI (with cross-fitting) procedure for SS mean estimation

Input: Data $\mathcal{D} = \mathcal{L} \cup \mathcal{U}$, K = the number of folds to use for CF, M = number of samples to draw from $\Pi_{\boldsymbol{\theta}}$ (the final posterior (9) from BDMI-CF), and the improper prior as in (5).

Output: Posterior samples $\theta_1 \ldots, \theta_M$ from $\Pi_{\boldsymbol{\theta}}$, the posterior mean $\widehat{\theta}_{\text{BDM}}(\widetilde{m}_{\text{CF}})$ as a point estimate of θ_0, and a $100 \times (1-\alpha)\%$ credible interval (CI) for θ_0, for a given $\alpha \in (0, 1)$.

Split $\mathcal{D}$ randomly into K disjoint sets: $(\mathcal{D}_k)_{k=1}^K \equiv (\mathcal{L}_k \cup \mathcal{U}_k)_{k=1}^K$, as in Section 3.3, and let $\mathcal{S}_k = \mathcal{L} \backslash \mathcal{L}_k$.

for $k = 1$ **to** K*:* **do**

- Pick any Bayesian (or frequentist) regression method to obtain a posterior $\Pi_{\mathbf{m}}^{(k)}$ for m based on $\mathcal{S}_k$.
- Draw *one* sample $\widetilde{m}^{(k)} \sim \Pi_{\mathbf{m}}^{(k)}$. Given $\widetilde{m}^{(k)}$, compute $\Pi_{\boldsymbol{\theta}}^{(k)}$ for θ based on $\mathcal{D}_k$ as in Proposition 3.2.
- Draw M many samples of θ from $\Pi_{\boldsymbol{\theta}}^{(k)}$: $\{\theta_1^{(k)}, \ldots, \theta_M^{(k)}\}$, for *each* $k = 1, \ldots, K$.

Obtain the *samples* $\theta_1, \ldots \theta_M \sim \Pi_{\boldsymbol{\theta}}$ as: $\theta_j := K^{-1} \sum_{k=1}^K \theta_j^{(k)}$ for $j = 1, \ldots, M$, using $\theta_j^{(k)}$ from Step 5.

Obtain $\widehat{\theta}_{\text{BDM}}(\widetilde{m}_{\text{CF}}) = \mu_n(\widetilde{m}_{\text{CF}}) + \mu_N(\widetilde{m}_{\text{CF}})$ as in (10) $\rightsquigarrow$ posterior mean of BDMI-CF *(point estimate)*.

Use the $(\alpha/2)^{th}$ and $(1-\alpha/2)^{th}$ sample quantiles of $\theta_1, \ldots \theta_M$ as a $(1-\alpha)$-level *CI* of θ_0 via BDMI-CF. (We use Monte Carlo (MC) approximations to calculate the posterior quantiles of θ using a sufficiently large number M of samples of θ so that the statistical error margin dominates the MC error.)

parameter K can be interpreted as a 'tuning parameter' that embodies the *variance-bias trade-off*. Specifically, as K increases, the training data size grows, leading to more stable nuisance estimation (reducing bias). However, this comes at the cost of smaller test data sizes, which may increase finite-sample variance. Thus, selecting K involves balancing these competing factors to achieve optimal performance. Based on extensive simulations under various settings (see Section 5), we observed that $K = 5$ or 10 generally provides (near-)optimal (and fairly robust) performance in terms of *both* estimation and inference. We therefore recommend such a K in practice.

Remark 3.4 (Choice of methods for the nuisance posterior $\Pi_{\mathbf{m}}$)**.** We conclude by discussing the choice of methods to obtain the nuisance posterior $\Pi_{\mathbf{m}}$. Firstly, BDMI is fully flexible in that it allows $\Pi_{\mathbf{m}}$ to be *any user-chosen off-the-shelf approach* that can be used *without* any modifications/adjustments to the posterior (or its prior). Therefore, it allows most standard Bayesian (or frequentist) regression approaches, *parametric* and *non-parametric*, provided they only satisfy some reasonable (and *high-level*) contraction conditions (formalized in Assumption 4.1). Parametric methods include traditional linear regression approaches such as Bayesian ordinary or ridge regression (corresponding to improper and Gaussian priors on the regression parameters), or their frequentist counterparts. Further, *sparsity* (or shrinkage) based parametric methods, commonly adopted in *high dimensional* settings can also be used, including sparse Bayesian linear regression based on spike-and-slab type priors (Mitchell and Beauchamp, 1988; George and McCulloch, 1993; Johnson and Rossell, 2012; Ročková and George, 2018) or continuous shrinkage priors (Carvalho et al., 2010; Bhattacharya et al., 2015), along with their frequentist counterparts such as LASSO (Hastie et al., 2015; Wainwright, 2019) or its variants. On the other hand, non-parametric methods may include Gaussian process regression (Williams, 1998), kernel smoothing-based methods (Tsybakov, 2009; Simonoff, 2012), reproducing kernel Hilbert space based methods (Berlinet and Thomas-Agnan, 2011), like smoothing splines (Green and Silverman, 1994), as well as modern black-box machine learning (ML) methods such as random forest (Breiman, 2001; Wager and Athey, 2018), Bayesian additive regression trees (BART) (Chipman et al., 2010), and neural networks (Specht, 1991; Farrell et al., 2021). These non-parametric methods are better suited for low dimensional (or fixed p) settings. Overall, BDMI affords notable flexibility to adapt to various modeling scenarios for $\Pi_{\mathbf{m}}$.

4 Theoretical properties of the BDMI procedure

In this section, we analyze in detail the theoretical underpinnings of our proposed BDMI procedure. Under mild regularity conditions, we show (in Theorems 4.1–4.2) that the BDMI posteriors $\{\Pi_{\boldsymbol{\theta}}^{(k)}\}_{k=1}^{K}$ (the 'one fold' versions) and $\Pi_{\boldsymbol{\theta}}$ (the final aggregated version via CF) all inherit BvM-type limiting behaviors with asymptotically Gaussian posteriors contracting around the true θ_0 at a $\sqrt{n}$-rate, along with various desirable properties on robustness, efficiency and nuisance insensitivity, which are all discussed in detail subsequently.

Assumption 4.1. We assume throughout that the number of folds K (for CF) is *fixed*. Further, we make the following *high-level* assumptions on the nuisance posterior $\Pi_{\mathbf{m}}$ (or its versions $\Pi_{\mathbf{m}}^{(k)}$ for any $k = 1, \ldots, K$):

(i) For any sample $\widetilde{m}_k \sim \Pi_{\mathbf{m}}^{(k)}(\cdot) \equiv \Pi_{\mathbf{m}}^{(k)}(\cdot; \mathcal{S}_k)$, we assume that $\|\widetilde{m}_k(\mathbf{X})\|_{\mathbb{L}_4(\mathbb{P}_{\mathbf{X}})} = O_{\mathbb{P}}(1)$ and $\|Y - \widetilde{m}_k(\mathbf{X})\|_{\mathbb{L}_4(\mathbb{P}_{\mathbf{Z}})} = O_{\mathbb{P}}(1)$, where $\mathbb{P}$ denotes the joint probability distribution $\Pi_{\mathbf{m}}^{(k)}(\mathcal{S}_k)$ for any $k = 1, \ldots, K$.

(ii) The posterior $\Pi_{\mathbf{m}}^{(k)}$ of m satisfies the *nuisance posterior contraction condition* (NPCC): $\Pi_{\mathbf{m}}^{(k)}$ contracts (at *some* rate a_n) around *some* non-random limiting function $m^*(\cdot) \in \mathbb{L}_2(\mathbb{P}_{\mathbf{X}})$ (with $m^*(\cdot)$ *not* necessarily equal to the true $m_0(\cdot)$). That is, for *some* (non-negative) sequence $a_n \to 0$, and for any $k = 1, \ldots, K$,

$$\Pi_{\mathbf{m}}^{(k)}\left[\{m : \|m(\mathbf{X}) - m^*(\mathbf{X})\|_{\mathbb{L}_2(\mathbb{P}_{\mathbf{X}})} > a_n\} \mid \mathcal{S}_k\right] \xrightarrow{\mathbb{P}} 0 \text{ under } \mathbb{P}_{\mathcal{S}_k}, \text{ as } n \to \infty. \tag{11}$$

Remark 4.1 (Discussion on Assumption 4.1)**.** The assumption on K and the condition (i) above are both fairly mild and reasonable. The condition (ii) is the *only* required assumption on the nuisance posterior $\Pi_{\mathbf{m}}^{(k)}$ for our Theorems 4.1–4.2. It embodies one of the key features of BDMI: it does *not* impose any restrictions on the distributional form or properties of $\Pi_{\mathbf{m}}^{(k)}$, nor the regression method (left entirely to the user's choice) used to obtain $\Pi_{\mathbf{m}}^{(k)}$. Typically, most of the existing Bayesian semi-parametric methods (Ray and van der Vaart, 2020; Luo et al., 2023; Breunig et al., 2025; Yiu et al., 2025) crucially rely on prior selection/modification or tailored posterior updates to mitigate nuisance estimation bias and achieve the $n^{-1/2}$ contraction rate for the target parameter. However, as Theorems 4.1–4.2 will demonstrate, the posterior convergence *rate* of θ and its *variability* are entirely unaffected by the posterior contraction rate and variability of $\Pi_{\mathbf{m}}^{(k)}$, or even the method used to obtain $\Pi_{\mathbf{m}}$, provided Assumption 4.1 holds (for a given m^*). This flexibility is largely due to our Bayesian debiasing approach presented in Section 3.1, and its exploitation under the Bayesian framework via targeted modeling of summary statistics, as in Section 3.2. It is worth noting that the condition (ii) is similar in spirit to $\mathbb{L}_2$-consistency conditions on nuisance estimators that (along with usage of CF) have become quite prevalent in the recent frequentist literature on debiased semi-parametric inference; see, e.g., Chernozhukov et al. (2018). The NPCC can be viewed as an appropriate (and suitable) *analogue* in the *Bayesian* framework.

Remark 4.2 (Examples of contraction rate a_n of the nuisance posterior $\Pi_{\mathbf{m}}$ and misspecification of $m_0(\cdot)$)**.** As detailed in Remark 3.4, Assumption 4.1 (ii) allows BDMI significant flexibility in accommodating a wide range of methods for estimating m. Specifically, $\Pi_{\mathbf{m}}^{(k)}$ can contract around a non-random function $m^*(\cdot)$, not necessarily equal to $m_0(\cdot)$, allowing misspecification. Further, regardless of $m^*(\cdot) = m_0(\cdot)$ or not (i.e., correctly specified or misspecified), the *posterior contraction rate* a_n of $\Pi_{\mathbf{m}}^{(k)}$ is *not* restricted, and it can be *any* rate that goes 0, potentially slower than the parametric rate (see Remark 4.3). For parametric methods in low-dimensional settings (p fixed or $p = o(n)$), contraction rates are typically $a_n = \sqrt{p/n}$. In high-dimensional settings ($p \gg n$), sparsity-based methods achieve rates of $a_n = \sqrt{s\log(p)/n}$, where s is the sparsity level of the regression parameter $\boldsymbol{\beta}$ (Wainwright, 2019). Non-parametric methods generally exhibit slower rates; for instance, kernel smoothing or smoothing splines achieve $a_n = n^{-q/(2q+p)}$, where q represents the smoothness level of $m_0(\cdot)$ (Tsybakov, 2009). Modern machine learning methods often achieve rates of $a_n = n^{-\alpha}$ for some $\alpha < 1/2$ (Chernozhukov et al., 2018). Finally, as noted above, BDMI remains robust even in misspecified cases, allowing for $\Pi_{\mathbf{m}}$ to contract around some function $m^*(\cdot) \neq m_0(\cdot)$. For instance, when $m_0(\cdot)$ is non-linear but a linear model is fitted, $\Pi_{\mathbf{m}}$ contracts around $m^*(\mathbf{X}) := \widetilde{\mathbf{X}}'\boldsymbol{\beta}^*$, where $\widetilde{\mathbf{X}} = (1, \mathbf{X}')'$ and $\boldsymbol{\beta}^* := \arg\min_{\boldsymbol{\beta}} \mathbb{E}\|Y - \widetilde{\mathbf{X}}'\boldsymbol{\beta}\|^2$ or equivalently, $\boldsymbol{\beta}^* = \{\mathbb{E}(\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}')\}^{-1}\mathbb{E}(\widetilde{\mathbf{X}}Y)$ and $m^*(\mathbf{X})$ is the *best linear predictor* of Y given $\mathbf{X}$, i.e., the $\mathbb{L}_2(\mathbb{P}_{\mathbf{X}})$-projection of $m_0(\cdot)$ onto the linear span of $\mathbf{X}$. This functional misspecification does not affect BDMI's ability to maintain $\sqrt{n}$-consistency/contraction for θ_0, as shown in Theorems 4.1–4.2.

Theorem 4.1. *Under Assumptions 2.1 and 4.1, the marginal posterior $\Pi_{\boldsymbol{\theta}}^{(k)}$ of θ (as in Proposition 3.2) obtained from one pair $(\mathcal{S}_k, \mathcal{D}_k)$ inherits a BvM-type limiting behavior as follows: for each $k = 1, \ldots, K$,*

$$\left\| \Pi_{\boldsymbol{\theta}}^{(k)} - \mathcal{N}\left(\widehat{\theta}_{\mathrm{BDM}}^{(k)}(m^*), \tau^2_{n_K, N_K}(m^*)\right) \right\|_{\mathrm{TV}} \xrightarrow{\mathbb{P}} 0 \text{ in probability under } \mathbb{P}_{\widetilde{\mathcal{D}}_k}, \text{ as } n, N \to \infty,$$

where, with $\sigma_1^2(m^*) := \mathrm{Var}_{\mathbf{Z}}\{Y - m^*(\mathbf{X})\}$ and $\sigma_2^2(m^*) := \mathrm{Var}_{\mathbf{X}}\{m^*(\mathbf{X})\}$, $\widehat{\theta}_{\mathrm{BDM}}^{(k)}(m^*)$ and $\tau^2_{n_K, N_K}(m^*)$ are:

$$\widehat{\theta}_{\mathrm{BDM}}^{(k)}(m^*) := \frac{1}{n_K} \sum_{i \in \mathcal{I}_k} \{Y_i - m^*(\mathbf{X}_i)\} + \frac{1}{N_K} \sum_{i \in \mathcal{J}_k} m^*(\mathbf{X}_i) \quad \text{and} \quad \tau^2_{n_K, N_K}(m^*) := \frac{\sigma_1^2(m^*)}{n_K} + \frac{\sigma_2^2(m^*)}{N_K}.$$

Further, let $h := \sqrt{n_K}(\theta - \theta_0)$ and $\Pi_{\mathbf{h}}^{(k)}$ be the posterior of h. Then, under Assumptions 2.1 and 4.1,

$$\left\| \Pi_{\mathbf{h}}^{(k)} - \mathcal{N}\left(\sqrt{n_K}\{\widehat{\theta}_{\mathrm{BDM}}^{(k)}(m^*) - \theta_0\}, n_K \tau^2_{n_K, N_K}(m^*)\right) \right\|_{\mathrm{TV}} \xrightarrow{\mathbb{P}} 0 \quad \text{in probability under } \mathbb{P}_{\widetilde{\mathcal{D}}_k}.$$

Theorem 4.2 (Main result)**.** *Under Assumptions 2.1 and 4.1, the final (aggregated) posterior $\Pi_{\boldsymbol{\theta}}$ of θ, as defined in (9), from the BDMI-CF procedure inherits a BvM-type limiting behavior as follows:*

$$\left\| \Pi_{\boldsymbol{\theta}} - \mathcal{N}(\widehat{\theta}_{\mathrm{BDM}}(m^*), \tau^2_{n,N}(m^*)) \right\|_{\mathrm{TV}} \xrightarrow{\mathbb{P}} 0 \quad \text{in probability w.r.t. } \mathbb{P}_{\mathcal{D}}, \text{ as } n, N \to \infty,$$

where $\widehat{\theta}_{\mathrm{BDM}}(m^*) := \mu_n(m^*) + \mu_N(m^*)$ as defined in (10) with $\widetilde{m}_{\mathrm{CF}}$ therein substituted by m^*, and $\tau^2_{n,N}(m^*) := \{\sigma_1^2(m^*)/n\} + \{\sigma_2^2(m^*)/N\}$ with $\sigma_1^2(m^*)$ and $\sigma_2^2(m^*)$ as defined in Theorem 4.1.

The BDMI-CF procedure provides the posterior $\Pi_{\boldsymbol{\theta}}$ with the posterior mean $\widehat{\theta}_{\mathrm{BDM}}(\widetilde{m}_{\mathrm{CF}})$ as defined in (10). Naturally, $\widehat{\theta}_{\mathrm{BDM}}(\widetilde{m}_{\mathrm{CF}})$ can be considered as a valid *SS point estimator* for θ_0. Beyond direct implications of Theorem 4.2, the asymptotic behavior of the SS estimator $\widehat{\theta}_{\mathrm{BDM}}(\widetilde{m}_{\mathrm{CF}})$ inherently is of separate interest. Towards that, in Corollary 4.1, we rigorously establish an asymptotically linear representation of $\widehat{\theta}_{\mathrm{BDM}}(\widetilde{m}_{\mathrm{CF}})$.

Corollary 4.1 (Asymptotically linear representation of the posterior mean $\widehat{\theta}_{\mathrm{BDM}}(\widetilde{m}_{\mathrm{CF}})$ of BDMI-CF)**.** *Under Assumptions 2.1 and 4.1, the posterior mean $\widehat{\theta}_{\mathrm{BDM}}(\widetilde{m}_{\mathrm{CF}})$ of $\Pi_{\boldsymbol{\theta}}$ as in (10) is asymptotically equivalent to the mean $\widehat{\theta}_{\mathrm{BDM}}(m^*)$ of the limiting distribution in Theorem 4.2 at a $1/\sqrt{n}$ rate. In particular,*

$$\begin{aligned} \sqrt{n}\,\{\,\widehat{\theta}_{\mathrm{BDM}}(\widetilde{m}_{\mathrm{CF}}) - \theta_0\,\} &= \sqrt{n}\,\{\,\widehat{\theta}_{\mathrm{BDM}}(m^*) - \theta_0\,\} + o_{\mathbb{P}_{\mathcal{D}}}(1) \\ &\equiv \sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}\{Y_i - m^*(\mathbf{X}_i)\} + \frac{1}{N}\sum_{i=n+1}^{n+N} m^*(\mathbf{X}_i) - \theta_0\right] + o_{\mathbb{P}_{\mathcal{D}}}(1). \end{aligned} \tag{12}$$

Remark 4.3 (Asymptotic properties of the posteriors $\Pi_{\boldsymbol{\theta}}^{(k)}$ and $\Pi_{\boldsymbol{\theta}}$)**.** Theorem 4.2 establishes a BvM-type result for the final BDMI-CF procedure presented in Section 3.3. Firstly, it shows that the posterior $\Pi_{\boldsymbol{\theta}}$ of θ behaves as Gaussian and concentrates around the true θ_0 at a rate $1/\sqrt{n}$ with $|\mathcal{L}| = n$. Importantly, while this rate is parametric in the labeled data size n, it is *non-standard* in the full data size $(n + N)$, particularly when $n/N \to 0$, making SS settings unique and their technical analyses substantially more challenging. Secondly, Theorem 4.2 demonstrates that for

large n, N, the posterior $\Pi_{\boldsymbol{\theta}}$ is approximately Normal with mean $\widehat{\theta}_{\text{BDM}}(m^*)$ and variance $\tau^2_{n,N}(m^*)$, which *matches* the asymptotic theory for corresponding existing frequentist approaches applied to the full data in recent SS inference literature (Zhang et al., 2019; Zhang and Bradic, 2022). Furthermore, it is important to note that all properties of the posterior $\Pi_{\boldsymbol{\theta}}$ discussed here, and all subsequent discussions in Section 4.1 below in the context of Theorem 4.2, also apply to Theorem 4.1 and $\Pi^{(k)}_{\boldsymbol{\theta}}$, with appropriate modifications for the one-fold data pair $(\mathcal{S}_k, \mathcal{D}_k)$ where $\mathcal{D}_k = \mathcal{L}_k \cup \mathcal{U}_k$ and $|\mathcal{L}_k| = n_K$. Since these extensions are straightforward and analogous, we refrain from restating them anywhere for brevity.

Remark 4.4 (Proof techniques and subtleties)**.** It is worth mentioning that while Theorems 4.1–4.2 have clear and strong implications, their proofs (deferred to the Supplement in the interest of space) are non-trivial, and involve a *synergy* of ideas and techniques from disparate literatures. Handling the theoretical underpinnings of BDMI and its key features: debiasing and the use of CF – both under a *Bayesian* framework – require bridging classical Bayesian tools/techniques for BvM-type results with those from the modern frequentist literature on debiased semi-parametric inference (Chernozhukov et al., 2018). Central to the proofs is the *interplay* between empirical process theory (along with CF), to handle the nuisance debiasing, and the *probabilistic structure* of Bayesian posteriors, to guarantee strong and nuisance-insensitive properties of BDMI while allowing $\Pi_{\mathbf{m}}$ to be generic throughout. In addition, the use of sample splitting and *posterior aggregation* via CF, though both crucial, introduce further technical subtleties that require novel adaptations under the Bayesian paradigm.

4.1 Robustness, efficiency and nuisance insensitivity of BDMI

Theorem 4.2 establishes that, under the SS setting, the posterior $\Pi_{\boldsymbol{\theta}}$ concentrates around the true parameter θ_0 at the parametric rate $1/\sqrt{n}$ (ensuring usage of the *full* data) and possesses *universal robustness* to the choice of the nuisance estimation method. This robustness manifests in two ways: (i) *global robustness* w.r.t. the limiting function $m^*(\cdot)$, ensuring that $\Pi_{\boldsymbol{\theta}}$ contracts around θ_0 at a rate $1/\sqrt{n}$ *regardless* of the contraction rate a_n of $\Pi_{\mathbf{m}}$ and *even if* $m^*(\cdot) \neq m_0(\cdot)$; and (ii) *insensitivity* to the nuisance estimation bias, as $\Pi_{\boldsymbol{\theta}}$ is *not* affected by slower convergence rates a_n of $\Pi_{\mathbf{m}}$, nor by $\Pi_{\mathbf{m}}$'s *own* first order properties like its shape, variability etc. (even after scaling by a_n). $\Pi_{\boldsymbol{\theta}}$ depends on $\Pi_{\mathbf{m}}$ *only* through its limit m^*, and validity/properties of $\Pi_{\boldsymbol{\theta}}$ as in Theorem 4.2 requires only $a_n \to 0$. Hence, BDMI effectively addresses the primary issue of the imputation approach (see Section 2.2), where nuisance estimation bias directly characterizes the first-order behavior/properties of the posterior for θ, and offers substantial *flexibility* in choosing regression methods to obtain $\Pi_{\mathbf{m}}$. In particular, it paves the way for using non-smooth or complex methods, like sparse regression (in high dimensions) or non-parametric ML methods, both of which may unavoidably have slow or unclear first order behaviors (refer to Remarks 3.4 and 4.2 for examples of these methods and their contraction rates). Moreover, BDMI-CF achieves *efficiency improvement* over the supervised approach based on $\mathcal{L}$, irrespective of whether $m^*(\cdot) = m_0(\cdot)$. While both $\Pi_{\boldsymbol{\theta}}$ and Π_{sup} converge to θ_0 at the parametric rate $1/\sqrt{n}$, the variance $\tau^2_{n,N}(m^*)$ of the limiting distribution is *always smaller* than the variance of the supervised approach as we will show in Remark 4.5, and further achieves the *semi-parametric efficiency bound* when $m^*(\cdot) = m_0(\cdot)$ (correctly specified case). These results align with frequentist asymptotic theory in recent SS inference literature (Zhang et al., 2019; Zhang and Bradic, 2022). Moreover, these desirable properties of $\Pi_{\boldsymbol{\theta}}$ also naturally extend to posterior summaries. In particular, the *posterior mean* $\widehat{\theta}_{\text{BDM}}(\widetilde{m}_{\text{CF}})$, as a valid SS point estimator

of θ_0, inherits these properties. As Corollary 4.1 shows, it remains $\sqrt{n}$-consistent, asymptotically Normal, and asymptotically linear regardless of the nuisance estimation method, and its expansion is unaffected by the estimation bias/error of the nuisance, showing its first-order insensitivity. Finally, its asymptotic variance also equals the posterior variance $\tau^2_{n,N}(m^*)$ (see Remark 4.5 below), ensuring *valid and accurate inference* for θ_0.

Remark 4.5 (Variance comparison)**.** Theorem 4.2 establishes that the posterior $\Pi_{\boldsymbol{\theta}}$ is asymptotically Normal with mean $\widehat{\theta}_{\text{BDM}}(m^*)$ and variance $\tau^2_{n,N}(m^*)$, which is also the variance of $\widehat{\theta}_{\text{BDM}}(m^*)$. Specifically, using the definition of $\widehat{\theta}_{\text{BDM}}(m^*)$ in Theorem 4.2, and due to the independence between $\mathcal{L}$ and $\mathcal{U}$, we have:

$$\text{Var}\{\widehat{\theta}_{\text{BDM}}(m^*)\} \;=\; \frac{\text{Var}\{Y - m^*(\mathbf{X})\}}{n} + \frac{\text{Var}\{m^*(\mathbf{X})\}}{N} \;\equiv\; \frac{\sigma_1^2(m^*)}{n} + \frac{\sigma_2^2(m^*)}{N} \;=\; \tau^2_{n,N}(m^*). \quad (13)$$

This equality is crucial for ensuring valid inference for θ_0. Using the asymptotic equivalence in Corollary 4.1, we can consider the asymptotic variance of $\widehat{\theta}_{\text{BDM}}(m^*)$ to *compare* the asymptotic variance of $\widehat{\theta}_{\text{BDM}}(\widetilde{m}_{\text{CF}})$ with the asymptotic variance of $\widehat{\theta}_{\text{sup}} \equiv \overline{Y}$ (based on $\mathcal{L}$). Further, for any non-random $g(\cdot) \in \mathbb{L}_2(\mathbb{P}_{\mathbf{X}})$, we have:

$$\begin{aligned} \sigma^2_{\text{sup}} \;&\equiv\; \lim_{n\to\infty} \text{Var}[\sqrt{n}\{\widehat{\theta}_{\text{sup}} - \theta_0\}] \;=\; \text{Var}(Y) \\ &=\; \text{Var}\{Y - g(\mathbf{X})\} \;+\; \text{Var}\{g(\mathbf{X})\} \;+\; 2\text{Cov}\{Y - g(\mathbf{X}), g(\mathbf{X})\}. \end{aligned} \quad (14)$$

Under Assumption 2.1 (i), where $\lim_{n,N\to\infty} n/N \to c \in [0, 1)$ and setting $g(\cdot) = m^*(\cdot)$ in (14), we obtain

$$\begin{aligned} \sigma^2_{\text{BDM}} \equiv \lim_{n,N\to\infty} \text{Var}\big[\sqrt{n}\{\widehat{\theta}_{\text{BDM}}(\widetilde{m}_{\text{CF}}) - \theta_0\}\big] &= \lim_{n,N\to\infty} \text{Var}\big[\sqrt{n}\{\widehat{\theta}_{\text{BDM}}(m^*) - \theta_0\}\big] = \lim_{n,N\to\infty} \tau^2_{n,N}(m^*) \\ = \text{Var}\{Y - m^*(\mathbf{X})\} + c\text{Var}\{m^*(\mathbf{X})\} \;&\equiv\; \sigma_1^2(m^*) + c\sigma_2^2(m^*) \le \sigma_1^2(m^*) + \sigma_2^2(m^*) = \sigma^2_{\text{sup}}. \end{aligned} \quad (15)$$

This inequality holds if either: (i) $m^*(\mathbf{X}) = m_0(\mathbf{X})$ (i.e., correctly specified model), or (ii) $m^*(\mathbf{X}) \neq m_0(\mathbf{X})$ (misspecified model) but $\text{Cov}\{Y - m^*(\mathbf{X}), m^*(\mathbf{X})\} = 0$. Moreover, the inequality in (15) is *strict* unless $m^*(\cdot)$ is a constant function. Hence, in either case, the SS estimator $\widehat{\theta}_{\text{BDM}}(\widetilde{m}_{\text{CF}})$ *outperforms* the supervised estimator $\widehat{\theta}_{\text{sup}}$ in terms of (asymptotic) variance and efficiency (see Table 1). Finally, note that the condition $\text{Cov}\{Y - m^*(\mathbf{X}), m^*(\mathbf{X})\} = 0$, represents a natural requirement on *orthogonality (in the population)* between the model-based predictions/target function $m^*(\mathbf{X})$ and the residuals $\{Y - m^*(\mathbf{X})\}$. This condition is satisfied by most reasonable regression procedures, including least squares-type methods, where the target functions (even if they are misspecified) $m^*(\cdot)$ *can be viewed as the* $\mathbb{L}_2(\mathbb{P}_{\mathbf{X}})$*-projection of* $m_0(\cdot)$ *onto the working model space*. For correctly specified models, i.e., $m^*(\cdot) = m_0(\cdot)$, this condition, of course, holds trivially.

Remark 4.6 (Adapting BDMI when $N < n$)**.** Our main focus is on scenarios where N is substantially larger than n, as reflected in Assumption 2.1 (i): $\lim_{n,N\to\infty} n/N = c \in [0, 1)$. BDMI – in its current form – requires $c < 1$ (i.e., $N > n$) to guarantee efficiency improvement, as Remark 4.5 shows. While it still applies when $N < n$, the improvement is not guaranteed. However, it is theoretically possible to *adapt* BDMI to guarantee it even if $c > 1$ (i.e., $N < n$) as well, by slightly modifying our modeling and likelihood construction (3)–(4) in Section 3.2. The primary reason behind this

Table 1: Full characterization of efficiency improvement with BDMI and its robustness in terms of rate and the pair $(\Pi_{\mathbf{m}}, m^*)$.

Comparison of the supervised versus SS estimators (BDMI) regarding efficiency and robustness			
Estimators	Rate of convergence	Limiting distributions	Asymptotic variance comparison
Supervised estimator: $\widehat{\theta}_{\text{sup}}$	$\frac{1}{\sqrt{n}}$	$\mathcal{N}(\theta_0, \frac{\sigma^2_{\text{sup}}}{n})$	$\sigma^2_{\text{sup}} = \sigma^2_1(m^*) + \sigma^2_2(m^*) + 2\text{Cov}[Y - m^*(\mathbf{X}), m^*(\mathbf{X})]$
SS estimator with BDMI: $\widehat{\theta}_{\text{BDM}}(\widetilde{m}_{\text{CF}})$	$\frac{1}{\sqrt{n}}$	$\mathcal{N}(\theta_0, \frac{\sigma^2_{\text{BDM}}}{n})$	$\sigma^2_{\text{BDM}} \equiv \sigma^2_1(m^*) + c\,\sigma^2_2(m^*) \leq \sigma^2_{\text{sup}}$ [see (15)] if either: **(i)** $m^*(\mathbf{X}) = m_0(\mathbf{X})$, or **(ii)** $m^*(\mathbf{X}) \neq m_0(\mathbf{X})$ and $\text{Cov}\{Y - m^*(\mathbf{X}), m^*(\mathbf{X})\} = 0$ hold. (**Note:** *Strict* inequality unless $m^*(\cdot)$ is constant.)

