Federated Learning for Sparse Bayesian Models with Applications to Electronic Health Records and Genomics

Brian Kidd[1], Kunbo Wang[2], Yanxun Xu[2], Yang Ni[1,†]
[1]Department of Statistics, Texas A&M University, College Station, Texas 77843, USA
[2]Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD 21218, USA
[†]Correspondence: yni@stat.tamu.edu

Federated learning is becoming increasingly more popular as the concern of privacy breaches rises across disciplines including the biological and biomedical fields. The main idea is to train models locally on each server using data that are only available to that server and aggregate the model (not data) information at the global level. While federated learning has made significant advancements for machine learning methods such as deep neural networks, to the best of our knowledge, its development in sparse Bayesian models is still lacking. Sparse Bayesian models are highly interpretable with natural uncertain quantification, a desirable property for many scientific problems. However, without a federated learning algorithm, their applicability to sensitive biological/biomedical data from multiple sources is limited. Therefore, to fill this gap in the literature, we propose a new Bayesian federated learning framework that is capable of pooling information from different data sources without breaching privacy. The proposed method is conceptually simple to understand and implement, accommodates sampling heterogeneity (i.e., non-iid observations) across data sources, and allows for principled uncertainty quantification. We illustrate the proposed framework with three concrete sparse Bayesian models, namely, sparse regression, Markov random field, and directed graphical models. The application of these three models is demonstrated through three real data examples including a multi-hospital COVID-19 study, breast cancer protein-protein interaction networks, and gene regulatory networks.

Keywords: Causal discovery; Distributed computation; Graphical models; Privacy; Sparse regression.

1. Introduction

Sparse models such as sparse regression and graphical models have been extensively studied and find numerous applications in biological and biomedical sciences such as biomarker identification for electronic health records data[1] and reverse-engineering gene regulatory networks for genomic data.[2] Sparse Bayesian models not only provide point estimation but also naturally quantify the estimation uncertainty, which facilitates interpretation especially for models that have moderate to large numbers of parameters. Shrinkage and variable selection priors have been developed for this purpose including the horseshoe prior,[3] the Bayesian lasso,[4] the spike-and-slab prior,[5] and the thresholding prior.[6] In this article, we study the sparse Bayesian

© 2022 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

models under the federated learning setting where data are distributed across multiple local sources (called local servers hereafter) and the goal is to perform global inference that pools information from local servers without breaching the local data privacy. Typical application includes privacy-preserved analyses of electronic health records data across multiple hospitals or medical centers where data may be limited in size in each site (hence independent analysis in each site would lack statistical power) but cannot be shared across sites due to the sensitivity of protected health information.

Federated learning is an emerging area and finds many applications especially in health.[7–9] Essentially, the idea is to train models locally on each server using data that are only available to that server and then send model information (instead of any private data) to a central server for aggregation. The central server subsequently sends the aggregated model information back to local servers. The exchange of information between the central and local servers can be an iterative process depending on the communication cost and the design of the federated learning algorithm.[10] Another interesting line of federated learning research considers heterogeneous scenarios where the data distributions may be different across local servers.[11] In general, methods developed for federated learning could be applied for distributing computational tasks on massive data, but the opposite is not true as distributed computing does not generally preserve privacy of the local data.

This article particularly focuses on Bayesian methods, which typically provide more natural uncertainty quantification than the frequentist counterpart. Bayesian inference, however, often requires running a long Markov chain Monte Carlo (MCMC) algorithm to achieve practical convergence, which can be time-consuming. Therefore, Bayesian distributed computing has been developed to improve the computational efficiency through parallelization. One such line of research is so-called consensus Monte Carlo for which MCMC is run on each local server without communication among the servers and the Monte Carlo samples are only aggregated at the end.[12–18] Intuitively, the idea is to divide the posterior into separate sub-posteriors to be computed on each local server; then the research question becomes how to effectively combine these local chains into a single posterior. However, in many situations (e.g. the local data being heterogeneous or highly non-Gaussian), consensus Monte Carlo may not have good empirical performance,[19] but work is continuing to attempt to overcome these issues.[20] There are also methods that run multiple chains with somewhat frequent communication during the course of MCMC.[19,21,22] These methods are potentially useful for federated learning but require carefully crafted MCMC methods to protect privacy. Another line of research involves using a distributed version of stochastic gradients within Langevin Dynamics (i.e., Langevin Monte Carlo),[23] which subsamples each local dataset for gradient approximation. In fact, multiple methods have applied the distributed stochastic gradients idea to federated learning.[24,25] However, gradient does not exist for discrete parameters such as variable selection indicators in sparse models, which is the main focus of this article. Lastly, Bayesian neural networks have seen recent advancements in the federated learning setting where the aggregation is achieved through fitting parametric or nonparametric models to local network parameters.[26,27] While useful for neural networks, it is not straightforward to extend their methods to other models including sparse models such as sparse regression and graphical models.

