Journal of Machine Learning Research 22 (2021) 1-44 Submitted 09/20; Revised 09/21; Published 10/21

Model Linkage Selection for Cooperative Learning

Jiaying Zhou ZHOU1054@UMN.EDU
School of Statistics
University of Minnesota
Minneapolis MN, 55455

Jie Ding DINGJ@UMN.EDU
School of Statistics
University of Minnesota
Minneapolis MN, 55455

Kean Ming Tan KEANMING@UMICH.EDU
Department of Statistics
University of Michigan
Ann Arbor MI, 48109

Vahid Tarokh VAHID.TAROKH@DUKE.EDU
Department of Electrical Engineering and Engineering
Duke University
Durham NC, 27708

Editor: Bert Huang

Abstract

We consider the distributed learning setting where each agent or learner holds a specific parametric model and a data source. The goal is to integrate information across a set of learners and data sources to enhance the prediction accuracy of a given learner. A natural way to integrate information is to build a joint model across a group of learners that shares common parameters of interest. However, the underlying parameter sharing patterns across a set of learners may not be known a priori. Misspecifying the parameter sharing patterns or the parametric model for each learner often yields a biased estimator that degrades the prediction accuracy. We propose a general method to integrate information across a set of learners that is robust against misspecification of both models and parameter sharing patterns. The main crux of our proposed method is to sequentially incorporate additional learners that can enhance the prediction accuracy of an existing joint model based on user-specified parameter sharing patterns across a set of learners. Theoretically, we show that the proposed method can data-adaptively select a parameter sharing pattern that enhances the predictive performance of a given learner. Extensive numerical studies are conducted to assess the performance of the proposed method.

Keywords: Data integration; decentralized learning; federated learning; model linkage selection; prediction efficiency.

1. Introduction

In recent years, there has been a growing interest in statistical learning problems with a set of decentralized learners, where each learner encompasses a specific data modality and a

statistical model built using domain-specific knowledge. The goal is to integrate information across decentralized learners to achieve higher statistical efficiency and predictive accuracy. Integrating information from different data sources is crucial in many scientific domains such as environmental science (Blangiardo et al., 2011; Xingjian et al., 2015), epidemiology (Yang et al., 2015; Guo et al., 2017), statistical machine learning problems (Ngiam et al., 2011; Kong et al., 2016; Cao and Jin, 2007; Ye et al., 2020; Xian et al., 2020), and computational biology (Simmonds and Higgins, 2007; Liu et al., 2009, 2015; Wen and Stephens, 2014). For instance, in the context of epidemiology, a considerable amount of online search data from different platforms are integrated and used to form accurate predictions of influenza epidemics (Yang et al., 2015; Guo et al., 2017). The aforementioned applications raise a critical question: how to reliably integrate information from different data sources to enhance statistical efficiency in a robust manner?

We consider the setting in which there are a set of learners, each consisting of a data set and a parametric statistical model. The learners may or may not share common parameters among themselves. The goal is to develop a framework to enhance the statistical efficiency of any particular learner, say $\mathcal{L}_1$, by integrating information from the other learners through parameter sharing. In general, learner $\mathcal{L}_1$ can be assisted explicitly or implicitly by building a joint model with the other learners with potentially different statistical models and different data sources. Explicit assistance could be achieved by joint modeling with a set of learners that share at least one parameter. On the other hand, implicit assistance could be achieved by joint modeling with learners whose parameters are not directly related to $\mathcal{L}_1$, but are related to learners who could explicitly assist $\mathcal{L}_1$. In principle, if the true underlying parameter sharing patterns among all learners are known a priori and that the parametric statistical model for each learner is correctly specified, then one can build a joint model with constraints on the shared parameters based on the joint likelihood function.

Many existing modeling methods can be formulated as special instances of the above setting. For example, when multiple learners employ the same parametric model across different data sources, it is usually studied in the context of data integration (Jensen et al., 2007; Vonesh et al., 2006; Liu et al., 2015; Lee et al., 2017b; Jordan et al., 2019; Danaher et al., 2014; Ma and Michailidis, 2016; Tang and Song, 2016; Li and Li, 2018; Tang et al., 2019; Maity et al., 2019; Shen et al., 2019), or in the context of distributed optimization (Boyd et al., 2011; Shi et al., 2014; Lee et al., 2017a; Li et al., 2019). A recent topic, referred to as *federated learning*, is also related to the above setting, where a central server (interpreted as the main learner) sends the current global model to a set of clients (interpreted as other learners), each client updates the model parameters with the local data source, and then returns them to the central server (Shokri and Shmatikov, 2015; Konečnỳ et al., 2016; McMahan et al., 2016).

Though it is often helpful to establish a joint model from multiple learners, naively combining data sources and performing joint modeling can lead to severely degraded statistical performance due to four possible reasons: (i) misspecified statistical models for some learners; (ii) misspecified parameter sharing patterns among learners; (iii) heterogeneity from different data sources (Simmonds and Higgins, 2007; Wen and Stephens, 2014; Liu et al., 2015); and (iv) distinct learning objectives for different learners (Zuech et al., 2015; Sivarajah et al., 2017; Xian et al., 2020; Wang et al., 2021). Most existing data integration and distributed computing methods are based on the assumption that the statistical model

for each learner is correctly specified and that the parameter sharing patterns are known a priori. A systematic statistical approach for decentralized learning robust against the four aspects mentioned above is relatively lacking.

In this paper, we propose a general approach to enhance the predictive performance of a specific learner $\mathcal{L}_1$ by integrating information from the other learners. We consider the setting where there are M learners and that the learners may have different statistical models and heterogeneous data sources. To characterize parameter sharing patterns among all learners, we introduce the notion of a model linkage graph. A model linkage graph $G = (V, E)$ consists of a set of M vertices V and a set of edges E, where each vertex represents a learner, and an edge between a pair of vertices encodes a parameter sharing pattern between the pair of learners. In addition, a pair of learners do not share any common parameters if there is no edge between them. A joint model that enhances the predictive performance of $\mathcal{L}_1$ can then be fit given a model linkage graph. However, the ground truth or the most suitable model linkage graph is not known a priori. Due to model misspecification within each learner and misspecified model linkages between pairs of learners, the prediction performance of $\mathcal{L}_1$ may degrade after incorporating information from other learners.

To enhance robustness against a misspecified model linkage graph, one could exhaustively build joint models for all possible sets of learners that are connected to $\mathcal{L}_1$ within a given model linkage graph. Then, the set of learners that yields the largest conditional marginal likelihood of $\mathcal{L}_1$ is selected. However, such an approach is computationally infeasible since the number of possible sets of learners grows exponentially with the number of learners. To address this challenge, we propose a greedy algorithm that is robust against a misspecified model linkage graph. Our proposed algorithm sequentially incorporates additional learners based on the user-specified model linkage graph, starting from learner $\mathcal{L}_1$. In each iteration of the algorithm, we utilize the joint model built with a group of learners from the previous iteration and search for the next learner to improve the marginal likelihood of the current group of learners. This process is continued until no more learners are included in the joint model. This approach approximately reduces the number of possible sets of model linkages from exponential to quadratic in the number of learners. Compared with the joint modeling approach, our proposed method is of distributed nature and does not require data sharing across learners.

To quantify the theoretical aspects of the proposed method, we introduce the notion of model linkage selection consistency and asymptotic prediction efficiency. They are different but conceptually parallel to asymptotic efficiency and selection consistency in the classical model selection literature (see, e.g., Ding et al., 2018). We show that the proposed algorithm achieves linkage selection consistency as long as the user-specified model linkage graph is a superset of the underlying model linkage graph. In other words, the proposed method selects data sources that are truly useful for enhancing the predictive performance of $\mathcal{L}_1$ in a data-adaptive manner as the sample size grows. In addition, we show that the proposed method achieves asymptotic prediction efficiency, meaning that the predictive performance of the selected model is asymptotically equivalent to that of the best joint model in hindsight.

The paper is outlined as follows. In Section 2, we provide the motivation, problem description, and definitions related to the model linkage graph. In Section 3, we propose a general method for integrating information from different learners to enhance the predictive performance of $\mathcal{L}_1$. The theoretical results for the proposed framework are provided in

Section 3.2.2. In Section 4, we perform numerical studies to evaluate the performance of the proposed method under different scenarios such as data contamination and model misspecification. We close with a discussion in Section 5, where we highlight some related literature on data integration and federated learning. The technical proofs and the regularity conditions needed for the main results are included in the Appendix.

2. Background and Motivation

Suppose that there are M learners, $\mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_M$. Each learner is a pair $\mathcal{L}_\kappa = (\mathbf{D}^{(\kappa)}, \mathcal{P}_\kappa)$, where $\mathbf{D}^{(\kappa)}$ is a set of n_κ observations from the sample space $\mathcal{D}^{(\kappa)}$, and $\mathcal{P}_\kappa$ is a user-specified class of parametric model for modeling $\mathbf{D}^{(\kappa)}$, parameterized by a p_κ-dimensional parameter $\boldsymbol{\theta}_\kappa \in \boldsymbol{\Theta}_\kappa$. Our goal is to develop a framework for enhancing the predictive performance of a particular learner, say $\mathcal{L}_1$, by integrating information from other learners, $\mathcal{L}_2, \ldots, \mathcal{L}_M$.

We will focus on the regression setting where $\mathcal{D}^{(\kappa)} = (\mathcal{Y}, \mathcal{X}^{(\kappa)})$, with $\mathcal{Y} = \mathbb{R}$ and $\mathcal{X}^{(\kappa)} = \mathbb{R}^{k_\kappa}$. The covariates are allowed to have different dimensions across the M learners due to different data sources. While we focus on the regression setting, the proposed approach can be applied more generally to data of different forms. Let $\mathbf{D}^{(\kappa)} = (\mathbf{y}^{(\kappa)}, \mathbf{X}^{(\kappa)}) \in \mathcal{D}^{(\kappa)}$, where $\mathbf{y}^{(\kappa)} \in \mathbb{R}^{n_\kappa}$ is an n_κ-dimensional vector of response and $\mathbf{X}^{(\kappa)} \in \mathbb{R}^{n_\kappa \times k_\kappa}$ is an $n_\kappa \times k_\kappa$ matrix of covariates. Let $\mathcal{P}_\kappa = \left\{ p^{(\kappa)}_{\boldsymbol{\theta}_\kappa}(\cdot|\boldsymbol{\theta}_k, \mathbf{X}^{(\kappa)}) : \boldsymbol{\theta}_\kappa \in \boldsymbol{\Theta}_\kappa, \mathbf{X}^{(\kappa)} \in \mathcal{X}^{(\kappa)} \right\}$ be a class of user-specified parametric model for modeling $\mathbf{y}^{(\kappa)}$ given $\mathbf{X}^{(\kappa)}$. For each learner, assume that the underlying response variable is generated independently according to the probability law $\mathcal{P}^*_\kappa$ described by a conditional density $p^*_\kappa(\cdot|\mathbf{x}^{(\kappa)})$, given the covariates $\mathbf{x}^{(\kappa)} \in \mathbb{R}^{k_\kappa}$. We start with providing several definitions that will serve as a foundation of the proposed framework: *model misspecification*, *model linkage*, and *model linkage misspecification*.

A class of user-specified model $\mathcal{P}_\kappa$ is said to be misspecified when it does not contain the underlying conditional density $p^*_\kappa(\cdot|\mathbf{x}^{(\kappa)})$. In practice, model misspecification often occurs due to an inappropriate functional form between the response and covariates, such as underfitting the model or neglecting dependent random noise (see, e.g., Domowitz and White, 1982; Clarke, 2005; Cawley and Talbot, 2010). We now provide a formal definition of model misspecification.

Definition 1 *A model* $\mathcal{P} = \{p_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ *is* well-specified *if there exists a* $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$ *such that* $p_{\boldsymbol{\theta}^*} = p^*$ *almost everywhere. A model* $\mathcal{P} = \{p_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ *is* misspecified *if* $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathcal{P}^*\{y : p_{\boldsymbol{\theta}}(y|\boldsymbol{\theta}, \mathbf{x}) = p^*(y|\mathbf{x})\} < 1$ *for some covariates* $\mathbf{x}$*, where* $\mathcal{P}^*$ *denotes the probability measure corresponding to* p^**.*

Figure 1 provides several examples of well-specified and misspecified models in the context of regression. The left panel of Figure 1 indicates that the learner $\mathcal{L}_1$ is well-specified, since the underlying linear model $p^*_1 \in \mathcal{P}_1$, where $\mathcal{P}_1$ is a class of user-specified linear model with two covariates. On the other hand, the middle panel of Figure 1 presents a case when the learner $\mathcal{L}_2$ is misspecified since $p^*_2 \notin \mathcal{P}_2$. Model misspecification also occurs when the user-specified model class is different from the underlying data-generating process, as illustrated in the right panel of Figure 1.

Next, we provide a new definition of model linkage. Suppose that there are two learners $\mathcal{L}_i$ and $\mathcal{L}_j$. A model linkage occurs between two learners $\mathcal{L}_i$ and $\mathcal{L}_j$ if they are restricted to share some common parameters. A formal definition is given in Definition 2.

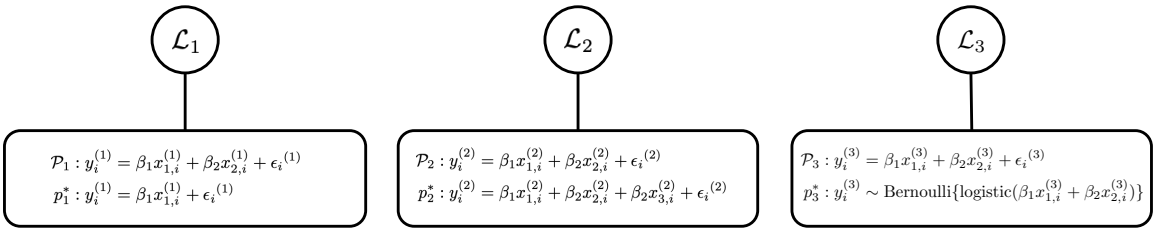

Figure 1: Three examples of well-specified and misspecified models. Left panel: well-specified model. Middle panel: misspecified model. Right panel: misspecified model.

Definition 2 (Model linkage) *Suppose that two learners $\mathcal{L}_i$ and $\mathcal{L}_j$ are well-specified. Let $\boldsymbol{\theta}_{i,\mathcal{S}_i}$ and $\boldsymbol{\theta}_{j,\mathcal{S}_j}$ be subvectors of $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$, indexed by the subsets $\mathcal{S}_i \subseteq \{1,\ldots,p_i\}$ and $\mathcal{S}_j \subseteq \{1,\ldots,p_j\}$, respectively. There exists a model linkage between $\mathcal{L}_i$ and $\mathcal{L}_j$ if $\boldsymbol{\theta}_{i,\mathcal{S}_i} = \boldsymbol{\theta}_{j,\mathcal{S}_j}$, also denoted by $\boldsymbol{\theta}_{\mathcal{S}_i,\mathcal{S}_j}$ for notational convenience. We also refer to $\boldsymbol{\theta}_{\mathcal{S}_i,\mathcal{S}_j}$ as the* shared common parameter *between $\mathcal{L}_i$ and $\mathcal{L}_j$.*

To put the idea of model linkage into perspective, we consider an example in the context of an epidemiological study with two learners, illustrated in Figure 2. A similar example was considered in (Plummer, 2015). Suppose that both learners are well-specified. Learner $\mathcal{L}_1$ concerns estimating the human papilomavirus (HPV) prevalence with data $\mathbf{D}^{(1)} = (\mathbf{y}^{(1)}, \mathbf{x}^{(1)})$, where $\mathbf{y}^{(1)}$ and $\mathbf{x}^{(1)}$ are both n-dimensional vectors recording the number of women infected with high-risk HPV and the population size in different states, respectively. A binomial model is specified for $y_i^{(1)}$ with a population size $x_i^{(1)}$ and HPV prevalence parameter $\theta_{1,i}$, with $i = 1, 2, \ldots, n$. Learner $\mathcal{L}_2$ models the relationship between the HPV prevalence $\theta_{1,i}$ and cancer incidence in the form of Poisson regression. In $\mathcal{L}_2$, let $\mathbf{D}_2 = (\mathbf{y}^{(2)}, \mathbf{x}^{(2)})$, where $\mathbf{y}^{(2)}$ is the number of cancer incidents and $\mathbf{x}^{(2)}$ is the women-years of follow up. Both $\mathcal{L}_1$ and $\mathcal{L}_2$ are restricted to share the same HPV prevalence parameters $\theta_{1,i},\ i = 1, 2, \ldots, n$, and thus there is a model linkage between $\mathcal{L}_1$ and $\mathcal{L}_2$.

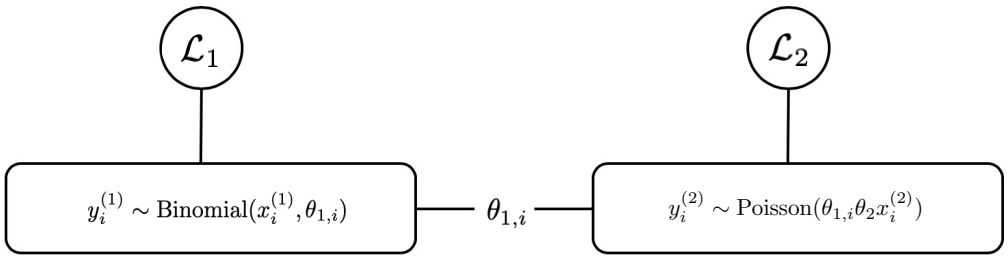

Figure 2: Learners $\mathcal{L}_1$ and $\mathcal{L}_2$ share common parameters $\theta_{1,i}$, $i = 1, \ldots, n$.

The definition and example above focus mainly on whether there is a model linkage between two learners. Such an idea can be generalized to a set of model linkages among a group of learners, which we refer to as *model linkage graph* in the following definition.

Definition 3 (Model linkage graph) *Let $G = (V, E)$ be an undirected* model linkage graph*, where V is a set of M vertices representing the M learners $\mathcal{L}_1, \ldots, \mathcal{L}_M$, and E is an edge set encoding model linkages between pairs of learners.*

In practice, the model linkage graph and linkages between pairs of learners are pre-specified by the user, usually based on domain-specific knowledge, before model fitting. Therefore, the prediction performance of a learner may not be improved after incorporating information from the model linkage graph due to the potential misspecification of models and model linkages. A concept in parallel with model misspecification in the context of model linkage misspecification is provided in Definition 4.

Definition 4 (Model linkage misspecification) *Suppose that there is a model linkage between two learners $\mathcal{L}_i$ and $\mathcal{L}_j$. In other words, a pair of subsets of parameters in two learners are restricted to be the same, say $\boldsymbol{\theta}_{i,\mathcal{S}_i} = \boldsymbol{\theta}_{j,\mathcal{S}_j}$. A model linkage is misspecified if either $\mathcal{L}_i$ or $\mathcal{L}_j$ is misspecified, or that $\boldsymbol{\theta}^*_{i,\mathcal{S}_i} \neq \boldsymbol{\theta}^*_{j,\mathcal{S}_j}$. More generally, a model linkage graph $G = (V, E)$ is misspecified if there exists a misspecified model linkage between a pair of learners.*

Recall that we are interested in enhancing the predictive performance of $\mathcal{L}_1$ by integrating information from other learners $\mathcal{L}_2, \ldots, \mathcal{L}_M$. Thus far, it is clear that a model linkage should exist between $\mathcal{L}_1$ and $\mathcal{L}_\kappa$ if the pair of learners shares common parameters. We now introduce the notion of *information flow* in which learners that do not share common parameters directly with $\mathcal{L}_1$ can also enhance its predictive performance by sharing common parameters with learners that can, in turn, assist $\mathcal{L}_1$. Consider an example with six learners as illustrated in Figure 3. For simplicity, assume that the statistical models for all learners are all well-specified and that the true model linkage graph is known. Learners $\mathcal{L}_2$ and $\mathcal{L}_4$ share common parameters with $\mathcal{L}_1$, and thus there exist model linkages between $\mathcal{L}_1$ and $\mathcal{L}_2$, and $\mathcal{L}_1$ and $\mathcal{L}_4$. In addition, learner $\mathcal{L}_3$ has a shared common parameter with $\mathcal{L}_2$, and thus in principle, $\mathcal{L}_3$ can help enhance the predictive performance of $\mathcal{L}_1$ implicitly. This process of positive feedback transmission is an *information flow* that enables implicit assistance to $\mathcal{L}_1$. In Figure 3, there are two such information flows, namely $\mathcal{L}_3 \to \mathcal{L}_2 \to \mathcal{L}_1$ and $\mathcal{L}_5 \to \mathcal{L}_4 \to \mathcal{L}_1$. Thus, in the model linkage graph, paths exist from $\mathcal{L}_3$ and $\mathcal{L}_5$ to $\mathcal{L}_1$. Learner $\mathcal{L}_6$ does not share common parameters with learners related to $\mathcal{L}_1$, and thus there is no model linkage between $\mathcal{L}_6$ and the others.

However, in practice, the true model linkage graph is not known a priori and needs to be specified. A misspecified model linkage graph may hamper the predictive performance of $\mathcal{L}_1$. The following section proposes a data-adaptive framework to identify an appropriate model linkage graph for prediction. The same idea can be used to improve parameter estimation accuracy.

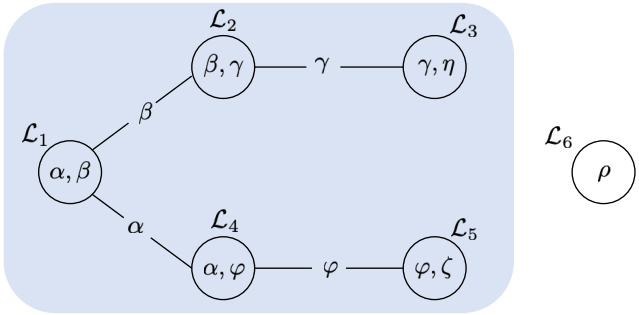

Figure 3: A model linkage graph with six learners. Learner $\mathcal{L}_1$ shares common parameters with $\mathcal{L}_2$ and $\mathcal{L}_4$, and is implicitly connected to $\mathcal{L}_3$ and $\mathcal{L}_5$ through $\mathcal{L}_2$ and $\mathcal{L}_4$.

3. General Bayesian Framework for Model Linkage Selection

3.1 Proposed Method

We propose a Bayesian framework to enhance the predictive performance of learner $\mathcal{L}_1$ by leveraging data from the other learners $\mathcal{L}_2, \ldots, \mathcal{L}_M$ according to a *user-specified model linkage graph*, $G = (V, E)$. Suppose that there exists a model linkage between two learners, say learners $\mathcal{L}_i$ and $\mathcal{L}_j$. Some of the elements in $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$ are restricted to be the same, say $\boldsymbol{\theta}_{i,\mathcal{S}_i} = \boldsymbol{\theta}_{j,\mathcal{S}_j} = \boldsymbol{\theta}_{\mathcal{S}_i,\mathcal{S}_j}$. From the Bayesian perspective, a natural way to integrate information is to compute the posterior distribution of the parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_{i,-\mathcal{S}_i}^{\mathrm{T}}, \boldsymbol{\theta}_{\mathcal{S}_i,\mathcal{S}_j}^{\mathrm{T}}, \boldsymbol{\theta}_{j,-\mathcal{S}_j}^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\theta}_{i,-\mathcal{S}_i}$ and $\boldsymbol{\theta}_{j,-\mathcal{S}_j}$ are obtained by removing the elements from $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$, indexed by the sets $\mathcal{S}_i$ and $\mathcal{S}_j$, respectively. Then, the posterior distribution of $\boldsymbol{\theta}$ can be computed as

$$\pi(\boldsymbol{\theta} \mid \mathbf{D}^{(i)}, \mathbf{D}^{(j)}) = \frac{p_{\boldsymbol{\theta}_i}^{(i)}(\mathbf{y}^{(i)}|\boldsymbol{\theta}_i, \mathbf{X}^{(i)}) p_{\boldsymbol{\theta}_j}^{(j)}(\mathbf{y}^{(j)}|\boldsymbol{\theta}_j, \mathbf{X}^{(j)}) \pi(\boldsymbol{\theta})}{p(\mathbf{y}^{(i)}, \mathbf{y}^{(j)}|\mathbf{X}^{(i)}, \mathbf{X}^{(j)})}$$

where $\pi(\boldsymbol{\theta})$ is a prior distribution on $\boldsymbol{\theta}$, and $p(\mathbf{y}^{(i)}, \mathbf{y}^{(j)}|\mathbf{X}^{(i)}, \mathbf{X}^{(j)})$ is the marginal likelihood obtained by integrating the numerator of the above equation with respect to $\boldsymbol{\theta}$.