'discontinuity' (in behavior w.r.t. c) is due to the second model in (3) for $\widetilde{m}(\mathbf{X}_i)$ being considered over $\mathbf{X}_i \in \mathcal{U}$ $(i = n+1, \ldots, n+N)$ only. One may alternatively consider this for $\mathbf{X}_i$'s over the *entire* $\mathcal{D} \equiv \mathcal{L} \cup \mathcal{U}$. Our current approach conveniently ensures that the two components (from the two models) forming the product in the likelihood (4) are actually based on *independent* sources of data, $\mathcal{L}$ and $\mathcal{U}$, ensuring the likelihood's probabilistic validity as a *joint* likelihood, and that θ and $b(\widetilde{m})$ can be learnt *simultaneously*. On the other hand, if the second component now includes all $\mathbf{X}_i \in \mathcal{D}$ $(i = 1, \ldots, n+N)$, then this product formulation is lost and one needs to consider an alternative *hierarchical* approach to learn the two parameters, as follows. For ease of exposition here, we keep the hyperparameters σ^2_1 and σ^2_2 implicit in the notations below. Let $L_1\{b(\widetilde{m}); \mathcal{L}\}$ denote the first component in the likelihood (4). Then, we *first* learn a posterior for $b(\widetilde{m})$ based on $L_1(\cdot)$, and then *given* a sample of $b(\widetilde{m})$, we learn $\theta \mid b(\widetilde{m})$ hierarchically using the 'conditional' likelihood $L_2\{\theta; \mathcal{D} \mid b(\widetilde{m})\}$, where $L_2(\cdot)$ is the *modified version* of the second component in (4) with *all* the $\mathbf{X}'_i s \in \mathcal{D}$ being now included (i.e., $i = 1, \ldots, n+N$). Collecting samples of $b(\widetilde{m})$ and $\theta \mid b(\widetilde{m})$ across this hierarchical approach eventually leads to the final posterior. Though technically more nuanced and also computation-intensive, this approach can be shown to have all the desirable properties of BDMI, while also allowing for $c > 1$. Nevertheless, given that our general focus is mostly on cases where $N \gg n$, we prefer to stick to our original BDMI formulation due to its simplicity, both technically and computationally.

4.2 A hierarchical variant of BDMI: h-BDMI

Recall that the original BDMI procedure, as described in Section 3, is constructed using a *single* random sample $\widetilde{m} \sim \Pi_{\mathbf{m}}$. Alternatively, a more traditional Bayesian approach can be adapted by considering multiple samples of $\widetilde{m}$ through a hierarchical construction, as briefly mentioned in Remark 3.1. This section presents this alternative version of BDMI, referred to as the *hierarchical-BDMI* (henceforth h-BDMI), which constructs a joint posterior of (θ, m) and then marginalizes over m to obtain the marginal posterior of θ. This differs from the original BDMI procedure, and h-BDMI aligns more closely with traditional hierarchical Bayesian modeling principles. For its

exposition, we focus on only one data fold, say $\widetilde{\mathcal{D}}_k := \mathcal{D}_k \cup S_k$, where $\mathcal{D}_k$ and $\mathcal{S}_k$ are as defined in Section 3.3 for some $k = 1, \ldots, K$. Following the conventional Bayesian idea of integrating out the nuisance parameter m, we proceed as follows. Using $\mathcal{S}_k$ as a training data, we obtain a posterior $\Pi_{\mathbf{m}}^{(k)} \equiv \Pi_{\mathbf{m}}^{(k)}(\cdot; \mathcal{S}_k)$ for m. By the conditional independence between $m \sim \Pi_{\mathbf{m}}^{(k)}$ and $\mathcal{D}_k$, the joint posterior of (θ, m) has the pdf $\pi(\theta, m \mid \widetilde{\mathcal{D}}_k) \;=\; \pi(\theta \mid m, \mathcal{D}_k)\, \pi_{\mathbf{m}}^{(k)}(m)$, where $\pi_{\mathbf{m}}^{(k)}(\cdot)$ is the pdf of the nuisance posterior $\Pi_{\mathbf{m}}^{(k)}$ of m. The pdfs $\pi(\theta \mid m, \mathcal{D}_k)$ and $\pi_{\mathbf{m}}^{(k)}(m)$ remain as defined in Section 3.3. By integrating out m, we obtain the marginal posterior of θ, denoted $\widetilde{\Pi}_{\boldsymbol{\theta}}^{(k)}$, with corresponding pdf $\pi_{\boldsymbol{\theta}}^{(k)}(\cdot)$, based on h-BDMI as follows:

$$\pi_{\boldsymbol{\theta}}^{(k)}(\theta) \;=\; \int \pi\big(\theta, m \mid \widetilde{\mathcal{D}}_k\big)\, \mathrm{d}m \;=\; \int \pi\big(\theta \mid m, \mathcal{D}_k\big)\, \pi_{\mathbf{m}}^{(k)}(m)\, \mathrm{d}m.$$

Estimation and inference on the true parameter $\theta_0 \equiv \mathbb{E}(Y)$ using h-BDMI can be performed based on this posterior $\widetilde{\Pi}_{\boldsymbol{\theta}}^{(k)}$, for any $k = 1, \ldots, K$. Using iterated expectations, the posterior mean $\widehat{\theta}_{\mathrm{hBDM}}^{(k)} \equiv \mathbb{E}_{\theta \sim \widetilde{\Pi}_{\boldsymbol{\theta}}^{(k)}}(\theta)$ of $\widetilde{\Pi}_{\boldsymbol{\theta}}^{(k)}$ can be expressed as $\widehat{\theta}_{\mathrm{hBDM}}^{(k)} \;\equiv\; \int \big\{ \int \theta\, \pi(\theta \mid m, \mathcal{D}_k)\, \mathrm{d}\theta \big\}\, \pi_{\mathbf{m}}^{(k)}\, \mathrm{d}m$, where the inner integral is the conditional mean of θ given m and $\mathcal{D}_k$, i.e., $\mathbb{E}(\theta \mid m, \mathcal{D}_k)$. Under the prior choice in (5), $\widehat{\theta}_{\mathrm{hBDM}}^{(k)}$ is explicitly given by:

$$\widehat{\theta}_{\mathrm{hBDM}}^{(k)} \;\equiv\; \frac{1}{n_K} \sum_{i \in \mathcal{I}_k} \big\{ Y_i - \widehat{m}^{(k)}(\mathbf{X}_i) \big\} + \frac{1}{N_K} \sum_{i \in \mathcal{J}_k} \widehat{m}^{(k)}(\mathbf{X}_i), \quad \text{where } \widehat{m}^{(k)} \text{ is the posterior mean of } \Pi_{\mathbf{m}}^{(k)}.$$

Also, it is easy to draw samples from the posterior $\widetilde{\Pi}_{\boldsymbol{\theta}}^{(k)}$ of θ to construct credible intervals. Specifically, for sufficiently large M, we first draw samples $\widetilde{m}_1, \ldots, \widetilde{m}_M$ from the posterior $\Pi_{\mathbf{m}}^{(k)}$, and for each sample, we draw a sample $\theta \mid \widetilde{m}_j, \mathcal{D}_k$ from the posterior $\Pi_{\boldsymbol{\theta}}^{(k)}$ as described in Proposition 3.2 for $j = 1, \ldots, M$. This process yields M samples of θ from the posterior $\widetilde{\Pi}_{\boldsymbol{\theta}}^{(k)}$. Finally, applying the h-BDMI procedure to each $\widetilde{\mathcal{D}}_1, \ldots, \widetilde{\mathcal{D}}_K$, we obtain the corresponding posteriors $\widetilde{\Pi}_{\boldsymbol{\theta}}^{(1)}, \ldots \widetilde{\Pi}_{\boldsymbol{\theta}}^{(K)}$. Following the aggregation approach detailed in Section 3.3, we can construct a CF-based aggregated posterior $\widetilde{\Pi}_{\boldsymbol{\theta}}$. These modifications can be incorporated into Algorithm 1, which we omit for brevity. We next present the result on the theoretical properties of h-BDMI on one data fold $\widetilde{\mathcal{D}}_k$, followed by a discussion on the differences between BDMI and h-BDMI.

Theorem 4.3. *Suppose Assumptions 2.1 and 4.1 hold, except that the NPCC (11) in Assumption 4.1 (ii) is replaced with a modified NPCC as follows: assume that the posterior $\Pi_{\mathbf{m}}^{(k)}$ of m satisfies the nuisance Bayes risk condition: $\mathbb{E}_{m \sim \Pi_{\mathbf{m}}^{(k)}} \{ \|m(\mathbf{X}) - m^*(\mathbf{X})\|_{\mathbb{L}_2(\mathbb{P}_{\mathbf{X}})}^2 \mid \mathcal{S}_k \} \xrightarrow{\mathbb{P}} 0$ under $\mathbb{P}_{\mathcal{S}_k}$. Then, the posterior $\widetilde{\Pi}_{\boldsymbol{\theta}}^{(k)}$ of θ from the h-BDMI procedure on any pair $(\mathcal{D}_k, \mathcal{S}_k)$ as above inherits a BvM-type limiting behavior as follows:*

$$\Big\| \widetilde{\Pi}_{\boldsymbol{\theta}}^{(k)} - \mathcal{N}\big(\widehat{\theta}_{\mathrm{BDM}}^{(k)}(m^*), \tau_{n_K, N_K}^2(m^*)\big) \Big\|_{\mathrm{TV}} \;\xrightarrow{\mathbb{P}}\; 0 \quad \text{in probability w.r.t. } \mathbb{P}_{\widetilde{\mathcal{D}}_k} \quad \text{as} \quad n, N \to \infty, \tag{16}$$

for any $k = 1, \ldots, K$, where $\widehat{\theta}_{\mathrm{BDM}}^{(k)}(m^*)$ and $\tau_{n_K, N_K}^2(m^*)$ are the same as defined in Theorem 4.1.

We conclude this section with a brief comparison between the BDMI and h-BDMI approaches. Firstly, Theorem 4.3 establishes a corresponding BvM-type result for h-BDMI, similar to Theorem 4.1 for the 'one data fold' version of the original BDMI procedure described in Section 3.3. While

both theorems demonstrate that the marginal posteriors of θ inherit a BvM-type limiting behavior with the same limiting posterior, they *do* have some important differences. Notably, Theorem 4.3 requires a *stronger* L_1-type (Bayes risk) convergence condition on the contraction of the posterior $\widetilde{\Pi}_{\mathbf{m}}^{(k)}$ around $m^*(\cdot)$, while Theorem 4.1 relies on the much weaker in-probability type condition (11). In practice, our simulation results in Section 5 reveal that the difference between BDMI and h-BDMI is less pronounced. In most cases, the two methods perform similarly in estimating θ_0, as illustrated in Table 2. Occasionally, h-BDMI tends to give slightly conservative coverages compared to BDMI (see Table 3), which is not unexpected since h-BDMI involves multiple samples (hence more noise) as it integrates out the nuisance parameter m rather than conditioning on a single draw.

A key advantage of BDMI lies in its simplicity and computational efficiency. Unlike h-BDMI, which requires multiple samples from the nuisance posterior $\Pi_{\mathbf{m}}$ of m, BDMI relies on only a single sample, reducing computation burden. Thus, we recommend the original BDMI approach for achieving *both* efficient estimation and reliable inference for the true parameter θ_0. For further details and discussions, we refer to Section 5.

Finally, while we have used a 'one fold' version of h-BDMI here for clarity, it also admits a CF-based full data version ('h-BDMI-CF', if we may) analogous to the BDMI-CF procedure in Section 3.3. This version inherits similar theoretical properties as Theorem 4.2 (with the same distinctions as above). In our simulations in Section 5, we implemented h-BDMI via this CF-based full data version to ensure a fair comparison with the BDMI-CF and supervised approaches. The notation 'h-BDMI' therein refers to this CF-based version.

5 Numerical studies

We conducted extensive simulation studies to investigate the finite sample performance, both in estimation and inference, for our proposed SS approach(es) and the supervised approach under various settings. In particular, as point estimators, we compare the supervised estimator $\widehat{\theta}_{\text{sup}} \equiv \overline{Y}$ based on $\mathcal{L}$, the posterior mean $\widehat{\theta}_{\text{BDM}} \equiv \widehat{\theta}_{\text{BDM}}(\widetilde{m}_{\text{CF}})$ of $\Pi_{\boldsymbol{\theta}}$ from the final BDMI-CF procedure (as in Algorithm 1) and the posterior mean $\widehat{\theta}_{\text{hBDM}}$ of $\widetilde{\Pi}_{\boldsymbol{\theta}}$ from the h-BDMI procedure (its CF based version) discussed in Section 4.2. We compare their estimation efficiencies based on the empirical mean squared error (**MSE**) and report their relative efficiencies (**RE**) compared to the supervised estimator $\widehat{\theta}_{\text{sup}}$. Further, for evaluating the accuracy of inference, we report the empirical coverage probabilities (**CovP**) and lengths (**Len**) of the 95% credible intervals (**CI**s) obtained from their respective posteriors. Finally, as a performance benchmark for estimation efficiency, we also report the maximum (oracle) asymptotic relative efficiency (**ORE**) relative to $\widehat{\theta}_{\text{sup}}$, given by $\text{Var}(\widehat{\theta}_{\text{sup}})/\tau_{n,N}^2(m^*)$, where $\tau_{n,N}^2(m^*) = \text{Var}\{Y - m^*(\mathbf{X})\}/n + \text{Var}\{m^*(\mathbf{X})\}/N$ with $m^*(\cdot) = m_0(\cdot)$. For the choice of the number of folds K, we consider $K = 5$ and 10. The reported simulation results are all based on 500 replications. We examine various true data generating mechanisms and different methods for nuisance parameter estimation, leading to both correctly specified and misspecified models for $m_0(\cdot)$. We discuss the correctly specified and misspecified model settings and their corresponding results in Sections 5.1– 5.2.

5.1 Simulation studies: Correctly specified models

Throughout, we set $n = 500$ and $N = 10000$, and considered $p = 50$ and $p = 166$ ($\approx n/3$), representing moderate and high dimensional settings (relative to n), respectively. We generated

$\mathbf{X} \sim \mathcal{N}_p(\mathbf{0}_p, I_p)$, and given $\mathbf{X} = \mathbf{x}$, we generated $Y \sim \mathcal{N}(m_0(\mathbf{x}), \sigma_0^2)$, where $m_0(\mathbf{x}) = \alpha_0 + \mathbf{x}'\boldsymbol{\beta}_0$ and $\sigma_0^2 = \text{Var}\{m_0(\mathbf{X})\}/5$, and we used $\alpha_0 = 5$ and $\boldsymbol{\beta}_0 = (\mathbf{1}'_{s/2}, \mathbf{0.5}'_{s/2}, \mathbf{0}'_{p-s})'$ (for different choices of s discussed below). Here, $\mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the d-variate ($d \geq 2$) Gaussian distribution with mean $\boldsymbol{\mu}_{d\times 1}$ and covariance matrix $\boldsymbol{\Sigma}_{d\times d}$, I_d denotes the identity matrix of order d, and the notation $\mathbf{a}_l$, for any positive integer l (e.g., $l = p$, $s/2$ or $p - s$, as above), denotes the vector $(a, \ldots, a)'_{l\times 1}$ for any $a \in \mathbb{R}$ (e.g., $a = 0$, 0.5 or 1, as above). The parameter s in $\boldsymbol{\beta}_0$ above denotes the *sparsity* of $\boldsymbol{\beta}_0$. For $p = 50$, we set $s = 7$ ($\approx \sqrt{p}$), or $s = 50 \equiv p$; while for $p = 166$, we set $s = 13$ ($\approx \sqrt{p}$), $s = 55$ ($\approx p/3$), $s = 83$ ($\approx p/2$), or $s = 166 \equiv p$. These choices of s span a variety of settings, including *sparse* ($s = \sqrt{p}$), *moderately dense* ($s = p/2$ or $p/3$), or *fully dense* ($s = p$) cases. Note that, except for the sparse case, none of these choices correspond to settings where s (or p) may be considered small or fixed relative to n, and therefore appropriate sparsity-friendly nuisance estimation methods may *still* fail to consistently estimate m_0. For illustrative purposes, we consider *three choices* (all parametric model based) for obtaining the nuisance posterior $\Pi_{\mathbf{m}}$: Bayesian ordinary linear regression (`Bols`), Bayesian ridge regression (`Bridge`), and a sparse Bayesian linear regression method (`Bsparse`) based on non-local priors (NLP) (Johnson and Rossell, 2012). In all cases, we consider the Gaussian linear regression model $Y_i \mid \mathbf{X}_i, \alpha, \boldsymbol{\beta}, \sigma \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\alpha + \mathbf{X}_i'\boldsymbol{\beta}, \sigma^2)$ for $i = 1, \ldots, n$. For `Bols`, we use a prior on $(\alpha, \boldsymbol{\beta}, \sigma^2)$ given by: $\pi(\alpha, \boldsymbol{\beta} \mid \sigma^2) \propto 1$ and $\pi(\sigma^2) \propto (\sigma^2)^{-1}$; and for `Bridge`, the prior employed on $(\alpha, \boldsymbol{\beta}, \sigma^2)$ is: $\pi(\alpha \mid \sigma^2) \propto 1$, $\boldsymbol{\beta} \mid \lambda, \sigma^2 \sim \mathcal{N}_p(\mathbf{0}_p, \lambda^{-1}\sigma^2 I_p)$, with α and β being independent, and $\pi(\sigma^2) \propto (\sigma^2)^{-1}$. We use an empirical Bayes approach to plug in a point estimate $\widehat{\lambda}$ for the prior precision (or ridge) parameter λ. The estimate $\widehat{\lambda}$ is obtained from the `R` package `glmnet` so that the posterior mean of $(\alpha, \boldsymbol{\beta}')' \in \mathbb{R}^{(p+1)}$ coincides with the cross-validated point estimate obtained from `cv.glmnet` in the `glmnet` package. For both these methods, we obtain that the posteriors of $(\alpha, \boldsymbol{\beta})$ are multivariate t-distributions. For the `Bsparse` method, we use the R package `mombf` to obtain posterior samples for $(\alpha, \boldsymbol{\beta})$. The implementation details of the `mombf` and `glmnet` packages are provided in Section S3 of the Supplementary Material.

Table 2 and Tables 3–4 present the results on estimation efficiency and inference, respectively, along with illustrations of the posteriors and their overall behaviors in Figures 1–2. As seen from Table 2 (as well as the box plots in Figures 1–2), the REs of $\widehat{\theta}_{\text{BDM}}$ and $\widehat{\theta}_{\text{hBDM}}$ w.r.t. $\widehat{\theta}_{\text{sup}}$, i.e., $\text{MSE}(\widehat{\theta}_{\text{sup}})/\text{MSE}(\widehat{\theta}_{\text{BDM}})$ and $\text{MSE}(\widehat{\theta}_{\text{sup}})/\text{MSE}(\widehat{\theta}_{\text{hBDM}})$, are consistently greater than 1, ranging roughly between 2 to 5 across most settings. This highlights the substantial efficiency improvement achieved by BDMI over the supervised approach. In addition, as illustrated in Figures 1–2, apart from point estimators, the *posteriors themselves* are consistently and significantly *tighter* than the supervised posteriors, while throughout resembling a Gaussian behavior centered at the true θ_0. These patterns hold generally *regardless* of the setting and/or the nuisance posterior.

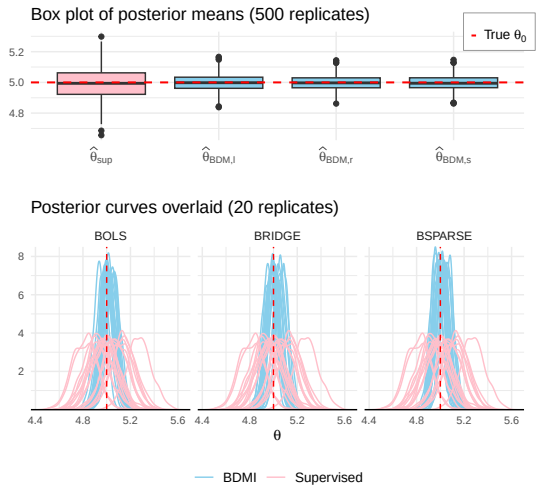
Furthermore, Table 2 illustrates that the efficiency improvement depends primarily on the dimensionality p and the sparsity level s. In moderate-dimensional settings ($p = 50$), BDMI achieves (near-)optimal efficiency gains, with RE values close to each other and approaching the ORE value, regardless of the sparsity levels (sparse $s = \sqrt{p}$ and fully dense $s = p$). This confirms that BDMI performs *optimally* when the model is correctly specified *and* estimated well enough. The impact of the sparsity level becomes particularly apparent in high-dimensional scenarios ($n = 500$ with $p = 166$), where finite-sample nuisance estimation *bias* introduces a *soft* form of misspecification. Specifically, sparsity-friendly nuisance models (e.g., `Bsparse`) struggle to consistently estimate $m_0(\cdot)$ in moderately dense ($s = p/2$) or fully dense ($s = p$) settings, leading to somewhat lower RE values. However, in sparse settings $s = \sqrt{p}$ ($s = 13$), `Bsparse` achieves RE values that are close to ORE by leveraging the underlying sparse structure, outperforming non-sparse methods such

Table 2: Relative efficiency (RE) of $\widehat{\theta}_{\text{BDM},i}$ and $\widehat{\theta}_{\text{hBDM},i}$ relative to $\widehat{\theta}_{\text{sup}}$, w.r.t. their empirical MSEs, for the settings in Section 5.1, where the methods (the subscript "i") used to obtain the nuisance posterior $\Pi_{\mathbf{m}}$ for BDMI are denoted as: $l =$ `Bols`, $r =$ `Bridge` and $s =$ `Bsparse`. **Settings:** $n = 500$, $N = 10000$, and: (i) $p = 50$, with $s = 7$ or 50; or (ii) $p = 166$, with $s = 13, 55, 83$ or 166. (As a performance benchmark, we also report the maximum (oracle) asymptotic relative efficiency (ORE) relative to $\widehat{\theta}_{\text{sup}}$.)

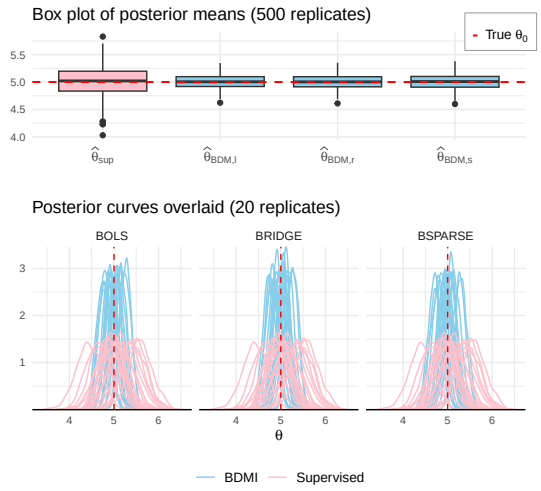
			$\widehat{\theta}_{\text{sup}}$		$\widehat{\theta}_{\text{BDM},l}$	$\widehat{\theta}_{\text{BDM},r}$	$\widehat{\theta}_{\text{BDM},s}$	$\widehat{\theta}_{\text{hBDM},l}$	$\widehat{\theta}_{\text{hBDM},r}$	$\widehat{\theta}_{\text{hBDM},s}$	
p	s	K	MSE	RE	RE	RE	RE	RE	RE	RE	**ORE**
50	7	5	0.01	1.00	3.99	4.38	5.00	4.61	4.69	5.06	4.80
		10	0.01	1.00	4.12	4.73	5.04	4.67	4.72	5.08	4.80
50	50	5	0.08	1.00	4.31	4.38	3.88	4.33	4.35	4.22	4.80
		10	0.08	1.00	4.35	4.41	4.02	4.37	4.42	4.30	4.80
166	13	5	0.02	1.00	2.84	3.46	4.56	3.30	3.62	4.75	4.80
		10	0.02	1.00	3.17	3.61	4.70	3.64	3.88	4.81	4.80
166	55	5	0.09	1.00	3.02	3.48	3.15	3.08	3.47	3.42	4.80
		10	0.09	1.00	3.45	3.83	3.41	3.49	3.81	3.78	4.80
166	83	5	0.13	1.00	3.01	3.28	1.40	3.03	3.31	1.49	4.80
		10	0.13	1.00	3.33	3.59	1.96	3.35	3.56	2.15	4.80
166	166	5	0.26	1.00	3.30	3.64	0.98	3.32	3.66	1.00	4.80
		10	0.26	1.00	3.60	3.81	0.98	3.58	3.82	1.00	4.80

as `Bols` and `Bridge`. Conversely, in fully dense settings ($p = s$, $s = 166$), `Bsparse` struggles to adapt and estimates a nearly constant function, resulting in RE values close to 1. In contrast, the non-sparse methods `Bols` and `Bridge` still target non-trivial approximations of m_0, yielding reasonably high RE values (approximately 3.70; see Table 2). These observations highlight that while BDMI remains robust under soft misspecification, the choice of a nuisance model can influence the *extent* of efficiency gain in dense settings, emphasizing the interplay among both p and s. Notably, even in high-dimensional (fully dense) cases, RE values remain still acceptable (RE > 1), albeit not optimal, and BDMI consistently provides correct coverage around 95% regardless of the nuisance parameter estimation methods. Moreover, Table 2 shows that the RE values of the SS estimators tend to be slightly higher for $K = 10$. But in general, the results – both for estimation and inference – seem to be fairly robust across both choices of K. We thus recommend either choice in practice.

Tables 3–4 exhibit that BDMI *consistently* achieves *correct* coverage probabilities for θ_0, maintaining approximately 95% coverage across all settings with various choices of p, s, K, as well as different methods for obtaining the nuisance posterior $\Pi_{\mathbf{m}}$. This highlights the *robustness* of BDMI in providing valid and accurate inference (correct coverage), as well as substantial *improvement* over supervised inference with *tighter* CIs (typically around 50% tighter) across settings – thereby validating its construction and our claimed theoretical properties. Figures 1–2 provide visual confirmation of these findings, showing that BDMI-based posteriors consistently exhibit always tighter spread than the supervised posterior, regardless of the setting or the nuisance posterior method. Additionally, the variability of the BDMI posteriors *remains* consistent within each setting, further emphasizing its robustness and nuisance-insensitivity across the different scenarios.



(a) Setting: $p = 50$ with $\mathbf{s = 7}$.



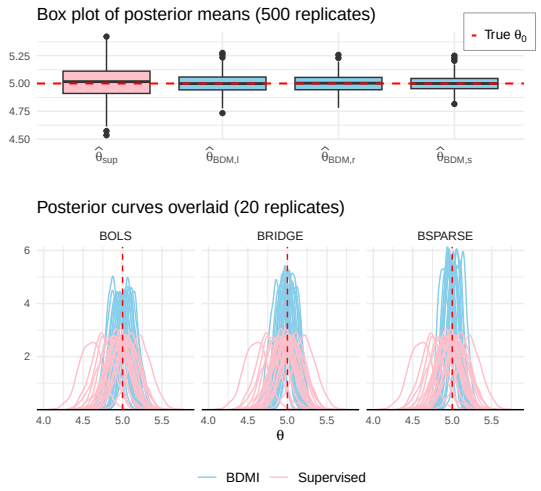
(b) Setting: $p = 50$ with $\mathbf{s = 50}$.

Figure 1: Box plots of posterior means (based on 500 replications) and plots of overlaid density curves (based on 20 iterations) for the posteriors Π_{sup} (pink) and $\Pi_{\boldsymbol{\theta}}$ (blue) of θ, with three different methods (`Bols`, `Bridge` and `Bsparse`) to obtain the nuisance posterior $\Pi_{\mathbf{m}}$ for BDMI. **Setting:** $n = 500$, $N = 10000$, $p = 50$, and $s = 7$ or 10. Each density curve is generated using 1000 posterior samples of θ. The red dashed vertical line indicates the true parameter of interest θ_0 and equals 5 for all settings.

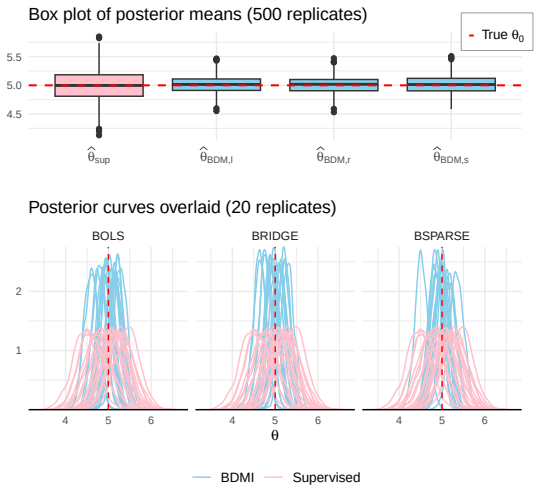
Table 3: Inference results for θ_0 based on the 95% CIs from the posteriors $\Pi_{\text{sup}}, \Pi_{\boldsymbol{\theta}}$ (BDM) and $\widetilde{\Pi}_{\boldsymbol{\theta}}$ (hBDM), for the settings in Section 5.1, with $n = 500$, $N = 10000$, $p = 50$, and $s = 7$ or 50. The methods used to obtain the nuisance posterior $\Pi_{\mathbf{m}}$ for BDM (or hBDM) are denoted as: $l =$ `Bols`, $r =$ `Bridge` and $s =$ `Bsparse`. The columns '**CovP**' and 'Len' respectively denote the average empirical coverage probability and the average length of the 95% CIs across the iterations.

		CI_{sup}		$\text{CI}_{\text{BDM},l}$		$\text{CI}_{\text{BDM},r}$		$\text{CI}_{\text{BDM},s}$		$\text{CI}_{\text{hBDM},l}$		$\text{CI}_{\text{hBDM},r}$		$\text{CI}_{\text{hBDM},s}$	
s	K	**CovP**	Len	**CovP**	Len	**CovP**	Len	**CovP**	Len	**CovP**	Len	**CovP**	Len	**CovP**	Len
7	5	0.95	0.42	0.94	0.21	0.95	0.21	0.95	0.20	0.96	0.23	0.96	0.21	0.95	0.20
	10	0.95	0.42	0.95	0.21	0.96	0.21	0.96	0.20	0.97	0.23	0.97	0.21	0.96	0.20
50	5	0.95	1.07	0.96	0.55	0.96	0.54	0.95	0.55	0.96	0.58	0.97	0.57	0.97	0.57
	10	0.95	1.07	0.96	0.55	0.96	0.54	0.95	0.55	0.97	0.57	0.97	0.57	0.97	0.57

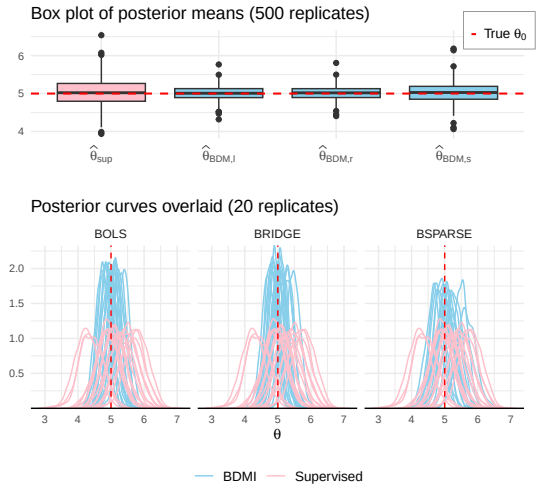
Lastly, comparing the BDMI and h-BDMI approaches, we observe that despite the former requiring only one sample from $\Pi_{\mathbf{m}}$, both methods perform similarly across most settings, which: (i) validates our earlier claims on their common theoretical properties, and (ii) also *reinforces* the crucial role of *debiasing* common to both, that *negates* any distinction between the use of one vs. many $\widetilde{m}$ samples. The point estimators $\widehat{\theta}_{\text{BDM}}$ and $\widehat{\theta}_{\text{hBDM}}$ show very similar efficiencies with h-BDMI marginally higher in some cases, while for inference, h-BDMI often tends to give slightly conservative coverages $> 95\%$ (likely due to more noise from its hierarchical nature). Overall, given



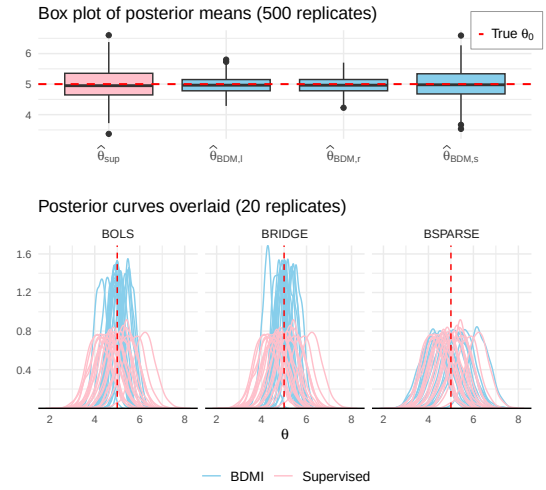
(a) Setting: $p = 166$ with $\mathbf{s = 13}$.