Our paper demonstrates how basic MCMC algorithms can be used within the federated learning setting by reformulating the model and adding an explicit layer for pooling the local models. As the order of MCMC updating steps can be interchanged, the communication between local servers and the global server can be reduced by running multiple local steps per global aggregation. Through multiple sparse models and real data examples, we show the simplicity and broad applicability of the proposed method.

2. Method

2.1. Overall Framework

We first introduce the proposed federated learning framework for Bayesian models. Later, we will provide several concrete examples illustrating the application of the proposed framework to three specific sparse Bayesian models – sparse regression, Markov random field, and directed graphical models.

Let $\boldsymbol{D}_1, \boldsymbol{D}_2, \ldots, \boldsymbol{D}_M$ denote M datasets and let $\boldsymbol{D} = \{\boldsymbol{D}_1, \ldots, \boldsymbol{D}_M\}$ be the collection of all datasets. If they are available on the same computing server (i.e., under the non-federated learning setting) and if they are independent and identically distributed (iid), then a single probability model can be used to model $\boldsymbol{D}$, $\boldsymbol{D} \sim P(\boldsymbol{D}|\boldsymbol{\theta}) = \prod_{k=1}^{M} P(\boldsymbol{D}_k|\boldsymbol{\theta})$, which is schematically represented by a directed acyclic graph in Figure 1(a). However, this model has two obvious downsides under the federated learning setting: (i) $\boldsymbol{D}_k$ is only available on the local server $k = 1, \ldots, M$ and cannot be shared with other servers due to privacy concerns, etc; and (ii) $\boldsymbol{D}_1, \ldots, \boldsymbol{D}_M$ may not be iid. A naive approach to address these two concerns is to consider M independent probability models (Figure 1(b)), one for each local server, $\boldsymbol{D}_k \sim P(\boldsymbol{D}_k|\boldsymbol{\theta}_k)$. This approach does not provide a joint inference across datasets, which can result in statistically inefficient inference and poor interpretation of model parameters. To provide joint inference while preserving privacy, federated learning approaches have been developed. For example, one may aggregate the estimates of $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M$ using some deterministic function $\boldsymbol{\theta} = f(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M)$ such as average for continuous parameters and majority vote for discrete parameters. Such deterministic approach is often ad hoc (e.g., lack of finite-sample theoretical justification) and generally does not propagate estimation uncertainty from local parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M$ to the global parameter $\boldsymbol{\theta}$. In this article, we will instead consider a probabilistic aggregation approach, which overcomes all the aforementioned limitations. The proposed approach is conceptually simple and natural for Bayesian models. Consider the following hierarchical model, for $k = 1, \ldots, M$,

$$\boldsymbol{D}_k \sim P(\boldsymbol{D}_k|\boldsymbol{\theta}_k), \quad \boldsymbol{\theta}_k \sim P(\boldsymbol{\theta}_k|\boldsymbol{\theta}), \quad \boldsymbol{\theta} \sim P(\boldsymbol{\theta}).$$

Given appropriate choices of $P(\boldsymbol{\theta}_k|\boldsymbol{\theta})$ and $P(\boldsymbol{\theta})$ (to be discussed later), this conceptually simple hierarchical model provides a principled recipe to probabilistically aggregate local information through the posterior distribution $P(\boldsymbol{\theta}|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M) \propto P(\boldsymbol{\theta}) \prod_{k=1}^{M} P(\boldsymbol{\theta}_k|\boldsymbol{\theta})$, which directly provides point and interval estimation of $\boldsymbol{\theta}$ through e.g., the posterior mean and the credible interval. Algorithmically, by exploiting the conditional independence of $\boldsymbol{\theta}_k$ and $\boldsymbol{D}_{-k}$ given $\boldsymbol{\theta}$ (subscript "$-k$" means removing $\boldsymbol{D}_k$ from $\boldsymbol{D}$), the computation is trivially parallelizable at the local level and no data ever need to be passed to the global server, hence preserving privacy;

see Figure 1(c). In Algorithm 1, we outline the federated learning MCMC pseudocode, which highlights the local parallelizability and privacy protection (there is no data sharing, and the shared parameters are not observation-level parameters).