More generally, one can compute the posterior distribution of the parameters according to the user-specified model linkage graph. Let $\mathcal{C}(G)$ be a set of indices recording the vertices that form a connected component with learner $\mathcal{L}_1$ in the user-specified graph G, including learner $\mathcal{L}_1$. That is, $\mathcal{C}(G)$ is a set containing all indices of learners that have at least a path to $\mathcal{L}_1$ and learner $\mathcal{L}_1$. For brevity, throughout the paper, let $\boldsymbol{\theta}$ be a vector obtained by concatenating the entries of $\boldsymbol{\theta}_\kappa$ for all $\kappa \in \mathcal{C}(G)$ without duplication. That is, the shared parameters between pairs of learners appear only once in $\boldsymbol{\theta}$. The posterior distribution for $\boldsymbol{\theta}$ under a specific model linkage can then be computed by

$$\pi(\boldsymbol{\theta} \mid \cup_{\kappa \in \mathcal{C}(G)} \mathbf{D}^{(\kappa)}) = \frac{\pi(\boldsymbol{\theta}) \prod_{\kappa \in \mathcal{C}(G)} p_{\boldsymbol{\theta}_\kappa}^{(\kappa)}(\mathbf{y}^{(\kappa)}|\boldsymbol{\theta}_\kappa, \mathbf{X}^{(\kappa)})}{p(\cup_{\kappa \in \mathcal{C}(G)} \mathbf{y}^{(\kappa)} | \cup_{\kappa \in \mathcal{C}(G)} \mathbf{X}^{(\kappa)})}$$

Then, the posterior predictive distribution for a new observation in learner $\mathcal{L}_1$ can be computed as

$$p(\widetilde{y} | \cup_{\kappa \in \mathcal{C}(G)} \mathbf{D}^{(\kappa)}, \widetilde{\mathbf{x}}) = \int_{\boldsymbol{\Theta}} p_{\boldsymbol{\theta}_1}^{(1)}(\widetilde{y}|\boldsymbol{\theta}_1, \widetilde{\mathbf{x}}) \pi(\boldsymbol{\theta} \mid \cup_{\kappa \in \mathcal{C}(G)} \mathbf{D}^{(\kappa)}) d\boldsymbol{\theta}, \tag{1}$$

Algorithm 1 Greedy Algorithm for Model Linkage Selection.

Input: User-specified graph G, data $\mathbf{D}^{(\kappa)}$, parameter vector $\boldsymbol{\theta}_\kappa$, parametric distribution $p^{(\kappa)}_{\boldsymbol{\theta}_\kappa}(\cdot \mid \boldsymbol{\theta}_\kappa, \mathbf{X}^{(\kappa)})$, prior distribution $\pi_\kappa(\cdot)$ on parameters $\boldsymbol{\theta}_\kappa$, for $\kappa = 1, \ldots, M$.

1: Initialize the index $\ell = 1$, linkage set $\zeta^{(1)} = \{1\}$.
2: **for** $\ell = 2, \ldots, M$ **do**
3: Let $N_G(\zeta^{(\ell-1)})$ denote the neighboring set of $\zeta^{(\ell-1)}$ within G, namely the set of learners in $G \backslash \zeta^{(\ell-1)}$ that have a model linkage with at least one learner in $\zeta^{(\ell-1)}$.
4: Calculate $j_{\text{opt}} = \text{argmax}_{j \in N_G(\zeta^{(\ell-1)})} p(\cup_{\kappa \in \zeta^{(\ell-1)}} \mathbf{y}^{(\kappa)} \mid \cup_{\kappa \in \zeta^{(\ell-1)}} \mathbf{X}^{(\kappa)}, \mathbf{D}^{(j)})$.
5: **if** $p(\cup_{\kappa \in \zeta^{(\ell-1)}} \mathbf{y}^{(\kappa)} | \cup_{\kappa \in \zeta^{(\ell-1)}} \mathbf{X}^{(\kappa)}) \geq p(\cup_{\kappa \in \zeta^{(\ell-1)}} \mathbf{y}^{(\kappa)} \mid \cup_{\kappa \in \zeta^{(\ell-1)}} \mathbf{X}^{(\kappa)}, \mathbf{D}^{(j_{\text{opt}})})$ **break**
6: Let $\zeta^{(\ell)} = \{j_{\text{opt}}\} \cup \zeta^{(\ell-1)}$.
7: **if** $\zeta^{(\ell)} = \mathcal{C}(G)$ **break**
8: **end for**
9: For a new predictor $\widetilde{\mathbf{x}}$, let $\widehat{p}^{(\ell)} = p(\cdot | \cup_{\kappa \in \zeta^{(\ell)}} \mathbf{D}^{(\kappa)}, \widetilde{\mathbf{x}})$ and $\widehat{\pi}^{(\ell)} = \pi(\cdot | \cup_{\kappa \in \zeta^{(\ell)}} \mathbf{D}^{(\kappa)})$.

Output: Predictive distribution $\widehat{p} = \widehat{p}^{(\ell)}$, posterior distribution $\widehat{\pi} = \widehat{\pi}^{(\ell)}$, and model linkage graph $\widehat{G}$.

where $\widetilde{\mathbf{x}}$ denotes the future observed covariates and $\boldsymbol{\Theta}$ is the parameter space. Note that the posterior predictive distribution has integrated information from learners in $\mathcal{C}(G)$ through the parameters in $\boldsymbol{\theta}$.

In practice, the true model linkage graph is not known a priori, and the user-specified model linkage graph G may be misspecified, as defined in Definition 4. A joint model as in (1) with a misspecified model linkage graph G can lead to severely biased parameter estimation, which affects the predictive quality of $\mathcal{L}_1$. Let $|\mathcal{C}(G)|$ be the cardinality of the set $\mathcal{C}(G)$. One way to address the aforementioned challenge is to exhaustively search all possible sets of model linkages over a graph with $|\mathcal{C}(G)|$ learners, and pick the set of model linkages that yields the largest marginal likelihood for $\mathcal{L}_1$ conditioned on other learners. However, the number of possible sets of model linkages grows exponentially with $|\mathcal{C}(G)|$, and it is computationally infeasible to evaluate all possible subgraphs of G.

To address the above challenge, we propose a greedy algorithm that is computationally feasible with theoretical guarantees. The proposed algorithm reduces the possible sets of model linkages from exponential in $|\mathcal{C}(G)|$ to quadratic in $|\mathcal{C}(G)|$. The main idea is to successively search for the next learner that will improve the conditional marginal likelihood of the current group of learners, starting from a singleton set $\mathcal{L}_1$. The algorithm will terminate and output an estimated model linkage graph $\widehat{G} = (V, \widehat{E})$ when adding any further learner no longer increases the marginal likelihood of the maintained learners. The greedy algorithm is outlined in Algorithm 1.

When the algorithm terminates at the ℓth iteration, $\mathcal{L}_1$ will integrate information from learners in $\zeta^{(\ell)}$, leading to the following posterior distribution of $\boldsymbol{\theta}$:

$$\pi(\boldsymbol{\theta} | \cup_{\kappa \in \zeta^{(\ell)}} \mathbf{D}^{(\kappa)}) = \frac{\pi(\boldsymbol{\theta}) \prod_{\kappa \in \zeta^{(\ell)}} p^{(\kappa)}_{\boldsymbol{\theta}_k}(\mathbf{y}^{(\kappa)} | \boldsymbol{\theta}_k, \mathbf{X}^{(\kappa)})}{p(\cup_{\kappa \in \zeta^{(\ell)}} \mathbf{y}^{(\kappa)} | \cup_{\kappa \in \zeta^{(\ell)}} \mathbf{X}^{(\kappa)})}.$$

Recall that the above $\boldsymbol{\theta}$ is the vector obtained by concatenating free parameters of $\boldsymbol{\theta}_\kappa$ for all $\kappa \in \zeta^{(\ell)}$ without duplication. The posterior predictive distribution for a new observation $\widetilde{\mathbf{x}}$ in learner $\mathcal{L}_1$ obtained from Algorithm 1 is computed as

$$p(\widetilde{y} | \cup_{\kappa \in \zeta^{(\ell)}} \mathbf{D}^{(\kappa)}, \widetilde{\mathbf{x}}) = \int_{\boldsymbol{\Theta}} p^{(1)}_{\boldsymbol{\theta}_1}(\widetilde{y} | \boldsymbol{\theta}_1, \widetilde{\mathbf{x}}) \pi(\boldsymbol{\theta} | \cup_{\kappa \in \zeta^{(\ell)}} \mathbf{D}^{(\kappa)}) d\boldsymbol{\theta}.$$

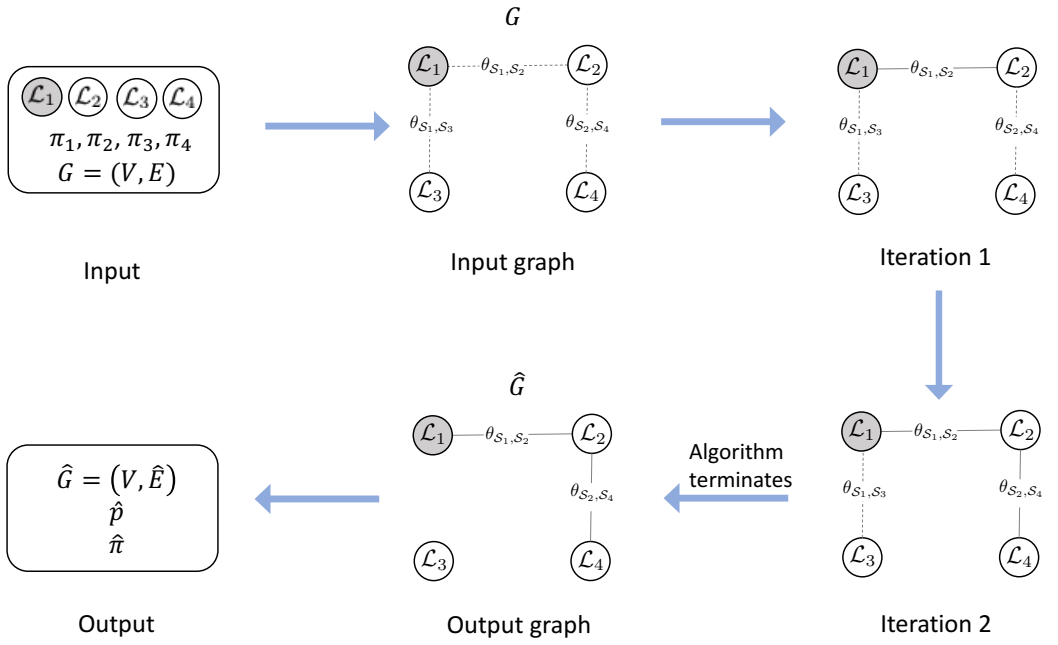

Figure 4: A schematic diagram illustrating Algorithm 1 with four learners. Each learner $\mathcal{L}_\kappa = (\mathbf{D}^{(\kappa)}, \mathcal{P}_\kappa)$ consists of the data and a user-specified class of parametric model. The algorithm starts with a user-specified graph that may or may not be misspecified. At each iteration, the algorithm selects a learner that maximizes the marginal likelihood of the current group of learners. The algorithm terminates when no such learners can be found. Finally, the algorithm outputs an estimated linkage graph $\widehat{G}$, predictive distribution $\widehat{p}$, and the posterior distribution $\widehat{\pi}$ for $\boldsymbol{\theta}$.

We now illustrate Algorithm 1 with a toy example in Figure 4. In this example, there are four learners, and the user-specified model linkage forms a connected component among all learners, i.e., there is a path from one learner to another learner. Suppose that the goal is to assist learner $\mathcal{L}_1$. It can be seen from the user-specified model linkage graph G that $\mathcal{L}_1$ is directly connected to $\mathcal{L}_2$ and $\mathcal{L}_3$, and implicitly connected with $\mathcal{L}_4$ through $\mathcal{L}_2$. At the first iteration of Algorithm 1, $\mathcal{L}_1$ computes its marginal likelihood $p(\mathbf{D}^{(1)})$, and conditional marginal likelihoods $p(\mathbf{D}^{(1)} \mid \mathbf{D}^{(2)})$ and $p(\mathbf{D}^{(1)} \mid \mathbf{D}^{(3)})$. If none of the conditional marginal likelihoods is larger than the marginal likelihood, Algorithm 1 will be terminated; otherwise, it includes the learner that produces the larger conditional marginal likelihood into the model linkage set. In this case, learner $\mathcal{L}_2$ is included in the model linkage set.

At the second iteration, the linkage set $\{\mathcal{L}_1, \mathcal{L}_2\}$ is treated as one learner since they are linked together during the first iteration. On the other hand, the candidate learners to establish a linkage are $\mathcal{L}_3$ and $\mathcal{L}_4$. In Figure 4, $\mathcal{L}_4$ is included in the linkage set following a similar argument as in the first iteration. Finally, $\mathcal{L}_3$ is not included in the linkage set based on our criterion and the algorithm terminates. Consequently, $\mathcal{L}_1$ will obtain a parameter estimation and predictive model that are trained from the union of $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_4$.

Remark 5 *Computation and Communication: In the above example, we note that $\mathcal{L}_1$ does not need to access the raw data of the other learners during the first iteration. In particular, $\mathcal{L}_1$ only requires $\mathcal{L}_2$ to share its likelihood function (e.g., in the form of an API) to calculate the (conditional) marginal likelihoods required for decision-making. Moreover, one could design more sophisticated protocols to increase the computation and communication efficiency in the proposed algorithm. For instance, when $\mathcal{L}_1$ decides whom to collaborate with, it only needs $\mathcal{L}_2$ and $\mathcal{L}_3$ to transmit their posterior distribution (in the form of, e.g., Monte Carlo samples) computed locally, to calculate the conditional marginal likelihood. Once a collaborator, say $\mathcal{L}_2$, is determined, the likelihood function of $\mathcal{L}_2$ will be sent to $\mathcal{L}_1$ to calculate the latest marginal likelihood and posterior.*

We briefly comment on the computation complexity of $\mathcal{L}_1$ in a general scenario. Suppose that at each iteration ℓ, each learner in $N_G(\zeta^{(\ell-1)})$ (following the notation in Algorithm 1) will send their likelihood functions to the linkage set $\zeta^{(\ell-1)}$. Let us consider the complexity of evaluating each candidate learner j in line 4 of Algorithm 1 as one unit. On the one hand, if learner $\mathcal{L}_1$ performs the computation alone, its total complexity units will be $O(\sum_{\ell=2,\ldots,M}(M-\ell+1)) = O(M^2)$ for large M. On the other hand, if everyone in the current collaboration set shares the computation cost, the complexity of $\mathcal{L}_1$ will be reduced to $O(\sum_{\ell=2,\ldots,M}(M-\ell+1)/(\ell-1)) = O(M\log M)$. The second setting is reasonable, because learners in $\zeta^{(\ell-1)}$ are already collaborators of $\mathcal{L}_1$ in the joint modeling.

3.2 Theoretical Results

We first provide definitions on *asymptotic prediction efficiency* and *model linkage selection consistency* in Section 3.2.1. Theoretical results for the proposed framework are presented in Section 3.2.2. Throughout the section, let $G = (V, E)$ and $\widehat{G} = (V, \widehat{E})$ be the user-specified and estimated model linkage graphs, respectively. The user-specified $G = (V, E)$ is usually specified based on prior scientific knowledge of practitioners, which may or may not be well-specified. Let $G^* = (V, E^*)$ be the largest subgraph of G, whose underlying model linkages are all well-specified after the statistical model for each learner is specified. That is, for any pair of learners $\mathcal{L}_i$ and $\mathcal{L}_j$ connected on G, their linkage is well-specified if and only if there exists an edge between them in G^*. Recall that more linkages imply a fewer number of free parameters in the joint model from G. Thus, intuitively speaking, G^* represents the most parsimonious parameterization of the underlying data-generating process. Ideally, the proposed algorithm can data-adaptively select $\widehat{G} = G^*$, thus identifying the correct model linkages and filtering out misspecified ones within G.

3.2.1 Definitions

Recall that the goal of the proposed framework is to enhance the predictive performance of $\mathcal{L}_1$ by borrowing information from other learners, $\mathcal{L}_2, \ldots, \mathcal{L}_M$. To evaluate the predictive performance of $\mathcal{L}_1$, we consider a general class of proper scoring rules (Gneiting and Raftery, 2007; Parry et al., 2012; Shao et al., 2019). Examples of proper scoring functions are the logarithmic score $s(p, y, \mathbf{x}) = -\log p(y|\mathbf{x})$ and the Brier score $s(p, y, \mathbf{x}) = -p(y|\mathbf{x}) + 0.5\int_{\mathbb{R}} p(\widetilde{y}|\mathbf{x})^2 d\widetilde{y}$ (Parry, 2016).

Definition 6 (Proper scoring function) *Let p^* be the true data-generating density function. A scoring function $s : (p, y, \mathbf{x}) \mapsto s(p, y, \mathbf{x})$ is proper if for any conditional density function p, we have $\int_{\mathcal{Y}} s(p, y, \mathbf{x}) p^*(y|\mathbf{x}) dy \geq \int_{\mathcal{Y}} s(p^*, y, \mathbf{x}) p^*(y|\mathbf{x}) dy$ almost surely.*

Let $\mathbb{E}$ be the expectation with respect to the data-generating distribution of y conditional on $\mathbf{x}$, denoted by p^*. Let $(\widetilde{y}, \widetilde{\mathbf{x}})$ be a new observation. The value $\mathbb{E}[s(p^*, \widetilde{y}, \widetilde{\mathbf{x}}) \mid \widetilde{\mathbf{x}}]$ is referred to as the *oracle score*, and $\mathbb{E}[s(\widehat{p}, \widetilde{y}, \widetilde{\mathbf{x}}) - s(p^*, \widetilde{y}, \widetilde{\mathbf{x}}) \mid \widetilde{\mathbf{x}}]$ is a non-negative expected prediction loss since $s(\cdot)$ is a proper scoring rule. It can be seen that the non-negativity of the Kullback-Leibler divergence from $p^*(\cdot|\mathbf{x})$ to $\widehat{p}(\cdot|\mathbf{x})$, defined as $D_{\mathrm{KL}}\{p^*(\cdot|\mathbf{x})||\widehat{p}(\cdot|\mathbf{x})\} = \int_{\mathcal{Y}} p^*(y|\mathbf{x})[\log\{p^*(y|\mathbf{x})\} - \log\{\widehat{p}(y|\mathbf{x})\}]dy$, implies that the logarithmic score is proper.

Recall that $\mathcal{C}(G)$ is a set of indices recording a set of learners that forms a connected component with $\mathcal{L}_1$ in a model linkage graph G. Next, we define linkage selection consistency.

Definition 7 (Linkage selection consistency) *Given a pre-specified model linkage graph G, suppose that $\psi : \{\mathbf{D}^{(\kappa)} : \kappa \in \mathcal{C}(G)\} \mapsto \widehat{G}$ is a linkage selection criterion in order to assist $\mathcal{L}_1$. Then, the linkage selection criterion ψ achieves linkage selection consistency if $\Pr(\widehat{E} = E^*) \to 1$ as $n \to \infty$.*

Here, the probability is defined over the observed data. In other words, a consistent linkage selection criterion ψ selects all the correct model linkages present in the user-specified linkage graphs. Next, we introduce the notion of asymptotic prediction efficiency. Suppose that C denotes a generic set of learners other than $\mathcal{L}_1$. Let $\widehat{p}_C$ denote the predictive distribution of $\mathcal{L}_1$ conditional all the learners in C.

Definition 8 (Asymptotic prediction efficiency) *Let $\widehat{p}$ be a constructed marginal predictive distribution for $\mathcal{L}_1$, and let $s(\cdot)$ be a proper scoring function. Then, $\widehat{p}$ is asymptotically prediction efficient if*

$$\frac{\mathbb{E}[s(\widehat{p}_{\mathcal{C}(G^*)}, \widetilde{y}, \widetilde{\mathbf{x}}) - s(p^*, \widetilde{y}, \widetilde{\mathbf{x}})]}{\mathbb{E}[s(\widehat{p}, \widetilde{y}, \widetilde{\mathbf{x}}) - s(p^*, \widetilde{y}, \widetilde{\mathbf{x}})]} \tag{2}$$

converges in probability to one as the number of observations $n_\kappa \to \infty$ for $\kappa = 1, \ldots, M$. Here, the expectation is taken over a new observation $(\widetilde{y}, \widetilde{\mathbf{x}})$. If $\widehat{p}$ is the posterior predictive distribution for $\mathcal{L}_1$ under a certain model linkage $\widehat{G}$, then $\widehat{G}$ is also referred to as an asymptotically efficient model linkage graph.

The ratio (2) contrasts the expected prediction loss of the constructed predictive density function $\widehat{p}$ and that of the predictive density function induced by G^*.

3.2.2 Theoretical Properties of Algorithm 1

We now proceed to study the theoretical properties of Algorithm 1. For technical convenience, we assume that learner $\mathcal{L}_1$ is well-specified. Note that, more generally, learner $\mathcal{L}_1$ may or may not be well-specified. In either case, the predictive performance can be evaluated by a proper scoring rule, e.g., the logarithmic rule. Throughout the theoretical studies, we consider the regime in which the number of learners M is fixed and the number of observations n_κ satifies $n_\kappa/n \to c_\kappa$, where $n = \sum_{\kappa=1}^{M} n_\kappa$ and c_κ is a positive constant.

Recall that $G = (V, E)$ is a user-specified graph. Moreover, recall that $G^* = (V, E^*)$ is the true model linkage graph, defined as the largest subgraph of G with correct model linkages between pairs of learners (given that each learner's model is specified). We denote the estimated model linkage graph from Algorithm 1 as $\widehat{G} = (V, \widehat{E})$.

Theorem 9 *Under some regularity conditions in Appendix A and given a user-specified graph G, the estimated model linkage edge set $\widehat{E}$ from Algorithm 1 achieves linkage selection consistency, namely* $\Pr(\widehat{E} = E^*) \to 1$ *as* $n \to \infty$.

Theorem 9 indicates that the estimated model linkage graph $\widehat{E}$ from Algorithm 1 is consistent in linkage selection, only selecting all the well-specified model linkages from G. Thus, the proposed approach is asymptotically robust against model linkage misspecification. We will verify the finite sample performance of Algorithm 1 using numerical studies in Section 4, by considering various settings such as model misspecification, model linkage misspecification, and data contamination. The following theorem guarantees that the predictive distribution constructed from Algorithm 1 is asymptotically prediction efficient.

Theorem 10 *Let $\widehat{p}$ be the constructed predictive distribution for $\mathcal{L}_1$ via Algorithm 1 based on its selected model linkage graph $\widehat{G}$. Let $s(\cdot)$ be a proper scoring function. Under the same conditions as in Theorem 9, $\widehat{p}$ is asymptotically prediction efficient.*

Note that Theorem 10 holds even when the statistical models for some learners are misspecified due to the definition of G^*. In other words, the proposed method yields a predictive distribution that is robust to model misspecification and model linkage graph misspecification.

4. Numerical Studies

4.1 Linear Regression Example

We consider a regression setting with six learners $\mathcal{L}_1, \ldots, \mathcal{L}_6$. The goal is to enhance the predictive performance of $\mathcal{L}_1$ by incorporating information from other learners. The data for the six learners are generated as follows:

$$y_i^{(\kappa)} = \begin{cases} \sum_{j=1}^{7} \beta_j^{(\kappa)} x_{ij}^{(\kappa)} + \epsilon_i^{(\kappa)} & \text{for } \kappa = 1, 3, 4, \\ \sum_{j=1}^{15} \beta_j^{(\kappa)} x_{ij}^{(\kappa)} + \epsilon_i^{(\kappa)} & \text{for } \kappa = 2, \\ \sum_{j=1}^{7} \beta_j^{(\kappa)} (x_{ij}^{(\kappa)} + 5)^2 + \epsilon_i^{(\kappa)} & \text{for } \kappa = 5, \\ \sum_{j=8}^{15} \beta_j^{(\kappa)} x_{ij}^{(\kappa)} + \epsilon_i^{(\kappa)} & \text{for } \kappa = 6, \end{cases}$$

where all regression coefficients are set to equal 0.3 except that $\beta_j^{(4)} = 0.6$ for $j = 1, \ldots, 7$. Each covariate $x_{ij}^{(\kappa)}$ and the random noise are generated from a standard Gaussian distribution for all the learners. Since learners $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_3$ share common parameters, and $\mathcal{L}_2$ shares common parameters with $\mathcal{L}_6$, there are model linkages among learners $\mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_3$, and $\mathcal{L}_6$. The true underlying model linkage graph G^* is illustrated on the right panel of Figure 5. Note that there is a model linkage between $\mathcal{L}_2$ and $\mathcal{L}_3$ since both share the same

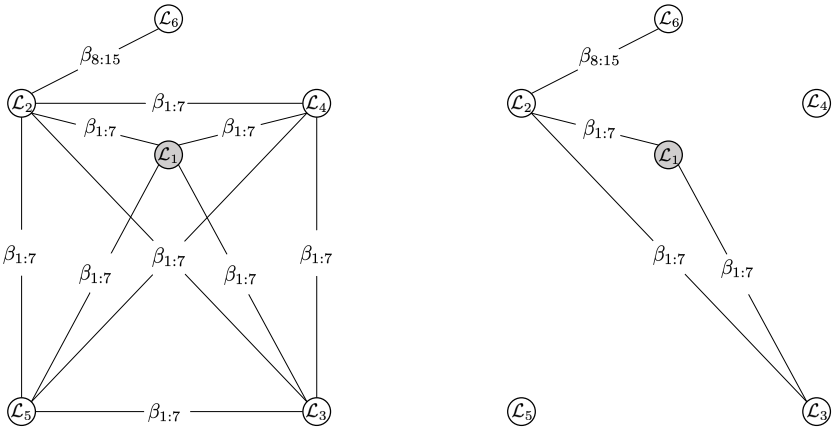

Figure 5: The user-specified model linkage graph G and the largest well-specified model linkage graph G^* within G are shown on the left and right panels, respectively.

common parameters with $\mathcal{L}_1$. For simplicity, we set the sample size for each learner to be n.