(b) Setting: $p = 166$ with $\mathbf{s = 55}$.



(c) Setting: $p = 166$ with $\mathbf{s = 83}$.



(d) Setting: $p = 166$ with $\mathbf{s = 166}$.

Figure 2: Box plots of posterior means and plots of overlaid density curves for the posteriors Π_{sup} (pink) and $\Pi_{\boldsymbol{\theta}}$ (blue) of θ. **Setting:** $n = 500$, $N = 10000$, $p = 166$, and $s = 13, 55, 83$ or 166. The rest of the caption details remain the same as in Figure 1.

its computational simplicity, we recommend the original BDMI approach.

5.2 Simulation studies: Misspecified models

Section 5.1 considered scenarios where the true model is linear, with Bayesian linear methods used to obtain the nuisance posterior $\Pi_{\mathbf{m}}$ of m. Although the models were technically "correctly" specified, high dimensional (and dense) settings do *not* necessarily guarantee consistent estimation of the true $m_0(\cdot)$, leading to a 'soft' form of misspecification. We now examine the functional form

Table 4: Inference results for θ_0 for the settings in Section 5.1, with $n = 500$, $N = 10000$, $p = 166$, and $s = 13, 55, 83$ or 166. The rest of the caption details remain the same as in Table 3.

		CI_{sup}		$\text{CI}_{\text{BDM},l}$		$\text{CI}_{\text{BDM},r}$		$\text{CI}_{\text{BDM},s}$		$\text{CI}_{\text{hBDM},l}$		$\text{CI}_{\text{hBDM},r}$		$\text{CI}_{\text{hBDM},s}$	
s	K	**CovP**	Len	**CovP**	Len	**CovP**	Len	**CovP**	Len	**CovP**	Len	**CovP**	Len	**CovP**	Len
13	5	0.95	0.56	0.94	0.35	0.96	0.32	0.94	0.26	0.98	0.36	0.96	0.32	0.95	0.27
	10	0.95	0.56	0.96	0.33	0.96	0.31	0.95	0.27	0.98	0.35	0.97	0.32	0.95	0.27
55	5	0.94	1.13	0.95	0.66	0.95	0.62	0.95	0.66	0.94	0.66	0.95	0.62	0.97	0.69
	10	0.94	1.13	0.95	0.64	0.96	0.62	0.95	0.64	0.95	0.64	0.95	0.62	0.97	0.67
83	5	0.95	1.38	0.96	0.80	0.95	0.77	0.95	1.16	0.96	0.81	0.95	0.76	0.97	1.12
	10	0.95	1.38	0.96	0.79	0.95	0.75	0.95	0.97	0.95	0.78	0.95	0.75	0.96	1.01
166	5	0.95	1.95	0.96	1.13	0.96	1.08	0.95	1.98	0.96	1.13	0.96	1.08	0.95	2.02
	10	0.95	1.95	0.96	1.10	0.96	1.06	0.93	2.00	0.96	1.10	0.96	1.06	0.95	2.04

of misspecification where the limiting function $m^*(\cdot)$ around which $\Pi_{\mathbf{m}}$ contracts, is *not* equal to $m_0(\cdot)$, i.e., $m_0(\cdot)$ is nonlinear but the *fitted* model for learning $\Pi_{\mathbf{m}}$ remains linear. Even in such cases, Theorems 4.2–4.3 ensure BDMI's validity with efficiency improvement persisting (see Table 1), though the improvements may not reach the optimal ORE.

Throughout, we set $N = 10000$ and $n = 500$. To illustrate our points, we specifically study non-linear, but low or moderate dimensional models with $p = 10$ (and sparsity $s = 10$ or 3) or $p = 50$ (and sparsity $s = 50$ or 7). We generated $\mathbf{X} \sim \mathcal{N}_p(\mathbf{0}, I_p)$ as in Section 5.1 and given $\mathbf{X} = \mathbf{x}$, we generate $Y \sim \mathcal{N}(m_0(\mathbf{x}), \sigma_0^2)$ with $m_0(\mathbf{x}) = \alpha_0 + \mathbf{x}'\boldsymbol{\beta}_0 + (\mathbf{x}'\boldsymbol{\gamma}_0)^2$ and $\sigma_0^2 = \text{Var}\{m_0(\mathbf{X})\}/5$. Here, $\alpha_0 = 5$, $\boldsymbol{\beta}_0 = (\mathbf{1}'_{s/2},\ \mathbf{0.5}'_{s/2},\ \mathbf{0}'_{p-s})'$ and $\boldsymbol{\gamma}_0$ is constructed to ensure $\sqrt{\mathbb{E}\{(\boldsymbol{\beta}_0'\mathbf{X})^2\}/\mathbb{E}\{(\boldsymbol{\gamma}_0'\mathbf{X})^4\}} = 3$, a reasonable balance between linear and quadratic signal parts. Despite the true $m_0(\cdot)$ being non-linear, we employed linear working models to update the nuisance posterior $\Pi_{\mathbf{m}}$ like `Bols`, `Bridge` and `Bsparse` methods as detailed in Section 5.1. Note that due to potential misspecification, $\Pi_{\mathbf{m}}$ now contracts around a non-random limiting function $m^*(\mathbf{X}) := \widetilde{\mathbf{X}}'\boldsymbol{\beta}^*$, where $\widetilde{\mathbf{X}} = (1, \mathbf{X}')'$ and $\boldsymbol{\beta}^* := \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}} \mathbb{E}(Y - \widetilde{\mathbf{X}}'\boldsymbol{\beta})^2$, i.e., $\boldsymbol{\beta}^* = \{\mathbb{E}(\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}')\}^{-1}\mathbb{E}(\widetilde{\mathbf{X}}Y)$, refer to Remark 4.2.

Unlike the settings in Section 5.1, where the theoretical ORE is attainable, it is *not* achievable here due to model misspecification. Instead, we calculated the *achievable* oracle asymptotic RE (ORE^*), defined as $\text{ORE}^* := \text{Var}(\widehat{\theta}_{\text{sup}})/\tau^2_{n,N}(m^*)$, where $\tau^2_{n,N}(m^*) = \text{Var}\{Y - m^*(\mathbf{X})\}/n + \text{Var}\{m^*(\mathbf{X})\}/N$ and $m^*(\cdot)$ is the possibly misspecified limit of $\Pi_{\mathbf{m}}$. However, both ORE and ORE^* are reported as performance benchmarks.

Table 5 and Tables 6–7 present the results on estimation and inference, respectively. Table 5 shows that the REs of the SS estimators $\widehat{\theta}_{\text{BDM}}$ and $\widehat{\theta}_{\text{hBDM}}$, compared to the supervised estimator $\widehat{\theta}_{\text{sup}}$, are substantially greater than 1, ranging roughly from 2.4 to 2.8 (matching the ORE^* closely) across most settings. Further, Tables 6–7 show that BDMI consistently achieves CovPs close to the nominal 95% level, and with significantly tighter CIs (typically 25–40% tighter) compared to the supervised approach across all settings and methods for $\Pi_{\mathbf{m}}$. All these findings highlight: (i) the *efficiency improvement* and (ii) *global robustness* that BDMI continues to enjoy *even under misspecification of* $\Pi_{\mathbf{m}}$, and further reinforces its first-order *insensitivity* to nuisance estimation bias. For more visual illustrations, see Figures S.1–S.2 in the Supplementary Material.

A notable aspect of the RE values in Table 5 is that the extent of the efficiency improvement is fairly *uniform* across the settings, and quite close to the *achievable* ORE^* in most cases – with a

Table 5: Relative efficiency (RE) of $\widehat{\theta}_{\text{BDM},i}$ and $\widehat{\theta}_{\text{hBDM},i}$ relative to $\widehat{\theta}_{\text{sup}}$, w.r.t. their empirical MSEs, for the settings in Section 5.2, with $n = 500$, $N = 10000$, and: (i) $p = 10$, with $s = 3$ or 10; or (ii) $p = 50$, with $s = 7$ or 50. The rest of the caption details remain the same as in Table 2. Further, apart from the ORE, as an additional benchmark appropriate for these misspecified settings considered in Section 5.2, we also report the oracle *achievable* asymptotic relative efficiency (ORE*) relative to $\widehat{\theta}_{\text{sup}}$.

			$\widehat{\theta}_{\text{sup}}$		$\widehat{\theta}_{\text{BDM},l}$	$\widehat{\theta}_{\text{BDM},r}$	$\widehat{\theta}_{\text{BDM},s}$	$\widehat{\theta}_{\text{hBDM},l}$	$\widehat{\theta}_{\text{hBDM},r}$	$\widehat{\theta}_{\text{hBDM},s}$		
p	s	K	MSE	RE	RE	RE	RE	RE	RE	RE	**ORE***	**ORE**
10	3	5	0.002	1.00	2.64	2.71	2.72	2.71	2.74	2.75	2.89	4.8
		10	0.002	1.00	2.65	2.71	2.82	2.72	2.74	2.82	2.89	4.8
10	10	5	0.017	1.00	2.62	2.67	2.60	2.63	2.66	2.63	2.81	4.8
		10	0.017	1.00	2.62	2.65	2.57	2.63	2.65	2.63	2.81	4.8
50	7	5	0.014	1.00	2.47	2.51	2.67	2.48	2.54	2.73	3.29	4.8
		10	0.014	1.00	2.47	2.54	2.72	2.52	2.57	2.75	3.29	4.8
50	50	5	0.093	1.00	2.44	2.49	1.70	2.45	2.49	1.92	2.78	4.8
		10	0.093	1.00	2.50	2.54	1.83	2.51	2.54	2.04	2.78	4.8

Table 6: Inference results for θ_0 for the settings in Section 5.2, with $n = 500$, $N = 10000$, $p = 10$, and $s = 3$ or 10. The rest of the caption details remain the same as in Table 3.

		CI_{sup}		$\text{CI}_{\text{BDM},l}$		$\text{CI}_{\text{BDM},r}$		$\text{CI}_{\text{BDM},s}$		$\text{CI}_{\text{hBDM},l}$		$\text{CI}_{\text{hBDM},r}$		$\text{CI}_{\text{hBDM},s}$	
s	K	**CovP**	Len	**CovP**	Len	**CovP**	Len	**CovP**	Len	**CovP**	Len	**CovP**	Len	**CovP**	Len
3	5	0.94	0.32	0.95	0.19	0.94	0.19	0.95	0.19	0.95	0.20	0.95	0.20	0.95	0.19
	10	0.94	0.32	0.95	0.19	0.94	0.19	0.96	0.19	0.96	0.20	0.95	0.19	0.96	0.19
10	5	0.95	0.53	0.95	0.32	0.94	0.32	0.95	0.32	0.94	0.32	0.95	0.32	0.95	0.33
	10	0.95	0.53	0.96	0.32	0.96	0.32	0.95	0.33	0.95	0.32	0.95	0.32	0.96	0.33

Table 7: Inference results for θ_0 for the settings in Section 5.2, with $n = 500$, $N = 10000$, $p = 50$, and $s = 7$ or 50. The rest of the caption details remain the same as in Table 3.

		CI_{sup}		$\text{CI}_{\text{BDM},l}$		$\text{CI}_{\text{BDM},r}$		$\text{CI}_{\text{BDM},s}$		$\text{CI}_{\text{hBDM},l}$		$\text{CI}_{\text{hBDM},r}$		$\text{CI}_{\text{hBDM},s}$	
s	K	**CovP**	Len	**CovP**	Len	**CovP**	Len	**CovP**	Len	**CovP**	Len	**CovP**	Len	**CovP**	Len
7	5	0.95	0.48	0.95	0.34	0.96	0.34	0.94	0.34	0.98	0.36	0.98	0.35	0.97	0.36
	10	0.95	0.48	0.94	0.34	0.96	0.34	0.95	0.34	0.97	0.36	0.98	0.35	0.97	0.36
50	5	0.94	1.18	0.94	0.75	0.95	0.75	0.94	0.88	0.94	0.77	0.94	0.75	0.96	0.92
	10	0.94	1.18	0.95	0.75	0.94	0.75	0.94	0.86	0.95	0.76	0.94	0.75	0.96	0.90

slight lowering, in general, for the higher $p = 50$ case, as expected. This indicates no substantial additional finite sample losses in estimating $m^*(\cdot)$ under the low/moderate dimensional settings here. On the other hand, the difference between the achievable ORE* and the optimal ORE indicate

the (unrecoverable) difference due to the $O(1)$ bias stemming from $\Pi_{\mathbf{m}}$ targeting $m^*(\cdot)$ and not the true $m_0(\cdot)$. One notable exception to the general uniform pattern in the REs is the case of `Bsparse` for $p = s = 50$, where the REs, while still high, are closer to 2. This arises since the dense setting introduces an *additional* layer of (soft) misspecification, making consistent estimation of even the $m^*(\cdot)$ more challenging for a sparsity-friendly method at such a choice of (p, s, n). Conversely, `Bols` and `Bridge`, which do not depend on sparsity, continue to provide higher REs around 2.5.

Finally, consistent with our findings in Section 5.1, BDMI and h-BDMI *still* perform similarly across all settings, with h-BDMI having slightly higher REs, while also exhibiting some conservativeness in CovPs, at least in some cases. Furthermore, similar to Section 5.1, the results (both for estimation and inference) remain fairly robust across $K = 5$ and $K = 10$. Thus, we continue to recommend either choice in practice.

Overall, as shown in Sections 5.1–5.2, BDMI *always* achieves significant efficiency improvements and valid inference, under both correctly specified and misspecified models, thus validating our theoretical results.

5.3 Real data analysis: Application to NHEFS data

In this section, we apply the proposed BDMI approach to a subset of data from the National Health and Nutrition Examination Survey Epidemiologic Follow-up Study (NHEFS), a longitudinal study jointly initiated by the National Center for Health Statistics and the National Institute on Aging in collaboration with other agencies of the United States Public Health Service (Hernán and Robins, 2020). The NHEFS was designed to investigate the effects of clinical, nutritional, demographic, and behavioral factors on various health outcomes, including morbidity and mortality. Data were collected during a baseline visit in 1971 and a follow-up visit in 1982. For our analysis, we focus on a cohort of 1425 individuals from this study. A detailed description of the dataset is available at https://hsph.harvard.edu/miguel-hernan/causal-inference-book. This dataset has been widely used in other studies for different purposes. For instance, Ertefaie et al. (2022) used this dataset to estimate causal parameters like the average treatment effect of quitting smoking on weight gain, and Chakrabortty et al. (2022) focused on quantile estimation under an SS framework.

Our primary goal is to estimate the *mean body weight*, θ_0, of the entire cohort in 1982 under a semi-supervised framework. Additionally, we aim to investigate whether there is a significant change in body weight within the cohort between 1971 and 1982. To achieve this, we compared the analysis results to the baseline measurement from 1971, which had a mean of 70.99 and a standard error of 0.41 for the 1425 individuals. To benchmark our results, we also consider a *gold standard* scenario where all 1425 observations (for the response) are available in 1982 (which is the case for this data). We take the mean weight $\widehat{\theta}_{\text{GS}} = 73.6$ of all 1425 individuals in the 1982 cohort as the gold standard (GS) estimator (i.e., a close 'proxy' of the truth).

To evaluate the performance of the proposed BDMI approach, we randomly select $n = 200$ observations as the labeled dataset $\mathcal{L}$, where body weight (response variable) is observed. For the unlabeled data $\mathcal{U}$, we randomly designate $N \in \{400, 800, 1220\}$ observations from the remaining data. This setup allows us to explore and compare the performance of BDMI under varying ratios of labeled and unlabeled data (n/N), particularly as this ratio approaches 0. In addition to body weight as the response variable, we considered a set of 20 important covariates in our analysis, including demographic, clinical, and behavioral factors (see Table S.1 in the Supplementary Material for their names and descriptions). These variables were also considered in other studies on this dataset, e.g., Chakrabortty et al. (2022) used them for SS quantile estimation.

The gold standard estimator $\widehat{\theta}_{\text{GS}}$ provides a benchmark for evaluating and comparing the performance of BDMI versus the supervised approach (based on the labeled data only). Given the labeled and unlabeled data, we calculated the supervised posteriors Π_{sup} (see Section 2.1) and $\Pi_{\boldsymbol{\theta}} \equiv \Pi_{\text{BDM}}$ based on the BDMI-CF approach (Algorithm 1) with $K = 5$. From each posterior distribution, 1000 samples were obtained to compute the point estimators $\widehat{\theta}_{\text{sup}}$ and $\widehat{\theta}_{\text{BDM}}$, along with the respective 95% credible intervals (CIs). Additionally, we calculated the *ratio* of the lengths (**RL**) of the 95% CIs from the supervised approach to those from BDMI. This RL serves as a natural measure of the relative efficiency of BDMI, where an RL greater than 1 indicates that BDMI provides tighter (and hence more efficient) CIs. Similar to our simulation study, 3 different methods are used to update the posterior $\Pi_{\mathbf{m}}$ of the nuisance parameter m, resulting in 3 distinct posteriors $\Pi_{\boldsymbol{\theta}} \equiv \Pi_{\text{BDM}}$ for the BDMI approach. Table 8 summarizes our findings from the data analysis.

Table 8: Results for the data analysis in Section 5.3. Estimation and inference for the mean weight of the cohort in 1982 based on the supervised (Π_{sup}) and BDMI (Π_{BDM}) approaches. Description of notations: **n**, the labeled data size; **N**, the unlabeled data size; **95% CI**, the 95% credible interval (CI); **RL**, the ratios of the lengths of the 95% CIs based on supervised approach versus BDMI; $\widehat{\theta}_{\text{GS}}$, the gold standard estimator (based on the entire cohort); $\widehat{\theta}_{\text{sup}}$, the supervised estimator; $\widehat{\theta}_{\text{BDM},i}$, the BDMI estimator where the subscript "i" denotes the method used to obtain the posterior of m: $l =$ `Bols`, $r =$ `Bridge`, $s =$ `Bsparse`.

			Π_{sup}		$\Pi_{\text{BDM},l}$			$\Pi_{\text{BDM},r}$			$\Pi_{\text{BDM},s}$		
n	N	$\widehat{\theta}_{\text{GS}}$	$\widehat{\theta}_{\text{sup}}$	95% CI	$\widehat{\theta}_{\text{BDM},l}$	95% CI	**RL**	$\widehat{\theta}_{\text{BDM},r}$	95% CI	**RL**	$\widehat{\theta}_{\text{BDM},s}$	95% CI	**RL**
	400	73.6	72.9	[70.6, 75.0]	73.6	[71.8, 75.4]	1.20	73.7	[72.0, 75.4]	1.28	73.8	[72.1, 75.6]	1.24
200	800	73.6	72.9	[70.6, 75.0]	73.7	[72.2, 75.2]	1.45	73.6	[72.1, 75.0]	1.53	73.7	[72.3, 75.1]	1.56
	1220	73.6	72.9	[70.6, 75.0]	73.2	[71.9, 74.5]	1.64	73.2	[71.9, 74.5]	1.74	73.3	[72.1, 74.7]	1.74

Table 8 highlights that BDMI demonstrates two key advantages over the supervised approach: *improved* accuracy and efficiency. First, the SS point estimates based on BDMI (across all versions) are consistently closer to the gold standard estimate, $\widehat{\theta}_{\text{GS}} = 73.6$, compared to the supervised estimate $\widehat{\theta}_{\text{sup}} = 72.6$ for all settings of N. Second, BDMI (across all versions) consistently produces significantly tighter 95% CIs than the supervised approach, with efficiency gains quantified by the ratio of CI lengths (RL), ranging from 1.2 to 1.7 across all settings. This corresponds to $20 - 70\%$ tighter intervals for BDMI. Notably, for a fixed number of labeled data, as the ratio n/N decreases (i.e., increasing the size of the unlabeled data), BDMI achieves substantial efficiency improvements by further reducing CI lengths compared to the supervised approach. For example, with $n = 200$, increasing N from 400 to 1220 improves the RL from around 1.24 to 1.74, reflecting a 40% reduction in CI length for BDMI. These results indicate that the posterior spread under BDMI becomes increasingly tighter as more unlabeled data are incorporated. Hence, these findings highlight BDMI's ability to *efficiently* leverage unlabeled data, providing strong empirical support for our theoretical framework regarding the importance of $\lim_{n,N\to\infty} n/N = c \in [0, 1)$. These results show that the BDMI procedure delivers both accurate point estimates (near identical to the GS version) and enhanced efficiency through shorter/tighter credible intervals, underscoring its advantage over the supervised approach. Finally, a notable feature of the BDMI based CIs for the mean weight of the 1982 cohort is that they consistently *exclude* the 1971 mean weight (70.99), indicating a significant weight gain, likely due to aging or quitting smoking (Ertefaie et al., 2022). In contrast, the supervised

approach *fails to detect* this change, as its 95% CI $(70.6, 75.0)$ includes the 1971 mean. These results highlight the improved efficiency and higher *power* of BDMI for detecting significant (and scientifically meaningful) differences in weights between the two cohorts.

6 Concluding discussions

We proposed the BDMI procedure for estimating the population mean $\theta_0 = \mathbb{E}(Y)$ under the SS setting. To the best of our knowledge, this is the first attempt to establish a Bayesian method that achieves desirable SS inference goals, including *efficiency improvement* and *global robustness*, while providing *rigorous* theoretical guarantees. Our methodology ensures that the posterior $\Pi_{\boldsymbol{\theta}}$ of the parameter of interest θ contracts around the true parameter θ_0 at the parametric rate $n^{-1/2}$ and is asymptotically Normal, *regardless* of the choice of method used to obtain a posterior for the nuisance parameter m, its contraction rate, or even potential misspecification of m. Moreover, the posterior mean of $\Pi_{\boldsymbol{\theta}}$, as an SS estimator of θ_0, *always* possesses $\sqrt{n}$-consistency, asymptotic normality, and first-order *insensitivity*, in addition to being at least as *efficient* as the supervised estimator (sample mean of Y). These theoretical results have been rigorously established in Section 4. One of the key contributions of BDMI lies in its ability to disentangle nuisance parameter estimation from inference on the target parameter by developing a novel debiasing approach under the Bayesian paradigm. It enables joint learning of the nuisance bias and the main parameter through targeted modeling of summary statistics, along with careful usage of sample splitting. We hope this research brings attention to the rarely used idea of modeling summary statistics within Bayesian inference and demonstrates its potential to address other Bayesian semi-parametric inference problems. While this work focuses on SS mean estimation, the underlying principles of BDMI can be extended to a broad range of problems, including missing data analysis, causal inference, and SS inference for other functionals. For instance, BDMI could be adapted to handle selection bias or distribution shifts between labeled and unlabeled data; this was recently explored in the frequentist SS literature (Zhang et al., 2023) but not yet addressed within a Bayesian framework. Further, extending BDMI to Bayesian SS inference for high dimensional target parameters (e.g., regression coefficients) poses additional theoretical and computational challenges, but also represents an important direction for future research. Finally, adapting BDMI's debiasing framework to causal inference or missing data settings offers exciting opportunities for advancing Bayesian semi-parametric methodologies. We hope this work generates interest in considering such Bayesian problems in the future.

Supplementary Material

Supplement to 'Bayesian Semi-supervised Inference via a Debiased Modeling Approach'. The supplement (Sections S1–S5) includes additional discussions, numerical results, and all technical materials (e.g., proofs) that could not be accommodated in the main paper: **(i)** additional figures and a table for the simulations and data analysis in Sections 5.2–5.3 (Section S1); **(ii)** additional discussion on the imputation approach introduced in Section 2.2, along with a detailed numerical study for comparison with BDMI (Section S2); **(iii)** implementation details of the `Bridge` and `Bsparse` methods used to obtain the nuisance posterior $\Pi_{\mathbf{m}}$ in our numerical studies (Section S3); **(iv)** proofs of all the main theoretical results (Section S4); and **(v)** proofs of preliminary results and intermediate lemmas utilized in the proofs of the main results (Section S5).

Acknowledgements

The authors would like to thank the Editor, the Associate Editor, and the three Reviewers for their constructive comments and suggestions that significantly helped improve the presentation and the content of the article.
This research was partially supported by the National Science Foundation grants: NSF-DMS 2113768 (to Abhishek Chakrabortty), and NSF-DMS 2210689 and NSF-DMS 1916371 (to Anirban Bhattacharya).

References

David Azriel, Lawrence D. Brown, Michael Sklar, Richard Berk, Andreas Buja, and Linda Zhao. Semi-supervised linear regression. *Journal of the American Statistical Association*, 117(540): 2238–2251, 2022.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer Science & Business Media, Dordrecht, 2011.

Anirban Bhattacharya, Debdeep Pati, Natesh S Pillai, and David B Dunson. Dirichlet–laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110(512):1479–1490, 2015.

Peter J. Bickel and Bart J. K. Kleijn. The Semiparametric Bernstein–von Mises Theorem. *The Annals of Statistics*, 40(1):206–237, 2012.

Dominique Bontemps. Bernstein von mises theorems for gaussian regression with increasing number of regressors. *The Annals of Statistics*, 39(5):2557–2584, 2011.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Christoph Breunig, Ruixuan Liu, and Zhengfei Yu. Double robust bayesian inference on average treatment effects. *Econometrica*, 93(2):539–568, 2025.

T. Tony Cai and Zijian Guo. Semisupervised inference for explained variance in high dimensional linear regression and its applications. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(2):391–419, 2020.

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

Ismaël Castillo and Judith Rousseau. A Bernstein–von Mises theorem for smooth functionals in semiparametric models. *The Annals of Statistics*, 43(5):2353–2383, 2015.

Abhishek Chakrabortty and Tianxi Cai. Efficient and adaptive linear regression in semi-supervised settings. *The Annals of Statistics*, 46(4):1541–1572, 2018.

Abhishek Chakrabortty, Guorong Dai, and Raymond J Carroll. Semi-supervised quantile estimation: Robust and efficient inference in high dimensional settings. *arXiv preprint arXiv:2201.10208*, 2022.

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, USA, 2006.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/Debiased Machine Learning for Treatment and Structural Parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

Bertrand Clarke and Jayanta K. Ghosh. Posterior convergence given the mean. *The Annals of Statistics*, 23(6):2116–2144, 1995.

Kjell A. Doksum and Albert Y. Lo. Consistent and robust Bayes procedures for location based on partial information. *The Annals of Statistics*, 18(1):443–453, 1990.

Christopher C. Drovandi, Anthony N. Pettitt, and Anthony Lee. Bayesian Indirect Inference Using a Parametric Auxiliary Model. *Statistical Science*, 30(1):72–95, 2015.

Ashkan Ertefaie, Nima S. Hejazi, and Mark J. van der Laan. Nonparametric inverse probability weighted estimators based on the highly adaptive lasso. *Biometrics*, 79(2):1029–1043, 2022.

Max H. Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021.

Paul Fearnhead and Dennis Prangle. Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 74(3):419–474, 2012.

Edward I. George and Robert E. McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

Peter J. Green and Bernard W. Silverman. *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman and Hall/CRC, 1994.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC, 2015.

Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, 2020.

Valen E. Johnson. Bayes factors based on test statistics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(5):689–701, 2005.

Valen E. Johnson and David Rossell. Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association*, 107(498):649–660, 2012.

Masanori Kawakita and Takafumi Kanamori. Semi-supervised learning with density-ratio estimation. *Machine Learning*, 91(2):189–209, 2013.

Bo'az Klartag. A central limit theorem for convex sets. *Inventiones mathematicae*, 168(1):91–131, 2007.

Isaac S. Kohane. Using electronic health records to drive discovery in disease genomics. *Nature Reviews Genetics*, 12(6):417–428, 2011.

John R. Lewis, Steven N. MacEachern, and Yoonkyung Lee. Bayesian Restricted Likelihood Methods: Conditioning on insufficient statistics in Bayesian regression (with discussion). *Bayesian Analysis*, 16(4):1393–1462, 2021.

Yu Luo, Daniel J Graham, and Emma J McCoy. Semiparametric Bayesian doubly robust causal estimation. *Journal of Statistical Planning and Inference*, 225:171–187, 2023.

Paul Marjoram, John Molitor, Vincent Plagnol, and Simon Tavaré. Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328, 2003.

Toby J. Mitchell and John J. Beauchamp. Bayesian Variable Selection in Linear Regression. *Journal of the American Statistical Association*, 83(404):1023–1032, 1988.

Alexander McFarlane Mood, Franklin A. Graybill, and Duane C. Boes. *Introduction to the Theory of Statistics*. McGraw-Hill, 1974.

Whitney K. Newey and James R. Robins. Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138*, 2018.

Andriy Norets. Bayesian regression with nonparametric heteroskedasticity. *Journal of Econometrics*, 185(2):409–419, 2015.

David Pollard. *A user's guide to measure theoretic probability*. Cambridge University Press, 2002.

John W. Pratt. Bayesian interpretation of standard inference statements. *Journal of the Royal Statistical Society: Series B (Statistical Methodological)*, 27(2):169–203, 1965.

Kolyan Ray and Botond Szabo. Debiased bayesian inference for average treatment effects. In *Advances in Neural Information Processing Systems*, volume 32, pages 11952–11962, 2019.

Kolyan Ray and Aad van der Vaart. Semiparametric Bayesian causal inference. *The Annals of Statistics*, 48(5):2999–3020, 2020.

Vincent Rivoirard and Judith Rousseau. Bernstein–von Mises theorem for linear functionals of the density. *The Annals of Statistics*, 40(3):1489–1523, 2012.

Veronika Ročková and Edward I George. The spike-and-slab lasso. *Journal of the American Statistical Association*, 113(521):431–444, 2018.

I. Richard Savage. Nonparametric Statistics: A Personal review. *Sankhyā: The Indian Journal of Statistics, Series A*, 31(2):107–144, 1969.

Steven L. Scott, Alexander W. Blocker, Fernando V. Bonassi, Hugh A. Chipman, George Edward I., and Robert E. McCulloch. Bayes and big data: the consensus monte carlo algorithm. *International Journal of Management Science and Engineering Management*, 11(2):78–88, 2016.

Steven L. Scott, Alexander W. Blocker, Fernando V. Bonassi, Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. Bayes and big data: The consensus Monte Carlo algorithm. In *Big Data and Information Theory*, pages 8–18. Routledge, 2022.

Matthias Seeger. Learning with labeled and unlabeled data. Technical Report EPFL-REPORT-161327, University of Edinburgh, UK, 2002.

Jeffrey S. Simonoff. *Smoothing Methods in Statistics*. Springer Science & Business Media, 2012.

Donald F. Specht. A general regression neural network. *IEEE Transactions on neural networks*, 2 (6):568–576, 1991.

Anastasios A. Tsiatis. *Semiparametric theory and missing data*. Springer, New York, 2006.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, New York, 2009.

Aad W. van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge University Press, 2000.

Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

Mike West. On scale mixtures of normal distributions. *Biometrika*, 74(3):646–648, 1987.

Christopher K. I. Williams. Prediction with Gaussian processes: From linear regression to linear prediction and beyond. In *Learning in graphical models*, pages 599–621. Springer, 1998.

Andrew Yiu, Edwin Fong, Chris Holmes, and Judith Rousseau. Semiparametric posterior corrections. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, pages 1–30, 2025.

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

Tong Zhang and Fernando J. Oles. The value of unlabeled data for classification problems. In *Proceedings of the Seventeenth ICML*, pages 1191–1198, 2000.

Yuqian Zhang and Jelena Bradic. High-dimensional semi-supervised learning: in search of optimal inference of the mean. *Biometrika*, 109(2):387–403, 2022.