The aggregation via the posterior distribution depends crucially on the choices of the prior distribution of local parameters given the global parameter $P(\boldsymbol{\theta}_k|\boldsymbol{\theta})$ and the prior distribution of the global parameter $P(\boldsymbol{\theta})$. Three properties are deemed desirable: (i) $P(\boldsymbol{\theta}_k|\boldsymbol{\theta})$ should encourage $\boldsymbol{\theta}_k$ to tightly concentrate around $\boldsymbol{\theta}$ so that $\boldsymbol{\theta}$ can be interpreted as a global version of local server-specific parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M$, (ii) $P(\boldsymbol{\theta}_k|\boldsymbol{\theta})$ should also allow occasional deviation of $\boldsymbol{\theta}_k$ from $\boldsymbol{\theta}$ if $\boldsymbol{D}_k$ strongly supports it, which accommodates non-iid scenarios, and (iii) $P(\boldsymbol{\theta})$ should encourage sparsity in $\boldsymbol{\theta}$ for better model interpretability. To make the discussion concrete, we now consider three specific sparse Bayesian models. For ease of exposition, we start with a sparse regression model.

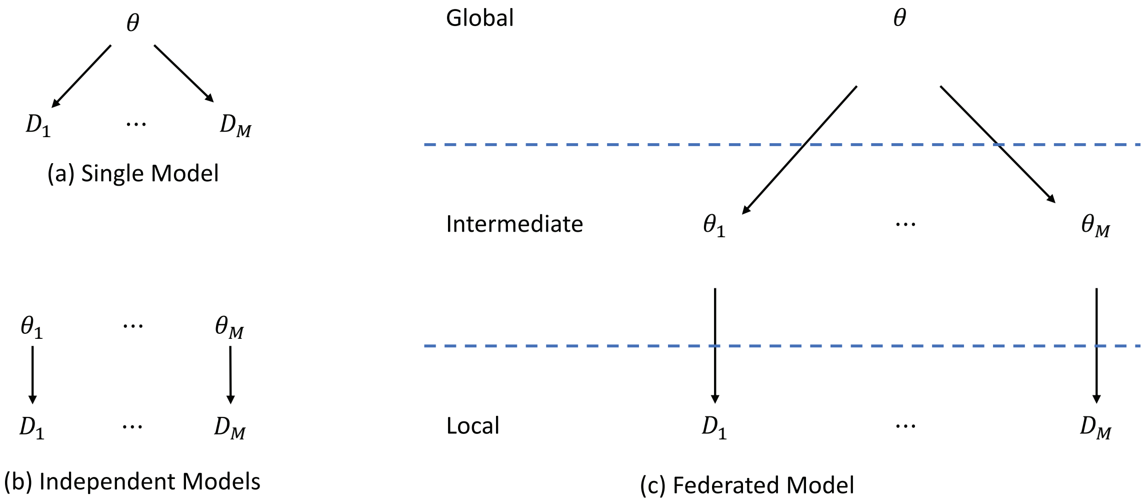


Fig. 1: Illustration of (a) a single model, (b) independent models, and (c) a federated model. The arrows represent the direct dependencies among the variables. The federated model has three levels: global, intermediate, and local. The parameters at the intermediate level are passed from local servers to the global server whereas the data never leave the local servers.

2.2. Example 1: Federated Sparse Regression

2.2.1. Sparse Regression

Let $\boldsymbol{D}_k = (\boldsymbol{X}_{ki}, Y_{ki})_{i=1}^{n_k}$ for $k = 1, \ldots, M$ denote the local server-specific dataset with n_k observations where $\boldsymbol{X}_{ki} = (X_{ki1}, \ldots, X_{kip})^T$ is p-dimensional covariate vector and Y_{ki} is the response variable for $i = 1, \ldots, n_k$. Consider the following server-specific regression model,

$$Y_{ki} = \boldsymbol{X}_{ki}^T \boldsymbol{\theta}_k + \epsilon_{ki}, \tag{1}$$

for $k = 1, \ldots, M$ and $i = 1 \ldots, n_k$, where $\boldsymbol{\theta}_k = (\theta_{k1}, \ldots, \theta_{kp})^T$ is the regression coefficient vector and $\epsilon_{ki} \sim N(0, \sigma_k^2)$ is a normal error term. For simplicity, we do not make joint inference on σ_k^2 as the parameter of interest of a regression model is typically the regression coefficient $\boldsymbol{\theta}_k$; but