In practice, the user needs to specify a model linkage graph for the six learners and a statistical model for each learner. For the numerical studies, we specify correct statistical models for $\kappa = \{1, 2, 3, 4, 6\}$, and we misspecify $\mathcal{L}_5$ by assuming that the covariates are linearly related to the response y. We further restrict $\beta_j^{(\kappa)}$ to be the same for $\kappa = 1, \ldots, 5$ and $j = 1, \ldots, 7$. Moreover, we restrict $\beta_j^{(2)} = \beta_j^{(6)}$ to be the same for $j = 8, \ldots, 15$. Hence, the model linkage graph is misspecified in the sense that we assume that there exist model linkages between $\mathcal{L}_4$ and $\{\mathcal{L}_1,\ \mathcal{L}_2,\ \mathcal{L}_3,\ \mathcal{L}_5\}$, and between $\mathcal{L}_5$ and $\{\mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_3\}$. The user-specified graph G is illustrated on the left panel of Figure 5. We apply Algorithm 1 with the aforementioned user-specified graph and impose a multivariate Gaussian distribution, $N_p(\mathbf{0}, 4\mathbf{I}_p)$, as the prior distribution for the regression coefficients for all learners.

We will compare the proposed Algorithm 1 to fitting the model using data only from $\mathcal{L}_1$, and using the combined data from $\mathcal{L}_1$ and $\mathcal{L}_4$. Recall that the actual regression coefficients in $\mathcal{L}_4$ are different from that of $\mathcal{L}_1$, and thus combining data in $\mathcal{L}_1$ and $\mathcal{L}_4$ can lead to severely biased estimates of regression coefficients.

To assess the model linkage selection accuracy, we calculate the selection accuracy as the proportion of times when the estimated model linkage graph $\widehat{G}$ from Algorithm 1 is equal to G^*. To evaluate the performance across different models, we generate 50 test data for $\mathcal{L}_1$, and calculate the mean squared error between the predicted response and the actual response in the test data. Note that the mean squared error is a surrogate of the predictive logarithmic score under the Gaussian noise assumption. The predicted response is obtained by taking the mean of the posterior predictive distribution for each model. In addition, we calculate the length of the 95% prediction interval obtained from the posterior predictive distribution. The results for a range of sample sizes $n \in \{50, 75, \ldots, 150\}$, averaged over 200 replications, are shown in Figure 6.

From the left panel of Figure 6, the selection accuracy from Algorithm 1 approaches one as the sample size increases. This is in line with the selection consistency result of Theorem 9. Notably, Algorithm 1 yields a consistent model linkage graph even though the user-specified model linkage graph G is misspecified as shown in Figure 5. From the middle panel of Figure 6, we see that the proposed greedy algorithm yields the lowest prediction mean squared error across a range of sample sizes n. The results indicate that combining data from $\mathcal{L}_1$ and $\mathcal{L}_4$ imprudently can lead to higher prediction mean squared errors. Meanwhile, properly integrating data can improve the predictive performance of a single learner $\mathcal{L}_1$. Finally, the average length of the 95% prediction interval for the different models is presented in the right panel of Figure 6. We see that the proposed algorithm yields the narrowest prediction interval across the range of n.

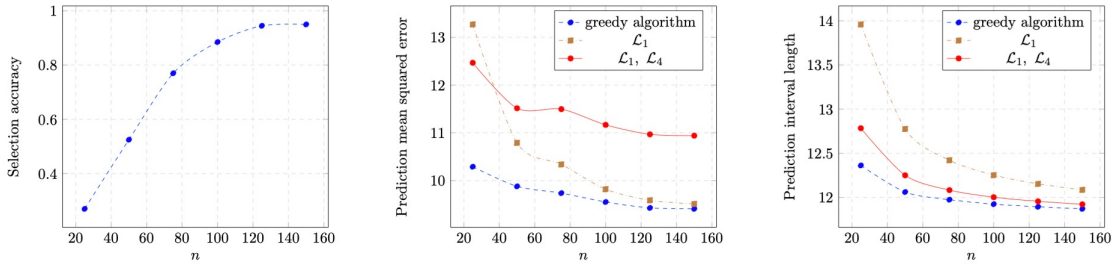

Figure 6: The results for modeling based on the proposed algorithm ("greedy algorithm"), modeling of the single agent $\mathcal{L}_1$ ("$\mathcal{L}_1$"), and combined modeling of $\mathcal{L}_1$ and $\mathcal{L}_4$ ("$\mathcal{L}_1, \mathcal{L}_4$") in the regression experiment. The evaluation uses the selection accuracy (left), prediction mean squared error (middle), and length of the 95% prediction interval (right), averaged over 200 replications.

4.2 Logistic Regression Example

In this section, we illustrate that the proposed framework can be employed for classification problems. We perform a numerical study with five learners $\mathcal{L}_1, \ldots, \mathcal{L}_5$ to enhance the prediction accuracy of $\mathcal{L}_1$. For each learner, we generate the covariates independently from a uniform distribution on the closed interval $[-1, 1]$. Then, the response variable is generated as the following:

$$\text{logit}(\Pr(y_i^{(\kappa)} \mid \mathbf{x}_i^{(\kappa)})) = \begin{cases} (\mathbf{x}_i^{(\kappa)})^{\mathrm{T}} \boldsymbol{\beta}^* & \text{for } \kappa = 1, 2, 3, 4, \\ 0 & \text{for } \kappa = 5, \end{cases}$$

where $\boldsymbol{\beta}^* = \{-0.8, -0.5, -0.2, 0.1, 0.4, 0.7, 1.0, 1.3, 1.6\}^{\mathrm{T}}$. Learners $\mathcal{L}_1$, $\mathcal{L}_2$, $\mathcal{L}_3$, and $\mathcal{L}_4$ share common parameters $\boldsymbol{\beta}^*$, and there are model linkages among $\mathcal{L}_1$, $\mathcal{L}_2$, $\mathcal{L}_3$, and $\mathcal{L}_4$. Learner $\mathcal{L}_5$ indicates that $y_i^{(5)}$ follows a Bernoulli distribution with probability 0.5 and is independent of the covariates. Thus, there are no model linkages between $\mathcal{L}_5$ and the other learners. We again set the sample sizes for all learners to the same n for simplicity.

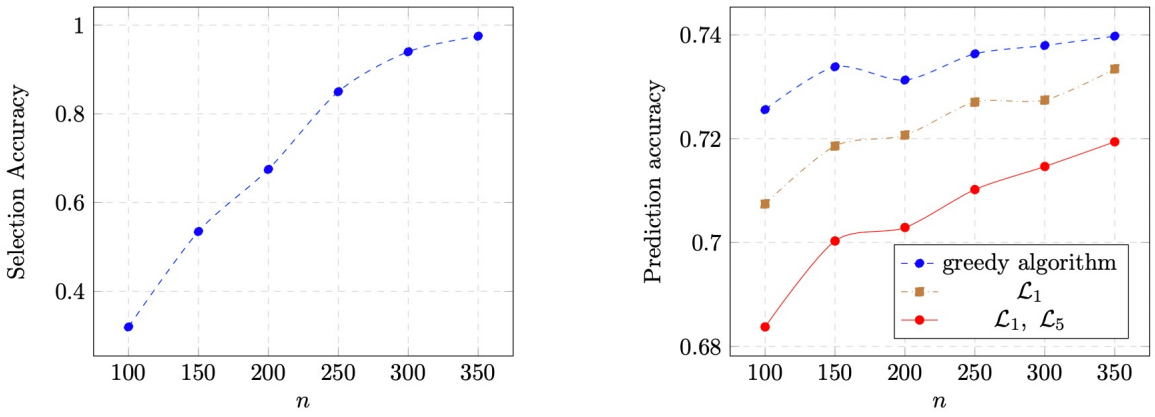

Figure 7: The results for modeling based on the proposed algorithm ("greedy algorithm"), modeling of the single agent $\mathcal{L}_1$ ("$\mathcal{L}_1$"), and combined modeling of $\mathcal{L}_1$ and $\mathcal{L}_5$ ("$\mathcal{L}_1, \mathcal{L}_5$") in the logistic regression experiment. The evaluation uses selection accuracy and prediction mean squared error, averaged over 200 replications.

We compare Algorithm 1 to models that are based on the data from $\mathcal{L}_1$, and combined data from $\mathcal{L}_1$ and $\mathcal{L}_5$. To evaluate the performance across different methods, we calculate the selection accuracy and prediction error. For Algorithm 1, we specify an incorrect model linkage graph where all learners are connected and fit the same logistic regression model for all learners. We set a multivariate normal distribution, $\mathcal{N}(\mathbf{0}, 4\mathbf{I})$, as the prior distribution for the regression coefficients for all learners. The results for a range of sample sizes $n = \{100, 150, \ldots, 350\}$, averaged over 200 replications, are shown in Figure 7.

From the left panel of Figure 7, we see that the selection accuracy converges to one as we increase the sample size n for each learner. That is, the greedy algorithm chooses not to include information from $\mathcal{L}_5$, even when the user-specified model linkage graph contains model linkages between $\mathcal{L}_5$ and the other learners. In addition, the proposed greedy algorithm yields the highest prediction error, whereas the modeling based on the joint data from $\mathcal{L}_1$ and $\mathcal{L}_5$ yields the lowest prediction error.

4.3 Logistic Regression with Data Contamination using Breast Cancer Data

Data contamination is an important issue when one decides whether to incorporate information from other learners. In practice, when several data sources are collected to have the same covariates, users tend to analyze the combined dataset to leverage more information. However, if some data sources are corrupted or contaminated, it is crucial to discriminate against them and avoid incorporating information from the contaminated learners. We illustrate that Algorithm 1 is robust against data contamination on some data sources.

We consider the Wisconsin Breast Cancer database (Mangasarian et al., 1995). The data consist of a response variable recording whether a cancer tissue is benign or malignant with 9 covariates from a total of 699 subjects. We randomly choose 100 samples as the test data

for evaluating prediction accuracy. Then, the data are randomly divided into 10 learners, in which each learner has n samples. Since the data from the 10 learners $\mathcal{L}_1,\ \mathcal{L}_2, \ldots,\ \mathcal{L}_{10}$ are subsamples of the original data set, we assume that the regression coefficients are the same across all learners. We then contaminate the data in $\mathcal{L}_{10}$ such that the binary response is flipped.

We apply Algorithm 1 with a misspecified model linkage graph by assuming that all learners are linked among each other. For each learner, we assume a logistic regression model with an intercept and 9 covariates. For simplicity, we impose the prior distribution $\mathcal{N}(0, 4^2)$ on all regression coefficients. The prediction accuracy for the proposed method, the method that uses $\mathcal{L}_1$, and the method that uses the combined data $\mathcal{L}_1$ and $\mathcal{L}_{10}$, averaged over 200 replications, are reported in Table 1.

From Table 1, we see that naively combining data from $\mathcal{L}_1$ and $\mathcal{L}_{10}$ will lead to a much lower prediction accuracy than the model using only data from $\mathcal{L}_1$. Our proposed method, on the other hand, chooses not to incorporate information from $\mathcal{L}_{10}$. Also, by combining data sources adaptively, our proposed method yields a prediction accuracy that is much higher than the model fit using data from $\mathcal{L}_1$ alone, for both cases of $n = \{25, 50\}$.

Table 1: Performance results of the proposed approach, model of $\mathcal{L}_1$ alone, and joint model of $\mathcal{L}_1$ and $\mathcal{L}_{10}$, as evaluated by the prediction accuracy. The results are averaged over 200 replications, with $n = \{25, 50\}$.

number of samples	**proposed method**	$\mathcal{L}_1$	$\mathcal{L}_1$ **and** $\mathcal{L}_{10}$
$n = 25$	0.928	0.844	0.500
$n = 50$	0.952	0.894	0.498

4.4 Integrating Information on Kidney Cancer Data

In this section, we analyze the kidney cancer data considered in Maity et al. (2019). The kidney cancer data consist of 33 different types of tumors, with a total of $n = 8108$ samples and up to $p = 198$ proteins for the different types of tumors. Each tumor type may have a different number of proteins. To study the association between patients' survival time and proteins, Maity et al. (2019) fit an accelerated failure time model with a log-normal assumption, from which they identified eight proteins that are most related to patients' survival time.

We now illustrate that integrating information from related cancer tumors using the proposed method can improve the prediction accuracy of patients' survival time. For simplicity, we consider only patients that are not alive at the observed survival time and fit a linear regression on the log-transformed survival time. We consider three types of tumors: (i) kidney renal clear cell carcinoma (KIRC), $\mathcal{L}_1$, (ii) kidney renal papillary cell carcinoma (KIRP), $\mathcal{L}_2$, and (iii) uterine corpus endometrial carcinoma (UCEC), $\mathcal{L}_3$, each of which has 146, 24, and 34 samples, respectively. Moreover, we pick up three proteins "PCADHERIN", "GAB2", and "HER3_pY1298" as the covariates, following (Maity et al., 2019) . The three proteins have been well studied and are all well-known for kidney tumor growth and in-

vasion (Blaschke et al., 2002; Duckworth et al., 2016; Akbani et al., 2014). In particular, the PCADHERIN has been considered as one of the most important proteins for kidney cancer (Maity et al., 2019). Both KIRC and KIRP originate from cells in the proximal convoluted tubules of the nephron (Chen et al., 2016). Thus, it is reasonable to assume that PCADHERIN has a similar effect on the log-transformed survival time of patients with KIRC and KIRP. On the other hand, PCADHERIN is expected to have a different effect on UCEC since UCEC is a type of uterine cancer. Inspired by the above domain knowledge, we set up a model linkage graph in the following way. We assume that there are linkages among KIRP, KIRC, and UCEC by sharing the effect of PCADHERIN on survival time across three tumor types.

We fit a linear regression model with a log-transformed response, namely $\log(y_i^{(\kappa)}) = \sum_{j=1}^{3} x_{ij}^{(\kappa)} \beta_j^{(\kappa)} + \epsilon_i^{(\kappa)}$, where the indices i, j, and κ denote the ith subject, the jth protein, and the κth tumor type. For simplicity, we assume that the random noise is normally distributed with different variances to account for heterogeneity across different tumor types, namely $\epsilon_i^{(\kappa)} \sim \mathcal{N}(0, \sigma_\kappa^2)$. Let $\beta_1^{(1)}, \beta_1^{(2)}, \beta_1^{(3)}$ be the regression coefficient for PCADHERIN across the three cancer types, which are linked on the specified linkage graph.

We compare the prediction accuracy of the proposed greedy algorithm with the model using only data from $\mathcal{L}_1$ (namely KIRC). To that end, we sample 20 data points from $\mathcal{L}_1$ such that the sample sizes across three tumor types are approximately the same. We treat the remaining data points for test purposes. The prior distributions of the regression coefficients are assumed to be standard normal, and the prior distributions for the three intercepts are assumed to follow a normal distribution with a mean of 10 and variance one. Moreover, we assume that $\sigma_\kappa^2 \sim \text{InvGamma}(2, 1)$ for $\kappa = 1, 2, 3$, where InvGamma denotes the inverse gamma distribution. We replicate the experiments above 100 times by sub-sampling 20 data observations from $\mathcal{L}_1$ for training the models.

The results indicate that Algorithm 1 connects $\mathcal{L}_2$ to $\mathcal{L}_1$ in 65% of the replications, and $\mathcal{L}_2$ to $\mathcal{L}_3$ 20% of the replications. In this example, Algorithm 1 selects different linkages across different replications due to randomness in the samples. The average prediction mean squared error (with its standard error) of the proposed method is 1.69(0.027). In comparison, the values are 1.74(0.031) if using data only from $\mathcal{L}_1$, and 1.66(0.024) if using the joint data from $\mathcal{L}_1$ and $\mathcal{L}_2$ throughout the 100 replications. The results imply that collaborating with $\mathcal{L}_2$ increases the predictive performance of $\mathcal{L}_1$, and Algorithm 1 tends to favor such a collaboration.

4.5 Epidemiological Data Study

In this experiment, we revisit the example illustrated in Figure 2 of Section 2. Recall that $\mathcal{L}_1$ employs n Binomial models $y_i^{(1)} \sim \text{Binomial}(x_i^{(1)}, \theta_{1,i})$, and learner $\mathcal{L}_2$ has poisson models $y_i^{(2)} \sim \text{Poisson}(\theta_{1,i}\theta_2 x_i^{(2)})$, with $i = 1, 2, \ldots, n$. The datasets for the two learners are denoted by $\mathbf{D}^{(1)} = (\mathbf{y}^{(1)}, \mathbf{x}^{(1)})$ and $\mathbf{D}^{(2)} = (\mathbf{y}^{(2)}, \mathbf{x}^{(2)})$. Note that the two learners are heterogeneous, and the parameters of $\mathcal{L}_2$ are not identifiable since $\theta_{1,i}\theta_2 = c\theta_{1,i} \cdot c^{-1}\theta_2$ for any positive constant c. With a joint modeling of $\mathcal{L}_1$ and $\mathcal{L}_2$, the parameters become identifiable. We first use the data in (Maucort-Boulch et al., 2008), where each learner has $n = 13$ data. For $\mathcal{L}_1$, the population size $x_i^{(1)}$ ranges from 37 to 700 and the empirical infection rate $y_i^{(1)}/x_i^{(1)}$

Table 2: Predictive performance of using a joint modeling of $\mathcal{L}_1$ and $\mathcal{L}_2$ ("Joint"), the proposed method ("Proposed"), and a single-agent modeling ("$\mathcal{L}_1$ or $\mathcal{L}_2$ alone"), evaluated for each learner. The evaluation is based on the expected log-predictive distribution, numerically computed from 1000 out-sample data and 100 replications. Standard errors are reported in the parentheses.

	$\mathcal{L}_1$ (Binomial)			$\mathcal{L}_2$ (Poisson)		
	Joint	Proposed	$\mathcal{L}_1$ alone	Joint	Proposed	$\mathcal{L}_2$ alone
Case 1	-3.76(0.006)	-3.86(0.016)	-4.05(0.009)	-3.76(0.002)	-3.76(0.002)	-3.76(0.002)
Case 2	-4.29(0.033)	-4.38(0.041)	-4.87(0.019)	-4.03(0.011)	-4.03(0.011)	-4.03(0.010)
Case 3	-8.37(0.069)	-4.58(0.135)	-4.04(0.016)	-3.62(0.009)	-3.63(0.009)	-3.63(0.009)
Case 4	-3.56(0.033)	-3.56(0.033)	-4.05(0.016)	-3.26(0.008)	-3.26(0.008)	-3.31(0.008)

ranges from 0 to 0.2. For $\mathcal{L}_2$, the women-years follow up $x_i^{(2)}$ ranges from 20000 to 550000, and the number of cancer incidences $y_i^{(2)}$ ranges from 10 to 700. In this experiment, we divided each $x_i^{(2)}$ by 1000 for computation efficiency, and note that such a scaling does not make an essential difference for parameter estimation. We used Uniform$(0, 1)$ as the prior distribution for each $\theta_{1,i}$ and Gamma$(5, 1)$ for θ_2. Algorithm 1 outputs that there is no linkage established between $\mathcal{L}_1$ and $\mathcal{L}_2$.

To develop more insights into the nature of the above models and data size, we perform four cases of simulated data experiments. We use $n = 13$ and the identical prior distributions as in the real-data experiment. In Case 1, we generated simulated data with $\mathbf{x}^{(1)} = (200, \overbrace{1000, ..., 1000}^{12})$ and $\mathbf{x}^{(2)} = (\overbrace{1000, ..., 1000}^{13})$, $\theta_2 = 10$, and $\theta_{1,i} = 0.1$ for $i = 1, 2, \ldots, n$. It serves as a case with large data, which tends to produce an accurate parameter estimation. Case 2 is similar to Case 1, except that $\mathbf{x}^{(1)} = (20, \overbrace{100, ..., 100}^{12})$ and $\mathbf{x}^{(2)} = (\overbrace{100, ..., 100}^{13})$. The latter two experimental cases are based on parameters estimated from the real data. Case 3 simulates the setting where there exists no underlying linkage between two learners. In particular, we simulate $y_i^{(1)}$ from the Binomial model with the original population size $x_i^{(1)}$ and the empirical rate $\theta_{1,i}$ that is calculated from the original data observations. We simulate $y_i^{(2)}$ from the Poisson model with the expectation that equals the original count observation. In contrast, Case 4 simulates the setting where there exists an underlying linkage between two learners. In particular, we estimate a joint model of the two learners, and use the posterior of $\theta_{1,i}$'s and θ_2 to generate $\mathbf{y}^{(1)}$ in Binomial models and $\mathbf{y}^{(2)}$ in Poisson models, with the original population sizes $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$.

To calculate the out-sample test performance of $\mathcal{L}_1$ or $\mathcal{L}_2$, whomever is being assisted, we generate 1000 test data based on the underlying data distributions. In particular, we numerically calculate $\mathbb{E}[\log p(\boldsymbol{y}^f \mid \boldsymbol{x}^f)]$ where p is the predictive distribution and $(\boldsymbol{y}^f, \boldsymbol{x}^f)$ denotes the test data. The results are shown in Table 2. Recall that in Cases 1, 2, and 4, there exists a model linkage between $\mathcal{L}_1$ and $\mathcal{L}_2$. In Case 3, there exists no underlying

model linkage. From the results, the test performance of the proposed method can data-adaptively approach the better performance of two options, namely with and without a linkage through the parameters $\theta_{1,i}$. Also, though $\mathcal{L}_2$ alone is not identifiable, it can improve $\mathcal{L}_1$'s performance in collaborative cases.

5. Conclusion and Further Remarks

With the rapid growth of low-cost data collection devices and decentralized learners, data analysts are faced with an important challenge to integrate information across a set of learners with diverse data sources. However, prudently combining data sources and fitting a joint model on all data sources can lead to biased estimates with a low prediction accuracy due to misspecified models or model linkages. We proposed a general approach to enhance the predictive performance of learner $\mathcal{L}_1$ by robustly integrating information across a set of learners. Since information are integrated through parameter linkages by sharing parameters, the data sources do not need to be transmitted across learners. As such, the proposed method is naturally compatible with decentralized learning that involves internet-of-things requiring low-energy consumption (Da Xu et al., 2014), smart sensors with limited hardware capacities (Zhou et al., 2016), and decentralized networks with limited communication bandwidths (Xiao and Luo, 2005). We showed that the proposed method is linkage selection-consistent and asymptotically prediction-efficient. The theoretical properties are established under the regime in which the number of learners M and the parameter dimensions are fixed. An interesting future problem is to study the theoretical properties of the proposed framework under the regime in which the number of learners or the dimension of parameters is allowed to diverge with the sample size.

In the following, we briefly describe the connection between the proposed framework and existing methods on data integration and distributed learning.

Data Integration. Integrating information from different data sources has been studied in the context of data integration (see, for instance, Tang and Song, 2016; Li and Li, 2018, and the references therein). When there is a unified model across multiple data sources, it is possible to improve statistical efficiency through parameter sharing or fitting one model using the combined data. For instance, Tang and Song (2016) employed a fused lasso approach to encourage the regression coefficients for different data sources to be similar. Li and Li (2018) developed an integrative linear discriminant analysis method by combining different data sources and showed that the classification accuracy could be improved compared with using a single data source. Some other work pre-specified constraints on latent variables to utilize heterogeneous data sources to address multiple parametric models. For example, in the study of gene regulatory networks, Jensen et al. (2007) proposed a Bayesian hierarchical model to integrate gene expression data, ChIP binding data, and promoter sequence data to infer statistical relationships between transcription factors and genes. The uniqueness of our work compared with existing methods is that our proposed method allows for a set of learners with diverse learning objectives and distinct statistical models. Furthermore, the set of learners share information only through linked parameters of interest. Therefore, the method in Section 3 can be used to help any learner efficiently identify cooperative learners when prior information is lacking.

Lunn et al. (2000) and Plummer (2015) also studied data integrations in the context of cut distributions, which can be seen as a probabilistic version of a two-step estimator. The main idea is to cut the propagation from uncertain models to precise models during joint learning to reduce biases propagated from incorrectly specified models (Lunn et al., 2009; Ogle et al., 2013). Jacob et al. (2017) proposed a predictive score principle for choosing the most appropriate joint modeling approach among the cut, full posterior, prior, and two-step approaches over a set of learners. It is possible to incorporate cut distribution into our proposed method to improve the finite-sample regime's performance. Nevertheless, the number of possible candidates exponentially increases with the number of learners, and thus the search space of each greedy selection step can be computationally prohibitive. Also, a systematic theoretical study of the cut distribution remains a challenging problem.