Yuqian Zhang, Abhishek Chakrabortty, and Jelena Bradic. Double robust semi-supervised inference for the mean: selection bias under MAR labeling with decaying overlap. *Information and Inference: A Journal of the IMA*, 12(3):2066–2159, 2023.

Xiaojin Zhu. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2008.

Supplement to "Bayesian Semi-supervised Inference via a Debiased Modeling Approach"

Gözde Sert, Abhishek Chakrabortty, and Anirban Bhattacharya
Department of Statistics, Texas A&M University [1]

This supplementary document (Sections S1–S5) includes additional discussions and numerical analyses, as well as technical details such as proofs and extended discussions that could not be accommodated in the main paper. Section S1 includes additional figures for the simulation results in Section 5.2 and a supplementary table for the data analysis in Section 5.3. In Section S2, we provide a detailed construction of the imputation approach, initially introduced in Section 2.2, and then present numerical studies to highlight its limitations, along with a comparative analysis with the BDMI approach. Section S3 outlines the implementation details of the methods used to obtain the nuisance posteriors in the numerical studies of Section 5. Section S4 presents the proofs of all the results in the main paper. Finally, Section S5 provides the proofs for all supporting lemmas or intermediate lemmas introduced in the course of the main proofs in Section S4.

S1 Additional figures and tables for numerical studies

Figures S.1–S.2 present additional plots for the simulation results in Section 5.2 for misspecified models.
Table S.1 lists the names and descriptions of the covariates used for the NHEFS data analysis in Section 5.3.

S2 The imputation approach and its limitations: A comparative analysis with BDMI

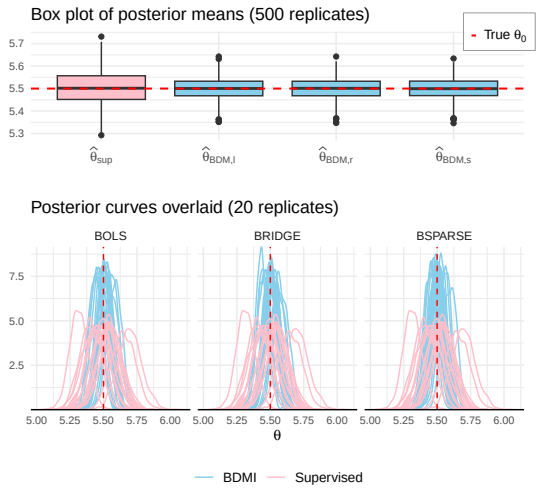
This section provides an extensive numerical comparison of the imputation-type approach (henceforth IMP) introduced in Section 2.2 with BDMI. For IMP, recall that one selects a Bayesian regression method to construct the nuisance posterior $\Pi_{\mathbf{m}}$ for m from the labeled data $\mathcal{L}$. Using the imputation (regression) representation $\theta_0 = \mathbb{E}_{\mathbf{X}}\{m_0(\mathbf{X})\}$, one can compute the induced posterior by approximating $\mathbb{E}_{\mathbf{X}}$ with an empirical average over $\mathcal{U}$. Specifically, given $\widetilde{m} \sim \Pi_{\mathbf{m}}$, we define a new random variable:

$$\theta_{\text{imp}} \equiv \theta_{\text{imp}}(\widetilde{m}) = \frac{1}{N}\sum_{i=n+1}^{n+N} \widetilde{m}(\mathbf{X}_i), \quad \text{and let } \Pi_{\text{imp}} \text{ be the (induced) posterior of } \theta_{\text{imp}}. \tag{S.1}$$

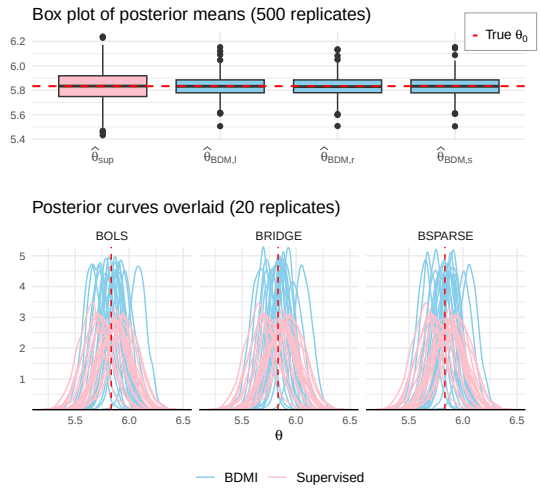
The posterior mean $\widehat{\theta}_{\text{imp}}$ of Π_{imp}, a point estimate of θ_0 under IMP, by linearity of expectation, is given by:

$$\widehat{\theta}_{\text{imp}} \equiv \widehat{\theta}_{\text{imp}}(\widehat{m}) := \frac{1}{N}\sum_{i=n+1}^{n+N} \widehat{m}(\mathbf{X}_i), \text{ where } \widehat{m}(\cdot) := \mathbb{E}_{m\sim\Pi_{\mathbf{m}}}\big\{m(\cdot)\,|\,\mathcal{L}\big\} \text{ is the posterior mean of } \Pi_{\mathbf{m}}.$$

[1] *Email addresses*: gozdesert@stat.tamu.edu (Gözde Sert), abhishek@stat.tamu.edu (Abhishek Chakrabortty), anirbanb@stat.tamu.edu (Anirban Bhattacharya)



(a) Setting for misspecified model: $p = 10$ with $\mathbf{s = 3}$



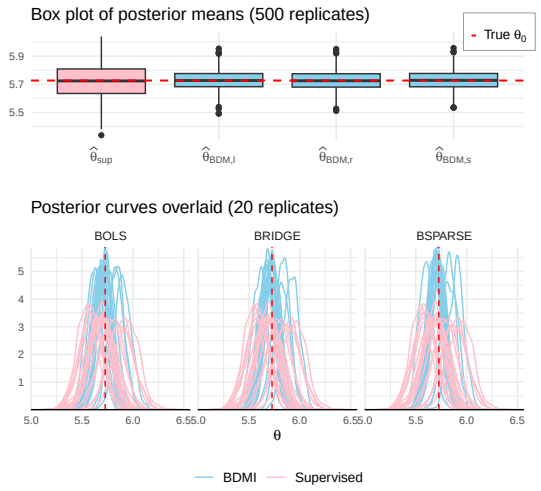
(b) Setting for misspecified model: $p = 10$ with $\mathbf{s = 10}$

Figure S.1: Box plots of posterior means (based on 500 replications) and plots of overlaid density curves (based on 20 iterations) for the posteriors Π_{sup} (pink) and $\Pi_{\boldsymbol{\theta}}$ (blue) of θ, with three different methods (`Bols`, `Bridge` and `Bsparse`) to obtain the nuisance posterior $\Pi_{\mathbf{m}}$ for BDMI. **Setting** (from Section 5.2): $n = 500$, $N = 10000$, $p = 10$, and $s = 3$ or $s = 10$. (Each density curve is generated using 1000 posterior samples of θ. The red dashed vertical line indicates the true parameter of interest θ_0.)

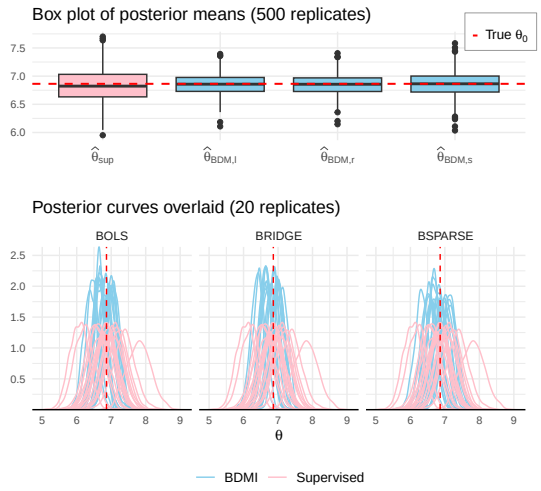
It is straightforward to sample from Π_{imp}; to generate B samples of θ from Π_{imp}, one first draws B samples $\widetilde{m}^{(1)}, \ldots, \widetilde{m}^{(B)}$ of m from the nuisance posterior Π_m, and then uses the construction in (S.1) to obtain the corresponding posterior samples $\theta^{(1)}, \ldots, \theta^{(B)}$. The posterior mean $\widehat{\theta}_{\text{imp}}$ is approximated by $B^{-1} \sum_{b=1}^{B} \theta^{(b)}$.

To enable a fair comparison, we compare IMP with h-BDMI, as both methods share a hierarchical structure, and thus differences in performance can be attributed to *debiasing*, which is the key distinction between these approaches. Specifically, we use the CF version of h-BDMI with $K = 10$ throughout. As shown in Section 5.1, the performance of our original BDMI (single-sample version) is nearly indistinguishable from h-BDMI, and we have observed the same trends we report below when comparing IMP with BDMI.

We adhere to the data generation setting described in Section 5.1. Specifically, we examine the case where $p = 166$ with four different sparsity levels: $s = 13$ (sparse), $s = 55$ or $s = 83$ (moderately dense), and $s = 166$ (fully dense). For $\Pi_{\mathbf{m}}$, we consider two different methods: `Bridge` and `Bsparse`, as described in Section 5.1. This yields two versions each for the induced posterior Π_{imp} under IMP and the aggregated posterior $\Pi_{\boldsymbol{\theta}}$ under BDMI, along with their respective posterior means ($\widehat{\theta}_{\text{imp}}$ and $\widehat{\theta}_{\text{BDM}}$). Figures S.3–S.10 display boxplots of the point estimators (based on 500 replications) and density plots of the posteriors across a random subset of 50 replications to improve visual clarity. The *odd-numbered* figures correspond to IMP and the *even-numbered* ones to BDMI. The posterior curves are based on 1000 posterior samples each. The left and right panels in each figure correspond to `Bridge` and `Bsparse`, respectively.

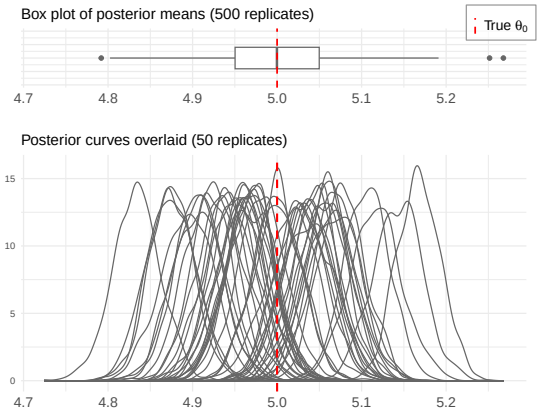


(a) Setting for misspecified model: $p = 50$ with $\mathbf{s = 7}$.

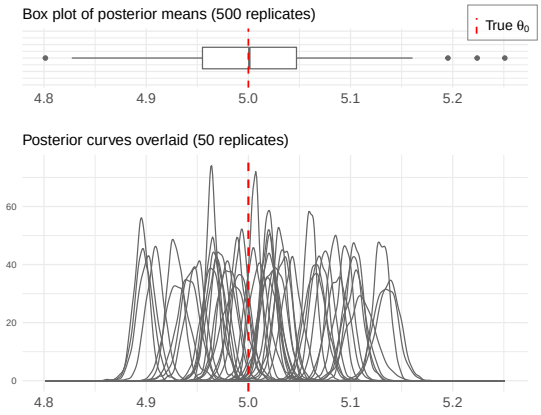


(b) Setting for misspecified model: $p = 50$ with $\mathbf{s = 50}$.

Figure S.2: Box plots of posterior means and plots of overlaid density curves for the posteriors Π_{sup} (pink) and $\Pi_{\boldsymbol{\theta}}$ (blue) of θ. **Setting (Section 5.2):** $n = 500$, $N = 10000$, $p = 50$, and $s = 7$ or $s = 50$. The rest of the caption details are the same as Figure S.1.



(a) `Bridge`



(b) `Bsparse`

Figure S.3: Box plot of posterior means (based on 500 replications) and the overlaid density curves (based on 50 iterations) of the posterior Π_{imp} of θ for the **imputation** approach (IMP) with two different methods (left: `Bridge`; right: `Bsparse`) to obtain the posterior $\Pi_{\mathbf{m}}$ of the nuisance parameter m. Each density curve is generated using 1000 posterior samples of $\theta \sim \Pi_{\text{imp}}$. **Setting:** $n = 500$, $N = 10000$, $p = 166$, and $\mathbf{s = 13}$. The coverage probabilities based on IMP are **56**% for the `Bridge` method and **23**% for the `Bsparse` method. (The red dashed vertical line indicates the true parameter of interest θ_0 and equals 5 for all settings.)

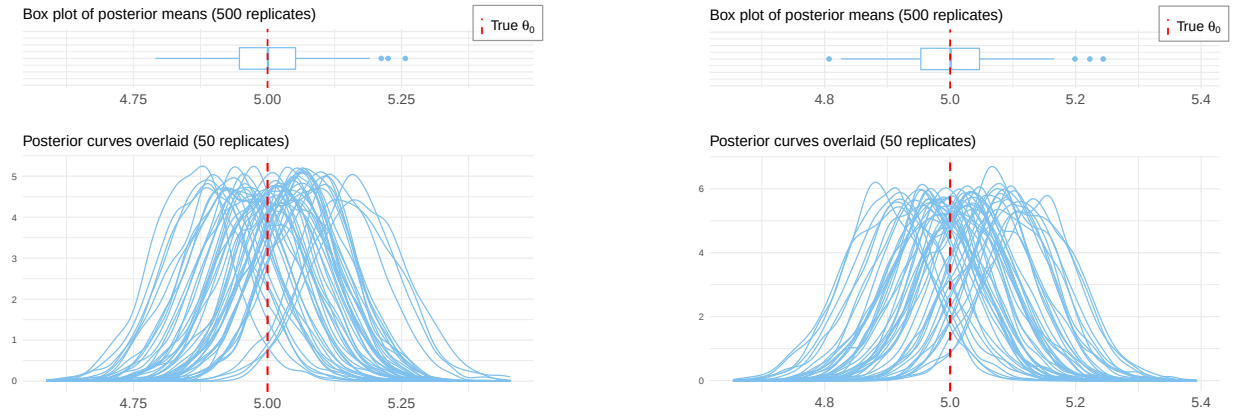
We first comment on the point estimators $\widehat{\theta}_{\text{imp}}$ and $\widehat{\theta}_{\text{BDM}}$. Figures S.3–S.10 show that both point

Table S.1: Covariates included for the NHEFS data analysis in Section 5.3.

Variable name	Description
`active`	On your usual day, how active were you in 1971?
`age`	Age in 1971
`alcoholfreq`	How often did you drink in 1971?
`allergies`	Use allergies medication in 1971
`asthma`	DX asthma in 1971
`cholesterol`	Serum cholesterol (mg/100ml) in 1971
`dbp`	Diastolic blood pressure in 1982
`education`	Amount of education by 1971
`exercise`	In recreation, how much exercise in 1971?
`ht`	Height in centimeters in 1971
`price71`	Average tobacco price in the state of residence 1971 (US$2008)
`price82`	Average tobacco price in the state of residence 1982 (US$2008)
`race`	White, black or other in 1971
`sbp`	Systolic blood pressure in 1982
`sex`	Male or female
`smokeintensity`	Number of cigarettes smoked per day in 1971
`smokeyrs`	Years of smoking
`tax71`	Tobacco tax in the state of residence 1971 (US$2008)
`tax82`	Tobacco tax in the state of residence 1971 (US$2008)
`wt71`	Weight in kilograms in 1971

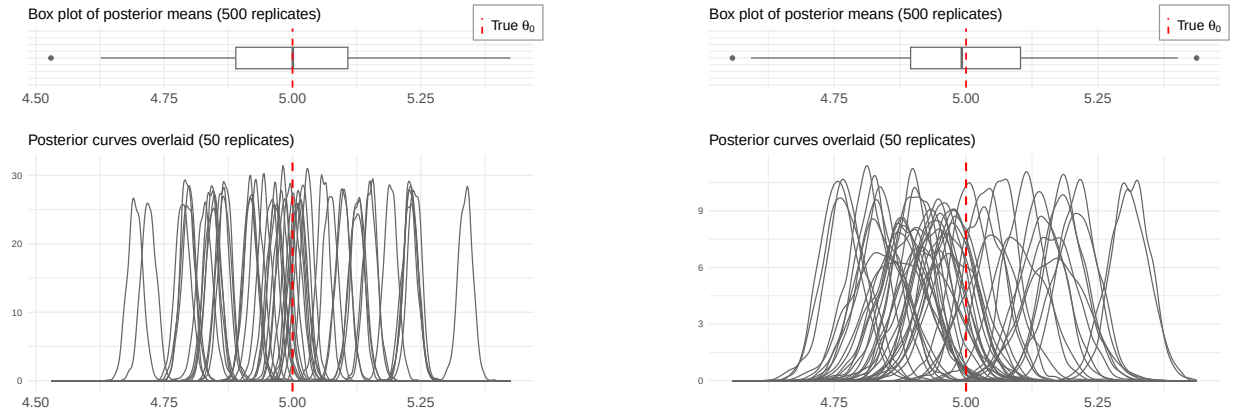
estimators appear unbiased across all settings (sparse, moderately dense, and dense), regardless of the method used (`Bridge` or `Bsparse`) to obtain the nuisance posterior $\Pi_{\mathbf{m}}$. Their medians are consistently centered around θ_0 with similar variability. While IMP performs comparably to BDMI in terms of point estimation, important differences emerge when examining the entire *posteriors*, Π_{imp} and $\Pi_{\boldsymbol{\theta}}$, themselves.

The posteriors from the imputation approach exhibit substantial variability *across* the two methods as well as the different sparsity settings, showcasing its *sensitivity* (in the first order) to nuisance estimation (both in method choice and the setting). Moreover, the imputation posteriors are often very narrow, especially in the more dense cases, and show considerable variation across simulation replicates, with their supports increasingly becoming disjoint. As a result, the imputation posteriors often fail to cover θ_0, leading to *severe undercoverage*. Across the two methods and the four different settings, the imputation posterior's coverage of the symmetric 95% credible interval ranges between $5\% - 56\%$. In stark contrast, the BDMI posteriors *remain* stable across methods and settings, maintaining a Gaussian shape, and vary smoothly across simulation replicates, with coverage *consistently* close to the nominal level, showcasing the superiority of BDMI over IMP. Its ability to provide provably valid inference and its stability (more generally, the overall posterior's smooth behavior) across settings and choices of nuisance models reinforces the importance of its *debiased* nature and *insensitivity* to nuisance estimation – an aspect that may be useful more generally in other settings as well.



(a) `Bridge` (b) `Bsparse`

Figure S.4: Box plot of posterior means (based on 500 replications) and the overlaid density curves (based on 50 iterations) of the posterior $\Pi_{\boldsymbol{\theta}}$ of θ for the **BDMI** approach with two different methods (left: `Bridge`; right: `Bsparse`) to obtain the nuisance posterior $\Pi_{\mathbf{m}}$ of m. Each density curve is generated using 1000 posterior samples of $\theta \sim \Pi_{\boldsymbol{\theta}}$. **Setting:** $n = 500$, $N = 10000$, $p = 166$, and $\mathbf{s = 13}$ (and $K = 10$). The coverage probabilities based on BDMI are 96% for `Bridge` and 95% for `Bsparse`.

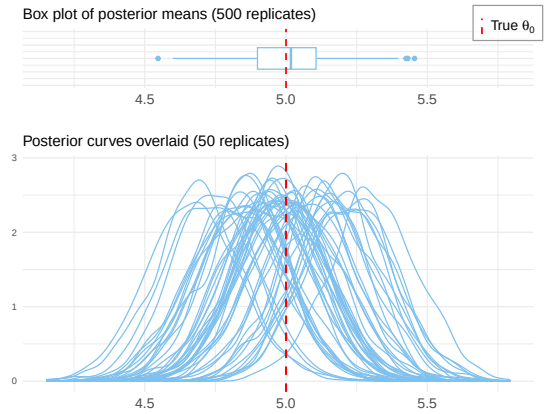


(a) `Bridge` (b) `Bsparse`

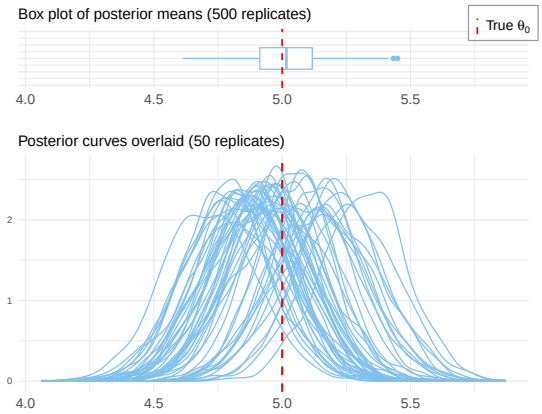
Figure S.5: Box plot of posterior means and the overlaid density curves of Π_{imp} for θ based on **IMP** with `Bridge` and `Bsparse` methods for $\mathbf{s = 55}$. The corresponding coverages are **12**% and **43**%. The rest of the caption remains the same as in Figure S.3.

S3 Implementation details of the `Bridge` and `Bsparse` methods to obtain $\Pi_{\mathbf{m}}$ in Section 5

In this section, we collect some technical details regarding implementations of two of the methods used to obtain the nuisance posterior $\Pi_{\mathbf{m}}$ in our numerical studies in Section 5: Bayesian ridge

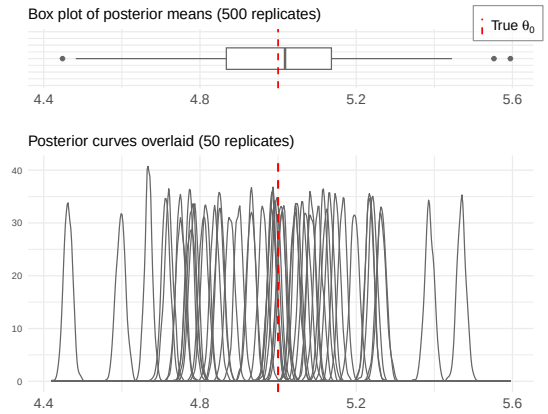


(a) `Bridge`

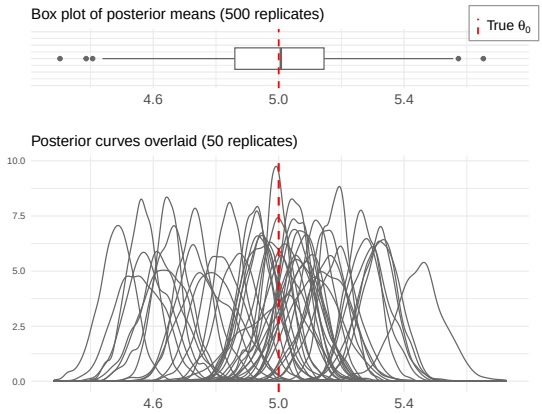


(b) `Bsparse`

Figure S.6: Box plot of posterior means and overlaid density curves of $\Pi_{\boldsymbol{\theta}}$ for θ based on **BDMI** with the `Bridge` and `Bsparse` methods for $\mathbf{s} = \mathbf{55}$. The corresponding coverages are 96% and 95%. The rest of the caption remains the same as in Figure S.4.



(a) `Bridge`



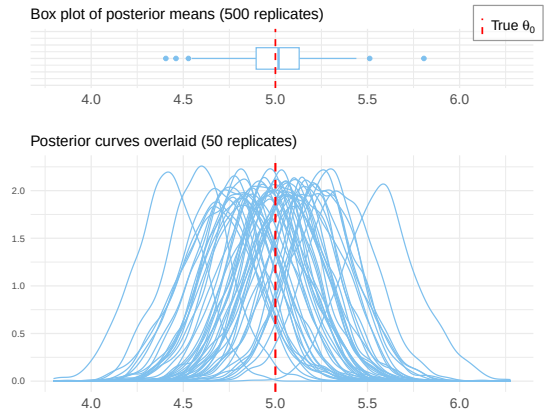
(b) `Bsparse`

Figure S.7: Box plot of posterior means and the overlaid density curves of Π_{imp} for θ based on **IMP** with `Bridge` and `Bsparse` methods for $\mathbf{s} = \mathbf{83}$. The corresponding coverages are **7**% and **45**%. The rest of the caption remains the same as in Figure S.3.

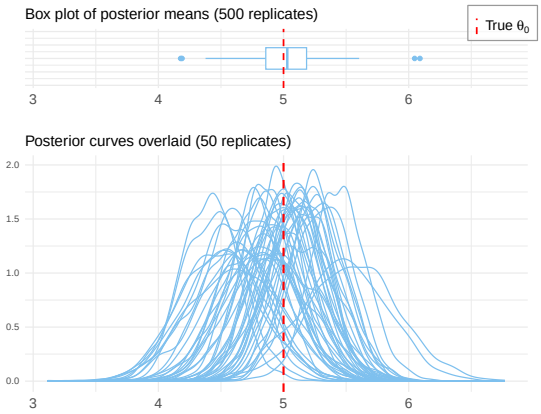
regression (`Bridge`) (in Section S3.1), and sparse Bayesian linear regression via non-local priors (`Bsparse`) (in Section S3.2).

S3.1 Implementation details for `Bridge`: Empirical Bayes approach for tuning parameter selection

For the `Bridge` method, we adopt an empirical Bayes approach to estimate the prior precision parameter (or the ridge tuning parameter, in frequentist terminology) λ, effectively bridging frequentist and Bayesian methodologies. The estimate $\widehat{\lambda}$ is obtained using the `R` package `glmnet`,

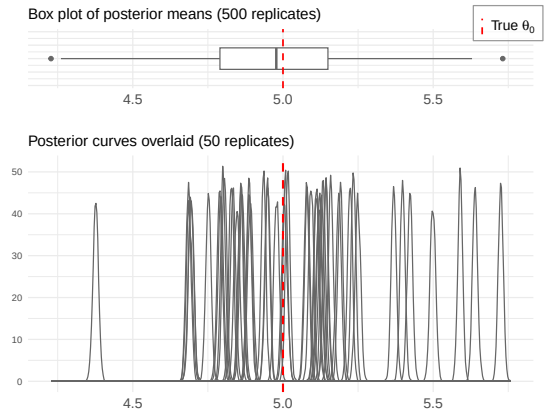


(a) `Bridge`

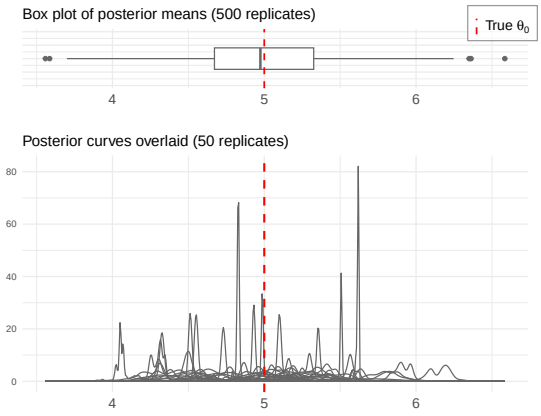


(b) `Bsparse`

Figure S.8: Box plot of posterior means and overlaid density curves of $\Pi_{\boldsymbol{\theta}}$ for θ based on **BDMI** with the `Bridge` and `Bsparse` methods for $\mathbf{s} = \mathbf{83}$. The corresponding coverages are 95% and 95%. The rest of the caption remains the same as in Figure S.4.



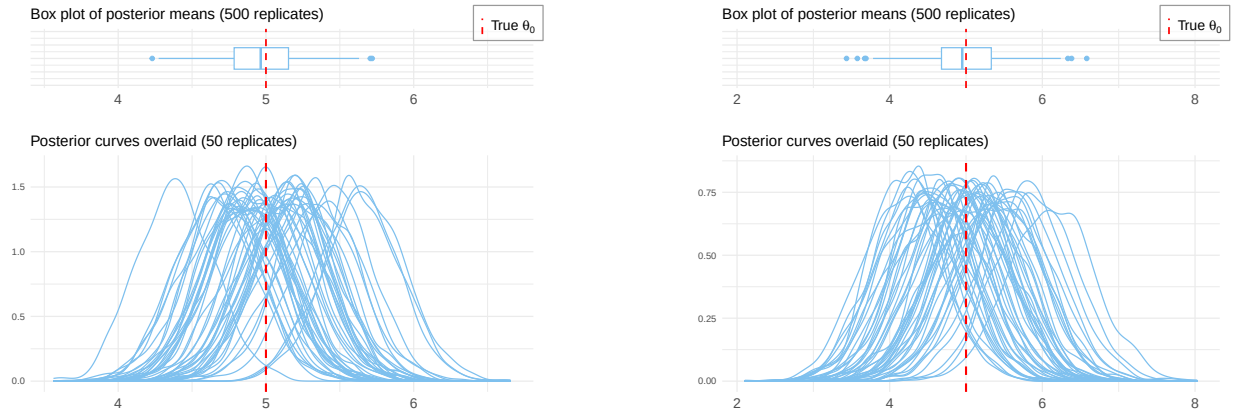
(a) `Bridge`



(b) `Bsparse`

Figure S.9: Box plot of posterior means and the overlaid density curves of Π_{imp} for θ based on **IMP** with `Bridge` and `Bsparse` methods for $\mathbf{s} = \mathbf{166}$. The corresponding coverages are **5**% and **25**%. The rest of the caption remains the same as in Figure S.3.

specifically its `cv.glmnet` function, along with a scale transformation thereafter. This approach ensures that the posterior mean of $(\alpha, \beta')' \in \mathbb{R}^{(p+1)}$ from our approach aligns with the cross-validated point estimate from `glmnet`, offering a data-driven approach for hyper-parameter selection. A notable aspect of `cv.glmnet` is its *standardization* (column-by-column) of the design matrix $\mathbb{X}_{n\times p} := (\mathbf{X}_1, \ldots, \mathbf{X}_n)'$, as well as *scaling* of the response vector $\mathbf{Y}_{n\times 1} := (Y_1, \ldots, Y_n)'$ to have unit standard deviation. Since penalized methods, in general, are not scale invariant, these adjustments are critical to ensure equal penalization of all predictors and mitigate any potential biases due to differences in scale (for both $\mathbb{X}$ and $\mathbf{Y}$). Writing $\mathbb{X}$ column-wise as $\mathbb{X}_{n\times p} \equiv (\mathbf{x}_1, \ldots, \mathbf{x}_p)$, let



(a) `Bridge` (b) `Bsparse`

Figure S.10: Box plot of posterior means and overlaid density curves of $\Pi_{\boldsymbol{\theta}}$ for θ based on **BDMI** with the `Bridge` and `Bsparse` methods for $\mathbf{s} = \mathbf{166}$. The corresponding coverages are 96% and 93%. The rest of the caption remains the same as in Figure S.4.

$\mathbb{Z}_{n\times p} \equiv (\mathbf{z}_1, \ldots, \mathbf{z}_p)$ denote the corresponding (column-by-column) standardized version of $\mathbb{X}$, i.e., $\mathbf{z}_j := (\mathbf{x}_j - \overline{x}_j \mathbf{J}_n)/s_j \in \mathbb{R}^n$, where $\overline{x}_j$ and s_j respectively denote the sample mean and sample standard deviation of $\mathbf{x}_j$, for $j = 1, \ldots, p$; and the vector $\mathbf{J}_n := (1, \ldots, 1) \in \mathbb{R}^n$. Further, let s_Y be the sample standard deviation of $\mathbf{Y}$.

Then, the objective function minimized by `cv.glmnet` is given by:

$$l(a, \mathbf{b}) := \frac{1}{2n} \left\| \frac{\mathbf{Y}}{s_Y} - a\mathbf{J}_n - \mathbb{Z}\mathbf{b} \right\|_2^2 + \frac{\lambda}{2} \|\mathbf{b}\|_2^2,$$

where $\mathbf{J}_n := (1, \ldots, 1)' \in \mathbb{R}^n$ and $\|\cdot\|_2$ is the L_2-vector norm. Note that the intercept parameter a is *not* penalized, as is the usual practice. It is also important to note that while `glmnet` performs the standardization and scaling *internally* (by default), the *final estimator* it returns is in the *original scale* of the data (for both $\mathbf{Y}$ and $\mathbb{X}$). That is, if $(\widehat{a}, \widehat{\mathbf{b}}')'$ denotes the minimizer from above (with optimally chosen λ) for fitting $\mathbf{Y}/s_Y \sim a\mathbf{J}_n + \mathbb{Z}\mathbf{b}$, then the final (ridge) estimator it returns is $(\widehat{\alpha}, \widehat{\boldsymbol{\beta}}')'$ (for fitting the model $\mathbf{Y} \sim \alpha\mathbf{J}_n + \mathbb{X}\boldsymbol{\beta}$), with $\widehat{\alpha}$ and $\widehat{\boldsymbol{\beta}}$ obtained from appropriately transforming back $\widehat{a}$ and $\widehat{\mathbf{b}}$.