Algorithm 1 General Algorithm

```
Input: $\boldsymbol{D}_k$ and hyperparameters
Output: Monte Carlo samples of $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M$ and $\boldsymbol{\theta}$
  Initialize $\boldsymbol{\theta}^{(0)}$ on the global server
  for $t$ in $1, \ldots, T$ do                                                    ▷ MCMC iterator
      parfor $k$ in $1 \ldots, M$ do                                              ▷ Parallel for-loop
          Send the global parameter $\boldsymbol{\theta}^{(t-1)}$ to local server $k$
          Sample $\boldsymbol{\theta}_k^{(t)} | \boldsymbol{D}_k, \boldsymbol{\theta}^{(t-1)} \sim P(\boldsymbol{\theta}_k | \boldsymbol{D}_k, \boldsymbol{\theta}^{(t-1)})$ on local server $k$      ▷ Local Update
          Send $\boldsymbol{\theta}_k^{(t)}$ to the global server
      end parfor
      Sample $\boldsymbol{\theta}^{(t)} \sim P(\boldsymbol{\theta} | \boldsymbol{\theta}_1^{(t)}, \ldots, \boldsymbol{\theta}_M^{(t)})$ on the global server      ▷ Global Aggregation
  end for
```

if desired, our method can be easily extended for joint inference of σ_k^2. In many applications, not all covariates are predictive of the response variable and, correspondingly, $\boldsymbol{\theta}_k$ is assumed to be sparse, i.e., most of the entries $\boldsymbol{\theta}_k$ are zero or very close to zero.

2.2.2. Prior

We now specify the prior distributions $P(\boldsymbol{\theta}_k|\boldsymbol{\theta})$, $P(\boldsymbol{\theta})$, and $P(\sigma_k^2)$. To achieve the fist two desired properties outlined at the end of Section 2.1, we impose an element-wise mean-shifted horseshoe prior for $\boldsymbol{\theta}_k$, which is centered around the global parameter $\boldsymbol{\theta}$,

$$\begin{aligned}\theta_{kj}|\theta_j &\sim N(\theta_j, \lambda_{kj}^2 \tau_j^2),\\ \lambda_{kj}, \tau_j &\sim C^+(0,1),\end{aligned}$$

where $C^+(0,1)$ is the standard half-Cauchy distribution. The mean-zero horseshoe prior[28,29] has been extensively studied in the sparse regression model, which is capable of shrinking small coefficients aggressively towards zero while leaving large coefficients untouched. Our use of mean-shifted horseshoe prior aggressively shrinks local parameter θ_{kj} towards the global parameter θ_j but still allows substantial deviation if data dictates so.

To encourage sparsity, we assume a spike-and-slab prior[5] on the global parameter with a beta-Bernoulli hyperprior,

$$\begin{aligned}\theta_j|\gamma_j &\sim \gamma_j N(0, \eta_j) + (1-\gamma_j) N(0, c_0 \eta_j),\\ \gamma_j &\sim \text{Bernoulli}(\rho), \quad \rho \sim \text{beta}(a_\rho, b_\rho),\end{aligned}$$

where c_0 is fixed small constant (e.g., 0.01) and γ_j is a binary indicator variable, which equals 1 if θ_j is significantly away from 0 and equals 0 if θ_j is so small that it can be safely treated as zero without affecting the model fit. The prior specification is completed with conjugate inverse-gamma priors for variance parameters $\sigma_k^2 \sim IG(a_\sigma, b_\sigma)$ and $\eta_j \sim IG(a_\eta, b_\eta)$.

In summary, the local horseshoe prior shrinks local parameters towards the global parameter (i.e., the aggregation) and the global spike-and-slab prior induces sparsity.

2.2.3. MCMC

We expand the “Local Update” and the “Global Aggregation” steps of Algorithm 1 for sparse regression model in Algorithms 2 and 3, respectively. Note that for the sampling of horseshoe-related parameters, we utilize the parameter expansion technique.[30] Also note that one can opt to run multiple local update steps per each global aggregation due to the standard Markov chain theory; see the for-loop in Algorithm 2.