Distributed Learning. Data privacy has gained much attention recently, especially in distributed learning, where data curators do not wish to share the original data. This motivates some recent advances in distributed learning such as federated learning, where a central server sends the current global statistical model to a set of selected clients, and each client updates the model parameter with local data and returns the updates to the central server (Shokri and Shmatikov, 2015; Konečnỳ et al., 2016; Diao et al., 2020b). The objective function for federated learning is typically formulated as

$$\min_{\boldsymbol{\theta}} \; F(\boldsymbol{\theta}) := \sum_{\kappa=1}^{M} F_\kappa(\boldsymbol{\theta}), \quad \text{where} \quad F_\kappa(\boldsymbol{\theta}) = \sum_{(\mathbf{x},y)\in\mathbf{D}^{(\kappa)}} f(\boldsymbol{\theta};\mathbf{x},y), \tag{3}$$

where f denotes a global loss function, $\boldsymbol{\theta}$ parameterizes a global model to learn, and $\mathbf{D}^{(\kappa)}$ is a labeled dataset of the κth client. To optimize over $\boldsymbol{\theta}$, each client locally takes several iterations of (stochastic) gradient descent on the current parameter using its local data. Then the server takes a weighted average of the resulting parameters. Within a similar context, Jordan et al. (2019) proposed a communication-efficient surrogate likelihood framework for solving distributed statistical estimation problems, which provably improves upon simple averaging schemes.

In the context of our proposed framework, consider a set of learners each holding a data source $\mathbf{D}^{(\kappa)}$ and personal objective function f_κ, and a set of optimization constraints $\mathcal{C}$. A frequentist counterpart of our Bayesian approach is to minimize a proper scoring function (Dawid and Musio, 2015), e.g., the negative log-likelihood function, added with some form of regularization. The unknown parameters can be estimated by solving the following optimization problem

$$\begin{aligned} \min_{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_M} \; F(\boldsymbol{\theta}) &:= \sum_{\kappa=1}^{M} F_\kappa(\boldsymbol{\theta}_\kappa) + R(\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_M), \quad \text{subject to } \mathcal{C}, \\ \text{where } F_\kappa(\boldsymbol{\theta}) &= \sum_{(\mathbf{x},y)\in\mathbf{D}^{(\kappa)}} f_\kappa(\boldsymbol{\theta};\mathbf{x},y), \end{aligned} \tag{4}$$

where R is a suitably chosen regularization function. The federated learning falls into the above formulation when the models are restricted to be the same among different learners, namely $f_\kappa = f, \boldsymbol{\theta}_\kappa = \boldsymbol{\theta}$ for all κ. Without the constraint $\mathcal{C}$ and regularization R, (4) is equivalent to optimizing M individual objectives separately.

Both the developed Bayesian formulation or the frequentist counterpart in (4) can also be regarded as forms of personalized federated learning, where decentralized and heterogeneous agents participate in joint learning with peer agents to boost local learning performance. A key characteristic of personalization is that each agent has a specific local task, loss, model, and data. As such, the order of establishing linkages and the selected set of collaborative learners may depend on whom to assist. This is reflected through the asymmetric nature of the proposed Algorithm 1 in *finite-sample regimes*. To illustrate this point, we provide a toy example that involves three learners, each with a Gaussian model $y^{(\kappa)} \sim \mathcal{N}(\mu_\kappa, 1)$, $\kappa = 1, 2, 3$. It can be regarded as a regression with $x^{(\kappa)} = 1$ and unknown parameters μ_κ's. Suppose that G is a fully connected graph, and the prior of μ_κ is $\mathcal{N}(0, 10^2)$. The observed data are $y^{(1)} = 2$, $y^{(2)} = -0.3$, $y^{(3)} = -2$, respectively. One can verify that if $\mathcal{L}_1$ is the one to be assisted, Algorithm 1 will link it with $\mathcal{L}_2$ at the first step; Then, $\mathcal{L}_1$ and $\mathcal{L}_2$ are not linked with $\mathcal{L}_3$, terminating the algorithm. On the other hand, if $\mathcal{L}_2$ will be assisted, the algorithm first links it with $\mathcal{L}_3$ and then stops. Consequently, $\mathcal{L}_1$ and $\mathcal{L}_2$ will not establish a linkage in that case. The above indicates the asymmetric nature of collaboration: $\mathcal{L}_1$ being assisted by $\mathcal{L}_2$ doest not mean that $\mathcal{L}_1$ can assist $\mathcal{L}_2$ in the presence of other learners. Nevertheless, it can be verified that there is no such asymmetry in the case of two learners, meaning that $\mathcal{L}_1$ selecting $\mathcal{L}_2$ is equivalent to $\mathcal{L}_2$ selecting $\mathcal{L}_1$ in Algorithm 1.

Even in the context of vanilla federated learning, considerations of user misspecification or adversarial attacks are relatively new (Bhagoji et al., 2019; Jere et al., 2020), and the proposed notion of prediction efficiency and selection consistency are readily applicable. Moreover, in a general learning scenario where parameters may lose interpretability, it is still possible to build linkages among learners to reduce the overall model complexity and generalization errors. For example, Diao et al. (2019, 2020a) recently showed that appropriately restricting deep neural network parameters can significantly improve the performance of multi-modal image generation, compared with state-of-the-art methods that train an image generator separately from each data modality. From a theoretical perspective, the risk bound can be reduced by restricting the size of the function spaces through $\mathcal{C}$. When each learner's predictive performance is not severely biased by other learners compared with its reduced variance, it is worth establishing a joint optimization in the form of (4).

Acknowledgments

The authors thank the action editor and two anonymous referees for their constructive comments. The authors thank Dr. Veera Baladandayuthapani for sharing the kidney cancer data in Maity et al. (2019). Jiaying Zhou was supported by the Army Research Laboratory and the Army Research Office under grant number W911NF-20-1-0222 and National Science Foundation under grant number ECCS-2038603. Jie Ding and Vahid Tarokh were supported by the Office of Naval Research under grant number N00014-18-1-2244. Kean Ming Tan was supported by National Science Foundation under grant numbers DMS-1949730 and DMS-2113346, and National Institutes of Health under grant number RF1-MH122833.

Appendix A. Notation and Regularity Conditions

We start with introducing some regularity conditions needed for the theoretical development. These regularity conditions are generalizations of those in Walker (1969) from scalar to multidimensional vector. Suppose that each observation $(y_i, \mathbf{x}_i)$, $1 \leq i \leq n$, is modeled via a joint distribution with a density function $p(y, \mathbf{x} \mid \boldsymbol{\theta})$ with respect to a σ-finite measure μ. Moreover, $\mathbf{x} \in \mathbb{R}^k$ is modeled using the density function $h(\mathbf{x})$ that is independent of the parameter $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)^{\mathrm{T}} \in \mathbb{R}^p$. The joint density can thus be written as $p(y, \mathbf{x} \mid \boldsymbol{\theta}) = p(y \mid \mathbf{x}, \boldsymbol{\theta})h(\mathbf{x})$. Let the true conditional density function of y given $\mathbf{x}$ be $p^*(y \mid \mathbf{x})$, and thus the true joint density of $(y, \mathbf{x})$ can be written as $p^*(y, \mathbf{x}) = p^*(y \mid \mathbf{x})h^*(\mathbf{x})$. Note that $h(\mathbf{x})$ and $h^*(\mathbf{x})$ are not necessarily the same due to potential model misspecification when modeling $\mathbf{x}$.

Let $\boldsymbol{\theta}^*$ be an interior value in the parameter space $\boldsymbol{\Theta}$ defined as the minimizer of the Kullback-Leibler divergence between $p(y \mid \mathbf{x}, \boldsymbol{\theta}^*)$ and $p^*(y \mid \mathbf{x})$:

$$\begin{aligned}
\boldsymbol{\theta}^* &= \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\operatorname{argmax}} \int_{\{\mathcal{Y}, \mathcal{X}\}} \log \frac{p(y, \mathbf{x} \mid \boldsymbol{\theta})}{p^*(y, \mathbf{x})} p^*(y, \mathbf{x}) d\mu \\
&= \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\operatorname{argmax}} \int_{\{\mathcal{Y}, \mathcal{X}\}} [\log p(y \mid \mathbf{x}, \boldsymbol{\theta}) + \log\{h(\mathbf{x})\}] p^*(y \mid \mathbf{x}) h^*(\mathbf{x}) d\mathbf{x} dy \\
&= \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\operatorname{argmax}} \int_{\{\mathcal{X}\}} h^*(\mathbf{x}) \int_{\{\mathcal{Y}\}} \log p(y \mid \mathbf{x}, \boldsymbol{\theta}) p^*(y \mid \mathbf{x}) dy d\mathbf{x} \\
&= \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\operatorname{argmax}} \int_{\{\mathcal{Y}\}} \log p(y \mid \mathbf{x}, \boldsymbol{\theta}) p^*(y \mid \mathbf{x}) dy. \qquad (5)
\end{aligned}$$

Note that when the parametric model $p(y \mid \mathbf{x}, \boldsymbol{\theta})$ is well-specified, $p(y \mid \mathbf{x}, \boldsymbol{\theta}^*) = p^*(y \mid \mathbf{x})$. Let $\ell(\boldsymbol{\theta}) = \sum_{i=1}^n \log p(y_i, \mathbf{x}_i \mid \boldsymbol{\theta})$ be the log-likelihood function for the n observations and let $\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\operatorname{argmax}} \ell(\boldsymbol{\theta})$ be the maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$. Let $\boldsymbol{I}_n(\boldsymbol{\theta})$ be the observed Fisher information matrix with $\{\boldsymbol{I}_n(\boldsymbol{\theta})\}_{i,j} = -\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}$, and let $\boldsymbol{I}(\boldsymbol{\theta})$ be the expected Fisher information matrix for a single observation $(y, \mathbf{x})$ with $\{\boldsymbol{I}(\boldsymbol{\theta})\}_{i,j} = -\mathbb{E}\left\{\frac{\partial^2 \log p(y, \mathbf{x}|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\right\}$. Let $\{\boldsymbol{J}(\boldsymbol{\theta})\}_{i,j} = \mathbb{E}\{\frac{\partial \log p(y, \mathbf{x}|\boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \log p(y, \mathbf{x}|\boldsymbol{\theta})}{\partial \theta_j}\}$. We define $T(n) = \Theta\{f(n)\}$, namely when n is large, there exist some fixed constants c_1 and c_2 such that $c_1 f(n) \leq |T(n)| \leq c_2 f(n)$.

Some regularity conditions that are needed in the theoretical development are listed in the following.

(I) The parameter space $\boldsymbol{\Theta} \subseteq \mathbb{R}^p$ is compact.

(II) The set of points $\{\mathcal{Y}, \mathcal{X}\} = \{(y, \mathbf{x}) : p(y, \mathbf{x} \mid \boldsymbol{\theta}) > 0\}$ is independent of $\boldsymbol{\theta}$.

(III) If $\boldsymbol{\theta}_\alpha \neq \boldsymbol{\theta}_\beta$, then $\mu\{(y, \mathbf{x}) : p(y, \mathbf{x} \mid \boldsymbol{\theta}_\alpha) \neq p(y, \mathbf{x} \mid \boldsymbol{\theta}_\beta)\} > 0$.

(IV) For all $(y, \mathbf{x}) \in \{\mathcal{Y}, \mathcal{X}\}$ and $\delta > 0$, we have $|\log p(y, \mathbf{x}|\boldsymbol{\theta}) - \log p(y, \mathbf{x}|\boldsymbol{\theta}')| < H_\delta(y, \mathbf{x}, \boldsymbol{\theta}')$ as long as $\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2 < \delta$. Here, function H_δ has the property that $\lim_{\delta \to 0} H_\delta(y, \mathbf{x}, \boldsymbol{\theta}') = 0$ and that

$$\lim_{\delta \to 0} \int_{\{\mathcal{Y}, \mathcal{X}\}} H_\delta(y, \mathbf{x}, \boldsymbol{\theta}') p^*(y, \mathbf{x}) d\mu = 0.$$

(V) If $\boldsymbol{\Theta}$ is not bounded, then for any $\boldsymbol{\theta}_M \in \boldsymbol{\Theta}$, and sufficiently large Δ, we have

$$\log p(y, \mathbf{x}|\boldsymbol{\theta}) - \log p(y, \mathbf{x}|\boldsymbol{\theta}_M) < K_\Delta(y, \mathbf{x}, \boldsymbol{\theta}_M),$$

where $\|\boldsymbol{\theta}\|_2 > \Delta$ and K_Δ has the property that

$$\lim_{\Delta \to \infty} \int_{\{\mathcal{Y}, \mathcal{X}\}} K_\Delta(y, \mathbf{x}, \boldsymbol{\theta}_M) p^*(y, \mathbf{x}) d\mu < 0.$$

(VI) The maximum likelihood estimator (MLE), denoted by $\widehat{\boldsymbol{\theta}}$, exists, and the matrix $\boldsymbol{I}_n(\widehat{\boldsymbol{\theta}})$ is positive definite almost surely.

(VII) The log-likelihood function $\log p(y, \mathbf{x}|\boldsymbol{\theta})$ is twice continuously differentiable with respect to $\boldsymbol{\theta}$ in some neighborhood of $\boldsymbol{\theta}^*$.

(VIII) The first and second derivatives with respect to $\boldsymbol{\theta}$, and the integral of $\log p(y, \mathbf{x}|\boldsymbol{\theta})$, are exchangeable.

(IX) There exists a $\delta > 0$ such that

$$\left| \frac{\partial^2 \log p(y, \mathbf{x} \mid \boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} - \frac{\partial^2 \log p(y, \mathbf{x} \mid \boldsymbol{\theta}^*)}{\partial \theta_i \partial \theta_j} \right| < M_\delta(y, \mathbf{x}, \boldsymbol{\theta}^*)$$

for any pair (i, j) and $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 < \delta$, where the function M_δ satisfies

$$\lim_{\delta \to 0} \int_{\{\mathcal{Y}, \mathcal{X}\}} M_\delta(y, \mathbf{x}, \boldsymbol{\theta}^*) p^*(y, \mathbf{x}) d\mu = 0.$$

(X) The prior density function is continuous at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ and $\pi(\boldsymbol{\theta}^*) > 0$.

(XI) If $\mathcal{L}_1$ has a well-specified model $p_{\boldsymbol{\theta}_1}$, and $\mathcal{L}_2$ has misspecified model $p_{\boldsymbol{\theta}_2}$, the model linkage (defined in Definition 2) between $\mathcal{L}_1$ and $\mathcal{L}_2$ is misspecified (Definition 4).

Conditions (I)–(V) ensure that when $\boldsymbol{\theta}^*$ is an interior value of $\boldsymbol{\Theta}$, $\ell(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}^*)$ is sufficiently small for values of $\boldsymbol{\theta}$ that are not in the vicinity of $\boldsymbol{\theta}^*$, with probability tending to one as $n \to \infty$. We note that Condition (I) is a stronger than necessary assumption made to simplify the technical arguments. Alternatively, one may remove this assumption and show that the parameter estimate falls into a compact set with the probability increasing to one as the sample size increases. Conditions (VI)–(IX) ensure that when $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, $n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ has a limiting distribution $\mathcal{N}\big(\mathbf{0}, \boldsymbol{I}(\boldsymbol{\theta}^*)^{-1}\boldsymbol{J}(\boldsymbol{\theta}^*)\boldsymbol{I}(\boldsymbol{\theta}^*)^{-1}\big)$. Conditions (VII)–(IX) assume that $\boldsymbol{I}_n(\boldsymbol{\theta})$ is smooth in the vicinity of $\boldsymbol{\theta}^*$. Condition (XI) is needed to guarantee that learners with misspecified models are not included in the joint model to enhance the statistical performance of $\mathcal{L}_1$.

Appendix B. Assumptions on the Scoring Rule

Let $\mathbf{y} = \{y_1, y_2, \ldots, y_n\}^{\mathrm{T}}$ be an n-dimensional vector of the response and $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}^{\mathrm{T}} \in \mathbb{R}^{n\times k}$ be the design matrix of covariates. In the following, we state some assumptions on the score function s defined in Definition 6.

(A1) $\sup_{p,y,\mathbf{x}} \mathbb{E}[s(p, y, \mathbf{x})] \in (0, \infty)$.

(A2) Assume that the model $p(y|\mathbf{x}, \boldsymbol{\theta})$ is well-specified. Let $p(\widetilde{y}|\widetilde{\mathbf{x}}, \mathbf{y}, \mathbf{X})$ be the Bayesian predictive distribution of $\widetilde{y}$ given the new predictor $\widetilde{\mathbf{x}}$ and the data $\mathbf{y}$, $\mathbf{X}$. As $n \to \infty$,

$$\mathbb{E}\left[s\{p(\widetilde{y}|\widetilde{\mathbf{x}}, \mathbf{y}, \mathbf{X}), \widetilde{y}, \widetilde{\mathbf{x}}\} - s\{p(\widetilde{y}|\widetilde{\mathbf{x}}, \boldsymbol{\theta}^*), \widetilde{y}, \widetilde{\mathbf{x}}\}\right] = \Theta(n^{-\frac{1}{2}}). \tag{6}$$

The first assumption indicates that the expected loss is bounded by some positive constant for any given density function and prediction point. The second assumption indicates that the difference between the prediction score and oracle score is in the order of $n^{-1/2}$. Many commonly used scoring rules such as the Kullback-Leibler divergence and cross-entropy satisfy the aforementioned assumptions.

Appendix C. Proof of Theorems 9–10

We start with some technical lemmas that will be helpful for the proof of Theorems 9–10. Let $L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}) = p(\mathbf{y}, \mathbf{X}|\boldsymbol{\theta}) = \prod_{i=1}^n p(y_i, \mathbf{x}_i|\boldsymbol{\theta})$ be the likelihood function for the n observations and let $\ell(\boldsymbol{\theta})$ be the log-likelihood function. Let $p(\mathbf{y}, \mathbf{X}) = \int p(\mathbf{y}, \mathbf{X}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$ be the marginal likelihood of $(\mathbf{y}, \mathbf{X})$. The following lemma is a multidimensional counterpart of Walker (1969) that provides limiting properties for the maximum likelihood estimator and marginal likelihood.

Lemma 11 *Assume that the regularity conditions in Appendix A hold. Let $\boldsymbol{\theta}^* \in \mathbb{R}^p$ be an interior value in the parameter space $\boldsymbol{\Theta}$, and assume that the data pair $(y_i, \mathbf{x}_i) \in \mathbb{R}^{k+1}$ has density $p(y_i, \mathbf{x}_i|\boldsymbol{\theta})$ for $i = 1, 2, \ldots, n$. Let $\widehat{\boldsymbol{\theta}}$ be the maximum likelihood estimator of $\boldsymbol{\theta}$. As $n \to \infty$, the following results hold:*

(i) Let $N(\delta) = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 < \delta\}$ be a neighborhood of $\boldsymbol{\theta}^*$ contained in $\boldsymbol{\Theta}$. For any positive δ, there exists a positive number $k(\delta)$ depending on δ such that

$$\lim_{n\to\infty} \Pr\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}\setminus N(\delta)} n^{-1}\{\ell(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}^*)\} < -k(\delta)\right] = 1.$$

(ii) Let $(\widehat{\xi^2})^{-1} = \det|\boldsymbol{I}_n(\widehat{\boldsymbol{\theta}})|$, we have $\lim_{n\to\infty} n^{-p}\left\{(\widehat{\xi^2})^{-1}\right\} = \det|\boldsymbol{I}(\boldsymbol{\theta}^*)|$.

(iii) $\ell(\boldsymbol{\theta}^*) - \ell(\widehat{\boldsymbol{\theta}}) = \Theta(1)$.

(iv) $\lim_{n\to\infty} \left\{p(\mathbf{y}, \mathbf{X}|\widehat{\boldsymbol{\theta}})\widehat{\xi}\right\}^{-1} p(\mathbf{y}, \mathbf{X}) = (2\pi)^{\frac{p}{2}}\pi(\boldsymbol{\theta}^*)$.

Lemma 11(i) indicates that the difference between $\ell(\boldsymbol{\theta})$ and $\ell(\boldsymbol{\theta}^*)$ will be large when $\boldsymbol{\theta}$ is not in the δ-neighborhood of $\boldsymbol{\theta}^*$. Lemma 11(ii) establishes that the determinant of the observed Fisher information matrix converges to the determinant of the expected Fisher information matrix. Lemma 11(iii) shows that the log-likelihood function evaluated at $\boldsymbol{\theta}^*$ and $\widehat{\boldsymbol{\theta}}$ are at the same order. The proof of Lemma 11 is provided in Appendix D.1.

Next, we present a key lemma that provides similar results as those of Lemma 11, but under the setting with non-identically distributed data that arise from two different data sources from two different learners. To this end, we define some notation. Without loss of generality, we consider two learners $\mathcal{L}_1$ and $\mathcal{L}_2$ with sample size n_1 and n_2, respectively. In particular, for each learner $\mathcal{L}_\kappa$ with $\kappa = 1, 2$, the user-specified density function of the data pair $(y_i^{(\kappa)}, \mathbf{x}_i^{(\kappa)}) \in \mathbb{R}^{k_\kappa+1}$ is denoted as $p^{(\kappa)}_{\boldsymbol{\theta}_\kappa}(y_i^{(\kappa)}, \mathbf{x}_i^{(\kappa)}|\boldsymbol{\theta}_\kappa)$. Let $p^*_\kappa(y_i^{(\kappa)}, \mathbf{x}_i^{(\kappa)})$ be the true underlying density function. Similar to Lemma 11, let $\boldsymbol{\theta}^*_\kappa \in \mathbb{R}^{p_\kappa}$ be an interior value in the parameter space $\boldsymbol{\Theta}_\kappa$.

As defined in Definition 2, let $\boldsymbol{\theta}_{1,\mathcal{S}_1} = \boldsymbol{\theta}_{2,\mathcal{S}_2} = \boldsymbol{\theta}_{\mathcal{S}_1,\mathcal{S}_2} \in \mathbb{R}^{p_s}$ be the shared parameter between $\mathcal{L}_1$ and $\mathcal{L}_2$. Moreover, let $\boldsymbol{\theta}_\mathcal{C} = (\boldsymbol{\theta}^{\mathrm{T}}_{1,-\mathcal{S}_1}, \boldsymbol{\theta}^{\mathrm{T}}_{\mathcal{S}_1,\mathcal{S}_2}, \boldsymbol{\theta}^{\mathrm{T}}_{2,-\mathcal{S}_2})^{\mathrm{T}} \in \mathbb{R}^{p_\mathcal{C}}$ be the vector obtained by concatenating entries of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ without duplication, which has a dimension of $p_\mathcal{C} = p_1 + p_2 - p_s$. Let $\boldsymbol{\theta}^*_\mathcal{C} \in \mathbb{R}^{p_\mathcal{C}}$ be its corresponding interior value in $\boldsymbol{\Theta}_\mathcal{C}$. We denote $\widetilde{\boldsymbol{\theta}}_1 = (\boldsymbol{\theta}^{\mathrm{T}}_{1,-\mathcal{S}_1}, \boldsymbol{\theta}^{\mathrm{T}}_{\mathcal{S}_1,\mathcal{S}_2})^{\mathrm{T}}$ and $\widetilde{\boldsymbol{\theta}}_2 = (\boldsymbol{\theta}^{\mathrm{T}}_{2,-\mathcal{S}_2}, \boldsymbol{\theta}^{\mathrm{T}}_{\mathcal{S}_1,\mathcal{S}_2})^{\mathrm{T}}$ as the parameters for $\mathcal{L}_1$ and $\mathcal{L}_2$ after incorporating information from the model linkage between the two learners. We note that $\widetilde{\boldsymbol{\theta}}_1$ and $\widetilde{\boldsymbol{\theta}}_2$ are equivalent to $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, respectively, after some reordering of the elements. The notation are simply defined to facilitate the proof of the theoretical results. Let $\mathbf{y}^{(\kappa)} = (y_1, y_2, \ldots, y_{n_\kappa})^{\mathrm{T}}$ and $\mathbf{X}^{(\kappa)} = (\mathbf{x}_1^{(\kappa)}, \mathbf{x}_2^{(\kappa)}, \ldots, \mathbf{x}_{n_\kappa}^{(\kappa)})^{\mathrm{T}}$. Denote $L(\widetilde{\boldsymbol{\theta}}_1|\mathbf{y}^{(1)}, \mathbf{X}^{(1)})$, $L(\widetilde{\boldsymbol{\theta}}_2|\mathbf{y}^{(2)}, \mathbf{X}^{(2)})$, and $L(\boldsymbol{\theta}_\mathcal{C}|\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \mathbf{X}^{(1)}, \mathbf{X}^{(2)})$ as the likelihood functions of $\mathcal{L}_1$, $\mathcal{L}_2$, and ($\mathcal{L}_1$, $\mathcal{L}_2$), where $L(\widetilde{\boldsymbol{\theta}}_1|\mathbf{y}^{(1)}, \mathbf{X}^{(1)}) = p^{(1)}_{\widetilde{\boldsymbol{\theta}}_1}(\mathbf{y}^{(1)}, \mathbf{X}^{(1)}|\widetilde{\boldsymbol{\theta}}_1)$, $L(\widetilde{\boldsymbol{\theta}}_2|\mathbf{y}^{(2)}, \mathbf{X}^{(2)}) = p^{(2)}_{\widetilde{\boldsymbol{\theta}}_2}(\mathbf{y}^{(2)}, \mathbf{X}^{(2)}|\widetilde{\boldsymbol{\theta}}_2)$, and

$$\begin{aligned} L(\boldsymbol{\theta}_\mathcal{C}|\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \mathbf{X}^{(1)}, \mathbf{X}^{(2)}) &= p^{(\mathcal{C})}_{\boldsymbol{\theta}_\mathcal{C}}(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \mathbf{X}^{(1)}, \mathbf{X}^{(2)}|\boldsymbol{\theta}_\mathcal{C}) \\ &= p^{(1)}_{\widetilde{\boldsymbol{\theta}}_1}(\mathbf{y}^{(1)}, \mathbf{X}^{(1)}|\widetilde{\boldsymbol{\theta}}_1) p^{(2)}_{\widetilde{\boldsymbol{\theta}}_2}(\mathbf{y}^{(2)}, \mathbf{X}^{(2)}|\widetilde{\boldsymbol{\theta}}_2). \end{aligned}$$

Let ℓ_1, ℓ_2, and $\ell_\mathcal{C}$ be their corresponding log-likelihood functions and let $\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2$, and $\widehat{\boldsymbol{\theta}}_\mathcal{C}$ be the MLEs obtained from maximizing ℓ_1, ℓ_2, and $\ell_\mathcal{C}$, respectively. Let $\boldsymbol{I}^{(\kappa)}_{n_\kappa}(\widetilde{\boldsymbol{\theta}}_\kappa)$ and $\boldsymbol{I}^{(\kappa)}(\widetilde{\boldsymbol{\theta}}_\kappa)$ be the observed Fisher information matrix and expected Fisher information matrix on single observation, respectively, for $\kappa = 1, 2$. Let $\boldsymbol{I}_n(\boldsymbol{\theta}_\mathcal{C})$ be the matrix with element $\{\boldsymbol{I}_n(\boldsymbol{\theta}_\mathcal{C})\}_{i,j} = -\frac{\partial^2 \ell_\mathcal{C}(\boldsymbol{\theta}_\mathcal{C})}{\partial\theta_{\mathcal{C},i}\partial\theta_{\mathcal{C},j}}$, where $\theta_{\mathcal{C},i}$ is the ith element of $\boldsymbol{\theta}_\mathcal{C}$. Next, let $\pi(\widetilde{\boldsymbol{\theta}}_1)$, $\pi(\widetilde{\boldsymbol{\theta}}_2)$, and $\pi(\boldsymbol{\theta}_\mathcal{C})$ be the prior density of $\widetilde{\boldsymbol{\theta}}_1$, $\widetilde{\boldsymbol{\theta}}_2$, and $\boldsymbol{\theta}_\mathcal{C}$, respectively. Let $p(\mathbf{y}^{(\kappa)}, \mathbf{X}^{(\kappa)}) = \int p^{(\kappa)}_{\widetilde{\boldsymbol{\theta}}_\kappa}(\mathbf{y}^{(\kappa)}, \mathbf{X}^{(\kappa)}|\widetilde{\boldsymbol{\theta}}_\kappa)\pi(\widetilde{\boldsymbol{\theta}}_\kappa)d\widetilde{\boldsymbol{\theta}}_\kappa$ be the marginal likelihood of $(\mathbf{y}^{(\kappa)}, \mathbf{X}^{(\kappa)})$, for $\kappa = 1, 2$. Moreover, the marginal likelihood of $(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \mathbf{X}^{(1)}, \mathbf{X}^{(2)})$ is expressed as

$$p(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \mathbf{X}^{(1)}, \mathbf{X}^{(2)}) = \int p^{(1)}_{\widetilde{\boldsymbol{\theta}}_1}(\mathbf{y}^{(1)}, \mathbf{X}^{(1)}|\widetilde{\boldsymbol{\theta}}_1) p^{(2)}_{\widetilde{\boldsymbol{\theta}}_2}(\mathbf{y}^{(2)}, \mathbf{X}^{(2)}|\widetilde{\boldsymbol{\theta}}_2)\pi(\boldsymbol{\theta}_\mathcal{C})d\boldsymbol{\theta}_\mathcal{C}.$$

We now present the results in the following lemma.