The optimal λ value, denoted as $\widetilde{\lambda} =$ `lambda.min`, from `cv.glmnet` is selected to minimize the cross-validation error of the above optimization (involving the scaled version of $\mathbf{Y}$). Therefore, to integrate this into our own Bayesian modeling framework (involving the original $\mathbf{Y}$), we apply the following transformation:

$$\widehat{\lambda} \;=\; \widetilde{\lambda} \cdot (n/s_y), \quad \text{and we use this transformed } \widehat{\lambda} \text{ as our Gaussian prior's precision parameter.}$$

This transformation ensures consistency between the frequentist and Bayesian approaches by aligning the ridge point estimator $(\widehat{\alpha}, \widehat{\boldsymbol{\beta}}')'$ from `cv.glmnet` with the posterior mean of $(\alpha, \boldsymbol{\beta}')'$ from our Bayesian modeling.

For the Bayesian component, we model the data $\widetilde{D}_n := (\mathbf{Y}_{n\times 1}, \mathbb{Z}_{n\times p})$, where $\mathbf{Y}$ is the original

response vector and $\mathbb{Z}_{n\times p}$ is the standardized design matrix, using a Gaussian likelihood and specify priors as follows:

$$\mathbf{Y} \mid \mathbb{Z}, \widetilde{\alpha}, \widetilde{\boldsymbol{\beta}}, \sigma^2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}_n\big(\widetilde{\alpha}\mathbf{J}_n + \mathbb{Z}\widetilde{\boldsymbol{\beta}}, \sigma^2 I_n\big), \quad \text{with}$$
$$\pi(\widetilde{\alpha} \mid \sigma^2) \propto 1, \ \widetilde{\boldsymbol{\beta}} \mid \sigma^2 \sim \mathcal{N}_p(\mathbf{0}_p, \widehat{\lambda}^{-1}\sigma^2 I_p) \ \text{ and } \ \pi(\sigma^2) \propto (\sigma^2)^{-1},$$

and we assume that $\widetilde{\alpha}$ and $\widetilde{\boldsymbol{\beta}}$ are independent. After obtaining samples from the posterior distribution $\Pi_{\widetilde{\gamma}}$ (a multivariate t-distribution) for the parameter $\widetilde{\gamma} := (\widetilde{\alpha}, \widetilde{\boldsymbol{\beta}}')' \in \mathbb{R}^{(p+1)}$, we transform these samples back to the *original scale,* i.e., the same scale of $(\mathbf{Y}, \mathbb{X})$, to obtain posterior samples of $(\alpha, \boldsymbol{\beta}')'$ as follows:

$$\alpha = \widetilde{\alpha} - \sum_{j=1}^{p} \frac{(\widetilde{\beta})_j \cdot \overline{x}_j}{s_j} \ \text{ and } \ (\boldsymbol{\beta})_j = \frac{(\widetilde{\boldsymbol{\beta}})_j}{s_j}, \quad \text{for } j = 1, \ldots, p, \ \ [\text{and } (\mathbf{v})_j \text{ denotes the } j^{th} \text{ entry of } \mathbf{v}],$$

where $\overline{x}_j$ and s_j respectively denote the sample mean and sample standard deviation of $\mathbf{x}_j$, for $j = 1, \ldots, p$.

S3.2 Implementation details for `Bsparse`: The `R` package `mombf`

For the `Bsparse` method of obtaining $\Pi_{\mathbf{m}}$, we used a sparse Bayesian linear regression based on non-local priors (NLP) (Johnson and Rossell, 2012). We implemented it using the R package `mombf`, which provides tools for Bayesian model selection and parameter estimation with a focus on NLP. In our implementation, we used the package's default options to ensure consistency and simplicity. The main function used from the package is `modelSelection`, which performs Bayesian variable selection for linear models using NLP.

This function has two key arguments that allow specification of prior distributions as follows:

- `priorCoef`: Determines the prior for the regression coefficients.
- `priorDelta`: Specifies a prior distribution for the model space.

For our implementation, we selected the default choices for these arguments:

- `priorCoef = momprior(tau = 0.348)`, where τ represents the prior dispersion parameter, controlling the strength of penalization applied to small regression coefficients.
- `priorDelta = modelbbprior(alpha.p = 1, beta.p = 1)`, which sets a Beta-Binomial prior for the model space.

Using these settings, we performed Bayesian model selection with `modelSelection`. To obtain posterior samples for the regression coefficients, we employed the `rnlp` function from the `mombf` package. The `rnlp` function includes two important parameters: `center` and `scale` that control whether pre-processing is applied to the response variable (Y) and covariates ($\mathbf{X}$). We used the following choices for each of these:

- `center = TRUE`: Centers Y and $\mathbf{X}$ by subtracting their means to remove potential biases caused by non-zero means in predictors.

- `scale = TRUE`: Scales $\mathbf{X}$ by dividing each covariate by its standard deviation to ensure fair penalization across all predictors.

By default, both parameters are set to `FALSE`, meaning no pre-processing is applied unless explicitly specified. However, since our data was not pre-standardized, we set both `center = TRUE` and `scale = TRUE` in our implementation. It is important to note that even when centering and scaling are applied (i.e., `center = TRUE` and `scale = TRUE`), the `rnlp` function provides posterior samples for regression coefficients on the *original scale* of the data. This is achieved since the `mombf` package internally stores information on the centering and scaling transformations applied during pre-processing. These stored values are used to transform posterior samples back to their original scale, ensuring that results remain interpretable in terms of the original data.

S4 Proofs of the main results

In this section, we present the proofs of all the results from the main paper. We begin by introducing some additional notations and some preliminary lemmas that will be used throughout the proofs of the main results. The rest of the section is organized as follows: **(i)** Section S4.1 enlists the preliminary lemmas; **(ii)** Section S4.2 presents the proof of Proposition 3.1; **(iii)** Section S4.3 presents the proof of Proposition 3.2; **(iv)** Section S4.4 provides the proof of Theorem 4.1; **(v)** Section S4.5 presents the proof of Theorem 4.2; **(vi)** Section S4.6 presents the proof of Corollary 4.1; and **(vii)** Section S4.7 provides the proof of Theorem 4.3.

Throughout this section, we will use the following *additional notations*, in addition to those introduced in the main paper. For any functions $f(\cdot) \in \mathbb{L}_2(\mathbb{P}_{\mathbf{Z}})$ and $g(\cdot) \in \mathbb{L}_2(\mathbb{P}_{\mathbf{X}})$, and for any $k \in \{1, \ldots, K\}$, define:

$$\text{(i)}\ \ \mathbb{E}_{n_K}^{(k)}\{f(\mathbf{Z})\} := n_K^{-1} \sum_{i \in \mathcal{I}_k} f(\mathbf{Z}_i) \ \text{ and } \ \mathbb{G}_{n_K}^{(k)}\{f(\mathbf{Z})\} := n_K^{1/2}\big[\mathbb{E}_{n_K}^{(k)}\{f(\mathbf{Z})\} - \mathbb{E}_{\mathbf{Z}}\{f(\mathbf{Z})\}\big]; \quad \text{and}$$
$$\text{(ii)}\ \ \mathbb{E}_{N_K}^{(k)}\{g(\mathbf{X})\} := N_K^{-1} \sum_{i \in \mathcal{J}_k} g(\mathbf{X}_i) \ \text{ and } \ \mathbb{G}_{N_K}^{(k)}\{g(\mathbf{X})\} := N_K^{1/2}\big[\mathbb{E}_{N_K}^{(k)}\{g(\mathbf{X})\} - \mathbb{E}_{\mathbf{X}}\{g(\mathbf{X})\}\big]. \tag{S.2}$$

Further, for any two absolutely continuous probability measures P and Q on $\mathbb{R}$ with corresponding densities $p(\cdot)$ and $q(\cdot)$, we will use the well-known identity for their TV distance: $\|P - Q\|_{\text{TV}} = (1/2) \int |p(x) - q(x)| \mathrm{d}x$ (Tsybakov, 2009, Lemma 2.1 (Scheffés' theorem)). Additionally, a Gamma distribution with shape and rate parameters $(\alpha, \beta) > 0$ and density $\frac{\beta^\alpha}{\Gamma(\alpha)} e^{-\beta x} x^{\alpha-1} \mathbb{1}_{(0,\infty)}(x)$ is denoted Gamma(α, β); and if $W \sim$ Gamma(α, β), then we denote $1/W \sim$ IG(α, β), i.e., the inverse Gamma (IG) distribution with parameters $(\alpha, \beta) > 0$.

Remark S4.1 (Empirical process notations)**.** The notations $\mathbb{E}_{n_K}^{(k)}(\cdot)$ and $\mathbb{E}_{N_K}^{(k)}(\cdot)$ in (S.2) simply denote the empirical mean operators on the data folds $\mathcal{L}_k$ (indexed by $\mathcal{I}_k$ and of size $n_K = n/K$) and $\mathcal{U}_k$ (indexed by $\mathcal{J}_k$ and of size $N_K = N/K$), respectively. Similarly, $\mathbb{G}_{n_K}^{(k)}(\cdot)$ denotes the corresponding $n_K^{1/2}$-scaled (and centered) empirical process on $\mathcal{L}_k$ indexed by $f(\cdot)$, and and $\mathbb{G}_{N_K}^{(k)}(\cdot)$ denotes the $N_K^{1/2}$-scaled (and centered) empirical process on $\mathcal{U}_k$ indexed by $g(\cdot)$. Notations of this flavor are fairly common in the modern semi-parametric inference literature, as well as SS inference literature, that require empirical process and/or sample-splitting techniques (Chernozhukov et al., 2018; van der Vaart, 2000; Zhang and Oles, 2000; Chakrabortty et al., 2022).

S4.1 Preliminaries

The following results will be used to prove the main results of the paper.

Lemma S4.1 (Invariance property of the TV distance (Pollard, 2002, Chapter 3)). *Let P and Q be absolutely continuous probability measures on $\mathbb{R}$ with the corresponding densities $p(\cdot)$ and $q(\cdot)$. For fixed $\mu \in \mathbb{R}$ and $\sigma > 0$, define $p^{\mu,\sigma}(t) := \sigma^{-1} p\{(t-\mu)/\sigma\}$ and $q^{\mu,\sigma}(t) := \sigma^{-1} q\{(t-\mu)/\sigma\}$ as the corresponding location-shifted and scaled version of $p(\cdot)$ and $q(\cdot)$, with the respective probability measures $P^{\mu,\sigma}$ and $Q^{\mu,\sigma}$. Then, $\|P - Q\|_{\mathrm{TV}} = \|P^{\mu,\sigma} - Q^{\mu,\sigma}\|_{\mathrm{TV}}$.*

Lemma S4.2 (An upper bound for the TV distance between two Gaussian distributions with the same variance). *Let $P = \mathcal{N}(\mu_1, \sigma^2)$ and $Q = \mathcal{N}(\mu_2, \sigma^2)$ be two Normal distributions. Then, $\|P - Q\|_{\mathrm{TV}} = 2\,\Phi\{|\mu_1 - \mu_2|/(2\sigma)\} - 1$, where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard Normal distribution $\mathcal{N}(0,1)$. This further implies that $\|P - Q\|_{\mathrm{TV}} \leq (2\pi)^{-1/2}|\mu_1 - \mu_2|/\sigma$.*

Lemma S4.3 (Adopted from Lemma 4.9 in Klartag (2007)). *Let $P = \mathcal{N}(0, \sigma_1^2)$ and $Q = \mathcal{N}(0, \sigma_2^2)$. Then, for some universal constant $C > 0$, $\|P - Q\|_{\mathrm{TV}} \leq C|(\sigma_2^2/\sigma_1^2) - 1|$.*

Lemma S4.4 (TV distance between a t–distribution and a Normal distribution). *Let $P = t_\nu(\mu, \sigma^2)$ and $Q = \mathcal{N}(\mu, \sigma^2)$. Then, for some constant $C_0 > 0$, we have $\|P - Q\|_{\mathrm{TV}} \leq C_0/\sqrt{\nu}$.*

Lemma S4.5 (TV distance for convolutions). *Let P, Q be two probability distributions with the pdfs $p(\cdot)$, $q(\cdot)$, respectively. Suppose $p(x) = (p_1 * p_2)(x)$ and $q(x) = (q_1 * q_2)(x)$, where $p_i(\cdot), q_i(\cdot)$ are the pdfs of the corresponding distributions P_i, Q_i for $i = 1, 2$, and $*$ is the convolution operator which is defined in Proposition 3.1. Then, $\|P - Q\|_{\mathrm{TV}} \leq \|P_1 - Q_1\|_{\mathrm{TV}} + \|P_2 - Q_2\|_{\mathrm{TV}}$.*

Lemma S4.6 (Conditional convergence $\Rightarrow$ unconditional convergence; adopted from Chernozhukov et al. (2018, Lemma 6.1)). *Let U_n and V_n be sequences of random variables (with a joint distribution). (a) If, for $\varepsilon_n \to 0$, $\mathbb{P}(|U_n| > \varepsilon_n \,|\, V_n) \xrightarrow{\mathbb{P}} 0$, then $\mathbb{P}(|U_n| > \varepsilon_n) \to 0$. (b) Let b_n be a sequence of positive constants. If $|U_n| = O_{\mathbb{P}}(b_n)$ conditional on V_n, namely, that for any $t_n \to \infty$, $\mathbb{P}(|U_n| > t_n b_n \,|\, V_n) \xrightarrow{\mathbb{P}} 0$, then $|U_n| = O_{\mathbb{P}}(b_n)$ unconditionally, namely, that for any $t_n \to \infty$, $\mathbb{P}(|U_n| > t_n b_n) \to 0$.*

Proofs of the preliminary results are presented in Section S5 of the Supplementary Material. We are now ready to present the proof of the main results in Sections 3 and 4.

S4.2 Proof of Proposition 3.1

For notational simplicity, we define $W(\mathbf{Z}; \widetilde{m}) := Y - \widetilde{m}(\mathbf{X})$ and $\delta(\widetilde{m}) := \theta - b(\widetilde{m})$. Then, the likelihood function in (4) becomes

$$
\begin{aligned}
L\big\{\delta(\widetilde{m}), b(\widetilde{m}), \sigma_1^2(\widetilde{m}), \sigma_2^2(\widetilde{m})\big\} \;\propto\; & \frac{1}{\{\sigma_1^2(\widetilde{m})\}^{n/2}} \exp\left[-\frac{1}{2\sigma_1^2(\widetilde{m})} \sum_{i=1}^{n} \big\{W(\mathbf{Z}_i; \widetilde{m}) - b(\widetilde{m})\big\}^2\right] \times \\
& \frac{1}{\{\sigma_2^2(\widetilde{m})\}^{N/2}} \exp\left[-\frac{1}{2\sigma_2^2(\widetilde{m})} \sum_{i=n+1}^{n+N} \big\{\widetilde{m}(\mathbf{X}_i) - \delta(\widetilde{m})\big\}^2\right] \\
:=\; & L\{b(\widetilde{m}), \sigma_1^2(\widetilde{m})\}\; L\{\delta(\widetilde{m}), \sigma_2^2(\widetilde{m})\}.
\end{aligned}
$$

Since the determinant of the Jacobian matrix is 1, the prior on the model parameters becomes

$$
\pi\big\{\delta(\widetilde{m}), b(\widetilde{m}), \sigma_1^2(\widetilde{m}), \sigma_2^2(\widetilde{m})\big\} \;\propto\; \big\{\sigma_1^2(\widetilde{m})\sigma_2^2(\widetilde{m})\big\}^{-1}. \tag{S.3}
$$

By Bayes' theorem, the joint posterior density of $\{\delta(\widetilde{m}), b(\widetilde{m})\}$ can be calculated by integrating out the parameters $\{\sigma_1^2(\widetilde{m}), \sigma_2^2(\widetilde{m})\}$. Then, we obtain that

$$\begin{aligned}\pi\{\delta(\widetilde{m}), b(\widetilde{m}) \mid \mathcal{D}\} \;&\propto \int \frac{L\{\delta(\widetilde{m}), \sigma_2^2(\widetilde{m})\}}{\sigma_2^2(\widetilde{m})}\, \mathrm{d}\sigma_2^2(\widetilde{m}) \times \int \frac{L\{b(\widetilde{m}), \sigma_1^2(\widetilde{m})\}}{\sigma_1^2(\widetilde{m})}\, \mathrm{d}\sigma_1^2(\widetilde{m}) \\ &= \pi\{\delta(\widetilde{m}) \mid \mathcal{D}\}\pi\{b(\widetilde{m}) \mid \mathcal{D}\}.\end{aligned}$$

Then the joint posterior density $\pi\{\delta(\widetilde{m}), b(\widetilde{m}) \mid \mathcal{D}\}$ is the product of the marginal posterior densities of $\delta(\widetilde{m})$ and $b(\widetilde{m})$. This implies that posterior distributions of $\delta(\widetilde{m})$ and $b(\widetilde{m})$ are independent. Since $\theta = \delta(\widetilde{m}) + b(\widetilde{m})$ by the construction of $\delta(\widetilde{m})$, the marginal posterior distribution of θ can be calculated as a convolution of posterior distributions of $\delta(\widetilde{m})$ and $b(\widetilde{m})$.

To conclude the proof of Proposition 3.1, it is enough to show the marginal posterior distribution of $b(\widetilde{m})$ is a t–distribution with degrees of freedom v_n, center $\mu_n(\widetilde{m})$ and scale $\widehat{\sigma}_{1,n}^2(\widetilde{m})/n$ as defined in (6), since the calculation of the posterior distribution of $\delta(\widetilde{m})$ follows the same steps.

Towards establishing this, we first observe that

$$b(\widetilde{m}) \mid \sigma_1^2(\widetilde{m}), \mathcal{D} \sim \mathcal{N}(\mu_n(\widetilde{m}), \sigma_1^2(\widetilde{m})/n) \text{ and } \sigma_1^2(\widetilde{m}) \mid \mathcal{D} \sim \mathrm{IG}\left(\frac{n-1}{2}, \frac{\sum_{i=1}^n \{W(\mathbf{Z}_i; \widetilde{m}) - \mu_n(\widetilde{m})\}^2}{2}\right),$$

where $\mu_n(\widetilde{m}) = n^{-1}\sum_{i=1}^n W(\mathbf{Z}_i; \widetilde{m})$. Since the t–distribution can be expressed as a scale mixture of a Normal distribution, we obtain that $b(\widetilde{m}) \mid \mathcal{D} \sim t_{\nu_n}(\mu_n(\widetilde{m}), \widehat{\sigma}_{1,n}^2(\widetilde{m})/n)$ where $\nu_n := n-1$ and

$$\frac{\widehat{\sigma}_{1,n}^2(\widetilde{m})}{n} \;:=\; \frac{\sum_{i=1}^n \{W(\mathbf{Z}_i; \widetilde{m}) - \mu_n(\widetilde{m})\}^2}{n(n-1)}.$$

By following the same steps, we obtain that $\delta(\widetilde{m}) \sim t_{\nu_N}(\mu_N(\widetilde{m}), \widehat{\sigma}_{2,N}^2(\widetilde{m})/N)$ where $\nu_N := N-1$,

$$\mu_N(\widetilde{m}) \;:=\; \frac{1}{N}\sum_{i=n+1}^{n+N} \widetilde{m}(\mathbf{X}_i) \quad \text{and} \quad \frac{\widehat{\sigma}_{2,N}^2(\widetilde{m})}{N} \;:=\; \frac{\sum_{i=n+1}^{n+N} \{\widetilde{m}(\mathbf{X}_i) - \mu_N(\widetilde{m})\}^2}{N(N-1)}.$$

Hence, the marginal posterior of θ is a convolution of two t-distributions: $t_{\nu_n}(\mu_n(\widetilde{m}), \widehat{\sigma}_{1,n}^2(\widetilde{m})/n)$ and $t_{\nu_N}(\mu_N(\widetilde{m}), \widehat{\sigma}_{2,N}^2(\widetilde{m})/N)$. This completes the proof. ■

S4.3 Proof of Proposition 3.2

To avoid repetition, we refer to the proof of Proposition 3.1 in Section S4.2. Consider the likelihood function in (7) and the prior density in (S.3). Then, by following the same steps as in the proof of Proposition 3.1 this time applied to the data fold $\widetilde{\mathcal{D}}_k$ instead of $\mathcal{D}$ as in Proposition 3.1, we obtain that the marginal posterior distribution $\Pi_{\boldsymbol{\theta}}^{(k)}$ is a convolution of two t-distributions with the desired parameters as given in the statement of Proposition 3.2. This concludes the proof. ■

S4.4 Proof of Theorem 4.1

For notational simplicity, we set $k = 1$ w.l.o.g. and present the proof for $k = 1$. By the triangle inequality, we first observe that

$$\begin{aligned} & \|\Pi_{\boldsymbol{\theta}}^{(1)} - \mathcal{N}(\widehat{\theta}_{\mathrm{BDM}}^{(1)}(m^*), \tau^2_{n_K,N_K}(m^*))\|_{\mathrm{TV}} \\ \leq\ & \|\Pi_{\boldsymbol{\theta}}^{(1)} - \mathcal{N}(\widehat{\theta}_{\mathrm{BDM}}^{(1)}(\widetilde{m}_1), \tau^2_{n_K,N_K}(\widetilde{m}_1))\|_{\mathrm{TV}} \\ & + \|\mathcal{N}(\widehat{\theta}_{\mathrm{BDM}}^{(1)}(\widetilde{m}_1), \tau^2_{n_K,N_K}(\widetilde{m}_1)) - \mathcal{N}(\widehat{\theta}_{\mathrm{BDM}}^{(1)}(m^*), \tau^2_{n_K,N_K}(m^*))\|_{\mathrm{TV}} \ :=\ T_1 + T_2, \end{aligned} \tag{S.4}$$

where $\widehat{\theta}_{\mathrm{BDM}}^{(k)}(\widetilde{m}_k) := \mu_{n_K}(\widetilde{m}_k) + \mu_{N_K}(\widetilde{m}_k) = n_K^{-1}\sum_{i\in\mathcal{I}_k}\{Y_i - \widetilde{m}_k(\mathbf{X}_i)\} + N_K^{-1}\sum_{i\in\mathcal{J}_k}\widetilde{m}_1(\mathbf{X}_i)$ and $\tau^2_{n_K,N_K}(\widetilde{m}_k) := \sigma_1^2(\widetilde{m}_k)/n_K + \sigma_2^2(\widetilde{m}_k)/N_K$ for any $k = 1,\ldots,K$ (we specifically set $k = 1$ here). Then, the problem reduces to showing both T_1 and T_2 converge to 0 in probability w.r.t. $\mathbb{P}_{\widetilde{\mathcal{D}}_1}$.

We first consider T_1 in (S.4). By Proposition 3.2 (refer to the Supplementary Material), we have that the posterior $\Pi_{\boldsymbol{\theta}}^{(1)}$ of θ is the convolution of two t-distributions $t_{\nu_{n_K}}(\mu_{n_K}(\widetilde{m}_1), \widehat{\sigma}^2_{1,n_K}(\widetilde{m}_1)/n_K)$ and $t_{\nu_{N_K}}(\mu_{N_K}(\widetilde{m}_1), \widehat{\sigma}^2_{2,N_K}(\widetilde{m}_1)/N_K)$, where the parameters are as defined in (8) (refer to the Supplementary Material) by setting $k = 1$. Also, we can always write a Normal distribution as a convolution of two independent Normal distributions. Further, by Lemma S4.5, we observe that

$$\begin{aligned} T_1 &= \|\Pi_{\boldsymbol{\theta}}^{(1)} - \mathcal{N}(\widehat{\theta}_{\mathrm{BDM}}^{(1)}(\widetilde{m}_1), \tau^2_{n_K,N_K}(\widetilde{m}_1))\|_{\mathrm{TV}} \\ &\leq \|t_{\nu_{n_K}}(\mu_{n_K}(\widetilde{m}_1), \widehat{\sigma}^2_{1,n_K}(\widetilde{m}_1)/n_K) - \mathcal{N}(\mu_{n_K}(\widetilde{m}_1), \sigma_1^2(\widetilde{m}_1)/n_K)\|_{\mathrm{TV}} \\ &\quad + \|t_{\nu_{N_K}}(\mu_{N_K}(\widetilde{m}_1), \widehat{\sigma}^2_{2,N_K}(\widetilde{m}_1)/N_K) - \mathcal{N}(\mu_{N_K}(\widetilde{m}_1), \sigma_2^2(\widetilde{m}_1)/N_K)\|_{\mathrm{TV}} \\ &= \|t_{\nu_{n_K}}(0, \widehat{\sigma}^2_{1,n_K}(\widetilde{m}_1)) - \mathcal{N}(0, \sigma_1^2(\widetilde{m}_1))\|_{\mathrm{TV}} + \|t_{\nu_{N_K}}(0, \widehat{\sigma}^2_{2,N_K}(\widetilde{m}_1)) - \mathcal{N}(0, \sigma_2^2(\widetilde{m}_1))\|_{\mathrm{TV}} \\ &:= T_{11} + T_{12}, \end{aligned} \tag{S.5}$$

where (S.5) is obtained from the invariance property of the TV distance from Lemma S4.1.

Next, we consider T_{11} in (S.5). By the triangle inequality and the construction of $\widehat{\sigma}^2_{1,n_K}(\widetilde{m}_1)$, we get

$$\begin{aligned} T_{11} &= \|t_{\nu_{n_K}}(0, \widehat{\sigma}^2_{1,n_K}(\widetilde{m}_1)) - \mathcal{N}(0, \sigma_1^2(\widetilde{m}_1))\|_{\mathrm{TV}} \\ &\leq \|t_{\nu_{n_K}}(0, \widehat{\sigma}^2_{1,n_K}(\widetilde{m}_1)) - \mathcal{N}(0, \widehat{\sigma}^2_{1,n_K}(\widetilde{m}_1))\|_{\mathrm{TV}} + \|\mathcal{N}(0, \widehat{\sigma}^2_{1,n_K}(\widetilde{m}_1)) - \mathcal{N}(0, \sigma_1^2(\widetilde{m}_1))\|_{\mathrm{TV}} \\ &\leq \frac{C_0}{\nu_{n_K}} + \left|\frac{\widehat{\sigma}^2_{1,n_K}(\widetilde{m}_1) - \sigma_1^2(\widetilde{m}_1)}{\sigma_1^2(\widetilde{m}_1)}\right|, \end{aligned}$$

where the last step follows from Lemma S4.4 (applied to the first TV distance in the second line above) and Lemma S4.3 (applied to the second TV distance in the second line above). By the definition of ν_{n_K} (i.e., $\nu_{n_K} = n_K - 1$), $1/\nu_{n_K} \to 0$ as $n \to \infty$ (since $n_K = n/K$ and K is fixed, refer to Section 3.3 for further notational clarification). Thus to show $T_{11} \xrightarrow{\mathbb{P}} 0$ under $\mathbb{P}_{\mathcal{D}_1,\widetilde{m}_1}$, it is enough to show

$$\Sigma_{\epsilon,n_K}(\mathcal{D}_1, \widetilde{m}_1) \ :=\ \left|\frac{\widehat{\sigma}^2_{1,n_K}(\widetilde{m}_1) - \sigma_1^2(\widetilde{m}_1)}{\sigma_1^2(\widetilde{m}_1)}\right| \ \to\ 0, \text{ in probability w.r.t. } \mathbb{P}_{\mathcal{D}_1,\widetilde{m}_1}.$$

This means that for any $t > 0$, $\mathbb{P}_{\mathcal{D}_1,\widetilde{m}_1}(\Sigma_{\epsilon,n_K}(\mathcal{D}_1,\widetilde{m}_1) > t) \to 0$. We further observe that

$$\begin{aligned}\mathbb{P}_{\mathcal{D}_1,\widetilde{m}_1}\{\Sigma_{\epsilon,n_K}(\mathcal{D}_1,\widetilde{m}_1) > t\} &= \mathbb{E}_{\mathcal{D}_1,\widetilde{m}_1}[\mathbf{1}\{\Sigma_{\epsilon,n_K}(\mathcal{D}_1,\widetilde{m}_1) > t\}] \\ &= \mathbb{E}_{\widetilde{m}_1}(\mathbb{E}_{\mathcal{D}_1}[\mathbf{1}\{\Sigma_{\epsilon,n_K}(\mathcal{D}_1,\widetilde{m}_1) > t\} \mid \widetilde{m}_1]) \\ &= \mathbb{E}_{\widetilde{m}_1}[\mathbb{P}_{\mathcal{D}_1}\{\Sigma_{\epsilon,n_K}(\mathcal{D}_1,\widetilde{m}_1) > t \mid \widetilde{m}_1\}],\end{aligned} \tag{S.6}$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function and the second step uses the fact that $\widetilde{m}_1 \perp\!\!\!\perp \mathcal{D}_1$ (refer to the construction of BDMI in Section 3.3 by taking $k = 1$). We note that $0 \leq \mathbb{P}_{\mathcal{D}_1}\{\Sigma_{\epsilon,n_K}(\mathcal{D}_1,\widetilde{m}_1) > t \mid \widetilde{m}_1\} \leq 1$, and it is random through $\widetilde{m}_1$. Then, by the dominated convergence theorem (DCT) (alternatively, refer to Lemma S4.6 here), it is sufficient to show $\mathbb{P}_{\mathcal{D}_1}\{\Sigma_{\epsilon,n_K}(\mathcal{D}_1,\widetilde{m}_1) > t \mid \widetilde{m}_1\} \xrightarrow{\mathbb{P}} 0$ under $\mathbb{P}_{\widetilde{m}_1}$ to conclude that $\mathbb{P}_{\mathcal{D}_1,\widetilde{m}_1}\{\Sigma_{\epsilon,n_K}(\mathcal{D}_1,\widetilde{m}_1) > t\} \to 0$. Next, we observe that for any $t > 0$,

$$\mathbb{P}_{\mathcal{D}_1}\{\Sigma_{\epsilon,n_K}(\mathcal{D}_1,\widetilde{m}_1) > t \mid \widetilde{m}_1\} \;=\; \mathbb{P}_{\mathcal{D}_1}\{|\widehat{\sigma}^2_{1,n_K}(\widetilde{m}_1) - \sigma^2_1(\widetilde{m}_1)| > \tilde{t} \mid \widetilde{m}_1\},$$

for $\tilde{t} \;:=\; t\,|\sigma^2_1(\widetilde{m}_1)| > 0$, where given $\widetilde{m}_1$, we can think of $\sigma^2_1(\widetilde{m}_1)$ as a fixed non-random quantity. Then, by Chebyshev's inequality, we obtain that

$$Z_{n_K}(\widetilde{m}_1) \;:=\; \mathbb{P}_{\mathcal{D}_1}\{|\widehat{\sigma}^2_{1,n_K}(\widetilde{m}_1) - \sigma^2_1(\widetilde{m}_1)| > \tilde{t} \mid \widetilde{m}_1\} \;\leq\; (\tilde{t})^{-2}\,\mathrm{Var}\{\widehat{\sigma}^2_{1,n_K}(\widetilde{m}_1) \mid \widetilde{m}_1\},$$

where the last inequality uses $\mathbb{E}_{\mathcal{D}_1}\{\widehat{\sigma}^2_{1,n_K}(\widetilde{m}_1) - \sigma^2_1(\widetilde{m}_1)\,|\,\widetilde{m}_1\} = 0$ that can be obtained by the construction of $\widehat{\sigma}^2_{1,n_K}(\widetilde{m}_1)$ (refer to (8) by setting $k = 1$). For notational simplicity, let $W(\mathbf{Z};\widetilde{m}_1) := Y - \widetilde{m}_1(\mathbf{X})$. Then, using Theorem 2 in Chapter VI of Mood et al. (1974), we obtain that

$$\mathrm{Var}\{\widehat{\sigma}^2_{1,n_K}(\widetilde{m}_1) \mid \widetilde{m}_1\} \;=\; \frac{\mu_4(\widetilde{m}_1)}{n_K} \;+\; \frac{(n_K - 3)\{\sigma^2_1(\widetilde{m}_1)\}^2}{n_K(n_K - 1)}, \tag{S.7}$$

where $\mu_4(\widetilde{m}_1) := \mathbb{E}_{\mathbf{Z}}([W(\mathbf{Z};\widetilde{m}_1) - \mathbb{E}_{\mathbf{Z}}\{W(\mathbf{Z};\widetilde{m}_1)\}]^4\,|\,\widetilde{m}_1)$ is the fourth central moment of $W(\mathbf{Z};\widetilde{m}_1)$ w.r.t. $\mathbf{Z}$ ($\perp\!\!\!\perp \widetilde{m}_1$) given $\widetilde{m}_1$. Now, by Assumption 4.1 (i) and the construction/definition of $W(\mathbf{Z};\widetilde{m}_1)$ as above, i.e., $W(\mathbf{Z};\widetilde{m}_1) \equiv Y - \widetilde{m}_1(\mathbf{X})$, we have that $\mu_4(\widetilde{m}_1) = O_{\mathbb{P}}(1)$ under the joint probability distribution $\Pi^{(1)}_{\mathbf{m}}(\mathcal{S}_1)$.