Algorithm 2 Local Update for Sparse Regression

```
for $\ell$ in $1, \dots, L$ do
    Sample $\nu_{kj} \sim IG(1, 1 + \lambda_{kj}^{-2})$                    ▷ Parameter Expansion[30]
    Sample $\lambda_{kj}^2 \sim IG\big[1, \nu_{kj}^{-1} + (\theta_{kj} - \theta_j)^2/(2\tau_j^2)\big]$
    Sample $\boldsymbol{\theta}_k \sim f(\boldsymbol{\theta}_k) \propto \prod_{i=1}^{n_k} N(Y_{ki}|\boldsymbol{X}_{ki}^T\boldsymbol{\theta}_k, \sigma_k^2) \prod_{j=1}^{p} N(\theta_{kj}|\theta_j, \lambda_{kj}^2\tau_j^2)$
    Sample $\sigma_k^2 \sim IG(a_\sigma + n_k/2, b_\sigma + \sum_{i=1}^{n_k}(Y_{ki} - \boldsymbol{X}_{ki}^T\boldsymbol{\theta}_k)^2/2)$
end for
```

Algorithm 3 Global Aggregation for Sparse Regression

```
Sample $\xi_j \sim IG(1, 1 + \tau_j^{-2})$                    ▷ Parameter Expansion[30]
Sample $\tau_j^2 \sim IG\big[(M+1)/2, \xi_j^{-1} + \sum_{k=1}^{M}(\theta_{kj} - \theta_j)^2/\lambda_{kj}^2\big]$
Sample $\boldsymbol{\theta} \sim f(\boldsymbol{\theta}) \propto \prod_{j=1}^{p}\big[N(\theta_j|0, c_0^{1-\gamma_j}\eta)\prod_{k=1}^{M} N(\theta_{kj}|\theta_j, \lambda_{kj}^2\tau_j^2)\big]$
Sample $\eta \sim IG\big[a_\eta + p/2, b_\eta + \sum_{j=1}^{p}\theta_j^2/c_0^{1-\gamma_j}\big]$
Sample $\gamma_j \sim \text{Bernoulli}(q_j)$ with $q_j = \frac{\rho N(\theta_j|0,\eta)}{\rho N(\theta_j|0,\eta) + (1-\rho)N(\theta_j|0,c_0\eta)}$
Sample $\rho \sim \text{beta}(a_\rho + \sum_{j=1}^{p}\gamma_j, b_\rho + p - \sum_{j=1}^{p}\gamma_j)$
```

2.3. Example 2: Federated Markov Random Field

The sparse regression model in Section 2.2 can be extended to the sparse Gaussian Markov random field model (also known as the Gaussian graphical model), which can also be worked out in a federated learning setting. Let $\boldsymbol{D}_k = (\boldsymbol{Y}_{ki})_{i=1}^{n_k}$ for $k = 1, \dots, M$ where $\boldsymbol{Y}_{ki} = (Y_{ki1}, \dots, Y_{kip})^T$ is a random vector whose conditional independence relationships are of interest. We assume a centered multivariate Gaussian distribution,

$$\boldsymbol{Y}_{ki} \sim N(0, \boldsymbol{\Omega}_k^{-1}), \tag{2}$$

with precision (inverse covariance) matrix $\boldsymbol{\Omega}_k = [\omega_{kjh}]_{j=1,h=1}^{p,p}$. If $\omega_{kjh} = 0$, then Y_{kj} and Y_{kh} are conditionally independent given all the other variables. Often, such conditional independence relationships are represented by an undirected graph/network where nodes represent the random variables and two nodes are connected $j - h$ by an undirected edge if and only if $\omega_{kjh} \neq 0$. Interestingly, Gaussian Markov random field is closely related to sparse regression, which leads to the so-called neighborhood selection method.[31] Note that the joint distribution

(2) implies the conditional distribution of Y_{kij} given all the other variables,

$$Y_{kij} = \boldsymbol{Y}_{ki,-j}^{T}\boldsymbol{\theta}_{kj} + \epsilon_{kij}, \tag{3}$$

with $\boldsymbol{\theta}_{kj} = -\boldsymbol{\Omega}_{k,-j,j}/\omega_{kjj}$ and $\epsilon_{kij} \sim N(0, \omega_{kjj}^{-1})$, which is exactly a regression model with response Y_{kij} and covariates $\boldsymbol{Y}_{ki,-j}$. Therefore, $\omega_{kjh} = 0$ if and only if $\theta_{kjh} = 0$. Consequently, estimating a sparse precision matrix $\boldsymbol{\Omega}_k$ reduces to estimating the set of sparse regression coefficient for p independent regressions. Hence, the proposed federated learning algorithm for sparse regression can be applied in parallel to (3) for $j = 1, \ldots, p$. One caveat is that the neighborhood selection method has no guarantee of the symmetry of $\boldsymbol{\Omega}_k$ but simple post-processing procedures based on union or intersection can be used to obtain a consensus undirected graph.[31]