Lemma 12 *Assume that the regularity conditions in Appendix A hold. Suppose that* $n = n_1 + n_2$, *and* $n_1/n \to c_1$, $n_2/n \to c_2$, *as* $n \to \infty$, *where* $c_1, c_2 \in (0,1)$. *Then, the following results hold:*

(i) Let $N_{\mathcal{C}}(\delta) = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}^*_{\mathcal{C}}\|_2 < \delta\}$ *be a neighborhood of* $\boldsymbol{\theta}^*_{\mathcal{C}}$ *contained in* $\boldsymbol{\Theta}_{\mathcal{C}}$. *For any* $\delta > 0$, *there exists a positive number* $k(\delta)$ *depending on* δ *such that*

$$\lim_{n\to\infty} \Pr\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_{\mathcal{C}}\setminus N_{\mathcal{C}}(\delta)} n^{-1}\left\{\ell_{\mathcal{C}}(\boldsymbol{\theta}) - \ell_{\mathcal{C}}(\boldsymbol{\theta}^*_{\mathcal{C}})\right\} < -k(\delta)\right] = 1.$$

(ii) Let $(\widehat{\xi}^2_{\mathcal{C}})^{-1} = \det|\boldsymbol{I}_n(\widehat{\boldsymbol{\theta}}_{\mathcal{C}})|$, *we have* $\lim_{n\to\infty} n^{-p_{\mathcal{C}}}\left\{(\widehat{\xi}^2_{\mathcal{C}})^{-1}\right\} = \Theta(1)$.

(iii) $\ell_{\mathcal{C}}(\boldsymbol{\theta}^*_{\mathcal{C}}) - \ell_{\mathcal{C}}(\widehat{\boldsymbol{\theta}}_{\mathcal{C}}) = \Theta(1)$.

(iv) $\lim_{n\to\infty}\left\{p^{(\mathcal{C})}_{\boldsymbol{\theta}^*_{\mathcal{C}}}(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \mathbf{X}^{(1)}, \mathbf{X}^{(2)}|\widehat{\boldsymbol{\theta}}_{\mathcal{C}})\widehat{\xi}_{\mathcal{C}}\right\}^{-1} p(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \mathbf{X}^{(1)}, \mathbf{X}^{(2)}) = (2\pi)^{\frac{p_{\mathcal{C}}}{2}}\pi(\boldsymbol{\theta}^*_{\mathcal{C}})$.

Lemmas 13–14 concern the magnitudes of the joint marginal likelihood and marginal likelihood for each learner, under the case when the model linkage is well-specified or misspecified. Similar results were established in the context of change point detection (Du et al., 2016). Both lemmas will be used to prove that Algorithm 1 will select the correct model linkage in each iteration of the algorithm.

Lemma 13 *Assume that the regularity conditions in Appendix A hold. Let* $\boldsymbol{\theta}_{\mathcal{S}_1,\mathcal{S}_2} \in \mathbb{R}^{p_s}$ *be the shared parameter between* $\mathcal{L}_1$ *and* $\mathcal{L}_2$ *as defined in Definition 2. Let the interior value of* $\boldsymbol{\theta}_{\mathcal{S}_1,\mathcal{S}_2}$ *in* $\mathcal{L}_1$ *and* $\mathcal{L}_2$ *be* $\boldsymbol{\theta}^*_{1,\mathcal{S}_1}$ *and* $\boldsymbol{\theta}^*_{2,\mathcal{S}_2}$, *respectively. If* $\boldsymbol{\theta}^*_{1,\mathcal{S}_1} = \boldsymbol{\theta}^*_{2,\mathcal{S}_2}$, *and* $n_1/n_2 = \Theta(1)$, *as* $m = \min\{n_1, n_2\} \to \infty$, *we have*

$$\frac{p(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}|\mathbf{X}^{(1)}, \mathbf{X}^{(2)})}{p(\mathbf{y}^{(1)}|\mathbf{X}^{(1)})p(\mathbf{y}^{(2)}|\mathbf{X}^{(2)})} \xrightarrow{p} \Theta(m^{\frac{p_s}{2}}).$$

Lemma 14 *Under the same conditions as in Lemma 13, if* $\boldsymbol{\theta}^*_{1,\mathcal{S}_1} \neq \boldsymbol{\theta}^*_{2,\mathcal{S}_2}$, *as* $\min\{n_1, n_2\} \to \infty$, *we have*

$$\frac{p(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}|\mathbf{X}^{(1)}, \mathbf{X}^{(2)})}{p(\mathbf{y}^{(1)}|\mathbf{X}^{(1)})p(\mathbf{y}^{(2)}|\mathbf{X}^{(2)})} \xrightarrow{p} 0. \tag{7}$$

C.1 Proof of Theorem 9

Proof The main idea of the proof is to show that in each iteration, the proposed algorithm will integrate information from one additional learner based on the linkage set E^*, and avoid incorporating information from the learner that is not in the linkage set E^*. Consequently, the final output of the algorithm $\widehat{E} = E^*$. We start the proof at the ℓth iteration.

At the ℓth iteration, let $\zeta^{(\ell-1)}$ be the set of learners that are already included in the joint model built from the previous $\ell - 1$ iterations. Let $\zeta^{(\ell-1)}_{\mathcal{C}(G)} \subseteq \{2, \ldots, M\}\backslash\zeta^{(\ell-1)}$ be a non-empty set of learners with at least a path to $\mathcal{L}_1$ as defined in the user-specified model

linkage graph G. Note that $\zeta^{(\ell-1)}_{\mathcal{C}(G)}$ is non-empty, otherwise, the algorithm would have been terminated at the $(\ell-1)$th iteration.

There are two cases: (i) there is at least a $j \in \zeta^{(\ell-1)}_{\mathcal{C}(G)}$ and an $i \in \zeta^{(\ell-1)}$ such that the model linkage between $\mathcal{L}_j$ and $\mathcal{L}_i$ is well-specified; and (ii) there are no well-specified linkages between pairs of learners in $\zeta^{(\ell-1)}_{\mathcal{C}(G)}$ and $\zeta^{(\ell-1)}$.

Case (i): Suppose that there is a well-specified model linkage between $\mathcal{L}_j$ for a $j \in \zeta^{(\ell-1)}_{\mathcal{C}(G)}$ and $\mathcal{L}_i$ for an $i \in \zeta^{(\ell-1)}$. By Lemma 13, Step 2 of Algorithm 1 will hold, and $\mathcal{L}_j$ will form a new set with $\zeta^{(\ell-1)}$, namely $\zeta^{(\ell)} = \zeta^{(\ell-1)} \cup \{j\}$. More generally, if there are more than one learner in $\zeta^{(\ell-1)}_{\mathcal{C}(G)}$ that have well-specified linkages with learners in $\zeta^{(\ell-1)}$, the algorithm will select $j_{\text{opt}} = \text{argmax}_{j \in \zeta^{(\ell-1)}_{\mathcal{C}(G)}} p(\cup_{\kappa \in \zeta^{(\ell-1)}} \mathbf{y}^{(\kappa)} | \mathbf{y}^{(j)})$ and form a new set $\zeta^{(\ell)} = \zeta^{(\ell-1)} \cup \{j_{\text{opt}}\}$.

Case (ii): On the other hand, if the model linkages between $\mathcal{L}_j$ for $j \in \zeta^{(\ell-1)}_{\mathcal{C}(G)}$ and $\mathcal{L}_i$ for $i \in \zeta^{(\ell-1)}$ are misspecified for all i and j, Lemma 14 ensures that Algorithm 1 will be terminated, and hence, the misspecified model linkages are not included into $\zeta^{(\ell-1)}$.

As a result, the algorithm terminates when all well-specified linkages are included, and no misspecified linkages and misspecified models will be included, namely $\widehat{E} = E^*$. This concludes the proof. ■

C.2 Proof of Theorem 10

Proof Recall that $\mathcal{C}(G)$ is the set of indices recording the vertices that form a connected component with learner $\mathcal{L}_1$ in the user-specified graph G, and G^* is the true underlying graph after the statistical models for all learners and G are specified. Let $\mathcal{G} = \{\bar{G} = (V, \bar{E}) : \bar{E} \in E,\ \bar{E} \neq E^*\}$ be a set of graphs that is a subgraph of G, but $\bar{E} \neq E^*$. Let $\widehat{p}_{\mathcal{C}(G)}$ be the posterior predictive distribution constructed based on learners in $\mathcal{C}(G)$ as defined in (1), and let $G_i = (V, E_i)$ be a graph with edge set E_i. We consider the prediction efficiency ratio defined as follows:

$$\frac{\mathbb{E}\{s(\widehat{p}_{\mathcal{C}(G^*)}, \widetilde{y}, \widetilde{\mathbf{x}}) - s(p^*, \widetilde{y}, \widetilde{\mathbf{x}})\}}{\Pr(\widehat{E} = E^*)\mathbb{E}\{s(\widehat{p}_{\mathcal{C}(G^*)}, \widetilde{y}, \widetilde{\mathbf{x}}) - s(p^*, \widetilde{y}, \widetilde{\mathbf{x}})\} + \sum_{G_i \in \mathcal{G}} \Pr(\widehat{E} = E_i)\mathbb{E}\{s(\widehat{p}_{\mathcal{C}(G_i)}, \widetilde{y}, \widetilde{\mathbf{x}}) - s(p^*, \widetilde{y}, \widetilde{\mathbf{x}})\}}.$$

To prove Theorem 10, it suffices to show that

$$\Pr(\widehat{E} = E^*) \to 1 \tag{8}$$

and

$$\frac{\Pr(\widehat{E} = E_i)\mathbb{E}\{s(\widehat{p}_{\mathcal{C}(G_i)}, \widetilde{y}, \widetilde{\mathbf{x}}) - s(p^*, \widetilde{y}, \widetilde{\mathbf{x}})\}}{\mathbb{E}\{s(\widehat{p}_{\mathcal{C}(G^*)}, \widetilde{y}, \widetilde{\mathbf{x}}) - s(p^*, \widetilde{y}, \widetilde{\mathbf{x}})\}} \to 0 \tag{9}$$

for all $G_i \in \mathcal{G}$. Equation (8) is a direct consequence of the result in Theorem 1. In the remaining of the proof, we focus on establishing (9).

To show (9), we consider two cases: (1) all learners in $\mathcal{C}(G_i)$ are well-specified and that the model linkages that form a connected component with $\mathcal{L}_1$ are well-specified; (2) there exist at least one learner with misspecified models or misspecified model linkages in $\mathcal{C}(G_i)$.

For case (1), Assumption (A2) in Section B indicates that $\mathbb{E}\{s(\widehat{p}_{\mathcal{C}(G_i)}, \widetilde{y}, \widetilde{\mathbf{x}}) - s(p^*, \widetilde{y}, \widetilde{\mathbf{x}})\} = \Theta(n_{\mathcal{C}(G_i)}^{-1/2})$, and $\mathbb{E}\{s(\widehat{p}_{\mathcal{C}(G^*)}, \widetilde{y}, \widetilde{\mathbf{x}}) - s(p^*, \widetilde{y}, \widetilde{\mathbf{x}})\} = \Theta(n_{\mathcal{C}(G^*)}^{-1/2})$, where $n_{\mathcal{C}(G_i)}$ and $n_{\mathcal{C}(G^*)}$ are the sample sizes of $\mathcal{C}(G_i)$ and $\mathcal{C}(G^*)$, respectively. Combining the above and the result from Theorem 9 that $\Pr(\widehat{E} = E_i) \to 0$ leads to (9).

For case (2), consider a learner $\mathcal{L}_w \in \mathcal{C}(G_i)$ that has misspecified model linkages with some other learners in $\mathcal{C}(G_i)$, denoted as $\mathcal{C}(G_i)_{\text{miss}}$. By definition, $\mathcal{C}(G_i)_{\text{miss}} \subset \mathcal{C}(G_i)$. If $\widehat{E} = E_i$, then G_i must include the path between $\mathcal{L}_w$ and the learners in $\mathcal{C}(G_i)_{\text{miss}}$, which further indicates that

$$\Pr(\widehat{E} = E_i) \leq \Pr\left\{\frac{p(\mathbf{y}^{(w)}, \cup_{\kappa\in\mathcal{C}(G_i)_{\text{miss}}}\mathbf{y}^{(\kappa)}|\mathbf{X}^{(w)}, \cup_{\kappa\in\mathcal{C}(G_i)_{\text{miss}}}\mathbf{X}^{(\kappa)})}{p(\mathbf{y}^{(w)}|\mathbf{X}^{(w)})p(\cup_{\kappa\in\mathcal{C}(G_i)_{\text{miss}}}\mathbf{y}^{(\kappa)}|\cup_{\kappa\in\mathcal{C}(G_i)_{\text{miss}}}\mathbf{X}^{(\kappa)})} > 1\right\}, \tag{10}$$

where the right side of the equation can be interpreted as the probability of including $\mathcal{L}_w$ in Step 2 of Algorithm 1 to build a joint model.

From the proof of Lemma 14, we have

$$\frac{p(\mathbf{y}^{(w)}, \cup_{\kappa\in\mathcal{C}(G_i)_{\text{miss}}}\mathbf{y}^{(\kappa)}|\mathbf{X}^{(w)}, \cup_{\kappa\in\mathcal{C}(G_i)_{\text{miss}}}\mathbf{X}^{(\kappa)})}{p(\mathbf{y}^{(w)}|\mathbf{X}^{(w)})p(\cup_{\kappa\in\mathcal{C}(G_i)_{\text{miss}}}\mathbf{y}^{(\kappa)}|\cup_{\kappa\in\mathcal{C}(G_i)_{\text{miss}}}\mathbf{X}^{(\kappa)})} = \Theta\{n_w^{p'/2}\exp(-n_w C_w)\} \tag{11}$$

for some positive finite constants C_w and p', where n_w is the number of samples in $\mathcal{L}_w$. By an application of the Markov's inequality and (11), we have

$$\Pr\left\{\frac{p(\mathbf{y}^{(w)}, \cup_{\kappa\in\mathcal{C}(G_i)_{\text{miss}}}\mathbf{y}^{(\kappa)}|\mathbf{X}^{(w)}, \cup_{\kappa\in\mathcal{C}(G_i)_{\text{miss}}}\mathbf{X}^{(\kappa)})}{p(\mathbf{y}^{(w)}|\mathbf{X}^{(w)})p(\cup_{\kappa\in\mathcal{C}(G_i)_{\text{miss}}}\mathbf{y}^{(\kappa)}|\cup_{\kappa\in\mathcal{C}(G_i)_{\text{miss}}}\mathbf{X}^{(\kappa)})} > 1\right\} \tag{12}$$

$$\leq \mathbb{E}\left\{\frac{p(\mathbf{y}^{(w)}, \cup_{\kappa\in\mathcal{C}(G_i)_{\text{miss}}}\mathbf{y}^{(\kappa)}|\mathbf{X}^{(w)}, \cup_{\kappa\in\mathcal{C}(G_i)_{\text{miss}}}\mathbf{X}^{(\kappa)})}{p(\mathbf{y}^{(w)}|\mathbf{X}^{(w)})p(\cup_{\kappa\in\mathcal{C}(G_i)_{\text{miss}}}\mathbf{y}^{(\kappa)}|\cup_{\kappa\in\mathcal{C}(G_i)_{\text{miss}}}\mathbf{X}^{(\kappa)})}\right\} \tag{13}$$

$$< n_w^{p'/2}\exp(-\frac{1}{2}n_w C_w), \tag{14}$$

implying $\Pr(\widehat{E} = E_i) < n_w^{p'/2}\exp(-0.5 n_w C_w)$.

Due to the existence of misspecified linkages in $\mathcal{C}(G_i)$, Assumption (A2) is no longer applicable to bound $\mathbb{E}\{s(\widehat{p}_{\mathcal{C}(G_i)}, \widetilde{y}, \widetilde{\mathbf{x}}) - s(p^*, \widetilde{y}, \widetilde{\mathbf{x}})\}$. We instead employ Assumption (A1) to bound $\mathbb{E}\{s(\widehat{p}_{\mathcal{C}(G_i)}, \widetilde{y}, \widetilde{\mathbf{x}}) - s(p^*, \widetilde{y}, \widetilde{\mathbf{x}})\}$ by a finite constant C'. Combining the above, we have

$$\Pr(\widehat{E} = E_i)\mathbb{E}\{s(\widehat{p}_{\mathcal{C}(G_i)}, \widetilde{y}, \widetilde{\mathbf{x}}) - s(p^*, \widetilde{y}, \widetilde{\mathbf{x}})\} < C' n_w^{p'/2}\exp(-\frac{1}{2}n_w C_w). \tag{15}$$

By Assumption (A2) in Section B and (15), we conclude that

$$\frac{\Pr(\widehat{E} = E_i)\mathbb{E}\{s(\widehat{p}_{\mathcal{C}(G_i)}, \widetilde{y}, \widetilde{\mathbf{x}}) - s(p^*, \widetilde{y}, \widetilde{\mathbf{x}})\}}{\mathbb{E}\{s(\widehat{p}_{\mathcal{C}(G^*)}, \widetilde{y}, \widetilde{\mathbf{x}}) - s(p^*, \widetilde{y}, \widetilde{\mathbf{x}})\}} \leq C'' n_w^{(1+p')/2}\exp(-\frac{1}{2}n_w C_w) \to 0 \tag{16}$$

as $n_w \to \infty$, where C'' is some positive finite constant. This concludes the proof.

■

Appendix D. Proof of Lemmas 11–14

D.1 Proof of Lemma 11

Proof Proof of Lemma 11(i): Let $\boldsymbol{\theta} \in \boldsymbol{\Theta} \setminus \boldsymbol{\theta}^*$ and let $Z_i = \log\{p(y_i, \mathbf{x}_i|\boldsymbol{\theta})/p(y_i, \mathbf{x}_i|\boldsymbol{\theta}^*)\}$ be the log ratio between two joint densities evaluated under $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$. Let $\mathbb{E}(Z_i)$ be the expectation of Z_i with respect to the density function $p(y_i, \mathbf{x}_i \mid \boldsymbol{\theta}^*)$. We start with proving the following intermediate result that is helpful for the proof of Lemma 11(i):

$$\lim_{n\to\infty} \Pr\left[\frac{1}{n}\{\ell(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}^*)\} < -c(\boldsymbol{\theta})\right] = 1, \tag{17}$$

where $c(\boldsymbol{\theta})$ is a positive finite number that may depend on $\boldsymbol{\theta}$.

We consider two cases when $\mathbb{E}(Z_i)$ is finite and infinite, respectively. When $\mathbb{E}(Z_i)$ is finite, it follows from the Jensen's inequality that

$$\mathbb{E}(Z_i) < \log \mathbb{E}\{\exp(Z_i)\} = 0. \tag{18}$$

Thus, by the law of large number, we have

$$n^{-1}\sum_{i=1}^{n} Z_i \xrightarrow{p} \mathbb{E}(Z_i) < 0,$$

implying $\Pr\{n^{-1}\sum_{i=1}^{n} Z_i \geq 0.5\mathbb{E}(Z_i)\} \to 0$. That is, $\Pr\{n^{-1}\sum_{i=1}^{n} Z_i \leq 0.5\mathbb{E}(Z_i)\} \to 1$. Pick $c(\boldsymbol{\theta}) = -0.5\mathbb{E}(Z_i)$, and (17) is satisfied. If $\mathbb{E}(Z_i)$ is not finite, then we have $\mathbb{E}(Z_i) = -\infty$. Let $Z_i^* = \max\{Z_i, k\}$, where $k < 0$. Then $\mathbb{E}(|Z_i^*|) < \infty$. By the strong law of large number, we obtain

$$\frac{1}{n}\sum_{i=1}^{n} Z_i^* \xrightarrow{a.s.} \mathbb{E}(Z_i^*). \tag{19}$$

As $k \to -\infty$, by the monotone convergence theorem, we obtain $\mathbb{E}(Z_i^*) \to \mathbb{E}(Z_i) = -\infty$. Moreover, (19) implies

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} Z_i = \lim_{n\to\infty}\sup_{n\geq m} \frac{1}{m}\sum_{i=1}^{m} Z_i \leq \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} Z_i^* = \mathbb{E}(Z_i^*) = -\infty \tag{20}$$

almost surely . Thus, we obtain $\limsup_{n\to\infty} n^{-1}\sum_{i=1}^{n} Z_i = -\infty$ almost surely. In other words, $n^{-1}\sum_{i=1}^{n} Z_i \xrightarrow{a.s.} -\infty$. Any positive finite number $c(\boldsymbol{\theta})$ guarantees that (17) will hold.