Consequently, we obtain that $Z_{n_K}(\widetilde{m}_1) = o_{\mathbb{P}_{\widetilde{m}_1}}(1)$. This equivalently gives that for some sequence $b_{n_K,s_K} \to 0$, $Z_{n_K}(\widetilde{m}_1) = O_{\mathbb{P}_{\widetilde{m}_1}}(b_{n_K,s_K})$, where $s_K = n - n/K$ (the size of $\mathcal{S}_1$). We note that the double index used in b_{n_K,s_K} indicates that the rate depends not only on the term $\widehat{\sigma}^2_{1,n_K}(\widetilde{m}_1)$ but also on the size of $\mathcal{S}_1$ which is used to obtain the distribution $\Pi^{(1)}_{\mathbf{m}}$ of $\widetilde{m}_1$. Then by applying Lemma S4.6 (b), we obtain that $\Sigma_{\epsilon,n_K}(\mathcal{D}_1,\widetilde{m}_1) = O_{\mathbb{P}_{\mathcal{D}_1,\widetilde{m}_1}}(b_{n_K,s_K})$ which implies that $\Sigma_{\epsilon,n_K}(\mathcal{D}_1,\widetilde{m}_1) = o_{\mathbb{P}_{\mathcal{D}_1,\widetilde{m}_1}}(1)$. Hence, we conclude that as $n, N \to \infty$, $T_{11} \xrightarrow{\mathbb{P}} 0$ under $\mathbb{P}_{\mathcal{D}_1,\widetilde{m}_1}$. ■

Similarly, we follow the same steps for the term T_{12} in (S.5) and finally obtain that

$$T_{12} \;= \|t_{\nu_{N_K}}(0, \widehat{\sigma}^2_{2,N_K}(\widetilde{m}_1)) - \mathcal{N}(0, \sigma^2_2(\widetilde{m}_1))\|_{\mathrm{TV}} \;\leq\; \frac{C_0}{\nu_{N_K}} + \left|\frac{\widehat{\sigma}^2_{2,N_K}(\widetilde{m}_1) - \sigma^2_2(\widetilde{m}_1)}{\sigma^2_2(\widetilde{m}_1)}\right|.$$

Then, it is enough to show that

$$\Sigma_{r,N_K}(\mathcal{D}_1,\widetilde{m}_1) \;:=\; \left|\frac{\widehat{\sigma}^2_{2,N_K}(\widetilde{m}_1) - \sigma^2_2(\widetilde{m}_1)}{\sigma^2_2(\widetilde{m}_1)}\right| \;\to\; 0 \;\text{ in probability w.r.t. } \mathbb{P}_{\mathcal{D}_1,\widetilde{m}_1}.$$

Next, by using $\mathcal{D}_1 \perp\!\!\!\perp \widetilde{m}_1$ and by following the same idea and steps in (S.6), we observe that for any $t > 0$,

$$\mathbb{P}_{\mathcal{D}_1,\widetilde{m}_1}\{\Sigma_{r,N_K}(\mathcal{D}_1,\widetilde{m}_1) > t\} \;=\; \mathbb{E}_{\widetilde{m}_1}[\mathbb{P}_{\mathcal{D}_1}\{\Sigma_{r,N_K}(\mathcal{D}_1,\widetilde{m}_1) > t \mid \widetilde{m}_1\}].$$

Then, by the DCT (or Lemma S4.6 (b)), it is enough to show $\mathbb{P}_{\mathcal{D}_1}\{\Sigma_{r,N_K}(\mathcal{D}_1,\widetilde{m}_1) > t \mid \widetilde{m}_1\} \xrightarrow{\mathbb{P}} 0$ under $\mathbb{P}_{\widetilde{m}_1}$ to conclude that $\mathbb{P}_{\mathcal{D}_1,\widetilde{m}_1}\{\Sigma_{r,N_K}(\mathcal{D}_1,\widetilde{m}_1) > t\} \to 0$.

Further, given $\widetilde{m}_1$, $\sigma_2^2(\widetilde{m}_1)$ is a fixed non-random quantity, for any $t > 0$, we have that

$$\mathbb{P}_{\mathcal{D}_1}\{\Sigma_{r,N_K}(\mathcal{D}_1,\widetilde{m}_1) > t \mid \widetilde{m}_1\} \;=\; \mathbb{P}_{\mathcal{D}_1}\left\{\left|\widehat{\sigma}^2_{2,N_K}(\widetilde{m}_1) - \sigma_2^2(\widetilde{m}_1)\right| > \tilde{t} \mid \widetilde{m}_1\right\},$$

for $\tilde{t} = t\ |\sigma_2^2(\widetilde{m}_1)| > 0$. Let $Z_{N_K}(\widetilde{m}_1) := \mathbb{P}_{\mathcal{D}_1}\{|\widehat{\sigma}^2_{2,N_K}(\widetilde{m}_1) - \sigma_2^2(\widetilde{m}_1)| > \tilde{t} \mid \widetilde{m}_1\}$. We note that $Z_{N_K}(\widetilde{m}_1)$ is a random variable where its randomness comes from $\widetilde{m}_1$ and $0 \le Z_{N_K}(\widetilde{m}_1) \le 1$ by its definition. Then, by applying the DCT (or directly using Lemma S4.6), it is sufficient to prove that $Z_{N_K}(\widetilde{m}_1) \to 0$ in probability w.r.t. $\mathbb{P}_{\widetilde{m}_1}$ to conclude that $T_{12} \to 0$ in probability w.r.t. $\mathbb{P}_{\widetilde{\mathcal{D}}_1}$. Towards that, since $\mathbb{E}_{\mathcal{D}_1}\{\widehat{\sigma}^2_{2,N_K}(\widetilde{m}_1) - \sigma_2^2(\widetilde{m}_1) \mid \widetilde{m}_1\} = 0$ by the constriction of $\widehat{\sigma}^2_{2,N_K}(\widetilde{m}_1)$ (refer to (8) in the Supplementary Material by setting $k = 1$), by Chebyshev's inequality and following the same algebraic calculations as those used to obtain (S.7) but applied to $\widehat{\sigma}^2_{2,N_K}(\widetilde{m}_1)$ this time, we have

$$\mathbb{P}_{\mathcal{D}_1}\{\Sigma_{r,N_K}(\mathcal{D}_1,\widetilde{m}_1) > t \mid \widetilde{m}_1\} \;\le\; \frac{\mathrm{Var}\{\widehat{\sigma}^2_{2,N_K}(\widetilde{m}_1)\}}{\tilde{t}^2} \;=\; \frac{\mu_4(\widetilde{m}_1)}{N_K} + \frac{(N_K-3)\{\sigma_2^2(\widetilde{m}_1)\}^2}{N_K(N_K-1)},$$

where $\mu_4(\widetilde{m}_1) := \mathbb{E}_{\mathbf{X}}([\widetilde{m}_1(\mathbf{X}) - \mathbb{E}\{\widetilde{m}_1(\mathbf{X})\}]^4 \,|\, \widetilde{m}_1)$ and the last step follows from Mood et al. (1974, Theorem 2 in Chapter VI). Then, by Assumption 4.1 (i), we have $\mu_4(\widetilde{m}_1) = O_{\mathbb{P}}(1)$ under the joint probability distribution $\Pi^{(1)}_{\mathbf{m}}(\mathcal{S}_1)$. Hence we obtain that $Z_{N_K}(\widetilde{m}_1) = o_{\mathbb{P}_{\widetilde{m}_1}}(1)$. This directly gives that for some sequence $d_{N_K,s_K} \to 0$, $Z_{N_K}(\widetilde{m}_1) = O_{\mathbb{P}_{\widetilde{m}_1}}(d_{N_K,s_K})$. We again note that the double index in d_{N_K,s_K} indicates the dependency of the rate not only on the term $\widehat{\sigma}^2_{2,N_K}(\widetilde{m}_1)$ but also on the size of $\mathcal{S}_1$ which is used to obtain the distribution $\Pi^{(1)}_{\mathbf{m}}$ of $\widetilde{m}_1$. Then applying Lemma S4.6 (b), we obtain that $\Sigma_{r,N_K}(\mathcal{D}_1,\widetilde{m}_1) = O_{\mathbb{P}_{\mathcal{D}_1,\widetilde{m}_1}}(d_{N_K,s_K})$ which implies that $\Sigma_{r,N_K}(\mathcal{D}_1,\widetilde{m}_1) = o_{\mathbb{P}_{\mathcal{D}_1,\widetilde{m}_1}}(1)$. This concludes that as $n, N \to \infty$, $T_{12} \xrightarrow{\mathbb{P}} 0$ w.r.t. $\mathbb{P}_{\mathcal{D}_1,\widetilde{m}_1}$ and so T_1 in (S.4) $\xrightarrow{\mathbb{P}} 0$ w.r.t. $\mathbb{P}_{\mathcal{D}_1,\widetilde{m}_1}$. ■

Next, we consider the term $T_2 = \|\mathcal{N}(\widehat{\theta}^{(1)}_{\mathrm{BDM}}(\widetilde{m}_1), \tau^2_{n_K,N_K}(\widetilde{m}_1)) - \mathcal{N}(\widehat{\theta}^{(1)}_{\mathrm{BDM}}(m^*), \tau^2_{n_K,N_K}(m^*))\|_{\mathrm{TV}}$ in (S.4) and show that T_2 goes to zero in probability w.r.t. $\mathbb{P}_{\widetilde{\mathcal{D}}_1}$. To make the proof of this part clearer and streamlined, we first present the following lemma:

Lemma S4.7. *Under Assumption 4.1 and the setup of Theorem 4.1, we have*

$$\|m^*(\mathbf{X}) - \widetilde{m}_1(\mathbf{X})\|_{\mathbb{L}_2(\mathbb{P}_{\mathbf{X}})} \;=\; o_{\mathbb{P}_{\widetilde{m}_1}}(1). \tag{S.8}$$

For brevity, the proof of Lemma S4.7 is presented in Section S5.5 of the Supplementary Material.

Suppose Lemma S4.7 holds. Then, by the triangle inequality and the invariance property of the TV distance from Lemma S4.1, we obtain that

$$\begin{aligned} T_2 \;&=\; \|\mathcal{N}(\widehat{\theta}^{(1)}_{\mathrm{BDM}}(\widetilde{m}_1), \tau^2_{n_K,N_K}(\widetilde{m}_1)) - \mathcal{N}(\widehat{\theta}^{(1)}_{\mathrm{BDM}}(m^*), \tau^2_{n_K,N_K}(m^*))\|_{\mathrm{TV}} \\ &\le\; \|\mathcal{N}(\widehat{\theta}^{(1)}_{\mathrm{BDM}}(\widetilde{m}_1), \tau^2_{n_K,N_K}(\widetilde{m}_1)) - \mathcal{N}(\widehat{\theta}^{(1)}_{\mathrm{BDM}}(\widetilde{m}_1), \tau^2_{n_K,N_K}(m^*))\|_{\mathrm{TV}} \end{aligned}$$

$$
\begin{aligned}
&+ \ \|\mathcal{N}(\widehat{\theta}^{(1)}_{\mathrm{BDM}}(\widetilde{m}_1), \tau^2_{n_K,N_K}(m^*)) - \mathcal{N}(\widehat{\theta}^{(1)}_{\mathrm{BDM}}(m^*), \tau^2_{n_K,N_K}(m^*))\|_{\mathrm{TV}} \\
&= \ \|\mathcal{N}(0, \tau^2_{n_K,N_K}(\widetilde{m}_1)) - \mathcal{N}(0, \tau^2_{n_K,N_K}(m^*))\|_{\mathrm{TV}} + \|\mathcal{N}(\alpha, 1) - \mathcal{N}(0,1)\|_{\mathrm{TV}} \\
&:= \ T_{21} + T_{22}, \quad \text{where } \alpha = \ \{\widehat{\theta}^{(1)}_{\mathrm{BDM}}(\widetilde{m}_1) - \widehat{\theta}^{(1)}_{\mathrm{BDM}}(m^*)\}/\tau_{n_K,N_K}(m^*).
\end{aligned}
$$

We first consider the TV distance T_{21}. By Lemma S4.3, we obtain the following bound for T_{21}:

$$
T_{21} \ \equiv \ \|\mathcal{N}(0, \tau^2_{n_K,N_K}(\widetilde{m}_1)) - \mathcal{N}(0, \tau^2_{n_K,N_K}(m^*))\|_{\mathrm{TV}} \ \leq \ C \left| \frac{\tau^2_{n_K,N_K}(\widetilde{m}_1) - \tau^2_{n_K,N_K}(m^*)}{\tau^2_{n_K,N_K}(m^*)} \right|, \quad \text{(S.9)}
$$

for some constant $C < \infty$. By using the definitions of the terms $\tau^2_{n_K,N_K}(\widetilde{m}_1)$ and $\tau^2_{n_K,N_K}(m^*)$ (refer to Theorem 4.1), we obtain that

$$
\|\mathcal{N}(0, \tau^2_{n_K,N_K}(\widetilde{m}_1)) - \mathcal{N}(0, \tau^2_{n_K,N_K}(m^*))\|_{\mathrm{TV}} \leq \frac{C\left\{|\sigma^2_1(\widetilde{m}_1) - \sigma^2_1(m^*)| + (n/N)|\sigma^2_2(\widetilde{m}_1) - \sigma^2_2(m^*)|\right\}}{\sigma^2_1(m^*) + (n/N)\sigma^2_2(m^*)}.
$$

Since the denominator (on the right-hand side (RHS) above) is greater than and bounded away from zero, it is enough to show that both of the terms $|\sigma^2_1(\widetilde{m}_1) - \sigma^2_1(m^*)|$ and $|\sigma^2_2(\widetilde{m}_1) - \sigma^2_2(m^*)|$ in the numerator (on the RHS above) converge to 0 in probability. By using $\widetilde{m}_1 \perp\!\!\!\perp (Y, \mathbf{X}) \in \mathcal{D}_1$ (refer to the construction of BDMI in Section 3.3 by taking $k = 1$), we observe that

$$
\begin{aligned}
|\sigma^2_1(\widetilde{m}_1) - \sigma^2_1(m^*)| \ &= \ |\mathrm{Var}_{\mathbf{Z}}[\{Y - \widetilde{m}_1(\mathbf{X})\}|\widetilde{m}_1] - \mathrm{Var}_{\mathbf{Z}}\{Y - m^*(\mathbf{X})\}| \\
&\leq \ |\mathbb{E}_{\mathbf{Z}}[\{m^*(\mathbf{X}) - \widetilde{m}_1(\mathbf{X})\}\{2Y - \widetilde{m}_1(\mathbf{X}) - m^*(\mathbf{X})\} \mid \widetilde{m}_1]| \\
&\quad + |\mathbb{E}_{\mathbf{Z}}[\{m^*(\mathbf{X}) - \widetilde{m}_1(\mathbf{X})\} \mid \widetilde{m}_1]\mathbb{E}_{\mathbf{Z}}[\{2Y - \widetilde{m}_1(\mathbf{X}) - m^*(\mathbf{X})\} \mid \widetilde{m}_1]|,
\end{aligned}
$$

where the last step uses the triangle inequality. By applying the Cauchy–Schwarz inequality and the triangle inequality for the $\mathbb{L}_2(\mathbb{P}_{\mathbf{X}})$-norm (and the underlying inner product), we have

$$
\begin{aligned}
|\sigma^2_1(\widetilde{m}_1) - \sigma^2_1(m^*)| \ &\leq \ 2\|m^*(\mathbf{X}) - \widetilde{m}_1(\mathbf{X})\|_{\mathbb{L}_2(\mathbb{P}_{\mathbf{X}})}\{\|2Y\|_{\mathbb{L}_2(\mathbb{P}_Y)} + \|2m^*(\mathbf{X})\|_{\mathbb{L}_2(\mathbb{P}_{\mathbf{X}})}\} \\
&\quad + \ 2\|m^*(\mathbf{X}) - \widetilde{m}_1(\mathbf{X})\|^2_{\mathbb{L}_2(\mathbb{P}_{\mathbf{X}})}.
\end{aligned}
$$

Since (S.8) holds by Lemma S4.7, and $\|Y\|_{\mathbb{L}_2(\mathbb{P}_Y)} < \infty$ and $\|m^*(\mathbf{X})\|_{\mathbb{L}_2(\mathbb{P}_{\mathbf{X}})} < \infty$ by Assumption 4.1 (ii), we conclude that $|\sigma^2_1(\widetilde{m}_1) - \sigma^2_1(m^*)| \xrightarrow{\mathbb{P}} 0$ under $\mathbb{P}_{\mathcal{D}_1,\widetilde{m}_1}$. Similarly, we have

$$
\begin{aligned}
|\sigma^2_2(\widetilde{m}_1) - \sigma^2_2(m^*)| \ &= \ |\mathrm{Var}\{\widetilde{m}_1(\mathbf{X})\,|\,\widetilde{m}_1\} - \mathrm{Var}\{m^*(\mathbf{X})\}| \\
&\leq \ 2\|\widetilde{m}_1(\mathbf{X}) - m^*(\mathbf{X})\|_{\mathbb{L}_2(\mathbb{P}_{\mathbf{X}})}\{\|2m^*(\mathbf{X})\|_{\mathbb{L}_2(\mathbb{P}_{\mathbf{X}})} + \|\widetilde{m}_1(\mathbf{X}) - m^*(\mathbf{X})\|_{\mathbb{L}_2(\mathbb{P}_{\mathbf{X}})}\},
\end{aligned}
$$

where the last step comes from the Cauchy-Schwarz inequality. Since $\|m^*(\mathbf{X})\|_{\mathbb{L}_2(\mathbb{P}_{\mathbf{X}})} < \infty$ and (S.8) holds (by Assumption 4.1 (ii)), we have $|\sigma^2_2(\widetilde{m}_1) - \sigma^2_2(m^*)| \xrightarrow{\mathbb{P}} 0$ under $\mathbb{P}_{\mathcal{D}_1,\widetilde{m}_1}$. Referring back to the inequality (S.9), and using all conclusions obtained above, and the fact that the denominator $\tau^2_{n_K,N_K}(m^*)$ in (S.9) is bounded away from zero, we now conclude that $T_{21} \to 0$ under $\mathbb{P}_{\mathcal{D}_1,\widetilde{m}_1}$. ■

Now, we consider $T_{22} = \|\mathcal{N}(\{\widehat{\theta}^{(1)}_{\mathrm{BDM}}(\widetilde{m}_1) - \widehat{\theta}^{(1)}_{\mathrm{BDM}}(m^*)\}/\tau_{n_K,N_K}(m^*), 1) - \mathcal{N}(0,1)\|_{\mathrm{TV}}$. Using

Lemma S4.2, we observe that:

$$T_{22} \;\leq\; \frac{|\sqrt{n_K}\{\widehat{\theta}^{(1)}_{\mathrm{BDM}}(\widetilde{m}_1) - \widehat{\theta}^{(1)}_{\mathrm{BDM}}(m^*)\}|}{\sqrt{2\pi n_K \tau^2_{n_K,N_K}(m^*)}}. \tag{S.10}$$

Since the denominator $2\,\pi\, n_K\, \tau^2_{n_K,N_K}(m^*) = \sigma^2_1(m^*) + (n/N)\sigma^2_2(m^*)$ is bounded below and away from zero, as $n, N \to \infty$, it converges to a non–random quantity which is bounded below and away from zero. This reduces the problem to showing the numerator $|\sqrt{n_K}\{\widehat{\theta}^{(1)}_{\mathrm{BDM}}(\widetilde{m}_1) - \widehat{\theta}^{(1)}_{\mathrm{BDM}}(m^*)\}| \xrightarrow{\mathbb{P}} 0$ under $\mathbb{P}_{\widetilde{\mathcal{D}}_1}$ (where recall that $\widetilde{\mathcal{D}}_k \equiv \mathcal{D}_k \cup \mathcal{S}_k$) thanks to the continuous mapping theorem (CMT) (van der Vaart, 2000, Theorem 2.3). We can define a continuous map $h(x,y) := xy^{-1}$ on $\mathbb{R} \times \mathbb{R}^+$ and then apply the CMT to argue that

$$\frac{|\sqrt{n_K}\{\widehat{\theta}^{(1)}_{\mathrm{BDM}}(\widetilde{m}_1) - \widehat{\theta}^{(1)}_{\mathrm{BDM}}(m^*)\}|}{\sqrt{2\pi n_K \tau^2_{n_K,N_K}(m^*)}} \;\xrightarrow{\mathbb{P}}\; 0,$$

which implies that $\|\mathcal{N}(\{\widehat{\theta}^{(1)}_{\mathrm{BDM}}(\widetilde{m}_1) - \widehat{\theta}^{(1)}_{\mathrm{BDM}}(m^*)\}/\tau^2_{n_K,N_K}(m^*), 1) - \mathcal{N}(0,1)\|_{\mathrm{TV}} \xrightarrow{\mathbb{P}} 0$ under $\mathbb{P}_{\widetilde{\mathcal{D}}_1}$.

Towards showing the numerator's convergence, by writing the terms explicitly, we observe that

$$\begin{aligned}
&|\sqrt{n_K}\,\{\widehat{\theta}^{(1)}_{\mathrm{BDM}}(\widetilde{m}_1) - \widehat{\theta}^{(1)}_{\mathrm{BDM}}(m^*)\}| \\
&= \left|\sqrt{n_K}\left[\frac{1}{n_K}\sum_{i\in\mathcal{I}_1}\{Y_i - \widetilde{m}_1(\mathbf{X}_i)\} + \frac{1}{N_K}\sum_{i\in\mathcal{J}_1}\widetilde{m}_1(\mathbf{X}_i) - \frac{1}{n_K}\sum_{i\in\mathcal{I}_1}\{Y_i - m^*(\mathbf{X}_i)\} - \frac{1}{N_K}\sum_{i\in\mathcal{J}_1} m^*(\mathbf{X}_i)\right]\right| \\
&= \left|\sqrt{n_K}\,\left[\frac{1}{n_K}\sum_{i\in\mathcal{I}_1}\{m^*(\mathbf{X}_i) - \widetilde{m}_1(\mathbf{X}_i)\} - \frac{1}{N_K}\sum_{i\in\mathcal{J}_1}\{m^*(\mathbf{X}_i) - \widetilde{m}_1(\mathbf{X}_i)\}\right]\right| \\
&= \left|\sqrt{n_K}\,[\mathbb{E}^{(1)}_{n_K}\{m^*(\mathbf{X}) - \widetilde{m}_1(\mathbf{X})\} - \mathbb{E}^{(1)}_{N_K}\{m^*(\mathbf{X}) - \widetilde{m}_1(\mathbf{X})\}]\right| \\
&= \left|\mathbb{G}^{(1)}_{n_K}\{m^*(\mathbf{X}) - \widetilde{m}_1(\mathbf{X})\} - \frac{\sqrt{n_K}}{\sqrt{N_K}}\,\mathbb{G}^{(1)}_{N_K}\{m^*(\mathbf{X}) - \widetilde{m}_1(\mathbf{X})\}\right| \\
&\leq |\mathbb{G}^{(1)}_{n_K}\{m^*(\mathbf{X}) - \widetilde{m}_1(\mathbf{X})\}| + \frac{\sqrt{n}}{\sqrt{N}}\,|\mathbb{G}^{(1)}_{N_K}\{m^*(\mathbf{X}) - \widetilde{m}_1(\mathbf{X})\}|,
\end{aligned}$$

where recall the notations $\mathbb{G}^{(k)}_{n_K}(\cdot)$ and $\mathbb{G}^{(k)}_{N_K}(\cdot)$ as defined at the beginning of Section S4 (and here we set $k=1$). We next want to show $\mathbb{G}^{(1)}_{n_K}\{m^*(\mathbf{X}) - \widetilde{m}_1(\mathbf{X})\} = o_{\mathbb{P}_{\mathcal{D}_1,\widetilde{m}_1}}(1)$. Since $\mathcal{D}_1 \perp\!\!\!\perp \widetilde{m}_1$ by the construction of BDMI (see Section 3.3 by setting $k=1$), we observe that

$$\begin{aligned}
\mathrm{Var}_{\mathbf{X}|\widetilde{m}_1}[\{m^*(\mathbf{X}) - \widetilde{m}_1(\mathbf{X})\} \mid \widetilde{m}_1] &= \mathrm{Var}_{\mathbf{X}}[\{m^*(\mathbf{X}) - \widetilde{m}_1(\mathbf{X})\} \mid \widetilde{m}_1] \quad [\text{by } \mathcal{D}_1 \perp\!\!\!\perp \widetilde{m}_1] \\
&= \mathbb{E}_{\mathbf{X}}\big[\{m^*(\mathbf{X}) - \widetilde{m}_1(\mathbf{X})\}^2 \,|\, \widetilde{m}_1\big] - (\mathbb{E}_{\mathbf{X}}[\{m^*(\mathbf{X}) - \widetilde{m}_1(\mathbf{X})\} \,|\, \widetilde{m}_1])^2 \\
&= o_{\mathbb{P}_{\widetilde{m}_1}}(1),
\end{aligned} \tag{S.11}$$

where the last step follows from (S.8). Then by Chebyshev's inequality, for any $t > 0$, we have

$$V_{n_K}(\widetilde{m}_1) \;:=\; \mathbb{P}_{\mathcal{D}_1}[\,|\mathbb{G}^{(1)}_{n_K}\{m^*(\mathbf{X}) - \widetilde{m}_1(\mathbf{X})\}| > t \mid \widetilde{m}_1\,]$$

$$\leq\ t^{-2}\, n_K \operatorname{Var}_{\mathbf{X}}[\mathbb{E}^{(1)}_{n_K}\{m^*(\mathbf{X}) - \widetilde{m}_1(\mathbf{X})\} \mid \widetilde{m}_1]$$
$$=\ t^{-2} \operatorname{Var}_{\mathbf{X}}[\{m^*(\mathbf{X}) - \widetilde{m}_1(\mathbf{X})\} \mid \widetilde{m}_1]\ =\ o_{\mathbb{P}_{\widetilde{m}_1}}(1),$$

where the last step uses that $\mathbb{E}^{(1)}_{n_K}\{m^*(\mathbf{X}) - \widetilde{m}_1(\mathbf{X})\}$ is a sum of *independent* random variables given $\widetilde{m}_1$, and the earlier conclusion obtained above in (S.11). Hence, we showed that $V_{n_K}(\widetilde{m}_1) \xrightarrow{\mathbb{P}} 0$ under $\mathbb{P}_{\widetilde{m}_1}$. This equivalently gives that $V_{n_K}(\widetilde{m}_1) = O_{\mathbb{P}_{\widetilde{m}_1}}(c_{n_K,s_K})$, for some $c_{n_K,s_K} \to 0$.

We here also note that double index in c_{n_K,s_K} signifies that the rate depends on both n_K and the size s_K of $\mathcal{S}_1$ which is used to obtain $\Pi^{(1)}_{\mathbf{m}}$. Then by applying Lemma S4.6 (b), we obtain that $\mathbb{G}^{(1)}_{n_K}\{m^*(\mathbf{X}) - \widetilde{m}_1(\mathbf{X})\} = O_{\mathbb{P}_{\mathcal{D}_1,\widetilde{m}_1}}(c_{n_K,s_K})$ which implies that $\mathbb{G}^{(1)}_{n_K}\{m^*(\mathbf{X}) - \widetilde{m}_1(\mathbf{X})\} = o_{\mathbb{P}_{\mathcal{D}_1,\widetilde{m}_1}}(1)$.