2.4. Example 3: Federated Directed Graphical Models

Markov random field is useful for investigating symmetric association but cannot be used to identify causal relationships, which are asymmetric (cause and effect are not exchangeable). Directed graphical models[32,33] are popular tools for discovering causality (i.e., generating plausible causal hypotheses in an exploratory fashion). Consider the following structural equation model,[34,35]

$$\boldsymbol{Y}_{ki} = \boldsymbol{Y}_{ki}\boldsymbol{\theta}_k + \boldsymbol{E}_{ki}, \tag{4}$$

where $\boldsymbol{\theta}_k = [\theta_{kjh}]_{j=1,h=1}^{p,p}$ is the causal effect matrix and $\boldsymbol{E}_{ki} = (\epsilon_{ki1}, \ldots, \epsilon_{kip})^T \sim N(0, \boldsymbol{\Sigma}_{ki})$ is the normally-distributed error vector with diagonal covariance $\boldsymbol{\Sigma}_{ki}$. Under the causal Markov assumption,[32,33] we say Y_{kh} is a direct cause of Y_{kj} if $\theta_{kjh} \neq 0$, which can be represented by an arrow $j \leftarrow h$ in a directed graph/network. The error distribution induce a distribution for $\boldsymbol{Y}_{ki}$,

$$\boldsymbol{Y}_{ki} \sim N(0, (\boldsymbol{I} - \boldsymbol{\theta}_k)^{-1}\boldsymbol{\Sigma}_{ki}(\boldsymbol{I} - \boldsymbol{\theta}_k)^{-T}),$$

where $\boldsymbol{I}$ is a $p \times p$ identity matrix. Note that for observational data, the causal relationships may not be identifiable due to Markov equivalence. To ensure identifiability, various methods have been developed. As an example, we take advantage of the non-Gaussianity for causal identifiability.[36] Specifically, we assume each diagonal entry of $\boldsymbol{\Sigma}_{ki}$ to be exponentially distributed, which induces a marginal Laplace distribution for ϵ_{kij} for $j = 1, \ldots, p$. We remark that the popular causal discovery method, Bayesian network, is a special case of the directed graphical model considered here by restricting the graph to be acyclic. Because biological systems tend to have feedback loops, we do not make such restriction. The price to pay is that we lose conjugacy but the proposed federated learning framework is still applicable with a minor tweak: replace the Gibbs sampling of $\boldsymbol{\theta}_k$ in Algorithm 2 by a Metropolis step. Specifically, we propose a new value $\boldsymbol{\theta}_k^\star$ from some proposal density $q(\cdot)$ such as normal, which could depend on the value of $\boldsymbol{\theta}_k$ from the last iteration. Then we accept $\boldsymbol{\theta}_k^\star$ with probability $\min(1, a)$ with

$$a = \frac{q(\boldsymbol{\theta}_k)N(0, (\boldsymbol{I} - \boldsymbol{\theta}_k^\star)^{-1}\boldsymbol{\Sigma}_{ki}(\boldsymbol{I} - \boldsymbol{\theta}_k^\star)^{-T})\prod_{j\neq h} N(\theta_{kjh}^\star|\theta_{jh}, \lambda_{kjh}^2\tau_{jh}^2)}{q(\boldsymbol{\theta}_k^\star)N(0, (\boldsymbol{I} - \boldsymbol{\theta}_k)^{-1}\boldsymbol{\Sigma}_{ki}(\boldsymbol{I} - \boldsymbol{\theta}_k)^{-T})\prod_{j\neq h} N(\theta_{kjh}|\theta_{jh}, \lambda_{kjh}^2\tau_{jh}^2)}.$$

3. Numerical Studies

We demonstrate the proposed methods with three real data examples. Simulation results are provided in the Supplementary Materials `https://www.dropbox.com/s/5cl1ag92otaos54/kidd_supp.pdf?dl=0`.