We now apply (17) to prove Lemma 11(i) holds in some open balls, where the union of these finite number of open balls covers $\boldsymbol{\Theta} \setminus N(\delta)$. Consider $\boldsymbol{\theta}_j \in \boldsymbol{\Theta}$ and let $N_j(\delta_j) = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_j\|_2 < \delta_j\}$ be a ball of size δ_j centered at $\boldsymbol{\theta}_j$. By the regularity condition (IV) in Appendix A, we have

$$\begin{aligned}\sup_{\boldsymbol{\theta}\in N_j(\delta_j)} \frac{1}{n}\{\ell(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}^*)\} &= \sup_{\boldsymbol{\theta}\in N_j(\delta_j)} \left[\frac{1}{n}\{\ell(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}_j)\} + \frac{1}{n}\{\ell(\boldsymbol{\theta}_j) - \ell(\boldsymbol{\theta}^*)\}\right] \\ &< \frac{1}{n}\sum_{i=1}^{n} H_\delta(y_i, \mathbf{x}_i, \boldsymbol{\theta}_j) + \frac{1}{n}\{\ell(\boldsymbol{\theta}_j) - \ell(\boldsymbol{\theta}^*)\}.\end{aligned} \tag{21}$$

By the weak law of large number, as $n \to \infty$, we have

$$\lim_{\delta \to 0} \frac{1}{n} \sum_{i=1}^{n} H_\delta(y_i, \mathbf{x}_i, \boldsymbol{\theta}_j) \xrightarrow{p} \mathbb{E}\{H_\delta(y_i, \mathbf{x}_i, \boldsymbol{\theta}_j)\} = 0 \tag{22}$$

Applying (17) with $\boldsymbol{\theta} = \boldsymbol{\theta}_j$, we obtain

$$\lim_{n \to \infty} \Pr\left[\frac{1}{n}\{\ell(\boldsymbol{\theta}_j) - \ell(\boldsymbol{\theta}^*)\} < -c_j\right] = 1, \tag{23}$$

where c_j is a positive constant that depends on $\boldsymbol{\theta}_j$. Applying (22) and (23), we get the upper bound in (21), which is shown below:

$$\lim_{n \to \infty} \Pr\left\{\sup_{\boldsymbol{\theta} \in N_j(\delta_j)} \frac{1}{n}\{\ell(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}^*)\} < \frac{1}{n}\{\ell(\boldsymbol{\theta}_j) - \ell(\boldsymbol{\theta}^*)\} + \frac{1}{n}\sum_{i=1}^{n} H_\delta(y_i, \mathbf{x}_i, \boldsymbol{\theta}_j) < -c_j\right\} = 1.$$

Then we get

$$\lim_{n \to \infty} \Pr\left\{\sup_{\boldsymbol{\theta} \in N_j(\delta_j)} \frac{1}{n}\{\ell(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}^*)\} < -c_j\right\} = 1. \tag{24}$$

thereafter.

If $\boldsymbol{\Theta}$ is bounded, the compact set $\boldsymbol{\Theta} \setminus N(\delta)$ can be covered by a finite number of balls, namely $N_1(\delta_1), N_2(\delta_2), \ldots, N_m(\delta_m)$, centered at $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_m$, respectively. Then Lemma 11(i) holds by (24) with

$$k(\delta) = \min\{c_1, c_2, \ldots, c_m\}.$$

If $\boldsymbol{\Theta}$ is unbounded, we apply the same argument to the bounded compact set $\boldsymbol{\Theta} \setminus \{N(\delta) \cup S(\Delta)\}$, where $S(\Delta) = \{\boldsymbol{\theta} : \|\boldsymbol{\theta}\|_2 > \Delta\}$ for sufficiently large Δ, from (V) in Appendix A, we have

$$\sup_{\boldsymbol{\theta} \in S(\Delta)} \frac{1}{n}\{\ell(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}^*)\} < \frac{1}{n}\sum_{i=1}^{n} K_\Delta(y_i, \mathbf{x}_i, \boldsymbol{\theta}^*). \tag{25}$$

If $\mathbb{E}\{K_\Delta(y_i, \mathbf{x}_i, \boldsymbol{\theta}^*)\} > -\infty$, by the weak law of large number, we get $n^{-1}\sum_{i=1}^{n} K_\Delta(y_i, \mathbf{x}_i, \boldsymbol{\theta}^*) \xrightarrow{p} \mathbb{E}\{K_\Delta(y_i, \mathbf{x}_i, \boldsymbol{\theta}^*)\} < 0$, thus we have

$$\lim_{n \to \infty} \Pr\left[\sup_{\boldsymbol{\theta} \in S(\Delta)} \frac{1}{n}\{\ell(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}^*)\} < \mathbb{E}\{K_\Delta(y_i, \mathbf{x}_i, \boldsymbol{\theta}^*)\}\right] = 1. \tag{26}$$

Under this case, Lemma 11(i) holds with $k(\delta) = \min\left[c_1, \ldots, c_m, -\mathbb{E}\{K_\Delta(y_i, \mathbf{x}_i, \boldsymbol{\theta}^*)\}\right]$.

If $\mathbb{E}\{K_\Delta(y_i, \mathbf{x}_i, \boldsymbol{\theta}^*)\} = -\infty$, using the similar argument in (19) and (20), we can derive the conclusion that

$$\frac{1}{n}\sum_{i=1}^{n} K_\Delta(y_i, \mathbf{x}_i, \boldsymbol{\theta}^*) \xrightarrow{a.s.} \mathbb{E}\{K_\Delta(y_i, \mathbf{x}_i, \boldsymbol{\theta}^*)\} = -\infty.$$

Lemma 11(i) still holds with $k(\delta) = \min\{c_1, \ldots, c_m\}$. ■

Proof of Lemma 11(ii): Since δ can be sufficiently small and $\lim_{n\to\infty} \Pr\left\{\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 < \delta\right\} = 1$, we have $n^{-1}|\{(\boldsymbol{I}_n(\widehat{\boldsymbol{\theta}}) - \boldsymbol{I}_n(\boldsymbol{\theta}^*)\}_{i,j}| < n^{-1}\sum_{i=1}^n M_\delta(y_i, \mathbf{x}_i, \boldsymbol{\theta}^*)$ from (IX) in Appendix A, the limiting property

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n M_\delta(y_i, \mathbf{x}_i, \boldsymbol{\theta}^*) = \mathbb{E}\left\{M_\delta(y_i, \mathbf{x}_i, \boldsymbol{\theta}^*)\right\} \to 0$$

and the weak law of large number imply that

$$\lim_{n\to\infty} \frac{1}{n}\left\{\boldsymbol{I}_n(\widehat{\boldsymbol{\theta}})\right\}_{i,j} = \lim_{n\to\infty} \frac{1}{n}\left\{\boldsymbol{I}_n(\boldsymbol{\theta}^*)\right\}_{i,j} \xrightarrow{p} \left\{\boldsymbol{I}(\boldsymbol{\theta}^*)\right\}_{i,j}, \tag{27}$$

Finally, by continuous mapping theorem, we obtain $n^{-p}\det|\boldsymbol{I}_n(\widehat{\boldsymbol{\theta}})| \xrightarrow{p} \det|\boldsymbol{I}(\boldsymbol{\theta}^*)|$.

Proof of Lemma 11(iii): Recall that $\widehat{\boldsymbol{\theta}}$ is the maximum likelihood estimator of $\boldsymbol{\theta}$. Thus, $\nabla\ell(\widehat{\boldsymbol{\theta}}) = \mathbf{0}$. By a second-order Taylor expansion, for $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$, there exists a $t \in [0,1]$ such that

$$\ell(\boldsymbol{\theta}^*) = \ell(\widehat{\boldsymbol{\theta}}) - \frac{1}{2}(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}})^{\mathrm{T}}\left[\boldsymbol{I}_n\{\widehat{\boldsymbol{\theta}} + t(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}})\}\right](\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}). \tag{28}$$

It suffices to show $(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}})^{\mathrm{T}}\left[\boldsymbol{I}_n\{\widehat{\boldsymbol{\theta}} + t(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}})\}\right](\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}) = \Theta(1)$. Since $\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}^*$, by (27) in the proof of Lemma 11(ii), we have $n^{-1}[\boldsymbol{I}_n\{\widehat{\boldsymbol{\theta}} + t(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}})\}]_{i,j} \xrightarrow{p} \{\boldsymbol{I}(\boldsymbol{\theta}^*)\}_{i,j}$. Also, we have $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}^* + \Theta(n^{-1/2})$. Consequently, we obtain

$$\begin{aligned}(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}})^{\mathrm{T}}\left[\boldsymbol{I}_n\{\widehat{\boldsymbol{\theta}} + t(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}})\}\right](\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}) &\xrightarrow{p} \left\{n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\right\}^{\mathrm{T}} \boldsymbol{I}(\boldsymbol{\theta}^*)\left\{n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\right\} \\ &= \Theta(1)\boldsymbol{I}(\boldsymbol{\theta}^*)\Theta(1) \\ &= \Theta(1).\end{aligned} \tag{29}$$

Proof of Lemma 11(iv): Recall that $p(\mathbf{y}, \mathbf{X}|\boldsymbol{\theta}) = \exp\{\ell(\boldsymbol{\theta})\}$. Thus, the marginal likelihood can be written as

$$\begin{aligned}p(\mathbf{y}, \mathbf{X}) &= \int_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \pi(\boldsymbol{\theta})\exp\{\ell(\boldsymbol{\theta})\}d\boldsymbol{\theta} \\ &= p(\mathbf{y}, \mathbf{X}|\widehat{\boldsymbol{\theta}})\int_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \pi(\boldsymbol{\theta})\exp\{\ell(\boldsymbol{\theta}) - \ell(\widehat{\boldsymbol{\theta}})\}d\boldsymbol{\theta} \\ &= p(\mathbf{y}, \mathbf{X}|\widehat{\boldsymbol{\theta}})\int_{\boldsymbol{\theta}\in\boldsymbol{\Theta}\setminus N(\delta)} \pi(\boldsymbol{\theta})\exp\{\ell(\boldsymbol{\theta}) - \ell(\widehat{\boldsymbol{\theta}})\}d\boldsymbol{\theta} + p(\mathbf{y}, \mathbf{X}|\widehat{\boldsymbol{\theta}})\int_{\boldsymbol{\theta}\in N(\delta)} \pi(\boldsymbol{\theta})\exp\{\ell(\boldsymbol{\theta}) - \ell(\widehat{\boldsymbol{\theta}})\}d\boldsymbol{\theta} \\ &:= I_1 + I_2.\end{aligned} \tag{30}$$

It suffices to show that $\left\{p(\mathbf{y}, \mathbf{X}|\widehat{\boldsymbol{\theta}})\widehat{\xi}\right\}^{-1} I_1 \xrightarrow{p} 0$ and $\left\{p(\mathbf{y}, \mathbf{X}|\widehat{\boldsymbol{\theta}})\widehat{\xi}\right\}^{-1} I_2 \xrightarrow{p} (2\pi)^{p/2}\pi(\boldsymbol{\theta}^*)$, respectively.

We first show $\left\{p(\mathbf{y},\mathbf{X}|\widehat{\boldsymbol{\theta}})\widehat{\xi}\right\}^{-1} I_1 \xrightarrow{p} 0$. Note that

$$I_1 = p(\mathbf{y},\mathbf{X}|\widehat{\boldsymbol{\theta}}) \exp\left\{\ell(\boldsymbol{\theta}^*) - \ell(\widehat{\boldsymbol{\theta}})\right\} \int_{\boldsymbol{\theta}\in\boldsymbol{\Theta}\setminus N(\delta)} \pi(\boldsymbol{\theta}) \exp\left\{\ell(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}^*)\right\} d\boldsymbol{\theta},$$

We start the proof by conditioning on the event $\exp\{\ell(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}^*)\} \leq \exp\{-nk(\delta)\}$. Thus,

$$\int_{\boldsymbol{\theta}\in\boldsymbol{\Theta}\setminus N(\delta)} \pi(\boldsymbol{\theta}) \exp\left\{\ell(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}^*)\right\} d\boldsymbol{\theta} \leq \exp\left\{-nk(\delta)\right\} \int_{\boldsymbol{\theta}\in\boldsymbol{\Theta}\setminus N(\delta)} \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \leq \exp\left\{-nk(\delta)\right\}. \tag{31}$$

Multiplying I_1 with $\left\{p(\mathbf{y},\mathbf{X}|\widehat{\boldsymbol{\theta}})\widehat{\xi}\right\}^{-1}$ and by (31), we obtain $\left\{p(\mathbf{y},\mathbf{X}|\widehat{\boldsymbol{\theta}})\widehat{\xi}\right\}^{-1} I_1 \leq \exp\{\ell(\boldsymbol{\theta}^*) - \ell(\widehat{\boldsymbol{\theta}})\}\widehat{\xi}^{-1} \exp\{-nk(\delta)\}$. By Lemma 11(iii), we have $\ell(\boldsymbol{\theta}^*) - \ell(\widehat{\boldsymbol{\theta}}) = \Theta(1)$. Moreover, by Lemma 11(ii) and Slutsky's Theorem, we obtain

$$\begin{aligned}\lim_{n\to\infty} \widehat{\xi}^{-1} \exp\left\{-nk(\delta)\right\} &= \lim_{n\to\infty} n^{-p/2} (\widehat{\xi}^{-2})^{1/2} n^{p/2} \exp\left\{-nk(\delta)\right\} \\ &= (\det|\boldsymbol{I}(\boldsymbol{\theta}^*)|)^{1/2} \lim_{n\to\infty} \left[\exp\left\{-nk(\delta) + p\log(n)/2\right\}\right] \\ &= 0.\end{aligned} \tag{32}$$

Finally, by Lemma 11(i), the event $\exp\{\ell(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}^*)\} \leq \exp\{-nk(\delta)\}$ holds with probability one as $n \to \infty$ for any $\boldsymbol{\theta} \in \boldsymbol{\Theta} \setminus N(\delta)$. Combining the above, we have

$$\left\{p(\mathbf{y},\mathbf{X}|\widehat{\boldsymbol{\theta}})\widehat{\xi}\right\}^{-1} I_1 \xrightarrow{p} 0.$$

Next, we show that $\left\{p(\mathbf{y},\mathbf{X}|\widehat{\boldsymbol{\theta}})\widehat{\xi}\right\}^{-1} I_2 \xrightarrow{p} (2\pi)^{p/2}\pi(\boldsymbol{\theta}^*)$. By a second-order Taylor expansion, we have

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= \ell(\widehat{\boldsymbol{\theta}}) - \frac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^{\mathrm{T}} \boldsymbol{I}_n\{\widehat{\boldsymbol{\theta}} + t(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})\}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}) \\ &= \ell(\widehat{\boldsymbol{\theta}}) - \frac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^{\mathrm{T}} \boldsymbol{I}_n(\widehat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}) - \frac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^{\mathrm{T}} \left[\boldsymbol{I}_n\{\widehat{\boldsymbol{\theta}} + t(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})\} - \boldsymbol{I}_n(\widehat{\boldsymbol{\theta}})\right](\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}).\end{aligned} \tag{33}$$

For notational simplicity, let $R_n = 0.5(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^{\mathrm{T}} \left[\boldsymbol{I}_n\{\widehat{\boldsymbol{\theta}} + t(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})\} - \boldsymbol{I}_n(\widehat{\boldsymbol{\theta}})\right](\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})$. By (33), I_2 can be rewritten as

$$\begin{aligned}I_2 =& p(\mathbf{y},\mathbf{X}|\widehat{\boldsymbol{\theta}}) \int_{\boldsymbol{\theta}\in N(\delta)} \pi(\boldsymbol{\theta}) \exp\left\{\ell(\boldsymbol{\theta}) - \ell(\widehat{\boldsymbol{\theta}})\right\} d\boldsymbol{\theta} \\ =& p(\mathbf{y},\mathbf{X}|\widehat{\boldsymbol{\theta}}) \int_{\boldsymbol{\theta}\in N(\delta)} \pi(\boldsymbol{\theta}) \exp\left[-\frac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^{\mathrm{T}} \boldsymbol{I}_n(\widehat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}) - R_n\right] d\boldsymbol{\theta}.\end{aligned} \tag{34}$$

Under (X) in Appendix A, given any $\epsilon' > 0$, let $\epsilon = 2\epsilon'/\pi(\boldsymbol{\theta}^*)$, since the prior function $\pi(\boldsymbol{\theta})$ is continuous around $\boldsymbol{\theta}^*$, thus, we can choose a δ such that

$$|\pi(\boldsymbol{\theta}) - \pi(\boldsymbol{\theta}^*)| < \epsilon' < \epsilon\pi(\boldsymbol{\theta}^*) \quad \text{if} \quad \boldsymbol{\theta} \in N(\delta). \tag{35}$$

Let

$$I_3 = \int_{\boldsymbol{\theta}\in N(\delta)} \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})^{\mathrm{T}}\boldsymbol{I}_n(\widehat{\boldsymbol{\theta}})(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}) - R_n\right\} d\boldsymbol{\theta}. \tag{36}$$

By (35), we obtain

$$(1-\epsilon)\pi(\boldsymbol{\theta}^*)I_3 < \left\{p(\mathbf{y},\mathbf{X}|\widehat{\boldsymbol{\theta}})\right\}^{-1} I_2 < (1+\epsilon)\pi(\boldsymbol{\theta}^*)I_3. \tag{37}$$

It suffices to obtain lower and upper bounds for $\widehat{\xi}^{-1}I_3$.

We divide the derivation of upper and lower bounds for $\widehat{\xi}^{-1}I_3$ into two parts, we first show that $\widehat{\xi}^{-1}\int_{\boldsymbol{\theta}\in N(\delta)} \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})^{\mathrm{T}}\boldsymbol{I}_n(\widehat{\boldsymbol{\theta}})(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})\right\} d\boldsymbol{\theta} \to (2\pi)^{p/2}$, and prove later that $|R_n| < \epsilon$ for some proper δ.

Recall that $\widehat{\xi}^{-1} = \det|\boldsymbol{I}_n(\widehat{\boldsymbol{\theta}})|^{1/2}$, and by the regularity condition (VI) in Appendix A, $\{\boldsymbol{I}_n(\widehat{\boldsymbol{\theta}})\}^{-1}$ exists since $\boldsymbol{I}_n(\widehat{\boldsymbol{\theta}})$ is non-singular. We have

$$\begin{aligned}
&\widehat{\xi}^{-1}\int_{\boldsymbol{\theta}\in N(\delta)} \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})^{\mathrm{T}}\boldsymbol{I}_n(\widehat{\boldsymbol{\theta}})(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})\right\} d\boldsymbol{\theta} \\
&= \sqrt{(2\pi)^p}\int_{\boldsymbol{\theta}\in N(\delta)} \frac{\exp\left\{-\frac{1}{2}(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})^{\mathrm{T}}\boldsymbol{I}_n(\widehat{\boldsymbol{\theta}})(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})\right\}}{\sqrt{(2\pi)^p \det|\{\boldsymbol{I}_n(\widehat{\boldsymbol{\theta}})\}^{-1}|}} d\boldsymbol{\theta}.
\end{aligned} \tag{38}$$

Because the part inside the integral is the density function of multivariate normal distribution $\boldsymbol{\theta} \sim N(\widehat{\boldsymbol{\theta}}, \{\boldsymbol{I}_n(\widehat{\boldsymbol{\theta}})\}^{-1})$, we have that (38) will be less than

$$\sqrt{(2\pi)^p}\int \frac{\exp\left\{-\frac{1}{2}(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})^{\mathrm{T}}\boldsymbol{I}_n(\widehat{\boldsymbol{\theta}})(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})\right\}}{\sqrt{(2\pi)^p \det|\{\boldsymbol{I}_n(\widehat{\boldsymbol{\theta}})\}^{-1}|}} d\boldsymbol{\theta} = (2\pi)^{\frac{p}{2}}. \tag{39}$$

Moreover, the symmetry property of $\boldsymbol{I}_n(\widehat{\boldsymbol{\theta}})$ indicates that there exists a $p\times p$ matrix $\boldsymbol{V}$ such that $\boldsymbol{I}_n(\widehat{\boldsymbol{\theta}}) = \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}$. Change variable $\boldsymbol{\theta}' = \boldsymbol{V}\boldsymbol{\theta}$, (38) will be greater than

$$\begin{aligned}
&\sqrt{(2\pi)^p}\int_{\boldsymbol{\theta}'\in N'(\delta')} \frac{\exp\left\{-\frac{1}{2}(\boldsymbol{\theta}'-\boldsymbol{V}\widehat{\boldsymbol{\theta}})^{\mathrm{T}}(\boldsymbol{\theta}'-\boldsymbol{V}\widehat{\boldsymbol{\theta}})\right\}}{\sqrt{(2\pi)^p \det|\{\boldsymbol{I}_n(\widehat{\boldsymbol{\theta}})\}^{-1}|}} \det|\boldsymbol{V}^{-1}| d\boldsymbol{\theta}' \\
&= \sqrt{(2\pi)^p}\int_{\boldsymbol{\theta}'\in N'(\delta')} \frac{\exp\left\{-\frac{1}{2}(\boldsymbol{\theta}'-\boldsymbol{V}\widehat{\boldsymbol{\theta}})^{\mathrm{T}}(\boldsymbol{\theta}'-\boldsymbol{V}\widehat{\boldsymbol{\theta}})\right\}}{\sqrt{(2\pi)^p}} d\boldsymbol{\theta}',
\end{aligned} \tag{40}$$

where $N'(\delta') = \{\boldsymbol{\theta}' : \|\boldsymbol{\theta}' - \boldsymbol{V}\boldsymbol{\theta}^*\|_2 < \delta'\}$, and δ' is determined by $\delta' = \min_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2=\delta} \|\boldsymbol{V}\boldsymbol{\theta} - \boldsymbol{V}\boldsymbol{\theta}^*\|_2$. We show that $\delta' = \min_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2=\delta} \|\boldsymbol{V}\boldsymbol{\theta} - \boldsymbol{V}\boldsymbol{\theta}^*\|_2 \to \infty$. Let $\delta(\boldsymbol{\theta}) = \|\boldsymbol{V}\boldsymbol{\theta} - \boldsymbol{V}\boldsymbol{\theta}^*\|_2$ under the condition that $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 = \delta$, if $\delta(\boldsymbol{\theta}) < \infty$, then $\boldsymbol{V}\boldsymbol{\theta} - \boldsymbol{V}\boldsymbol{\theta}^*$ and $\boldsymbol{\theta} - \boldsymbol{\theta}^*$ are both elementwise finite, and their lengths are both finite number p, we conclude from above that $\det|(\boldsymbol{V}\boldsymbol{\theta} - \boldsymbol{V}\boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^{\mathrm{T}}| < \infty$. However, $\det|\boldsymbol{V}| = (\det|\boldsymbol{I}_n(\widehat{\boldsymbol{\theta}})|)^{\frac{1}{2}} \to \infty$ from Lemma 11(ii), thus $\det|(\boldsymbol{V}\boldsymbol{\theta} - \boldsymbol{V}\boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^{\mathrm{T}}| = \det|\boldsymbol{V}| \cdot \det|(\boldsymbol{\theta} - \boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^{\mathrm{T}}| \to \infty$, which is a contradiction. All the above indicates that $\delta(\boldsymbol{\theta}) \to \infty$ for any $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 = \delta$, which

concludes $\delta' = \min_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2=\delta} \|\boldsymbol{V}\boldsymbol{\theta}-\boldsymbol{V}\boldsymbol{\theta}^*\|_2 \to \infty$. Therefore, we have (40) converging to $(2\pi)^{p/2}$ in probability. The conclusion $\widehat{\xi}^{-1}\int_{\boldsymbol{\theta}\in N(\delta)} \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})^{\mathrm{T}}\boldsymbol{I}_n(\widehat{\boldsymbol{\theta}})(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})\right\} d\boldsymbol{\theta} \to (2\pi)^{p/2}$ is then derived.