By following similar steps as above (for $\mathbb{G}^{(1)}_{n_K}\{m^*(\mathbf{X}) - \widetilde{m}_1(\mathbf{X})\}$), we have the same conclusion for $\mathbb{G}^{(1)}_{N_K}\{m^*(\mathbf{X}) - \widetilde{m}_1(\mathbf{X})\}$, i.e., $\mathbb{G}^{(1)}_{N_K}\{m^*(\mathbf{X}) - \widetilde{m}_1(\mathbf{X})\} = o_{\mathbb{P}_{\mathcal{D}_1,\widetilde{m}_1}}(1)$. To be precise, since the condition (S.8) holds and $\mathcal{D}_1 \perp\!\!\!\perp \widetilde{m}_1$ (in particular, $\mathbf{X} \in \mathcal{D}_1 \perp\!\!\!\perp \widetilde{m}_1$), by using (S.11), we first obtain that

$$W_{N_K}(\widetilde{m}_1)\ :=\ \mathbb{P}_{\mathcal{D}_1}[\,|\mathbb{G}^{(1)}_{N_K}\{m^*(\mathbf{X}) - \widetilde{m}_1(\mathbf{X})\}| > t \mid \widetilde{m}_1\,] \quad \text{for any } t > 0,$$
$$\leq\ t^{-2} \operatorname{Var}[\{m^*(\mathbf{X}) - \widetilde{m}_1(\mathbf{X})\} \mid \widetilde{m}_1]\ =\ o_{\mathbb{P}_{\widetilde{m}_1}}(1).$$

This implies that for some $h_{N_K,s_K} \to 0$, $W_{N_K}(\widetilde{m}_1) = O_{\mathbb{P}_{\widetilde{m}_1}}(h_{N_K,s_K})$. We note that the double index in h_{N_K,s_K} reveals the dependency of the rate on both N_K and the size $\mathcal{S}_1$ (used to obtain $\Pi^{(1)}_{\mathbf{m}}$ of $\widetilde{m}_1$). Then by applying Lemma S4.6 (b), we obtain that $\mathbb{G}^{(1)}_{N_K}\{m^*(\mathbf{X}) - \widetilde{m}_1(\mathbf{X})\} = O_{\mathbb{P}_{\mathcal{D}_1,\widetilde{m}_1}}(h_{N_K,s_K})$ which implies that $\mathbb{G}^{(1)}_{N_K}\{m^*(\mathbf{X}) - \widetilde{m}_1(\mathbf{X})\} = o_{\mathbb{P}_{\mathcal{D}_1,\widetilde{m}_1}}(1)$. Referring back to the inequality (S.10), and using all conclusions above along with the fact that the denominator $n_K \tau^2_{n_K,N_K}(m^*)$ in (S.10) is bounded away from zero, we now conclude that $T_{22} \xrightarrow{\mathbb{P}} 0$ under $\mathbb{P}_{\mathcal{D}_1,\widetilde{m}_1}$. Hence, the entire proof of the first part of Theorem 4.1 is now completed (assuming Lemma S4.7 holds, as shown later in Section S5.5). ■

Lastly, the second part of Theorem 4.1 immediately follows from the invariance property of the TV distance (refer to Lemma S4.1), by setting $h = \sqrt{n_K}(\theta - \theta_0)$. Specifically, let $\Pi^{(k)}_{\mathbf{h}}$ be the posterior of h. Then, using the invariance property of the TV distance from Lemma S4.1, we have

$$\big\|\Pi^{(k)}_{\boldsymbol{\theta}} - \mathcal{N}(\widehat{\theta}^{(k)}_{\mathrm{BDM}}(m^*), \tau^2_{n_K,N_K}(m^*))\big\|_{\mathrm{TV}} = \big\|\Pi^{(k)}_{\mathbf{h}} - \mathcal{N}(\sqrt{n_K}\{\widehat{\theta}^{(k)}_{\mathrm{BDM}}(m^*) - \theta_0\}, n_K \tau^2_{n_K,N_K}(m^*))\big\|_{\mathrm{TV}}.$$

We already showed that the left-hand side of the equality above converges to 0 in probability under $\mathbb{P}_{\mathcal{D}_k}$. This directly gives the second claim of Theorem 4.1 and completes the proof of the entire result. ■

S4.5 Proof of Theorem 4.2

Let $\widetilde{\theta}_{\mathrm{BDM}}$ be a new random variable defined as $\widetilde{\theta}_{\mathrm{BDM}} := K\,\theta_{\mathrm{BDM}} = \sum_{k=1}^{K} \theta_k$ where $\theta_1, \ldots, \theta_k$ are independent random variables from the corresponding posteriors $\Pi^{(1)}_{\boldsymbol{\theta}}, \ldots, \Pi^{(K)}_{\boldsymbol{\theta}}$ (refer to (9)). Let $\widetilde{\Pi}_{\boldsymbol{\theta}}$ be the posterior distribution of $\widetilde{\theta}_{\mathrm{BDM}}$. Then, by using the invariance property of the TV distance from Lemma S4.1, to prove Theorem 4.2, it suffices to show the following:

$$\Big\|\widetilde{\Pi}_{\boldsymbol{\theta}} - \mathcal{N}(K\,\widehat{\theta}_{\mathrm{BDM}}(m^*), K^2\,\tau^2_{n,N}(m^*))\Big\|_{\mathrm{TV}}\ \to\ 0 \ \text{ in probability w.r.t. } \mathbb{P}_{\mathcal{D}}.$$

By using Lemma S4.5 and the construction of $\widetilde{\theta}_{\mathrm{BDM}}$, we first obtain that

$$\begin{aligned} T &:= \left\|\widetilde{\Pi}_{\boldsymbol{\theta}} - \mathcal{N}(K\widehat{\theta}_{\mathrm{BDM}}(m^*), K^2\tau^2_{n,N}(m^*))\right\|_{\mathrm{TV}} \\ &\leq \left\|\Pi^{(1)}_{\boldsymbol{\theta}} - \mathcal{N}(\widehat{\theta}^{(1)}_{\mathrm{BDM}}(m^*), \tau^2_{n_K,N_K}(m^*))\right\|_{\mathrm{TV}} + \cdots + \left\|\Pi^{(K)}_{\boldsymbol{\theta}} - \mathcal{N}(\widehat{\theta}^{(K)}_{\mathrm{BDM}}(m^*), \tau^2_{n_K,N_K}(m^*))\right\|_{\mathrm{TV}} \\ &:= T_1 + \cdots + T_K, \end{aligned}$$

where $n_K = n/K, N_K = N/K$ and for $k = 1, \ldots, K$, $T_k = \left\|\Pi^{(k)}_{\boldsymbol{\theta}} - \mathcal{N}(\widehat{\theta}^{(k)}_{\mathrm{BDM}}(m^*), \tau^2_{n_K,N_K}(m^*))\right\|_{\mathrm{TV}}$,

$$\widehat{\theta}^{(k)}_{\mathrm{BDM}}(m^*) = \frac{1}{n_K}\sum_{i\in\mathcal{I}_k}\{Y_i - m^*(\mathbf{X}_i)\} + \frac{1}{N_K}\sum_{i\in\mathcal{J}_k} m^*(\mathbf{X}_i) \quad \text{and} \quad \tau^2_{n_K,N_K}(m^*) = \frac{\sigma^2_1(m^*)}{n_K} + \frac{\sigma^2_2(m^*)}{N_K}.$$

Under the assumptions of Theorem 4.2, we can apply Theorem 4.1 to each of the TV distances $T_1, \ldots, T_K$ defined above. This gives that $T_k \xrightarrow{\mathbb{P}} 0$ under $\mathbb{P}_{\mathcal{D}}$ for $k = 1, \ldots, K$. Since each TV distance converges to 0 in probability and K is *fixed*, we finally obtain that $T \xrightarrow{\mathbb{P}} 0$ under $\mathbb{P}_{\mathcal{D}}$. Hence, by using the invariance property of the TV distance from Lemma S4.1, we first observe that

$$\left\|\Pi_{\boldsymbol{\theta}} - \mathcal{N}(\widehat{\theta}_{\mathrm{BDM}}(m^*), \tau^2_{n,N}(m^*))\right\|_{\mathrm{TV}} = \left\|\widetilde{\Pi}_{\boldsymbol{\theta}} - \mathcal{N}(K\,\widehat{\theta}_{\mathrm{BDM}}(m^*), K^2\,\tau^2_{n,N}(m^*))\right\|_{\mathrm{TV}}.$$

Since we already proved that the RHS of the equality above converges to 0 in probability w.r.t. $\mathbb{P}_{\mathcal{D}}$, we immediately conclude that $\left\|\Pi_{\boldsymbol{\theta}} - \mathcal{N}(\widehat{\theta}_{\mathrm{BDM}}(m^*), \tau^2_{n,N}(m^*))\right\|_{\mathrm{TV}} \to 0$ in probability w.r.t. $\mathbb{P}_{\mathcal{D}}$. ■

S4.6 Proof of Corollary 4.1

We first define a new random variable $Z_n \equiv \sqrt{n}(\,\theta_{\mathrm{BDM}} - \theta_0\,)$ with corresponding posterior distribution $\mathcal{P}_n(\mathcal{D})$ where θ_{BDM} is as defined in (9). Let $\mathcal{P}(\mathcal{D})$ be the corresponding limiting Normal distribution with mean $\sqrt{n}\{\,\widehat{\theta}_{\mathrm{BDM}}(m^*) - \theta_0\,\}$ and variance $n\,\tau^2_{n,N}(m^*)$ obtained from Theorem 4.2 (after applying the appropriate scaling and location shifts, in particular, using Lemma S4.1) for the posterior $\mathcal{P}_n(\mathcal{D})$ and let $Z \sim \mathcal{P}(\mathcal{D})$. Note that the distributions $\mathcal{P}_n \equiv \mathcal{P}_n(\mathcal{D})$ and $\mathcal{P} \equiv \mathcal{P}(\mathcal{D})$ are random through the data $\mathcal{D}$. Next, observe that

$$\begin{aligned} &\sqrt{n}\{\,\widehat{\theta}_{\mathrm{BDM}}(\widetilde{m}_{\mathrm{CF}}) - \theta_0\,\} - \sqrt{n}\{\,\widehat{\theta}_{\mathrm{BDM}}(m^*) - \theta_0\,\} = o_{\mathbb{P}_{\mathcal{D}}}(1) \\ &\quad\Leftrightarrow\ |\sqrt{n}\,\{\,\widehat{\theta}_{\mathrm{BDM}}(\widetilde{m}_{\mathrm{CF}}) - \theta_0\,\} - \sqrt{n}\,\{\,\widehat{\theta}_{\mathrm{BDM}}(m^*) - \theta_0\,\}| \to 0 \ \text{ in probability under } \mathbb{P}_{\mathcal{D}} \\ &\quad\Leftrightarrow\ |\mathbb{E}_{Z_n\sim\mathcal{P}_n(\mathcal{D})}(Z_n \mid \mathcal{D}) - \mathbb{E}_{Z\sim\mathcal{P}(\mathcal{D})}(Z \mid \mathcal{D})| \to 0 \ \text{ in probability under } \mathbb{P}_{\mathcal{D}}, \end{aligned}$$

where the last line uses the constructions of the random variables Z_n and Z. To use the exact formula of the TV distance between two Normal distributions with the same variance (refer to (S.12), specifically, see Lemma S4.2), we now consider the following Normal distribution: $\widetilde{\mathcal{P}} \equiv \widetilde{\mathcal{P}}(\mathcal{D}) := \mathcal{N}(\sqrt{n}\,\{\widehat{\theta}_{\mathrm{BDM}}(\widetilde{m}_{\mathrm{CF}}) - \theta_0\}, n\,\tau^2_{n,N}(m^*))$. We note that the distributions $\widetilde{\mathcal{P}}$ and $\mathcal{P}_n$ have the same mean $\sqrt{n}\,\{\,\widehat{\theta}_{\mathrm{BDM}}(\widetilde{m}_{\mathrm{CF}}) - \theta_0\,\}$, while the distributions $\widetilde{\mathcal{P}}$ and $\mathcal{P}$ have the same variance $n\,\tau^2_{n,N}(m^*)$. Let $\widetilde{Z} \sim \widetilde{\mathcal{P}}(\mathcal{D})$. Then, we observe that

$$\begin{aligned} |\mathbb{E}_{Z_n\sim\mathcal{P}_n}(Z_n|\mathcal{D}) - \mathbb{E}_{Z\sim\mathcal{P}}(Z|\mathcal{D})| &= |\mathbb{E}_{Z_n\sim\mathcal{P}_n}(Z_n|\mathcal{D}) - \mathbb{E}_{\widetilde{Z}\sim\widetilde{\mathcal{P}}}(\widetilde{Z}|\mathcal{D}) + \mathbb{E}_{\widetilde{Z}\sim\widetilde{\mathcal{P}}}(\widetilde{Z}|\mathcal{D}) - \mathbb{E}_{Z\sim\mathcal{P}}(Z|\mathcal{D})| \\ &= |\mathbb{E}_{\widetilde{Z}\sim\widetilde{\mathcal{P}}}(\widetilde{Z}|\mathcal{D}) - \mathbb{E}_{Z\sim\mathcal{P}}(Z|\mathcal{D})|, \end{aligned}$$

where the last step uses the fact $\mathbb{E}_{Z_n\sim\mathcal{P}_n}(Z_n|\mathcal{D}) = \mathbb{E}_{\widetilde{Z}\sim\widetilde{\mathcal{P}}}(\widetilde{Z}|\mathcal{D})$. Since both distributions $\widetilde{\mathcal{P}}$ and $\mathcal{P}$ are Normal with the same variance $n\,\tau^2_{n,N}(m^*)$, by the invariance property of the TV distance from Lemma S4.1, we have:

$$\begin{aligned}\|\mathcal{P}-\widetilde{\mathcal{P}}\|_{\mathrm{TV}} &\equiv \|\mathcal{P}(\mathcal{D})-\widetilde{\mathcal{P}}(\mathcal{D})\|_{\mathrm{TV}}\\ &= \big\|\mathcal{N}(\sqrt{n}\{\widehat{\theta}_{\mathrm{BDM}}(\widetilde{m}_{\mathrm{CF}})-\theta_0\}, n\tau^2_{n,N}(m^*)) - \mathcal{N}(\sqrt{n}\{\widehat{\theta}_{\mathrm{BDM}}(m^*)-\theta_0\}, n\tau^2_{n,N}(m^*))\big\|_{\mathrm{TV}}\\ &= \big\|\mathcal{N}(\alpha,1)-\mathcal{N}(0,1)\big\|_{\mathrm{TV}},\end{aligned}$$

where $\alpha = [\,\sqrt{n}\{\widehat{\theta}_{\mathrm{BDM}}(\widetilde{m}_{\mathrm{CF}})-\theta_0\} - \sqrt{n}\{\widehat{\theta}_{\mathrm{BDM}}(m^*)-\theta_0\}\,]/\sqrt{n\tau^2_{n,N}(m^*)}$. Then, by Lemma S4.2,

$$\|\mathcal{P}(\mathcal{D})-\widetilde{\mathcal{P}}(\mathcal{D})\|_{\mathrm{TV}} \;=\; 2\,\Phi\left(\frac{|\sqrt{n}\{\widehat{\theta}_{\mathrm{BDM}}(\widetilde{m}_{\mathrm{CF}})-\theta_0\} - \sqrt{n}\{\widehat{\theta}_{\mathrm{BDM}}(m^*)-\theta_0\}|}{2\sqrt{n\,\tau^2_{n,N}(m^*)}}\right) - 1, \qquad \text{(S.12)}$$

where $\Phi(\cdot)$ is the CDF of the standard Normal distribution $\mathcal{N}(0,1)$. Since the term $n\,\tau^2_{n,N}(m^*) = \sigma^2_1(m^*)+(n/N)\sigma^2_2(m^*)$ in the denominator (on the RHS in (S.12)) is greater than and away from zero, (S.12) implies that to complete the proof of Corollary 4.1, it is enough to show $\|\mathcal{P}(\mathcal{D})-\widetilde{\mathcal{P}}(\mathcal{D})\|_{\mathrm{TV}} \xrightarrow{\mathbb{P}} 0$ under $\mathbb{P}_{\mathcal{D}}$. Towards that, we first use the same approach used in the proof of Theorem 4.2 in Section S4.5. More explicitly, we write both of the Normal distributions $\widetilde{\mathcal{P}}(\mathcal{D})$ and $\mathcal{P}(\mathcal{D})$ as convolutions of K Normal distributions as follows:

$$\begin{aligned}\widetilde{\mathcal{P}}(\mathcal{D}) &= \mathcal{N}(\sqrt{n}\{\,\widehat{\theta}_{\mathrm{BDM}}(\widetilde{m}_{\mathrm{CF}})-\theta_0\,\}, n\tau^2_{n,N}(m^*)) \;=\; \widetilde{\mathcal{P}}^{(1)}(\widetilde{\mathcal{D}}_1) * \cdots * \widetilde{\mathcal{P}}^{(K)}(\widetilde{\mathcal{D}}_K), \quad \text{and}\\ \mathcal{P}(\mathcal{D}) &= \mathcal{N}(\sqrt{n}\{\,\widehat{\theta}_{\mathrm{BDM}}(m^*)-\theta_0\,\}, n\tau^2_{n,N}(m^*)) \;=\; \mathcal{P}^{(1)}(\widetilde{\mathcal{D}}_1) * \cdots * \mathcal{P}^{(K)}(\widetilde{\mathcal{D}}_K),\end{aligned}$$

where for $k=1,\ldots,K$, $\widetilde{\mathcal{P}}^{(k)}(\widetilde{\mathcal{D}}_k) := \mathcal{N}(\sqrt{n}\,\{\widehat{\theta}^{(k)}_{\mathrm{BDM}}(\widetilde{m}_k)-\theta_0\}/K, n\,\tau^2_{n_K,N_K}(m^*)/K^2)$ and $\mathcal{P}^{(k)}(\widetilde{\mathcal{D}}_k) := \mathcal{N}(\sqrt{n}\,\{\widehat{\theta}^{(k)}_{\mathrm{BDM}}(m^*)-\theta_0\}/K, n\,\tau^2_{n_K,N_K}(m^*)/K^2)$ with the parameters are as given in (8) and in Theorem 4.1.

Then, by applying Lemma S4.5, we obtain that

$$\big\|\,\mathcal{P}(\mathcal{D})-\widetilde{\mathcal{P}}(\mathcal{D})\,\big\|_{\mathrm{TV}} \;\le\; \big\|\,\mathcal{P}^{(1)}(\widetilde{\mathcal{D}}_1)-\widetilde{\mathcal{P}}^{(1)}(\widetilde{\mathcal{D}}_1)\,\big\|_{\mathrm{TV}} \;+\; \cdots \;+\; \big\|\,\mathcal{P}^{(K)}(\widetilde{\mathcal{D}}_K)-\widetilde{\mathcal{P}}^{(K)}(\widetilde{\mathcal{D}}_K)\,\big\|_{\mathrm{TV}}.$$

Furthermore, in the proof Theorem 4.1 (refer to Section S4.4), we have already established a result showing that $T_{22} = \big\|\mathcal{N}(\widehat{\theta}^{(1)}_{\mathrm{BDM}}(\widetilde{m}_1), \tau^2_{n_K,N_K}(m^*)) - \mathcal{N}(\widehat{\theta}^{(1)}_{\mathrm{BDM}}(m^*), \tau^2_{n_K,N_K}(m^*))\big\|_{\mathrm{TV}} \xrightarrow{\mathbb{P}} 0$ under $\mathbb{P}_{\widetilde{\mathcal{D}}_1}$.

Therefore, by using the invariance property of the TV distance from Lemma S4.1, for each $k=1,\ldots,K$, we obtain that $\big\|\mathcal{P}^{(k)}(\widetilde{\mathcal{D}}_k)-\widetilde{\mathcal{P}}^{(k)}(\widetilde{\mathcal{D}}_k)\big\|_{\mathrm{TV}} \xrightarrow{\mathbb{P}} 0$ in under $\mathbb{P}_{\widetilde{\mathcal{D}}_k}$. This immediately leads to the conclusion that $\big\|\mathcal{P}(\mathcal{D})-\widetilde{\mathcal{P}}(\mathcal{D})\big\|_{\mathrm{TV}} \xrightarrow{\mathbb{P}} 0$ under $\mathbb{P}_{\mathcal{D}}$. Hence, by using the equality in (S.12) (and that the denominator on the RHS in (S.12) is bounded away from zero), we conclude that $|\sqrt{n}\,\{\,\widehat{\theta}_{\mathrm{BDM}}(\widetilde{m}_{\mathrm{CF}})-\theta_0\,\} - \sqrt{n}\,\{\,\widehat{\theta}_{\mathrm{BDM}}(m^*)-\theta_0\,\}| \xrightarrow{\mathbb{P}} 0$ under $\mathbb{P}_{\mathcal{D}}$, which completes the proof of the result. ■

S4.7 Proof of Theorem 4.3

For notational simplicity, we set $k=1$ w.l.o.g. and present the proof below for $k=1$.

Let $Q = \mathcal{N}\big(\widehat{\theta}^{(1)}_{\mathrm{BDM}}(m^*), \tau^2_{n_K,N_K}(m^*)\big)$. Then let $\pi^{(1)}_{\boldsymbol{\theta}}(\cdot)$ and $q(\cdot)$ be the pdfs of the distributions $\widetilde{\Pi}^{(1)}_{\boldsymbol{\theta}}$ and Q, respectively. Then, by the integral representation of TV distance, we have

$$\begin{aligned}\big\|\,\widetilde{\Pi}^{(1)}_{\boldsymbol{\theta}} - Q\,\big\|_{\mathrm{TV}} &= \frac{1}{2}\int \big|\pi^{(1)}_{\boldsymbol{\theta}}(\theta) - q(\theta)\big|\,\mathrm{d}\theta = \frac{1}{2}\int \left|\int \pi(\theta \mid \widetilde{m}_1, \mathcal{D}_1)\,\pi^{(1)}_{\mathbf{m}}(\widetilde{m}_1)\,\mathrm{d}\widetilde{m}_1 - q(\theta)\right|\mathrm{d}\theta \\ &\leq \frac{1}{2}\int\int \big|\pi(\theta \mid \widetilde{m}_1, \mathcal{D}_1)\,\pi^{(1)}_{\mathbf{m}}(\widetilde{m}_1) - q(\theta)\big|\,\mathrm{d}\widetilde{m}_1\,\mathrm{d}\theta.\end{aligned}$$

Then, by using Fubini's theorem and the integral representation of the TV distance, we obtain that

$$\begin{aligned}\big\|\,\widetilde{\Pi}^{(1)}_{\boldsymbol{\theta}} - Q\,\big\|_{\mathrm{TV}} &\leq \frac{1}{2}\int\int \big|\pi(\theta \mid \widetilde{m}_1, \mathcal{D}_1) - q(\theta)\big|\,\mathrm{d}\theta\,\pi^{(1)}_{\mathbf{m}}(\widetilde{m}_1)\,\mathrm{d}\widetilde{m}_1 \\ &= \int \big\|\,\Pi_{(\boldsymbol{\theta}|\widetilde{m}_1,\mathcal{D}_1)} - Q\,\big\|_{\mathrm{TV}}\,\pi^{(1)}_{\mathbf{m}}(\widetilde{m}_1)\,\mathrm{d}\widetilde{m}_1 \\ &= \mathbb{E}_{\widetilde{m}_1\sim\Pi^{(1)}_{\mathbf{m}}}\big\{\,\|\,\Pi_{(\boldsymbol{\theta}|\widetilde{m}_1,\mathcal{D}_1)} - Q\,\|_{\mathrm{TV}} \mid \mathcal{S}_1\,\big\} := \mathbb{E}_{\widetilde{m}_1\sim\Pi^{(1)}_{\mathbf{m}}}\{\,T \mid \mathcal{S}_1\,\}.\end{aligned}$$

Then, let P be a Normal distribution with mean $\widehat{\theta}^{(1)}_{\mathrm{BDM}}(\widetilde{m}_1)$ and variance $\tau^2_{n_K,N_K}(\widetilde{m}_1)$, denoted as $P \equiv \mathcal{N}(\widehat{\theta}^{(1)}_{\mathrm{BDM}}(\widetilde{m}_1), \tau^2_{n_K,N_K}(\widetilde{m}_1))$. Then, by applying the triangle inequality, we observe that

$$\mathbb{E}_{\widetilde{m}_1\sim\Pi^{(1)}_{\mathbf{m}}}(T \mid \mathcal{S}_1) \leq \mathbb{E}_{\widetilde{m}_1\sim\Pi^{(1)}_{\mathbf{m}}}(\|\Pi_{(\boldsymbol{\theta}|\widetilde{m}_1,\mathcal{D}_1)} - P\|_{\mathrm{TV}} \mid \mathcal{S}_1) + \mathbb{E}_{\widetilde{m}_1\sim\Pi^{(1)}_{\mathbf{m}}}(\|P - Q\|_{\mathrm{TV}} \mid \mathcal{S}_1) := T_1 + T_2.$$

Thus, it is enough to show that both T_1 and T_2 converge to 0 in probability w.r.t. $\mathbb{P}_{\widetilde{\mathcal{D}}_1}$ to complete the proof.

We first consider T_1. By Proposition 3.2, we have $\Pi_{(\boldsymbol{\theta}|\widetilde{m}_1,\mathcal{D}_1)}$ is a convolution of two t-distributions $t_{\nu_{n_K}}(\mu_{n_K}(\widetilde{m}_1), \widehat{\sigma}^2_{1,n_K}(\widetilde{m}_1)/n_K)$ and $t_{\nu_{N_K}}(\mu_{N_K}(\widetilde{m}_1), \widehat{\sigma}^2_{2,N_K}(\widetilde{m}_1)/N_K)$, where the parameters are as defined in (8) (by setting $k = 1$ therein). By using the invariance property of the TV distance (see Lemma S4.1), we observe the following:

$$\begin{aligned}T_1 &= \mathbb{E}_{\widetilde{m}_1\sim\Pi^{(1)}_{\mathbf{m}}}\big\{\,\|\,\Pi_{(\boldsymbol{\theta}|\widetilde{m}_1,\mathcal{D}_1)} - P\,\|_{\mathrm{TV}} \mid \mathcal{S}_1\,\big\} \\ &\leq \mathbb{E}_{\widetilde{m}_1\sim\Pi^{(1)}_{\mathbf{m}}}\big\{\,\big\|\,t_{\nu_{n_K}}(\mu_{n_K}(\widetilde{m}_1), \widehat{\sigma}^2_{1,n_K}(\widetilde{m}_1)/n_K) - \mathcal{N}(\mu_{n_K}(\widetilde{m}_1), \sigma^2_1(\widetilde{m}_1)/n_K)\,\big\|_{\mathrm{TV}} \mid \mathcal{S}_1\,\big\} \\ &\quad + \mathbb{E}_{\widetilde{m}_1\sim\Pi^{(1)}_{\mathbf{m}}}\big\{\,\big\|\,t_{\nu_{N_K}}(\mu_{N_K}(\widetilde{m}_1), \widehat{\sigma}^2_{2,N_K}(\widetilde{m}_1)/N_K) - \mathcal{N}(\mu_{N_K}(\widetilde{m}_1), \sigma^2_2(\widetilde{m}_1)/N_K)\,\big\|_{\mathrm{TV}} \mid \mathcal{S}_1\,\big\} \\ &= \mathbb{E}_{\widetilde{m}_1\sim\Pi^{(1)}_{\mathbf{m}}}\big\{\,\big\|\,t_{\nu_{n_K}}(0, \widehat{\sigma}^2_{1,n_K}(\widetilde{m}_1)) - \mathcal{N}(0, \sigma^2_1(\widetilde{m}_1))\,\big\|_{\mathrm{TV}} \mid \mathcal{S}_1\,\big\} \\ &\quad + \mathbb{E}_{\widetilde{m}_1\sim\Pi^{(1)}_{\mathbf{m}}}\big\{\,\big\|\,t_{\nu_{N_K}}(0, \widehat{\sigma}^2_{2,N_K}(\widetilde{m}_1)) - \mathcal{N}(0, \sigma^2_2(\widetilde{m}_1))\,\big\|_{\mathrm{TV}} \mid \mathcal{S}_1\,\big\} \\ &:= \mathbb{E}_{\widetilde{m}_1\sim\Pi^{(1)}_{\mathbf{m}}}\{\,T_{11} \mid \mathcal{S}_1\,\} + \mathbb{E}_{\widetilde{m}_1\sim\Pi^{(1)}_{\mathbf{m}}}\{\,T_{12} \mid \mathcal{S}_1\,\}, \quad \text{where}\end{aligned}$$

$T_{11} := \big\|t_{\nu_{n_K}}(0, \widehat{\sigma}^2_{1,n_K}(\widetilde{m}_1)) - \mathcal{N}(0, \sigma^2_1(\widetilde{m}_1))\big\|_{\mathrm{TV}}$ and $T_{12} := \big\|t_{\nu_{N_K}}(0, \widehat{\sigma}^2_{2,N_K}(\widetilde{m}_1)) - \mathcal{N}(0, \sigma^2_2(\widetilde{m}_1))\big\|_{\mathrm{TV}}$. By the definition of the TV distance, we have $0 \leq T_{11} \leq 1$ and $0 \leq T_{12} \leq 1$. Thus if we show both T_{11} and T_{12} converge to 0 in probability w.r.t. $\mathbb{P}_{\mathcal{S}_1}$, then by applying the DCT (or Lemma S4.6 (b)) we conclude that T_1 converges to 0 in probability under $\mathbb{P}_{\widetilde{\mathcal{D}}_1}$. We recall that in the proof of Theorem 4.1 (see Section S4.4) we have already established that both T_{11} and T_{12} converge to 0 in probability $\mathbb{P}_{\widetilde{\mathcal{D}}_1}$. Therefore, by following the same steps in the proof of Theorem 4.1 for the analysis of T_{11} and T_{12}, we can conclude that both T_{11} and T_{12} converge to 0 in probability $\mathbb{P}_{\widetilde{\mathcal{D}}_1}$ which implies that T_1 converges to 0 in probability w.r.t. $\mathbb{P}_{\widetilde{\mathcal{D}}_1}$. ■

Next, we consider T_2. By using the triangle inequality, we first observe that

$$\begin{aligned} T_2 &= \mathbb{E}_{\widetilde{m}_1 \sim \Pi_{\mathbf{m}}^{(1)}} \big\{ \big\| \mathcal{N}(\widehat{\theta}_{\mathrm{BDM}}^{(1)}(\widetilde{m}_1), \tau_{n_K,N_K}^2(\widetilde{m}_1)) - \mathcal{N}(\widehat{\theta}_{\mathrm{BDM}}^{(1)}(m^*), \tau_{n_K,N_K}^2(m^*)) \big\|_{\mathrm{TV}} \mid \mathcal{S}_1 \big\} \\ &\leq \mathbb{E}_{\widetilde{m}_1 \sim \Pi_{\mathbf{m}}^{(1)}} \big\{ \big\| \mathcal{N}(\widehat{\theta}_{\mathrm{BDM}}^{(1)}(\widetilde{m}_1), \tau_{n_K,N_K}^2(\widetilde{m}_1)) - \mathcal{N}(\widehat{\theta}_{\mathrm{BDM}}^{(1)}(\widetilde{m}_1), \tau_{n_K,N_K}^2(m^*)) \big\|_{\mathrm{TV}} \mid \mathcal{S}_1 \big\} \\ &\quad + \mathbb{E}_{\widetilde{m}_1 \sim \Pi_{\mathbf{m}}^{(1)}} \big\{ \big\| \mathcal{N}(\widehat{\theta}_{\mathrm{BDM}}^{(1)}(\widetilde{m}_1), \tau_{n_K,N_K}^2(m^*)) - \mathcal{N}(\widehat{\theta}_{\mathrm{BDM}}^{(1)}(m^*), \tau_{n_K,N_K}^2(m^*)) \big\|_{\mathrm{TV}} \mid \mathcal{S}_1 \big\} \end{aligned}$$

Further, by using the invariance property of the TV distance in Lemma S4.1, we obtain that

$$\begin{aligned} T_2 &\leq \mathbb{E}_{\widetilde{m}_1 \sim \Pi_{\mathbf{m}}^{(1)}} \big\{ \big\| \mathcal{N}(0, \tau_{n_K,N_K}^2(\widetilde{m}_1)) - \mathcal{N}(0, \tau_{n_K,N_K}^2(m^*)) \big\|_{\mathrm{TV}} \mid \mathcal{S}_1 \big\} \\ &\quad + \mathbb{E}_{\widetilde{m}_1 \sim \Pi_{\mathbf{m}}^{(1)}} \big\{ \big\| \mathcal{N}(\alpha, 1) - \mathcal{N}(0, 1) \big\|_{\mathrm{TV}} \mid \mathcal{S}_1 \big\} \\ &:= T_{21} + T_{22}, \quad \text{where } \alpha = \sqrt{n_K} \{ \widehat{\theta}_{\mathrm{BDM}}^{(1)}(\widetilde{m}_1) - \widehat{\theta}_{\mathrm{BDM}}^{(1)}(m^*) \} / \tau_{n_K,N_K}(m^*). \end{aligned}$$

We first consider $T_{21} := \mathbb{E}_{\widetilde{m}_1 \sim \Pi_{\mathbf{m}}^{(1)}} \big\{ \big\| \mathcal{N}(0, \tau_{n_K,N_K}^2(\widetilde{m}_1)) - \mathcal{N}(0, \tau_{n_K,N_K}^2(m^*)) \big\|_{\mathrm{TV}} | \mathcal{S}_1 \big\}$. Then by following the same algebraic steps in the proof of Theorem 4.1 (see Section S4.4), we have that

$$\big\| \mathcal{N}(0, \tau_{n_K,N_K}^2(\widetilde{m}_1)) - \mathcal{N}(0, \tau_{n_K,N_K}^2(m^*)) \big\|_{\mathrm{TV}} \leq C \big\| \widetilde{m}_1(\mathbf{X}) - m^*(\mathbf{X}) \big\|_{\mathbb{L}_2(\mathbb{P}_{\mathbf{X}})},$$

for some fixed constant $C < \infty$. This implies that $T_{21} \leq \mathbb{E}_{\widetilde{m}_1 \sim \Pi_{\mathbf{m}}^{(1)}} (C \, \| \widetilde{m}_1(\mathbf{X}) - m^*(\mathbf{X}) \|_{\mathbb{L}_2(\mathbb{P}_{\mathbf{X}})} | \mathcal{S}_1)$. Hence, using the nuisance Bayes risk condition in Theorem 4.3, we conclude $T_{21} \xrightarrow{\mathbb{P}} 0$ w.r.t. $\mathbb{P}_{\widetilde{\mathcal{D}}_1}$. ■

We now consider $T_{22} = \mathbb{E}_{\widetilde{m}_1 \sim \Pi_{\mathbf{m}}^{(1)}} (\| \mathcal{N}(\alpha, 1) - \mathcal{N}(0, 1) \|_{\mathrm{TV}} \mid \mathcal{S}_1)$, where $\alpha = \sqrt{n_K} \{ \widehat{\theta}_{\mathrm{BDM}}^{(1)}(\widetilde{m}_1) - \widehat{\theta}_{\mathrm{BDM}}^{(1)}(m^*) \} / \tau_{n_K,N_K}(m^*)$. We first observe that by using Lemma S4.2, we obtain that

$$T_{22} \leq \mathbb{E}_{\widetilde{m}_1 \sim \Pi_{\mathbf{m}}^{(1)}} \left(\frac{| \sqrt{n_K} \{ \widehat{\theta}_{\mathrm{BDM}}^{(1)}(\widetilde{m}_1) - \widehat{\theta}_{\mathrm{BDM}}^{(1)}(m^*) \} |}{\sqrt{2 \pi \, n_K \, \tau_{n_K,N_K}^2(m^*)}} \, | \, \mathcal{S}_1 \right).$$

Since the denominator $\sqrt{2\pi \, n_K \, \tau_{n_K,N_K}^2(m^*)}$ (on the RHS of the inequality above) is a non–random quantity which is greater than and away from 0, for some constant $0 < C < \infty$, we obtain that

$$T_{22} \leq C \, \mathbb{E}_{\widetilde{m}_1 \sim \Pi_{\mathbf{m}}^{(1)}} [| \sqrt{n_K} \, \{ \widehat{\theta}_{\mathrm{BDM}}^{(1)}(\widetilde{m}_1) - \widehat{\theta}_{\mathrm{BDM}}^{(1)}(m^*) \} \mid \mathcal{S}_1].$$

By writing the terms $\widehat{\theta}_{\mathrm{BDM}}^{(1)}(\widetilde{m}_1)$ and $\widehat{\theta}_{\mathrm{BDM}}^{(1)}(m^*)$ explicitly and using the triangle inequality, we obtain the following bound for T_{22}:

$$T_{22} \leq C \left(\mathbb{E}_{\widetilde{m}_1 \sim \Pi_{\mathbf{m}}^{(1)}} [| \mathbb{G}_{n_K}^{(1)} \{ \widetilde{m}_1(\mathbf{X}) - m^*(\mathbf{X}) \} | \, | \mathcal{S}_1] + \sqrt{\frac{n}{N}} \, \mathbb{E}_{\widetilde{m}_1 \sim \Pi_{\mathbf{m}}^{(1)}} [| \mathbb{G}_{N_K}^{(1)} \{ \widetilde{m}_1(\mathbf{X}) - m^*(\mathbf{X}) \} | \, | \mathcal{S}_1] \right),$$

where we recall the notations $\mathbb{G}_{n_K}^{(k)}(\cdot)$ and $\mathbb{G}_{N_K}^{(k)}(\cdot)$ as defined in Section S4 (and here take $k = 1$). Further, it is clear that the analyses of $\mathbb{E}_{\widetilde{m}_1 \sim \Pi_{\mathbf{m}}^{(1)}} [| \mathbb{G}_{n_K}^{(1)} \{ \widetilde{m}_1(\mathbf{X}) - m^*(\mathbf{X}) \} | \mid \mathcal{S}_1]$ and $\mathbb{E}_{\widetilde{m}_1 \sim \Pi_{\mathbf{m}}^{(1)}} [| \mathbb{G}_{N_K}^{(1)} \{ \widetilde{m}_1(\mathbf{X}) - m^*(\mathbf{X}) \} | \mid \mathcal{S}_1]$ will follow the same steps by their definitions. Therefore, it suffices to show that $\mathbb{E}_{\widetilde{m}_1 \sim \Pi_{\mathbf{m}}^{(1)}} [| \mathbb{G}_{n_K}^{(1)} \{ \widetilde{m}_1(\mathbf{X}) - m^*(\mathbf{X}) \} | \, | \mathcal{S}_1] \to 0$ in probability w.r.t. $\mathbb{P}_{\widetilde{\mathcal{D}}_1}$ to conclude that $T_{22} \to 0$ in probability w.r.t. $\mathbb{P}_{\widetilde{\mathcal{D}}_1}$.