3.1. Johns Hopkins COVID-19 Data - Federated Sparse Regression

COVID-19 (a coronavirus) has been a recent pandemic receiving a great amount of attention worldwide. We analyze the COVID-19 clinical data electronically recorded in four Johns Hopkins' hospitals (i.e., $M = 4$). Each hospital provides 100-150 patients, leading to a total sample size of 552. Due to the sensitive protected health information, data cannot be easily shared across hospitals for the purpose of statistical analyses but local computation within each hospital is feasible. Therefore, this data provide an excellent opportunity to illustrate the practical utility of the proposed federated learning method.

An important marker for COVID-19 is the arterial oxygen saturation ($\mathrm{S_aO_2}$, our response variable), which, unfortunately, is difficult to measure. Instead, because of its non-invasiveness, the peripheral oxygen saturation ($\mathrm{S_pO_2}$, our main covariate) is often used as a proxy measurement for $\mathrm{S_aO_2}$. We will apply the federated sparse regression model to the Johns Hopkins data to examine the association between $\mathrm{S_pO_2}$ and $\mathrm{S_aO_2}$ in COVID-19 patients while adjusting for eight variables commonly collected at doctors visits: temperature in Celsius (Temp_C), mean arterial pressure (MAP), gender, age, and race, hemoglobin count (HGB), bilirubin levels, and creatinine levels. Dummy variable coding is used for gender (Male) and race (Race_b (Black), Race_h (Hispanic), Race_a (Asian)).

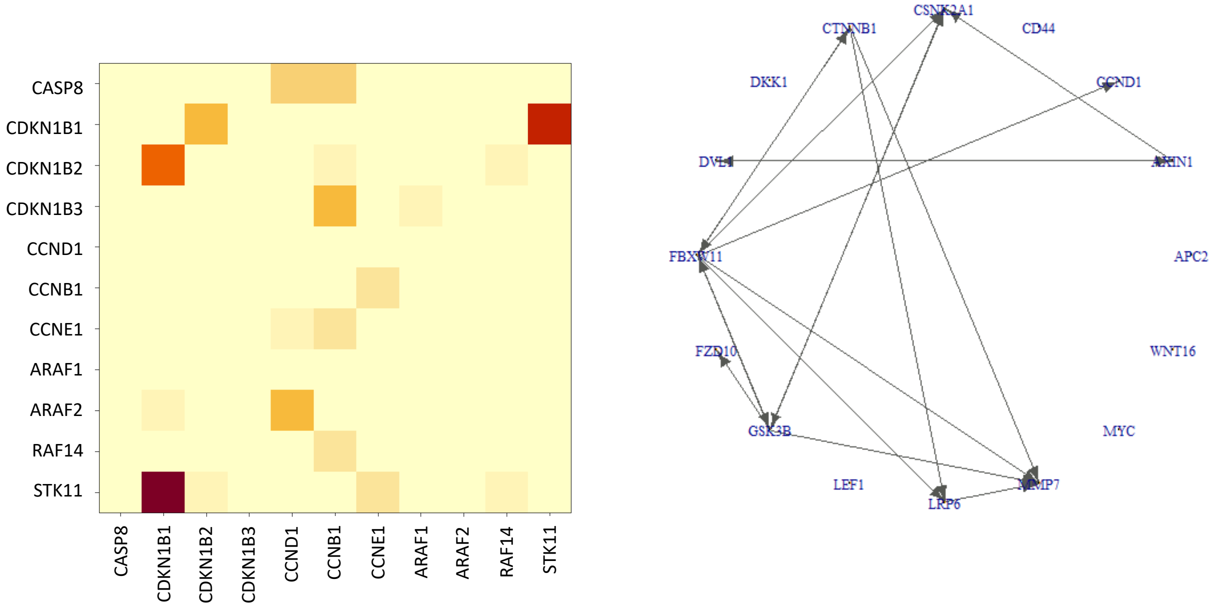
We run the federated learning algorithm with $T = 1000$ global aggregation and $L = 100$ local updates per each global aggregation. We report the posterior mean of $\boldsymbol{\theta}$ and posterior inclusion probability (PIP) in Table 1. PIP is defined as the posterior mean of γ_j and a large value indicates high significance of X_j. As expected, $\mathrm{S_pO_2}$ is the most significant predictor of $\mathrm{S_aO_2}$ with PIP=0.777, which demonstrates that the proposed federated sparse regression has the potential to identify important variable by pooling information from multiple local servers without breaching privacy.