We next derive the upper bound of $|R_n|$. By the triangle inequality, we have

$$\begin{aligned}\frac{1}{n}|R_n| &= \frac{1}{n}\left|\frac{1}{2}(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})^{\mathrm{T}}[\boldsymbol{I}_n\{\widehat{\boldsymbol{\theta}}+t(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})\}-\boldsymbol{I}_n(\widehat{\boldsymbol{\theta}})](\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})\right| \\ &\le \frac{1}{2n}\left|(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})^{\mathrm{T}}[\boldsymbol{I}_n\{\widehat{\boldsymbol{\theta}}+t(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})\}-\boldsymbol{I}_n(\boldsymbol{\theta}^*)](\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})\right| \\ &\quad + \frac{1}{2n}\left|(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})^{\mathrm{T}}\left\{\boldsymbol{I}_n(\boldsymbol{\theta}^*)-\boldsymbol{I}_n(\widehat{\boldsymbol{\theta}})\right\}(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})\right|.\end{aligned} \tag{41}$$

To further derive the upper bound of (41), consider the length p vector $\boldsymbol{b} = (b_1, b_2, \ldots, b_p)^{\mathrm{T}}$ and $p \times p$ matrix $\mathbf{A}$ with $\{\mathbf{A}\}_{i,j} = a_{i,j}$, $i, j \in \{1, 2, \ldots, p\}$. Let $g(\mathbf{A}, \boldsymbol{b}) = \mathrm{tr}(\boldsymbol{b}^{\mathrm{T}}\mathbf{A}\boldsymbol{b})$, this function can be formalized as $g(\mathbf{A}, \boldsymbol{b}) = \mathrm{tr}(\boldsymbol{b}\boldsymbol{b}^{\mathrm{T}}\mathbf{A}) = \sum_{j=1}^{p}\sum_{i=1}^{p} b_i b_j a_{i,j}$. Let $\boldsymbol{b} = \boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}$ and $\mathbf{A} = \boldsymbol{I}_n\{\widehat{\boldsymbol{\theta}}+t(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})\}-\boldsymbol{I}_n(\boldsymbol{\theta}^*)$, we have $\|\boldsymbol{b}\|_2 = \|(\boldsymbol{\theta}-\boldsymbol{\theta}^*)+(\boldsymbol{\theta}^*-\widehat{\boldsymbol{\theta}})\|_2 \le \|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2+\|\boldsymbol{\theta}^*-\widehat{\boldsymbol{\theta}}\|_2 \le \delta$ in probability, because $\boldsymbol{\theta} \in N(\delta)$ and $\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^* = \Theta(n^{-1/2})$. Thus, $|b_i| \le \delta$ for $i = 1, 2, \ldots, p$. We can get the inequality that $|g(\mathbf{A}, \boldsymbol{b})| \le \delta^2 \sum_{j=1}^{p}\sum_{i=1}^{p}|a_{i,j}|$. From triangle inequality we have $\|\widehat{\boldsymbol{\theta}}+t(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})-\boldsymbol{\theta}^*\|_2 \le t\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2+(1-t)\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_2 \xrightarrow{p} t\delta \le \delta$, thus $\widehat{\boldsymbol{\theta}}+t(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}) \in N(\delta)$ in probability. From (IX) in Appendix A, we have $|a_{i,j}| = |[\boldsymbol{I}_n\{\widehat{\boldsymbol{\theta}}+t(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})\}-\boldsymbol{I}_n(\boldsymbol{\theta}^*)]_{i,j}| \le \sum_{i=1}^{n} M_\delta(y_i, \mathbf{x}_i, \boldsymbol{\theta}^*)$. Thus we have $\boldsymbol{b}^{\mathrm{T}}\mathbf{A}\boldsymbol{b} \le \delta^2 \sum_{j=1}^{p}\sum_{k=1}^{p}\sum_{i=1}^{n} M_\delta(y_i, \mathbf{x}_i, \boldsymbol{\theta}^*)$. Note that this inequality also holds when $\mathbf{A} = \boldsymbol{I}_n(\boldsymbol{\theta}^*)-\boldsymbol{I}_n(\widehat{\boldsymbol{\theta}})$, since $\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^* = \Theta(n^{-1/2})$, which indicates $\widehat{\boldsymbol{\theta}} \in N(\delta)$ almost surely, thus (IX) in Appendix A can be applied and get the same conclusion as well. Based on (41) and the weak law of large number, $n^{-1}|R_n|$ is less than

$$\frac{1}{n}\delta^2 \sum_{j=1}^{p}\sum_{k=1}^{p}\sum_{i=1}^{n} M_\delta(y_i, \mathbf{x}_i, \boldsymbol{\theta}^*) \xrightarrow{p} p^2\delta^2 \mathbb{E}\left\{M_\delta(y_i, \mathbf{x}_i, \boldsymbol{\theta}^*)\right\} \quad \text{when} \quad n \to \infty. \tag{42}$$

Under (IX) in Appendix A, $\lim_{\delta\to 0}\mathbb{E}\left\{M_\delta(y_i, \mathbf{x}_i, \boldsymbol{\theta}^*)\right\} = 0$, given the condition that δ is chosen to make (35) hold, then for any $\epsilon > 0$, if δ is also chosen such that

$$\mathbb{E}\left\{M_\delta(y_i, \mathbf{x}_i, \boldsymbol{\theta}^*)\right\} < \frac{\epsilon}{2np^2\delta^2}.$$

Therefore

$$\lim_{n\to\infty} \Pr\left\{\sup_{\boldsymbol{\theta}\in N(\delta)} |R_n| < \epsilon\right\} = 1. \tag{43}$$

Hence, we get the conclusion

$$\lim_{n\to\infty} \Pr\left\{(2\pi)^{\frac{p}{2}}\exp(-\epsilon) < \widehat{\xi}^{-1} I_3 < (2\pi)^{\frac{p}{2}}\exp(\epsilon)\right\} = 1. \tag{44}$$

Since δ can be chosen so that (44) and (37) both hold for arbitrary small ϵ, we deduce the result

$$\lim_{n\to\infty} \Pr\left[(2\pi)^{\frac{p}{2}}\pi(\boldsymbol{\theta}^*)(1-\epsilon)\exp(-\epsilon) < \left\{p(\mathbf{y}, \mathbf{X}|\widehat{\boldsymbol{\theta}})\widehat{\xi}\right\}^{-1} I_2 < (2\pi)^{\frac{p}{2}}\pi(\boldsymbol{\theta}^*)(1+\epsilon)\exp(\epsilon)\right] = 1,$$

which leads to $\left\{p(\mathbf{y}, \mathbf{X}|\widehat{\boldsymbol{\theta}})\widehat{\xi}\right\}^{-1} I_2 \xrightarrow{p} (2\pi)^{p/2}\pi(\boldsymbol{\theta}^*)$ when $n \to \infty$. Then we get the conclusion

$$\lim_{n\to\infty} \left\{(\mathbf{y}, \mathbf{X}|\widehat{\boldsymbol{\theta}})\widehat{\xi}\right\}^{-1} p(\mathbf{y}, \mathbf{X}) = (2\pi)^{\frac{p}{2}}\pi(\boldsymbol{\theta}^*). \tag{45}$$

D.2 Proof of Lemma 12

Proof We start with the proof of Lemma 12(i). Recall that $\boldsymbol{\Theta}_{\mathcal{C}}$ is the parameter space of $\boldsymbol{\theta}_{\mathcal{C}}$, and $\widetilde{\boldsymbol{\theta}}_1 = (\boldsymbol{\theta}_{1,-\mathcal{S}_1}^{\mathrm{T}}, \boldsymbol{\theta}_{\mathcal{S}_1,\mathcal{S}_2}^{\mathrm{T}})^{\mathrm{T}}$ and $\widetilde{\boldsymbol{\theta}}_2 = (\boldsymbol{\theta}_{2,-\mathcal{S}_2}^{\mathrm{T}}, \boldsymbol{\theta}_{\mathcal{S}_1,\mathcal{S}_2}^{\mathrm{T}})^{\mathrm{T}}$ as the parameters for $\mathcal{L}_1$ and $\mathcal{L}_2$ after incorporating information from the model linkage between the two learners. Let $N_\kappa(\delta_\kappa) = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_\kappa^*\|_2 < \delta_\kappa\}$ be the neighborhood of $\boldsymbol{\theta}_\kappa^*$, and let $\boldsymbol{\Theta}_{1,\delta_1} = \{\boldsymbol{\theta}_{\mathcal{C}} : \widetilde{\boldsymbol{\theta}}_1 \in N_1(\delta_1), \boldsymbol{\theta}_{2,-\mathcal{S}_2} = \boldsymbol{\theta}_{2,-\mathcal{S}_2}^*\}$ and $\boldsymbol{\Theta}_{2,\delta_2} = \{\boldsymbol{\theta}_{\mathcal{C}} : \widetilde{\boldsymbol{\theta}}_2 \in N_2(\delta_2), \boldsymbol{\theta}_{1,-\mathcal{S}_1} = \boldsymbol{\theta}_{1,-\mathcal{S}_1}^*\}$. The parameter space $\boldsymbol{\Theta}_{\mathcal{C}}$ can be divided into $\boldsymbol{\Theta}_{1,\delta_1}$, $\boldsymbol{\Theta}_{2,\delta_2}$, and $\boldsymbol{\Theta}_{\mathcal{C}} \setminus (\boldsymbol{\Theta}_{1,\delta_1} \cup \boldsymbol{\Theta}_{2,\delta_2})$. For a fixed δ, there exist δ_1 and δ_2, such that $\boldsymbol{\Theta}_{\kappa,\delta_\kappa} \in N_{\mathcal{C}}(\delta)$, $\kappa = 1, 2$. Lemma 11(i) indicates that $\sup_{\widetilde{\boldsymbol{\theta}}_\kappa \in \boldsymbol{\Theta}_\kappa \setminus N_\kappa(\delta_\kappa)} n_\kappa^{-1}\left\{\ell_\kappa(\widetilde{\boldsymbol{\theta}}_\kappa) - \ell_\kappa(\boldsymbol{\theta}_\kappa^*)\right\} < -k_\kappa(\delta_\kappa)$ for some positive functions $k_\kappa(\delta_\kappa)$, $\kappa = 1, 2$. Let $k(\delta) = c_1 \cdot k_1(\delta_1) + c_2 \cdot k_2(\delta_2)$. Then,

$$\begin{aligned}
&\sup_{\boldsymbol{\theta}_{\mathcal{C}} \in \boldsymbol{\Theta}_{\mathcal{C}} \setminus N_{\mathcal{C}}(\delta)} n^{-1}\left\{\ell_{\mathcal{C}}(\boldsymbol{\theta}_{\mathcal{C}}) - \ell_{\mathcal{C}}(\boldsymbol{\theta}_{\mathcal{C}}^*)\right\} && (46)\\
&\le \sup_{\boldsymbol{\theta}_{\mathcal{C}} \in \boldsymbol{\Theta}_{\mathcal{C}} \setminus (\boldsymbol{\Theta}_{1,\delta_1} \cup \boldsymbol{\Theta}_{2,\delta_2})} n^{-1}\left\{\ell_{\mathcal{C}}(\boldsymbol{\theta}_{\mathcal{C}}) - \ell_{\mathcal{C}}(\boldsymbol{\theta}_{\mathcal{C}}^*)\right\} && (47)\\
&\le \sup_{\widetilde{\boldsymbol{\theta}}_1 \in \boldsymbol{\Theta}_1 \setminus N_1(\delta_1)} n^{-1}\left\{\ell_1(\widetilde{\boldsymbol{\theta}}_1) - \ell_1(\boldsymbol{\theta}_1^*)\right\} + \sup_{\widetilde{\boldsymbol{\theta}}_2 \in \boldsymbol{\Theta}_2 \setminus N_2(\delta_2)} n^{-1}\left\{\ell_2(\widetilde{\boldsymbol{\theta}}_2) - \ell_2(\boldsymbol{\theta}_2^*)\right\} && (48)\\
&= c_1 \sup_{\widetilde{\boldsymbol{\theta}}_1 \in \boldsymbol{\Theta}_1 \setminus N_1(\delta_1)} n_1^{-1}\left\{\ell_1(\widetilde{\boldsymbol{\theta}}_1) - \ell_1(\boldsymbol{\theta}_1^*)\right\} + c_2 \sup_{\widetilde{\boldsymbol{\theta}}_2 \in \boldsymbol{\Theta}_2 \setminus N_2(\delta_2)} n_2^{-1}\left\{\ell_2(\widetilde{\boldsymbol{\theta}}_2) - \ell_2(\boldsymbol{\theta}_2^*)\right\} && (49)\\
&\le c_1 \times \{-k_1(\delta_1)\} + c_2 \times \{-k_2(\delta_2)\} && (50)\\
&= -k(\delta), && (51)
\end{aligned}$$

where (48) holds by the fact that $\ell_{\mathcal{C}}(\boldsymbol{\theta}_{\mathcal{C}}) - \ell_{\mathcal{C}}(\boldsymbol{\theta}_{\mathcal{C}}^*) = \{\ell_1(\widetilde{\boldsymbol{\theta}}_1) - \ell_1(\boldsymbol{\theta}_1^*)\} + \{\ell_2(\widetilde{\boldsymbol{\theta}}_2) - \ell_2(\boldsymbol{\theta}_2^*)\}$ and (50) holds by applications of Lemma 11(i). Therefore, Lemma 12(i) is proved.

For Lemma 12(ii), we prove that $n^{-1}\boldsymbol{I}_n(\widehat{\boldsymbol{\theta}}_{\mathcal{C}})$ is positive definite by showing that $n^{-1}\boldsymbol{I}_n(\widehat{\boldsymbol{\theta}}_{\mathcal{C}})$ can be rewritten as a sum of two matrices, namely $n^{-1}\boldsymbol{I}_n(\widehat{\boldsymbol{\theta}}_{\mathcal{C}}) = \boldsymbol{M}_1 + \boldsymbol{M}_2$, where $\boldsymbol{M}_1$, $\boldsymbol{M}_2$ are positive definite matrices. The proof can then be concluded by the fact that all elements in $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ are bounded by some finite constants.

Recall that $\boldsymbol{\theta}_{\mathcal{C}} = (\boldsymbol{\theta}_{1,-\mathcal{S}_1}^{\mathrm{T}}, \boldsymbol{\theta}_{\mathcal{S}_1,\mathcal{S}_2}^{\mathrm{T}}, \boldsymbol{\theta}_{2,-\mathcal{S}_2}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{p_{\mathcal{C}}}$. Let $\widehat{\boldsymbol{\theta}}_{\mathcal{C},\kappa}$ be the MLE of $\widetilde{\boldsymbol{\theta}}_\kappa$ obtained by maximizing $\ell_{\mathcal{C}}$, $\kappa = 1, 2$. Let λ_κ be the smallest eigenvalue in $n_\kappa^{-1}\boldsymbol{I}_{n_\kappa}^{(\kappa)}(\widehat{\boldsymbol{\theta}}_{\mathcal{C},\kappa})$. By Condition VI in Appendix A (VI, the positive definite property of $\boldsymbol{I}_{n_\kappa}^{(\kappa)}(\widehat{\boldsymbol{\theta}}_{\mathcal{C},\kappa})$ indicates that $\lambda_\kappa > 0$, $\kappa = 1, 2$. In the following proof, we will focus on constructing $\boldsymbol{M}_1$. Construction of $\boldsymbol{M}_2$ is similar as $\boldsymbol{M}_1$ and is omitted.

We now define the elements in $\boldsymbol{M}_1$. For $i, j \le p_1$, let

$$\{\boldsymbol{M}_1\}_{i,j} = c_1 \frac{1}{n_1}\{\boldsymbol{I}_{n_1}^{(1)}(\widehat{\boldsymbol{\theta}}_{\mathcal{C},1})\}_{i,j} - \frac{c_1}{2}\lambda_1 \tag{52}$$

for $i = j$ and $i \leq p_1 - p_s$, and let

$$\{\boldsymbol{M}_1\}_{i,j} = c_1 \frac{1}{n_1} \{\boldsymbol{I}_{n_1}^{(1)}(\widehat{\boldsymbol{\theta}}_{\mathcal{C},1})\}_{i,j} \tag{53}$$

for the other elements.

When $p_1 + 1 \leq i \leq p_1 + p_2 - p_s$ or $p_1 + 1 \leq j \leq p_1 + p_2 - p_s$, all the elements are zeros except when $i = j$, we set

$$\{\boldsymbol{M}_1\}_{i,j} = \frac{c_2}{2}\lambda_2. \tag{54}$$

By construction, $\boldsymbol{M}_1$ is a 2×2 block diagonal matrix, where each block matrix is positive definite. The upper left diagonal block is the difference between $c_1 n_1^{-1}\{\boldsymbol{I}_{n_1}^{(1)}(\widehat{\boldsymbol{\theta}}_{\mathcal{C},1})\}$ and a diagonal matrix, with the first $p_1 - p_s$ diagonal entries are set to equal $0.5c_1\lambda_1$. The bottom right block matrix is a diagonal matrix with all diagonal entries equaling $0.5c_2\lambda_2$. Thus $\boldsymbol{M}_1$ is positive definite. The matrix $\boldsymbol{M}_2$ is constructed in a similar fashion and can be shown to be positive definite.

We now proceed to prove the limiting property of $n_1^{-1}\boldsymbol{I}_{n_1}^{(1)}(\widehat{\boldsymbol{\theta}}_{\mathcal{C},1})$. We have

$$\frac{1}{n_1}|\{\boldsymbol{I}_{n_1}^{(1)}(\widehat{\boldsymbol{\theta}}_{\mathcal{C},1}) - \boldsymbol{I}_{n_1}^{(1)}(\boldsymbol{\theta}_1^*)\}_{i,j}| < \frac{1}{n_1}\sum_{i=1}^{n_1} M_{\delta_1}(y_i^{(1)}, \mathbf{x}_i^{(1)}, \boldsymbol{\theta}_1^*) \tag{55}$$

from (IX) in Appendix A. Moreover, the limiting property

$$\lim_{n_1\to\infty} \frac{1}{n_1}\sum_{i=1}^{n_1} M_{\delta_1}(y_i^{(1)}, \mathbf{x}_i^{(1)}, \boldsymbol{\theta}_1^*) = \mathbb{E}\left\{M_{\delta_1}(y_i^{(1)}, \mathbf{x}_i^{(1)}, \boldsymbol{\theta}_1^*)\right\} \to 0 \tag{56}$$

implies that

$$\lim_{n_1\to\infty} \frac{1}{n_1}\left\{\boldsymbol{I}_{n_1}^{(1)}(\widehat{\boldsymbol{\theta}}_{\mathcal{C},1})\right\}_{i,j} = \lim_{n_1\to\infty} \frac{1}{n_1}\left\{\boldsymbol{I}_{n_1}^{(1)}(\boldsymbol{\theta}_1^*)\right\}_{i,j} = \left\{\boldsymbol{I}^{(1)}(\boldsymbol{\theta}_1^*)\right\}_{i,j}. \tag{57}$$

Equations (52)–(57) imply that $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ are positive definite and elementwise finite. Combining this with the continuous mapping theorem, we obtain the conclusion that $\det|\boldsymbol{M}_1 + \boldsymbol{M}_2| = \det|n^{-1}\boldsymbol{I}_n(\widehat{\boldsymbol{\theta}}_{\mathcal{C}})| = \Theta(1)$, which concludes Lemma 12(ii).

Lemma 12(iii) is implied by Lemma 11(iii). The proof of Lemma 12(iv) is similar to the proof of Lemma 11(iv), and is omitted.

■

D.3 Proof of Lemma 13

Proof Recall from Lemma 11(iv) and Lemma 12(iv) that $(\widehat{\xi}_1^2)^{-1} = \det|\boldsymbol{I}_{n_1}^{(1)}(\widehat{\boldsymbol{\theta}}_1)|$, $(\widehat{\xi}_2^2)^{-1} = \det|\boldsymbol{I}_{n_2}^{(2)}(\widehat{\boldsymbol{\theta}}_2)|$, and $(\widehat{\xi}_{\mathcal{C}}^2)^{-1} = \det|\boldsymbol{I}_n(\widehat{\boldsymbol{\theta}}_{\mathcal{C}})|$ with $n = n_1 + n_2$. Let $h_1(\mathbf{X}^{(1)})$ and $h_2(\mathbf{X}^{(2)})$ be the density function of $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$, respectively. Since the covariates between two learners

are independent, as $m \to \infty$, we have

$$\begin{aligned}\frac{p(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}|\mathbf{X}^{(1)}, \mathbf{X}^{(2)})}{p(\mathbf{y}^{(1)}|\mathbf{X}^{(1)})p(\mathbf{y}^{(2)}|\mathbf{X}^{(2)})} &= \frac{p(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \mathbf{X}^{(1)}, \mathbf{X}^{(2)})/h_1(\mathbf{X}^{(1)})h_2(\mathbf{X}^{(2)})}{p(\mathbf{y}^{(1)}, \mathbf{X}^{(1)})/h_1(\mathbf{X}^{(1)}) \times p(\mathbf{y}^{(2)}, \mathbf{X}^{(2)})/h_2(\mathbf{X}^{(2)})} \\ &= \frac{p(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \mathbf{X}^{(1)}, \mathbf{X}^{(2)})}{p(\mathbf{y}^{(1)}, \mathbf{X}^{(1)})p(\mathbf{y}^{(2)}, \mathbf{X}^{(2)})} \\ &= \Theta\left\{\frac{p^{(\mathcal{C})}_{\boldsymbol{\theta}_\mathcal{C}}(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \mathbf{X}^{(1)}, \mathbf{X}^{(2)})|\widehat{\boldsymbol{\theta}}_\mathcal{C})}{p^{(1)}_{\widetilde{\boldsymbol{\theta}}_1}(\mathbf{y}^{(1)}, \mathbf{X}^{(1)}|\widehat{\boldsymbol{\theta}}_1)p^{(2)}_{\widetilde{\boldsymbol{\theta}}_2}(\mathbf{y}^{(2)}, \mathbf{X}^{(2)}|\widehat{\boldsymbol{\theta}}_2)} \cdot \frac{\widehat{\xi}_\mathcal{C}}{\widehat{\xi}_1\widehat{\xi}_2}\right\}, \end{aligned} \tag{58}$$

where the third equality holds by an application of Lemma 11(iv) and Lemma 12(iv).

Since the model linkages are well-specified, the joint density can be factored as

$$p^{(\mathcal{C})}_{\boldsymbol{\theta}_\mathcal{C}}(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \mathbf{X}^{(1)}, \mathbf{X}^{(2)}|\boldsymbol{\theta}^*_\mathcal{C}) = p^{(1)}_{\widetilde{\boldsymbol{\theta}}_1}(\mathbf{y}^{(1)}, \mathbf{X}^{(1)}|\boldsymbol{\theta}^*_1)p^{(2)}_{\widetilde{\boldsymbol{\theta}}_2}(\mathbf{y}^{(2)}, \mathbf{X}^{(2)}|\boldsymbol{\theta}^*_2).$$

Thus, we have

$$\begin{aligned}&\frac{p^{(\mathcal{C})}_{\boldsymbol{\theta}_\mathcal{C}}(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \mathbf{X}^{(1)}, \mathbf{X}^{(2)}|\widehat{\boldsymbol{\theta}}_\mathcal{C})}{p^{(1)}_{\widetilde{\boldsymbol{\theta}}_1}(\mathbf{y}^{(1)}, \mathbf{X}^{(1)}|\widehat{\boldsymbol{\theta}}_1)p^{(2)}_{\widetilde{\boldsymbol{\theta}}_2}(\mathbf{y}^{(2)}, \mathbf{X}^{(2)}|\widehat{\boldsymbol{\theta}}_2)} \\ &= \frac{p^{(\mathcal{C})}_{\boldsymbol{\theta}_\mathcal{C}}(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \mathbf{X}^{(1)}, \mathbf{X}^{(2)}|\widehat{\boldsymbol{\theta}}_\mathcal{C})\left\{p^{(\mathcal{C})}_{\boldsymbol{\theta}_\mathcal{C}}(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \mathbf{X}^{(1)}, \mathbf{X}^{(2)}|\boldsymbol{\theta}^*_\mathcal{C})\right\}^{-1}}{p^{(1)}_{\widetilde{\boldsymbol{\theta}}_1}(\mathbf{y}^{(1)}, \mathbf{X}^{(1)}|\widehat{\boldsymbol{\theta}}_1)\left\{p^{(1)}_{\widetilde{\boldsymbol{\theta}}_1}(\mathbf{y}^{(1)}, \mathbf{X}^{(1)}|\boldsymbol{\theta}^*_1)\right\}^{-1} p^{(2)}_{\widetilde{\boldsymbol{\theta}}_2}(\mathbf{y}^{(2)}, \mathbf{X}^{(2)}|\widehat{\boldsymbol{\theta}}_2)\left\{p^{(2)}_{\widetilde{\boldsymbol{\theta}}_2}(\mathbf{y}^{(2)}, \mathbf{X}^{(2)}|\boldsymbol{\theta}^*_2)\right\}^{-1}} \\ &= \Theta(1), \end{aligned} \tag{59}$$

where the second equality holds by Lemma 11(iii).