By using the definition of convergence in probability, we want to show that: for any constant $t > 0$,

$$\mathbb{P}_{\widetilde{\mathcal{D}}_1}\left(\mathbb{E}_{\widetilde{m}_1 \sim \Pi_{\mathbf{m}}^{(1)}}\big[\,|\mathbb{G}_{n_K}^{(1)}\{\widetilde{m}_1(\mathbf{X}) - m^*(\mathbf{X})\}| \mid \mathcal{S}_1\,\big] > t\right) \;\to\; 0.$$

Towards this, for notational convenience, let us define $\mathcal{Z}_{n_K}(\widetilde{m}_1, \mathcal{S}_1) := [\mathbb{G}_{n_K}^{(1)}\{\widetilde{m}_1(\mathbf{X}) - m^*(\mathbf{X})\} \mid \mathcal{S}_1]$. Since $\mathcal{S}_1 \perp\!\!\!\perp \mathcal{D}_1$ by the construction of the h-BDMI procedure (see Section 4.2), we observe that

$$\begin{aligned}
\mathbb{P}_{\widetilde{\mathcal{D}}_1}[\mathbb{E}_{\widetilde{m}_1 \sim \Pi_{\mathbf{m}}^{(1)}}\{\,|\mathcal{Z}_{n_K}(\widetilde{m}_1, \mathcal{S}_1)|\,\} > t] \;&=\; \mathbb{E}_{\widetilde{\mathcal{D}}_1}[\mathbf{1}(\mathbb{E}_{\widetilde{m}_1 \sim \Pi_{\mathbf{m}}^{(1)}}\{\,|\mathcal{Z}_{n_K}(\widetilde{m}_1, \mathcal{S}_1)|\,\} > t)] \\
&=\; \mathbb{E}_{\mathcal{S}_1}(\mathbb{E}_{\mathcal{D}_1}[\mathbf{1}(\mathbb{E}_{\widetilde{m}_1 \sim \Pi_{\mathbf{m}}^{(1)}}\{\,|\mathcal{Z}_{n_K}(\widetilde{m}_1, \mathcal{S}_1)|\,\} > t) \mid \mathcal{S}_1\,]) \\
&=\; \mathbb{E}_{\mathcal{S}_1}\left(\mathbb{P}_{\mathcal{D}_1}[\mathbb{E}_{\widetilde{m}_1 \sim \Pi_{\mathbf{m}}^{(1)}}\{\,|\mathcal{Z}_{n_K}(\widetilde{m}_1, \mathcal{S}_1)|\,\} > t \mid \mathcal{S}_1\,]\right), \quad \text{(S.13)}
\end{aligned}$$

where $\mathbf{1}(\cdot)$ denotes the indicator function. Since $0 \le \mathbb{P}_{\mathcal{D}_1}[\mathbb{E}_{\widetilde{m}_1 \sim \Pi_{\mathbf{m}}^{(1)}}\{\,|\mathcal{Z}_{n_K}(\widetilde{m}_1, \mathcal{S}_1)|\,\} > t \mid \mathcal{S}_1] \le 1$, if we show that $\mathbb{P}_{\mathcal{D}_1}[\mathbb{E}_{\widetilde{m}_1 \sim \Pi_{\mathbf{m}}^{(1)}}\{\,|\mathcal{Z}_{n_K}(\widetilde{m}_1, \mathcal{S}_1)|\,\} > t \,|\, \mathcal{S}_1\,] \xrightarrow{\mathbb{P}} 0$ w.r.t. $\mathbb{P}_{\mathcal{S}_1}$, then by applying the DCT (or Lemma S4.6), we obtain that $\mathbb{E}_{\mathcal{S}_1}\left(\mathbb{P}_{\mathcal{D}_1}[\mathbb{E}_{\widetilde{m}_1 \sim \Pi_{\mathbf{m}}^{(1)}}\{\,|\mathcal{Z}_{n_K}(\widetilde{m}_1, \mathcal{S}_1)|\,\} > t \mid \mathcal{S}_1]\right) \xrightarrow{\mathbb{P}} 0$ under $\mathbb{P}_{\widetilde{\mathcal{D}}_1}$.

Further, we also observe that

$$\begin{aligned}
\mathbb{P}_{\mathcal{D}_1}\left[\mathbb{E}_{\widetilde{m}_1 \sim \Pi_{\mathbf{m}}^{(1)}}\{\,|\mathcal{Z}_{n_K}(\,\widetilde{m}_1, \mathcal{S}_1\,)|\,\} > t \mid \mathcal{S}_1\right] \;&=\; \mathbb{P}_{\mathcal{D}_1}\left(\left[\mathbb{E}_{\widetilde{m}_1 \sim \Pi_{\mathbf{m}}^{(1)}}\{\,|\mathcal{Z}_{n_K}(\widetilde{m}_1, \mathcal{S}_1)|\,\}\right]^2 > t^2 \mid \mathcal{S}_1\right) \\
&\le\; \mathbb{P}_{\mathcal{D}_1}\left[\mathbb{E}_{\widetilde{m}_1 \sim \Pi_{\mathbf{m}}^{(1)}}\{\,\mathcal{Z}_{n_K}^2(\widetilde{m}_1, \mathcal{S}_1)\,\} > t^2 \mid \mathcal{S}_1\right] \\
&\le\; t^{-2}\,\mathbb{E}_{\mathcal{D}_1}\left[\mathbb{E}_{\widetilde{m}_1 \sim \Pi_{\mathbf{m}}^{(1)}}\{\,\mathcal{Z}_{n_K}^2(\widetilde{m}_1, \mathcal{S}_1)\,\} \mid \mathcal{S}_1\right] \\
&=\; t^{-2}\,\mathbb{E}_{\widetilde{m}_1 \sim \Pi_{\mathbf{m}}^{(1)}}\left[\mathbb{E}_{\mathcal{D}_1}\{\,\mathcal{Z}_{n_K}^2(\,\widetilde{m}_1, \mathcal{S}_1)\,\} \mid \mathcal{S}_1\right],
\end{aligned}$$

where the last three steps come from Cauchy–Schwarz inequality, Markov's inequality, and Fubini's theorem, respectively. Next, by the construction of $\mathcal{Z}_{n_K}(\widetilde{m}_1, \mathcal{S}_1) \equiv [\mathbb{G}_{n_K}^{(1)}\{\widetilde{m}_1(\mathbf{X}) - m^*(\mathbf{X})\} \mid \mathcal{S}_1]$, we note that *given* $\widetilde{m}_1$, it is a sum of centered $\sqrt{n_K}$-scaled *independent* random variables. This implies that $\mathbb{E}_{\mathcal{D}_1}\{\mathcal{Z}_{n_K}(\widetilde{m}_1, \mathcal{S}_1) \,|\, \widetilde{m}_1\} = 0$, and further, $\mathrm{Var}_{\mathcal{D}_1}\{\mathcal{Z}_{n_K}(\widetilde{m}_1, \mathcal{S}_1) \,|\, \widetilde{m}_1\} = \mathbb{E}_{\mathcal{D}_1}\{\mathcal{Z}_{n_K}^2(\widetilde{m}_1, \mathcal{S}_1) \,|\, \widetilde{m}_1\}$. Next, by utilizing the definition of $\mathcal{Z}_{n_K}(\widetilde{m}_1, \mathcal{S}_1)$, we calculate that $\mathrm{Var}_{\mathcal{D}_1}\{\mathcal{Z}_{n_K}(\widetilde{m}_1, \mathcal{S}_1) \,|\, \widetilde{m}_1\} = \mathrm{Var}_{\mathbf{X}}[\{\widetilde{m}_1(\mathbf{X}) - m^*(\mathbf{X})\} \,|\, \widetilde{m}_1]$.

Therefore, by using the observations above, we finally obtain that

$$\begin{aligned}
\mathbb{P}_{\mathcal{D}_1}[\mathbb{E}_{\widetilde{m}_1 \sim \Pi_{\mathbf{m}}^{(1)}}\{\,|\mathcal{Z}_{n_K}(\widetilde{m}_1, \mathcal{S}_1)|\,\} > t \mid \mathcal{S}_1\,] \;&\le\; t^{-2}\,\mathbb{E}_{\widetilde{m}_1 \sim \Pi_{\mathbf{m}}^{(1)}}\left(\mathrm{Var}_{\mathbf{X}}[\{\widetilde{m}_1(\mathbf{X}) - m^*(\mathbf{X})\} \mid \widetilde{m}_1] \mid \mathcal{S}_1\right) \\
&\le\; t^{-2}\,\mathbb{E}_{\widetilde{m}_1 \sim \Pi_{\mathbf{m}}^{(1)}}\{\,\|\widetilde{m}_1(\mathbf{X}) - m^*(\mathbf{X})\|_{\mathbb{L}_2(\mathbb{P}_{\mathbf{X}})}^2 \mid \mathcal{S}_1\,\}.
\end{aligned}$$

Then, by using the nuisance Bayes risk condition given in Theorem 4.3, we directly have that the RHS of the inequality above converges to zero in probability w.r.t. $\mathbb{P}_{\mathcal{S}_1}$ which implies that $\mathbb{P}_{\mathcal{D}_1}[\mathbb{E}_{\widetilde{m}_1 \sim \Pi_{\mathbf{m}}^{(1)}}\{|\mathcal{Z}_{n_K}(\widetilde{m}_1, \mathcal{S}_1)|\} > t \mid \mathcal{S}_1\,] \xrightarrow{\mathbb{P}} 0$ under $\mathbb{P}_{\mathcal{S}_1}$. Next, by using the DCT (or Lemma S4.6 (b)), we obtain that

$$\mathbb{E}_{\mathcal{S}_1}\left(\mathbb{P}_{\mathcal{D}_1}[\mathbb{E}_{\widetilde{m}_1 \sim \Pi_{\mathbf{m}}^{(1)}}\{|\mathcal{Z}_{n_K}(\widetilde{m}_1, \mathcal{S}_1)|\} > t \mid \mathcal{S}_1\,]\right) \;\xrightarrow{\mathbb{P}}\; 0 \;\text{ under }\; \mathbb{P}_{\widetilde{\mathcal{D}}_1}.$$

By using the equality in (S.13) and recalling the definition of $\mathcal{Z}_{n_K}(\widetilde{m}_1, \mathcal{S}_1)$ therein, this equivalently implies that $\mathbb{E}_{\widetilde{m}_1 \sim \Pi^{(1)}_{\mathbf{m}}}[\,|\mathbb{G}^{(1)}_{n_K}\{\widetilde{m}_1(\mathbf{X}) - m^*(\mathbf{X})\}| \mid \mathcal{S}_1\,] \to 0$ in probability under $\mathbb{P}_{\widetilde{\mathcal{D}}_1}$, as $n \to \infty$.

Similarly, by following the same steps and calculations above, but this time for the second term $\mathbb{E}_{\widetilde{m}_1 \sim \Pi^{(1)}_{\mathbf{m}}}[\,|\mathbb{G}^{(1)}_{N_K}\{\widetilde{m}_1(\mathbf{X}) - m^*(\mathbf{X})\}| \mid \mathcal{S}_1\,]$, we obtain that $\mathbb{E}_{\widetilde{m}_1 \sim \Pi^{(1)}_{\mathbf{m}}}[\,|\mathbb{G}^{(1)}_{N_K}\{\widetilde{m}_1(\mathbf{X}) - m^*(\mathbf{X})\}| \mid \mathcal{S}_1\,] \xrightarrow{\mathbb{P}} 0$ under $\mathbb{P}_{\widetilde{\mathcal{D}}_1}$. Recalling the bound obtained for T_{22} above, and since as $n, N \to \infty$, $n/N \to c \in [0, 1)$, we then conclude that $T_{22} \xrightarrow{\mathbb{P}} 0$ under $\mathbb{P}_{\widetilde{\mathcal{D}}_1}$ as $n, N \to \infty$. Hence, this completes the entire proof. ■

S5 Proofs of the remaining results: Preliminary and intermediate lemmas

In this section, we provide proofs of the preliminary results (Lemmas S4.1–S4.6) employed in the proofs of the main results in Section S4, as well as proof of Lemma S4.7 introduced in the course of proving Theorem 4.1. Note also that two of the preliminary Lemmas S4.3 and S4.6 are directly adopted from existing papers.

S5.1 Proof of Lemma S4.1

We give a proof for the sake of completeness. We have

$$\begin{aligned}\|P^{\mu,\sigma} - Q^{\mu,\sigma}\|_{\mathrm{TV}} &= \frac{1}{2}\int |p^{\mu,\sigma}(x) - q^{\mu,\sigma}(x)|\,\mathrm{d}x = \frac{1}{2}\int \left|\frac{1}{\sigma}p\left(\frac{x-\mu}{\sigma}\right) - \frac{1}{\sigma}q\left(\frac{x-\mu}{\sigma}\right)\right|\mathrm{d}x \\ &= \frac{1}{2}\int |p(t) - q(t)|\,\mathrm{d}t = \|P - Q\|_{\mathrm{TV}},\end{aligned}$$

where going from the first to the second line, we make a change of variable $t = (x - \mu)/\sigma$. ■

S5.2 Proof of Lemma S4.2

By using the invariance property of the TV distance from Lemma S4.1, $\|P - Q\|_{\mathrm{TV}} = \|\mathcal{N}(\alpha, 1) - \mathcal{N}(0, 1)\|_{\mathrm{TV}}$, where $\alpha = (\mu_1 - \mu_2)/\sigma$. The result now follows from Lemma 4 of Bontemps (2011). ■

S5.3 Proof of Lemma S4.4

Since the TV distance is invariant under scaling and location-shift (see Lemma S4.1) and both the t-distribution $t_\nu(\mu, \sigma^2)$ and the Normal distribution $\mathcal{N}(\mu, \sigma^2)$ belong to a location-scale family, we have that $\|t_\nu(\mu, \sigma^2) - \mathcal{N}(\mu, \sigma^2)\|_{\mathrm{TV}} = \|t_\nu - \mathcal{N}(0, 1)\|_{\mathrm{TV}}$, where $t_\nu \equiv t_\nu(0, 1)$. This implies that the TV distance is free of the parameters μ and σ^2. Further, we note that the t-distribution $t_\nu(\mu, \sigma^2)$ can be expressed as a precision mixture of a Gaussian distribution (West, 1987). In particular, suppose $X \mid W \sim \mathcal{N}(\mu, W^{-1}\sigma^2)$, and $W \sim \mathrm{Gamma}(\nu/2, \nu/2)$ with the pdf $f_W(\cdot)$, then, $X \sim t_\nu(\mu, \sigma^2)$. Since $f_W(\cdot)$ is a pdf, we can write the TV distance above as follows:

$$\begin{aligned}\|t_\nu - \mathcal{N}(0, 1)\|_{\mathrm{TV}} &= \frac{1}{2}\int |t_\nu(x; 0, 1) - \mathcal{N}(x; 0, 1)|\,\mathrm{d}x \\ &= \frac{1}{2}\int \left|\int \{f_{X|W}(x) - \mathcal{N}(x; 0, 1)\}\, f_W(w)\,\mathrm{d}w\right|\mathrm{d}x,\end{aligned}$$

where $f_{X|W}(\cdot)$ is the pdf of the conditional random variable $X\,|\,W$ having a Normal $\mathcal{N}\big(0, W^{-1}\big)$ distribution.

Further, we obtain that

$$\begin{aligned}\|t_\nu - \mathcal{N}(0,1)\|_{\mathrm{TV}} \;&\leq\; \frac{1}{2}\int\int |f_{X|W}(x) - \mathcal{N}(x;0,1)| f_W(w)\,\mathrm{d}w\,\mathrm{d}x \\ &=\; \frac{1}{2}\int\int |f_{X|W}(x) - \mathcal{N}(x;0,1)|\,\mathrm{d}x\, f_W(w)\,\mathrm{d}w,\end{aligned}$$

where the last step uses Fubini's theorem to change the order of the integrals.

Next, by following the integral representation of the TV distance, we further obtain that

$$\|t_\nu - \mathcal{N}(0,1)\|_{\mathrm{TV}} \;\leq\; \int \|\mathcal{N}(0,w^{-1}) - \mathcal{N}(0,1)\|_{\mathrm{TV}}\, f_W(w)\,\mathrm{d}w.$$

Then, by using Lemma S4.3, there is a constant $C > 0$ such that $\|\mathcal{N}(0,w^{-1}) - \mathcal{N}(0,1)\|_{\mathrm{TV}} \leq C|w-1|$. Since $W \sim \mathrm{Gamma}(\nu/2, \nu/2)$, by using the Cauchy–Schwarz inequality, for some $C_0 > 0$, we have $\|t_\nu - \mathcal{N}(0,1)\|_{\mathrm{TV}} \;\leq\; C\int |w-1|\, f_W(w)\,\mathrm{d}w \;\leq\; C\,[\mathbb{E}_W(W-1)^2]^{1/2} \;=\; C_0/\sqrt{\nu}$. ■

S5.4 Proof of Lemma S4.5

The proof follows from the definition of the convolution operator and the triangle inequality. By the triangle inequality and Fubini's theorem, we observe that

$$\begin{aligned}\|P - Q\|_{\mathrm{TV}} \;&=\; \frac{1}{2}\int |p(z) - q(z)|\,\mathrm{d}z \;=\; \frac{1}{2}\int \left|\int p_1(x)p_2(z-x)\,\mathrm{d}x - \int q_1(x)q_2(z-x)\,\mathrm{d}x\right|\mathrm{d}z \\ &\leq\; \frac{1}{2}\int\int \{|p_1(x)p_2(z-x) - q_1(x)p_2(z-x)| \;+\; |q_1(x)p_2(z-x) - q_1(x)q_2(z-x)|\}\,\mathrm{d}x\,\mathrm{d}z \\ &=\; \frac{1}{2}\int\int |p_1(x) - q_1(x)|p_2(z-x)\,\mathrm{d}z\,\mathrm{d}x \;+\; \frac{1}{2}\int\int q_1(x)|p_2(z-x) - q_2(z-x)|\,\mathrm{d}z\,\mathrm{d}x \\ &=\; \frac{1}{2}\int |p_1(x) - q_1(x)|\,\mathrm{d}x + \frac{1}{2}\int |p_2(w) - q_2(w)|\,\mathrm{d}w \;=\; \|P_1 - Q_1\|_{\mathrm{TV}} \;+\; \|P_2 - Q_2\|_{\mathrm{TV}}.\end{aligned}$$

Hence, we have obtained the desired inequality. This completes the proof. ■

S5.5 Proof of Lemma S4.7

We start with writing the distribution $\mathbb{P}_{\widetilde{m}_1}$ explicitly. Since the randomness of $\widetilde{m}_1$ comes from both the data $\mathcal{S}_1$ and the posterior distribution $\Pi^{(1)}_{\mathbf{m}}(\cdot) \equiv \Pi^{(1)}_{\mathbf{m}}(\cdot;\mathcal{S}_1)$ itself, we have that $\mathbb{P}_{\widetilde{m}_1} = \mathbb{P}_{\mathcal{S}_1} \otimes \Pi^{(1)}_{\mathbf{m}}$. Then, to establish the result, by the definition of convergence in probability, we need to show that for any $t > 0$, $\mathbb{P}_{\widetilde{m}_1}\{\|m^*(\mathbf{X}) - \widetilde{m}_1(\mathbf{X})\|_{\mathbb{L}_2(\mathbb{P}_{\mathbf{X}})} > t\} \to 0$ as $n \to \infty$. Towards this goal, we first observe that

$$\begin{aligned}\mathbb{P}_{\widetilde{m}_1}\{\,\|m^*(\mathbf{X}) - \widetilde{m}_1(\mathbf{X})\|_{\mathbb{L}_2(\mathbb{P}_{\mathbf{X}})} > t\,\} \;&=\; \mathbb{E}_{\widetilde{m}_1}\{\mathbf{1}(\,\|m^*(\mathbf{X}) - \widetilde{m}_1(\mathbf{X})\|_{\mathbb{L}_2(\mathbb{P}_{\mathbf{X}})} > t\,)\} \\ &=\; \mathbb{E}_{\mathcal{S}_1}[\,\mathbb{E}_{\mathbf{m}|\mathcal{S}_1}\{\,\mathbf{1}(\,\|m^*(\mathbf{X}) - \widetilde{m}_1(\mathbf{X})\|_{\mathbb{L}_2(\mathbb{P}_{\mathbf{X}})} > t\,) \mid \mathcal{S}_1\,\}\,] \\ &=\; \mathbb{E}_{\mathcal{S}_1}[\,\Pi^{(1)}_{\mathbf{m}}\{\,\|m^*(\mathbf{X}) - \widetilde{m}_1(\mathbf{X})\|_{\mathbb{L}_2(\mathbb{P}_{\mathbf{X}})} > t \mid \mathcal{S}_1\,\}\,] \;:=\; \mathbb{E}_{\mathcal{S}_1}[\,T\{\,\mathcal{S}_1(n), t\,\}\,],\end{aligned} \tag{S.14}$$

$$\text{where } \; T\{\mathcal{S}_1(n), t\} \; := \; \Pi^{(1)}_{\mathbf{m}}\{\, \|m^*(\mathbf{X}) - \widetilde{m}_1(\mathbf{X})\|_{\mathbb{L}_2(\mathbb{P}_{\mathbf{X}})} > t \mid \mathcal{S}_1 \,\}, \tag{S.15}$$

and $\mathbf{1}(\cdot)$ denotes the indicator function, and the second step uses the law of iterated expectation. We note that the notation $T\{\mathcal{S}_1(n), t\}$ above indicates the dependence on the size of $\mathcal{S}_1$ (note that its size $n - n/K$ is of the same order as n) and the given $t > 0$. Further, since $T\{\mathcal{S}_1(n), t\}$ itself is a probability as in (S.15), $T\{\mathcal{S}_1(n), t\} \in [0, 1]$, and moreover, (S.15) implies that: for any $0 < t_1 < t_2$,

$$T\{\mathcal{S}_1(n), t_1\} \; \geq \; T\{\mathcal{S}_1(n), t_2\}, \;\; \text{so that} \;\; \mathbb{P}_{\mathcal{S}_1}[\, T\{\mathcal{S}_1(n), t_1\} \geq T\{\mathcal{S}_1(n), t_2\}\,] \; = \; 1 \;\; \forall \; t_1 < t_2. \tag{S.16}$$

Then, since $T\{\mathcal{S}_1(n), t\}$ is bounded, an application of the DCT (or Lemma S4.6) ensures that it suffices to show $T\{\mathcal{S}_1(n), t\} \xrightarrow{\mathbb{P}} 0$ under $\mathbb{P}_{\mathcal{S}_1}$. Now, recall the NPCC in Assumption 4.1 (ii): for some $a_n \to 0$, $\Pi^{(k)}_{\mathbf{m}}\{m : \|m^*(\mathbf{X}) - m(\mathbf{X})\|_{\mathbb{L}_2(\mathbb{P}_{\mathbf{X}})} > a_n \mid \mathcal{S}_1\} \xrightarrow{\mathbb{P}} 0$ under $\mathbb{P}_{\mathcal{S}_1}$, as $n \to \infty$ (we here take $k = 1$).

Then, using (S.15), we can *rewrite* the NPCC as follows: as $n \to \infty$, for some $a_n \to 0$,

$$T\{\mathcal{S}_1(n), a_n\} \; = \; o_{\mathbb{P}_{\mathcal{S}_1}}(1), \;\; \text{or equivalently,} \;\; \mathbb{P}_{\mathcal{S}_1}[\, T\{\mathcal{S}_1(n), a_n\} > \gamma\,] \to 0 \;\; \text{for any} \;\; \gamma > 0. \tag{S.17}$$

The RHS above equivalently says that for any $\delta > 0$ there exists a $n_{\gamma,\delta}$ such that for any $n \geq n_{\gamma,\delta}$, $\mathbb{P}_{\mathcal{S}_1}[\, T\{\mathcal{S}_1(n), a_n\} > \gamma\,] < \delta$. Note that the double index in $n_{\gamma,\delta}$ indicates the dependence on both γ and δ.

Also, we observe that for the given $t > 0$, since $a_n \to 0$, there exists n_t such that for all $n \geq n_t$, $a_n < t$ (and recall that $a_n \geq 0$ by definition), almost surely (a.s.) w.r.t. $\mathbb{P}_{\mathcal{S}_1}$ (i.e., $\mathbb{P}_{\mathcal{S}_1}[\, T\{\mathcal{S}_1(n), a_n\} \geq T\{\mathcal{S}_1(n), t\}\,] = 1$) for all $n \geq n_t$.

Now, by using the definition of convergence in probability, we ultimately need to show that for any $\varepsilon > 0$, $\mathbb{P}_{\mathcal{S}_1}[\, T\{S_1(n), t\} > \varepsilon\,] \to 0$ as $n \to \infty$, or equivalently, for any $\delta > 0$, there exists a $n^*_{\varepsilon,\delta}$ such that for any $n \geq n^*_{\varepsilon,\delta}$, $\mathbb{P}_{\mathcal{S}_1}[\, T\{S_1(n), t\} > \varepsilon\,] < \delta$.

Toward showing this, we let $n^*_{\varepsilon,\delta} := \max\{n_{\gamma,\delta}, n_t\}$ (by recalling the terms from the observations above and also setting $\gamma = \varepsilon$). Then, for any $n \geq n^*_{\varepsilon,\delta}$, by the total law of probability, we have that

$$\begin{aligned}
\mathbb{P}_{\mathcal{S}_1}[T\{\mathcal{S}_1(n), t\} > \varepsilon] \; &= \; \mathbb{P}_{\mathcal{S}_1}[\, T\{\mathcal{S}_1(n), t\} > \varepsilon, T\{\mathcal{S}_1(n), a_n\} > \varepsilon] \\
&\quad + \mathbb{P}_{\mathcal{S}_1}[T\{\mathcal{S}_1(n), t\} > \varepsilon, \;\; T\{\mathcal{S}_1(n), a_n\} \leq \varepsilon\,] \\
&\leq \; \mathbb{P}_{\mathcal{S}_1}[\, T\{\mathcal{S}_1(n), a_n\} > \varepsilon\,] + \mathbb{P}_{\mathcal{S}_1}[\, T\{\mathcal{S}_1(n), a_n\} \leq T\{\mathcal{S}_1(n), t\}\,] \; < \; \delta,
\end{aligned}$$

where the last step uses the following observations: $\mathbb{P}_{\mathcal{S}_1}[\, T\{\mathcal{S}_1(n), a_n\} > \varepsilon\,] < \delta$, since $n \geq n^*_{\varepsilon,\delta} \geq n_{\gamma,\delta}$ (with $\gamma \equiv \epsilon$) using (S.17), and $\mathbb{P}_{\mathcal{S}_1}[\, T\{\mathcal{S}_1(n), a_n\} \leq T\{\mathcal{S}_1(n), t\}\,] = 0$ using (S.16), since $n \geq n^*_{\varepsilon,\delta} \geq n_t$ (referring to the discussions and implications above). Hence, for the given $\varepsilon > 0$ and $t, \delta > 0$, we have: $\mathbb{P}_{\mathcal{S}_1}[\, T\{\mathcal{S}_1(n), t\} > \varepsilon\,] < \delta$ for any $n \geq n^*_{\varepsilon,\delta}$, which equivalently gives $T\{\mathcal{S}_1(n), t\} = o_{\mathbb{P}_{\mathcal{S}_1}}(1)$.

Finally, using the equality in (S.14), we can now apply the DCT (or Lemma S4.6) to conclude that $\mathbb{P}_{\widetilde{m}_1}\{\|m^*(\mathbf{X}) - \widetilde{m}_1(\mathbf{X})\|_{\mathbb{L}_2(\mathbb{P}_{\mathbf{X}})} > t\} \to 0$ for all $t > 0$, i.e., $\|m^*(\mathbf{X}) - \widetilde{m}_1(\mathbf{X})\|_{\mathbb{L}_2(\mathbb{P}_{\mathbf{X}})} = o_{\mathbb{P}_{\widetilde{m}_1}}(1)$, as claimed. ■