Also, Lemma 11(ii) and Lemma 12(ii) imply that

$$\frac{\widehat{\xi}_\mathcal{C}}{\widehat{\xi}_1\widehat{\xi}_2} = \frac{\widehat{\xi}_\mathcal{C}(n_1+n_2)^{\frac{p_1+p_2-p_s}{2}}}{\widehat{\xi}_1 n_1^{\frac{p_1}{2}} \cdot \widehat{\xi}_2 n_2^{\frac{p_2}{2}}} \cdot \frac{n_1^{\frac{p_1}{2}} n_2^{\frac{p_2}{2}}}{(n_1+n_2)^{\frac{p_1+p_2-p_s}{2}}} \xrightarrow{p} \Theta(1) \cdot \frac{n_1^{\frac{p_1}{2}} n_2^{\frac{p_2}{2}}}{(n_1+n_2)^{\frac{p_1+p_2-p_s}{2}}} = \Theta(m^{\frac{p_s}{2}}). \tag{60}$$

Substituting (59) and (60) into (58) concludes the proof. ■

D.4 Proof of Lemma 14

Proof From the proof of Lemma 13, we have

$$\frac{p(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}|\mathbf{X}^{(1)}, \mathbf{X}^{(2)})}{p(\mathbf{y}^{(1)}|\mathbf{X}^{(1)})p(\mathbf{y}^{(2)}|\mathbf{X}^{(2)})} = \frac{p(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \mathbf{X}^{(1)}, \mathbf{X}^{(2)})}{p(\mathbf{y}^{(1)}, \mathbf{X}^{(1)})p(\mathbf{y}^{(2)}, \mathbf{X}^{(2)})},$$

and it remains to show that

$$\frac{p(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \mathbf{X}^{(1)}, \mathbf{X}^{(2)})}{p(\mathbf{y}^{(1)}, \mathbf{X}^{(1)})p(\mathbf{y}^{(2)}, \mathbf{X}^{(2)})} \xrightarrow{p} 0. \tag{61}$$

By definition, $p(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \mathbf{X}^{(1)}, \mathbf{X}^{(2)}) = \int_{\mathbf{\Theta}_\mathcal{C}} p^{(1)}_{\widetilde{\boldsymbol{\theta}}_1}(\mathbf{y}^{(1)}, \mathbf{X}^{(1)}|\widetilde{\boldsymbol{\theta}}_1) p^{(2)}_{\widetilde{\boldsymbol{\theta}}_2}(\mathbf{y}^{(2)}, \mathbf{X}^{(2)}|\widetilde{\boldsymbol{\theta}}_2)\pi(\boldsymbol{\theta}_\mathcal{C})d\boldsymbol{\theta}_\mathcal{C}$. Choose a $\delta > 0$ such that $N_1(\delta)$ and $N_2(\delta)$ are non-overlapping neighborhoods of $\boldsymbol{\theta}_1^*$ and $\boldsymbol{\theta}_2^*$, respectively. We split $p(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \mathbf{X}^{(1)}, \mathbf{X}^{(2)})$ into three integrals, I_1, I_2, and I_3, taken on sets $\mathbf{\Theta}_{1,\delta}$, $\mathbf{\Theta}_{2,\delta}$, and $\mathbf{\Theta}_\mathcal{C} \setminus (\mathbf{\Theta}_{1,\delta} \cup \mathbf{\Theta}_{2,\delta})$, where $\mathbf{\Theta}_{1,\delta} = \{\boldsymbol{\theta}_\mathcal{C} : \widetilde{\boldsymbol{\theta}}_1 \in N_1(\delta)\}$ and $\mathbf{\Theta}_{2,\delta} = \{\boldsymbol{\theta}_\mathcal{C} : \widetilde{\boldsymbol{\theta}}_2 \in N_2(\delta)\}$. Note that $\widetilde{\boldsymbol{\theta}}_1 \subseteq \boldsymbol{\theta}_\mathcal{C}$ and $\widetilde{\boldsymbol{\theta}}_2 \subseteq \boldsymbol{\theta}_\mathcal{C}$

For the first integral, we have

$$\begin{aligned} I_1 &= \int_{\mathbf{\Theta}_{1,\delta}} p^{(1)}_{\widetilde{\boldsymbol{\theta}}_1}(\mathbf{y}^{(1)}, \mathbf{X}^{(1)}|\widetilde{\boldsymbol{\theta}}_1) p^{(2)}_{\widetilde{\boldsymbol{\theta}}_2}(\mathbf{y}^{(2)}, \mathbf{X}^{(2)}|\widetilde{\boldsymbol{\theta}}_2)\pi(\boldsymbol{\theta}_\mathcal{C})d\boldsymbol{\theta}_\mathcal{C} \\ &= p^{(2)}_{\widetilde{\boldsymbol{\theta}}_2}(\mathbf{y}^{(2)}, \mathbf{X}^{(2)}|\widehat{\boldsymbol{\theta}}_2)\widehat{\xi}_2 \exp\left\{\ell_2(\boldsymbol{\theta}_2^*) - \ell_2(\widehat{\boldsymbol{\theta}}_2)\right\} \\ &\times \int_{\mathbf{\Theta}_{1,\delta}} \widehat{\xi}_2^{-1} \exp\left\{\ell_2(\widetilde{\boldsymbol{\theta}}_2) - \ell_2(\boldsymbol{\theta}_2^*)\right\} p^{(1)}_{\widetilde{\boldsymbol{\theta}}_1}(\mathbf{y}^{(1)}, \mathbf{X}^{(1)}|\widetilde{\boldsymbol{\theta}}_1)\pi(\boldsymbol{\theta}_\mathcal{C})d\boldsymbol{\theta}_\mathcal{C}. \end{aligned} \tag{62}$$

Since $\widetilde{\boldsymbol{\theta}}_2 \notin N_1(\delta)$ in (62), according to Lemma 11(i), the integral on the right hand side in (62) is less than

$$\begin{aligned} &\int_{\mathbf{\Theta}_{1,\delta}} \widehat{\xi}_2^{-1} \exp\left\{\ell_2(\widetilde{\boldsymbol{\theta}}_2) - \ell_2(\boldsymbol{\theta}_2^*)\right\} p^{(1)}_{\widetilde{\boldsymbol{\theta}}_1}(\mathbf{y}^{(1)}, \mathbf{X}^{(1)}|\widetilde{\boldsymbol{\theta}}_1)\pi(\boldsymbol{\theta}_\mathcal{C})d\boldsymbol{\theta}_\mathcal{C} \\ &\le \widehat{\xi}_2^{-1} \exp\{-n_2 k_2(\delta)\} \int_{\mathbf{\Theta}_{1,\delta}} p^{(1)}_{\widetilde{\boldsymbol{\theta}}_1}(\mathbf{y}^{(1)}, \mathbf{X}^{(1)}|\widetilde{\boldsymbol{\theta}}_1)\pi(\boldsymbol{\theta}_\mathcal{C})d\boldsymbol{\theta}_\mathcal{C} \\ &\le \widehat{\xi}_2^{-1} \exp\{-n_2 k_2(\delta)\} \int_{\mathbf{\Theta}_1} p^{(1)}_{\widetilde{\boldsymbol{\theta}}_1}(\mathbf{y}^{(1)}, \mathbf{X}^{(1)}|\widetilde{\boldsymbol{\theta}}_1)\pi(\widetilde{\boldsymbol{\theta}}_1)d\widetilde{\boldsymbol{\theta}}_1 \\ &= \{n_2^{p_2}\widehat{\xi}_2^2\}^{-\frac{1}{2}} n_2^{\frac{p_2}{2}} \exp\{-n_2 k_2(\delta)\} p(\mathbf{y}^{(1)}, \mathbf{X}^{(1)}) \end{aligned} \tag{63}$$

with probability tending to 1 as $n_2 \to \infty$. From Lemmas 11(ii)–(iv), as $n_2 \to \infty$, we have

$$\begin{aligned} &\{n_2^{p_2}\widehat{\xi}_2^2\}^{-\frac{1}{2}} \xrightarrow{p} (\det|\boldsymbol{I}^{(2)}(\boldsymbol{\theta}_2^*)|)^{\frac{1}{2}}; \\ &\exp\left\{\ell_2(\boldsymbol{\theta}_2^*) - \ell_2(\widehat{\boldsymbol{\theta}}_2)\right\} \to \Theta(1); \\ &\{p^{(2)}_{\widetilde{\boldsymbol{\theta}}_2}(\mathbf{y}^{(2)}, \mathbf{X}^{(2)}|\widehat{\boldsymbol{\theta}}_2)\widehat{\xi}_2\}^{-1} p(\mathbf{y}^{(2)}, \mathbf{X}^{(2)}) \xrightarrow{p} (2\pi)^{\frac{p_2}{2}}\pi(\boldsymbol{\theta}_2^*). \end{aligned} \tag{64}$$

It follows that

$$\frac{I_1}{p(\mathbf{y}^{(1)}, \mathbf{X}^{(1)})p(\mathbf{y}^{(2)}, \mathbf{X}^{(2)})} = \Theta(n_2^{\frac{p_2}{2}} \exp\{-n_2 k_2(\delta)\}) \xrightarrow{p} 0. \tag{65}$$

Using a similar argument for I_2, as $n_1 \to \infty$, we have

$$\frac{I_2}{p(\mathbf{y}^{(1)}, \mathbf{X}^{(1)})p(\mathbf{y}^{(2)}, \mathbf{X}^{(2)})} = \Theta(n_1^{\frac{p_1}{2}} \exp\{-n_1 k_1(\delta)\}) \xrightarrow{p} 0. \tag{66}$$

For the integral I_3, we apply a similar argument as in the proof of I_1 and I_2. Specifically,

$$\begin{aligned}
I_3 &= \int_{\mathbf{\Theta}_{\mathcal{C}} \setminus \{\mathbf{\Theta}_{1,\delta} \cup \mathbf{\Theta}_{2,\delta}\}} p^{(1)}_{\widetilde{\boldsymbol{\theta}}_1}(\mathbf{y}^{(1)}, \mathbf{X}^{(1)} | \widetilde{\boldsymbol{\theta}}_1) p^{(2)}_{\widetilde{\boldsymbol{\theta}}_2}(\mathbf{y}^{(2)}, \mathbf{X}^{(2)} | \widetilde{\boldsymbol{\theta}}_2) \pi(\boldsymbol{\theta}_{\mathcal{C}}) d\boldsymbol{\theta}_{\mathcal{C}} \\
&= p^{(1)}_{\widetilde{\boldsymbol{\theta}}_1}(\mathbf{y}^{(1)}, \mathbf{X}^{(1)} | \widehat{\boldsymbol{\theta}}_1) \widehat{\xi}_1 p^{(2)}_{\widetilde{\boldsymbol{\theta}}_2}(\mathbf{y}^{(2)}, \mathbf{X}^{(2)} | \widehat{\boldsymbol{\theta}}_2) \widehat{\xi}_2 \exp\left\{\ell_1(\boldsymbol{\theta}_1^*) - \ell_1(\widehat{\boldsymbol{\theta}}_1)\right\} \exp\left\{\ell_2(\boldsymbol{\theta}_2^*) - \ell_2(\widehat{\boldsymbol{\theta}}_2)\right\} \\
&\times \int_{\mathbf{\Theta}_{\mathcal{C}} \setminus \{\mathbf{\Theta}_{1,\delta} \cup \mathbf{\Theta}_{2,\delta}\}} \widehat{\xi}_1^{-1} \widehat{\xi}_2^{-1} \exp\left\{\ell_1(\widetilde{\boldsymbol{\theta}}_1) - \ell_1(\boldsymbol{\theta}_1^*)\right\} \exp\left\{\ell_2(\widetilde{\boldsymbol{\theta}}_2) - \ell_2(\boldsymbol{\theta}_2^*)\right\} \pi(\boldsymbol{\theta}_{\mathcal{C}}) d\boldsymbol{\theta}_{\mathcal{C}}, \\
&\leq p^{(1)}_{\widetilde{\boldsymbol{\theta}}_1}(\mathbf{y}^{(1)}, \mathbf{X}^{(1)} | \widehat{\boldsymbol{\theta}}_1) \widehat{\xi}_1 p^{(2)}_{\widetilde{\boldsymbol{\theta}}_2}(\mathbf{y}^{(2)}, \mathbf{X}^{(2)} | \widehat{\boldsymbol{\theta}}_2) \widehat{\xi}_2 \exp\left\{\ell_1(\boldsymbol{\theta}_1^*) - \ell_1(\widehat{\boldsymbol{\theta}}_1)\right\} \exp\left\{\ell_2(\boldsymbol{\theta}_2^*) - \ell_2(\widehat{\boldsymbol{\theta}}_2)\right\} \\
&\times \widehat{\xi}_1^{-1} \widehat{\xi}_2^{-1} \int_{\mathbf{\Theta}_{\mathcal{C}}} \exp\left\{\ell_1(\widetilde{\boldsymbol{\theta}}_1) - \ell_1(\boldsymbol{\theta}_1^*)\right\} \exp\left\{\ell_2(\widetilde{\boldsymbol{\theta}}_2) - \ell_2(\boldsymbol{\theta}_2^*)\right\} \pi(\boldsymbol{\theta}_{\mathcal{C}}) d\boldsymbol{\theta}_{\mathcal{C}}.
\end{aligned}$$

Since the region $\mathbf{\Theta}_{\mathcal{C}} \setminus \{\mathbf{\Theta}_{1,\delta} \cup \mathbf{\Theta}_{2,\delta}\}$ contains neither the neighborhood of $\boldsymbol{\theta}_1^*$ nor the neighborhood of $\boldsymbol{\theta}_2^*$, by an application of Lemma 11(i), we have

$$\begin{aligned}
I_3 &\leq p^{(1)}_{\widetilde{\boldsymbol{\theta}}_1}(\mathbf{y}^{(1)}, \mathbf{X}^{(1)} | \widehat{\boldsymbol{\theta}}_1) \widehat{\xi}_1 p^{(2)}_{\widetilde{\boldsymbol{\theta}}_2}(\mathbf{y}^{(2)}, \mathbf{X}^{(2)} | \widehat{\boldsymbol{\theta}}_2) \widehat{\xi}_2 \exp\left\{\ell_1(\boldsymbol{\theta}_1^*) - \ell_1(\widehat{\boldsymbol{\theta}}_1)\right\} \exp\left\{\ell_2(\boldsymbol{\theta}_2^*) - \ell_2(\widehat{\boldsymbol{\theta}}_2)\right\} \\
&\times \widehat{\xi}_1^{-1} \widehat{\xi}_2^{-1} \exp\left\{-n_1 k_1(\delta)\right\} \exp\left\{-n_2 k_2(\delta)\right\}.
\end{aligned}$$

Using an argument similar to that of I_1 and Lemmas 12(ii)–(iv), we have

$$\frac{I_3}{p(\mathbf{y}^{(1)}, \mathbf{X}^{(1)}) p(\mathbf{y}^{(2)}, \mathbf{X}^{(2)})} = \Theta(n_1^{\frac{p_1}{2}} n_2^{\frac{p_2}{2}} \exp\{-n_1 k_1(\delta) - n_2 k_2(\delta)\}) \xrightarrow{p} 0. \tag{67}$$

Combining the above, we have

$$\frac{p(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \mathbf{X}^{(1)}, \mathbf{X}^{(2)})}{p(\mathbf{y}^{(1)}, \mathbf{X}^{(1)}) p(\mathbf{y}^{(2)}, \mathbf{X}^{(2)})} = \frac{I_1 + I_2 + I_3}{p(\mathbf{y}^{(1)}, \mathbf{X}^{(1)}) p(\mathbf{y}^{(2)}, \mathbf{X}^{(2)})} \xrightarrow{p} 0. \tag{68}$$

This concludes the proof of Lemma 14. ■

References

Rehan Akbani, Kwok shing Ng, Henrica Werner, Maria Shahmoradgoli, Fan Zhang, Zhenlin Ju, Wenbin Liu, Ji-Yeon Yang, Kosuke Yoshihara, Jun Li, Shiyun Ling, Elena Seviour, Prahlad Ram, John Minna, Lixia Diao, Pan Tong, John Heymach, Steven Hill, Frank Dondelinger, and Gordon Mills. A pan-cancer proteomic perspective on The Cancer Genome Atlas. *Nature Communication*, 5(1):1–15, 05 2014. doi: 10.17863/CAM.25567.

Arjun Nitin Bhagoji, Supriyo Chakraborty, Prateek Mittal, and Seraphin Calo. Analyzing federated learning through an adversarial lens. In *International Conference on Machine Learning*, pages 634–643. PMLR, 2019.

Marta Blangiardo, Anna Hansell, and Sylvia Richardson. A Bayesian model of time activity data to investigate health effect of air pollution in time series studies. *Atmospheric Environment*, 45(2):379–386, 2011.

Sabine Blaschke, Claudia Müller, Jasmina Markovic-Lipkovski, Sabine Puch, Nicolai Miosge, Volker Becker, Gerhard Mueller, and Gerd Klein. Expression of cadherin-8 in renal cell carcinoma and fetal kidney. *International Journal of Cancer*, 101(4):327–334, 10 2002. doi: 10.1002/ijc.10623.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

Guangzhen Cao and Ya Qiu Jin. A hybrid algorithm of the BP-ANN/GA for classification of urban terrain surfaces with fused data of Landsat ETM+ and ERS-2 SAR. *International Journal of Remote Sensing*, 28(2):293–305, 2007.

GC Cawley and NLC Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11(Jul):2079–2107, July 2010.

Fengju Chen, Yiqun Zhang, Yasin Şenbabaoğlu, Giovanni Ciriello, Lixing Yang, Ed Reznik, Brian Shuch, Goran Micevic, Guillermo De Velasco, and Eve Shinbrot. Multilevel genomics-based taxonomy of renal cell carcinoma. *Cell Reports*, 14(10):2476–2489, 2016.

Kevin A. Clarke. The phantom menace: omitted variable bias in econometric research. *Conflict Management and Peace Science*, 22(4):341–352, 2005. doi: 10.1080/07388940500339183.

Li Da Xu, Wu He, and Shancang Li. Internet of things in industries: a survey. *IEEE Transactions on Industrial Informatics*, 10(4):2233–2243, 2014.

Patrick Danaher, Pei Wang, and Daniela M Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B*, 76(2):373–397, 2014.

A Philip Dawid and Monica Musio. Bayesian model selection based on proper scoring rules. *Bayesian Analysis*, 10(2):479–499, 2015.

Enmao Diao, Jie Ding, and Vahid Tarokh. Restricted recurrent neural networks. *2019 IEEE International Conference on Big Data*, 2019.

Enmao Diao, Jie Ding, and Vahid Tarokh. Multimodal controller for generative models. *arXiv preprint arXiv:2002.02572*, 2020a.

Enmao Diao, Jie Ding, and Vahid Tarokh. HeteroFL: Computation and communication efficient federated learning for heterogeneous clients. In *International Conference on Learning Representations*, 2020b.

Jie Ding, Vahid Tarokh, and Yuhong Yang. Model selection techniques–an overview. *IEEE Signal Processing Magazine* , 35(6):16–34, 2018.

Ian Domowitz and Halbert White. Misspecified models with dependent observations. *Journal of Econometrics*, 20(1):35–58, October 1982.

Chao Du, Chu-Lan Michael Kao, and S. C. Kou. Stepwise signal extraction via marginal likelihood. *Journal of the American Statistical Association*, 111(513):314–330, 2016.

C Duckworth, L Zhang, Steven Carroll, S Ethier, and H Cheung. Overexpression of GAB2 in ovarian cancer cells promotes tumor growth and angiogenesis by upregulating chemokine expression. *Oncogene*, 35(31):4036–4047, 2016.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Pi Guo, Jianjun Zhang, Li Wang, Shaoyi Yang, Ganfeng Luo, Changyu Deng, Ye Wen, and Qingying Zhang. Monitoring seasonal influenza epidemics by using internet search data with an ensemble penalized regression model. *Scientific Reports*, 7(1):46469, 2017.

Pierre E Jacob, Lawrence M Murray, Chris C Holmes, and Christian P Robert. Better together? statistical learning in models made of modules. *arXiv preprint arXiv:1708.08719*, 2017.

Shane T Jensen, Guang Chen, and Christian J Stoeckert Jr. Bayesian variable selection and data integration for biological regulatory networks. *The Annals of Applied Statistics*, 1(2):612–633, 2007.

Malhar S Jere, Tyler Farnan, and Farinaz Koushanfar. A taxonomy of attacks on federated learning. *IEEE Security & Privacy*, 19(2):20–28, 2020.

Michael I Jordan, Jason D Lee, and Yun Yang. Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 114(526):668–681, 2019.

Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

Fanjie Kong, Xiaobing Li, Hong Wang, Dengfeng Xie, Xiang Li, and Yunxiao Bai. Land cover classification based on fused data from GF-1 and MODIS NDVI time series. *Remote Sensing*, 8(9):741, 2016.

Jason D Lee, Qihang Lin, Tengyu Ma, and Tianbao Yang. Distributed stochastic variance reduced gradient methods by sampling extra data with replacement. *Journal of Machine Learning Research*, 18(1):4404–4446, 2017a.

Jason D Lee, Qiang Liu, Yuekai Sun, and Jonathan E Taylor. Communication-efficient sparse regression. *Journal of Machine Learning Research*, 18(1):115–144, 2017b.

Boyue Li, Shicong Cen, Yuxin Chen, and Yuejie Chi. Communication-efficient distributed optimization in networks with gradient tracking. *arXiv preprint arXiv:1909.05844*, 2019.

Quefeng Li and Lexin Li. Integrative linear discriminant analysis with guaranteed error rate improvement. *Biometrika*, 105(4):917–930, 2018.

Dungang Liu, Regina Y Liu, and Minge Xie. Multivariate meta-analysis of heterogeneous studies using only summary statistics: efficiency and robustness. *Journal of the American Statistical Association*, 110(509):326–340, 2015.

Fei Liu, MJ Bayarri, and JO Berger. Modularization in Bayesian analysis, with emphasis on analysis of computer models. *Bayesian Analysis*, 4(1):119–150, 2009.

David Lunn, Nicky Best, David Spiegelhalter, Gordon Graham, and Beat Neuenschwander. Combining MCMC with 'sequential' PKPD modeling. *Journal of Pharmacokinetics and Pharmacodynamics*, 36(1):19, 2009.

David J Lunn, Andrew Thomas, Nicky Best, and David Spiegelhalter. WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10(4):325–337, 2000.

Jing Ma and George Michailidis. Joint structural estimation of multiple graphical models. *Journal of Machine Learning Research*, 17(1):5777–5824, 2016.

Arnab Maity, Anirban Bhattacharya, Bani Mallick, and Veerabhadran Baladandayuthapani. Bayesian data integration and variable selection for pan-cancer survival prediction using protein expression data. *Biometrics*, 76(1):316–325, 08 2019. doi: 10.1111/biom.13132.

Olvi L. Mangasarian, W. Nick Street, and William H. Wolberg. Breast cancer diagnosis and prognosis via linear programming. *Operations Research*, 43(4):570–577, 1995. doi: 10.1287/opre.43.4.570.

Delphine Maucort-Boulch, Silvia Franceschi, and Martyn Plummer. International correlation between human papillomavirus prevalence and cervical cancer incidence. *Cancer Epidemiology and Prevention Biomarkers*, 17(3):717–720, 2008.

H Brendan McMahan, Eider Moore, Daniel Ramage, and Seth Hampson. Communication-efficient learning of deep networks from decentralized data. *arXiv preprint arXiv:1602.05629*, 2016.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *International Conference on Machine Learning*, pages 689–696, 2011.

Kiona Ogle, Jarrett Barber, and Karla Sartor. Feedback and modularization in a Bayesian meta–analysis of tree traits affecting forest dynamics. *Bayesian Analysis*, 8(1):133–168, 2013.

Matthew Parry. Linear scoring rules for probabilistic binary classification. *Electronic Journal of Statistics*, 10(1):1596–1607, 2016.

Matthew Parry, A Philip Dawid, and Steffen Lauritzen. Proper local scoring rules. *Annals of Statistics*, 40(1):561–592, 2012.

Martyn Plummer. Cuts in Bayesian graphical models. *Statistics and Computing*, 25(1): 37–43, Jan 2015. ISSN 1573-1375. doi: 10.1007/s11222-014-9503-z.

Stephane Shao, Pierre E Jacob, Jie Ding, and Vahid Tarokh. Bayesian model comparison with the hyvärinen score: Computation and consistency. *Journal of the American Statistical Association*, 114(528):1826–1837, 2019.

Jieli Shen, Regina Y Liu, and Min-Ge Xie. *i*Fusion: Individualized fusion learning. *Journal of the American Statistical Association*, 115(531):1251–1267, 2019.

Wei Shi, Qing Ling, Kun Yuan, Gang Wu, and Wotao Yin. On the linear convergence of the ADMM in decentralized consensus optimization. *IEEE Transactions on Signal Processing*, 62(7):1750–1761, 2014.

Reza Shokri and Vitaly Shmatikov. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1310–1321. ACM, 2015.

MC Simmonds and JPT Higgins. Covariate heterogeneity in meta-analysis: criteria for deciding between meta-regression and individual patient data. *Statistics in Medicine*, 26 (15):2982–2999, 2007.

Uthayasankar Sivarajah, Muhammad Mustafa Kamal, Zahir Irani, and Vishanth Weerakkody. Critical analysis of big data challenges and analytical methods. *Journal of Business Research*, 70:263–286, 2017.

Lu Tang and Peter XK Song. Fused lasso approach in regression coefficients clustering: learning parameter heterogeneity in data integration. *Journal of Machine Learning Research*, 17(1):3915–3937, 2016.

Lu Tang, Ling Zhou, and Peter XK Song. Fusion learning algorithm to combine partially heterogeneous cox models. *Computational Statistics*, 34(1):395–414, 2019.

Edward F Vonesh, Tom Greene, and Mark D Schluchter. Shared parameter models for the joint analysis of longitudinal data and event times. *Statistics in Medicine*, 25(1):143–163, 2006.

AM Walker. On the asymptotic behaviour of posterior distributions. *Journal of the Royal Statistical Society: Series B*, 31(1):80–88, 1969.

Xinran Wang, Yu Xiang, Jun Gao, and Jie Ding. Information laundering for model privacy. In *International Conference on Learning Representations*, 2021.

Xiaoquan Wen and Matthew Stephens. Bayesian methods for genetic association analysis with heterogeneous subgroups: from meta-analyses to gene-environment interactions. *The Annals of Applied Statistics*, 8(1):176, 2014.

Xun Xian, Xinran Wang, Jie Ding, and Reza Ghanadan. Assisted learning: A framework for multi-organization learning. In *Advances in Neural Information Processing Systems*, pages 14580–14591, 2020.

Jin-Jun Xiao and Zhi-Quan Luo. Universal decentralized detection in a bandwidth-constrained sensor network. *IEEE Transactions on Signal Processing*, 53(8):2617–2624, 2005.

Shi Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-Chun Woo. Convolutional LSTM network: a machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems*, pages 802–810, 2015.

Shihao Yang, Mauricio Santillana, and SC Kou. Accurate estimation of influenza epidemics using Google search data via ARGO. *Proceedings of the National Academy of Sciences of the United States of America*, 112(47):14473–14478, 2015.

Chenglong Ye, Jie Ding, and Reza Ghanadan. Meta clustering for collaborative learning. *arXiv preprint arXiv:2006.00082*, 2020.

Qing Zhou, Di Li, Soummya Kar, Lauren M Huie, H Vincent Poor, and Shuguang Cui. Learning-based distributed detection-estimation in sensor networks with unknown sensor defects. *IEEE Transactions on Signal Processing*, 65(1):130–145, 2016.

Richard Zuech, Taghi M Khoshgoftaar, and Randall Wald. Intrusion detection and big heterogeneous data: a survey. *Journal of Big Data*, 2(1):3, 2015.