3.2. Breast Cancer Protein-Protein Interaction Networks - Federated Markov Random Field

Breast cancer is one of the most prevalent types of cancer, affecting over 5% of women in the United States throughout their lives. Since cancer is a genetic disease, modern treatment of breast cancer relies heavily on the fundamental understanding of genetic architecture of breast cancer tissues. Therefore, it is crucial to understand genetic networks at different levels such as gene and protein levels. In this section, to demonstrate federated Markov random field, we consider a Reverse Phase Protein Array data from the The Cancer Proteome Atlas.[37] Protein expression data are extracted from 7 sites with over 50 observations (the biggest site has 149 observations). We focus our analysis on $p = 11$ breast cancer-related proteins.[38] We reported PIP of all pairs of proteins in Figure 2(a) with darker color corresponding to higher PIP. The most significant interaction, STK11 and CDKN1B, is biologically plausible as STK11 is known

Table 1: COVID-19 data

Covariate	$\boldsymbol{\theta}$	PIP
S_pO_2	0.673	0.777
Age	-0.002	0.032
MAP	0.006	0.048
Temp_C	0.152	0.306
HGB	0.029	0.123
Bilirubin	-0.013	0.157
Creatinine	-0.017	0.090
Male	0.013	0.142
Race_b	-0.003	0.216
Race_h	-0.007	0.178
Race_a	-0.037	0.217



(a) Protein-protein interaction network

(b) Gene regulatory network

Fig. 2: Breast cancer genetic networks.

to phosphorylate CDKN1A[39] and CDKN1A and CDKN1B belong to the same family of CDK inhibitor. The next most significant association is between CDKN1B1 and CDKN1B2, which is also not surprising given they are the variants of the same protein CDKN1B. As we noted before, Figure 2(a) is not symmetric due to the artifact of neighborhood selection.[31] It can be symmetrized if desired by taking the maximum or minimum of PIP for each pair of pairs.

3.3. Breast Cancer Gene Regulatory Networks - Federated Directed Graphical Models

To demonstrate federated directed graphical models, we consider the breast cancer gene expression data obtained from the Genomic Data Commons project of the National Cancer Institute.[40] The consortium hosts data generated from over 45 different sites. We restrict our analysis to the 10 sites with over 50 observations, leading to total sample size 901 with the largest site having 227 observations and two others having over 100 observations each.

We focus our analysis on the WNT/β-catenin signaling pathway known to be critical for breast cancer development.[41] Particularly, $p = 16$ genes emphasized in the recent review paper[41] are considered. We present the estimated gene regulatory network in Figure 2(b) where Bayesian false discovery rate control[42] is used to threshold the PIP to obtain the sparse network.

Some feedback loops are interesting. For example, DVL1 is known to inactivate AXIN1, but our analysis also shows a direct feedback from AXIN1 to DVL1, which requires further experimental validation. In addition, the regulatory relationship from CTNNB1 to MMP7 also matches the existing biological knowledge that MMP7 is a downstream effect of CTNNB1.[43]

4. Discussion

We have brought sparse Bayesian models into the realm of federated learning. The proposed method is conceptually simple and allows for data heterogeneity (i.e., non-iid observations) and proper uncertainty quantification. By switching the MCMC order and updating local models multiple times between global server updates, we manage to the limit the communication cost while maintaining theoretical convergence (as MCMC eventually converges regardless of the update order). Through real data examples, we show the applicability of the proposed method for sparse regression, Markov random field, directed graphical models.

There are several future directions. First, we have only considered linear models for both regression and graphical models. Nonlinearity can be incorporated by spline basis expansion.[44] Second, some variables may not be measured in certain sites. By pooling the covariance information together through federated learning, one can impute these missing variables under the missing at random assumption. Preliminary simulations (not shown) support this idea. Third, we have focused on the federated learning setting where there is a central server. It would be interesting to extend our current approach to the scenarios where there is no central server and only pairwise direct communication among local serves is possible.

References

1. Y. Ni, F. C. Stingo, M. J. Ha, R. Akbani and V. Baladandayuthapani, Bayesian hierarchical varying-sparsity regression models with application to cancer proteogenomics, Journal of the American Statistical Association 114, 48 (2019).
2. J. Choi, R. Chapkin and Y. Ni, Bayesian causal structural learning with zero-inflated Poisson Bayesian networks, Advances in Neural Information Processing Systems 33, 5887 (2020).
3. C. M. Carvalho, N. G. Polson and J. G. Scott, The horseshoe estimator for sparse signals, Biometrika 97, 465 (2010).
4. T. Park and G. Casella, The bayesian lasso, Journal of the American Statistical Association 103, 681 (2008).

5. E. I. George and R. E. McCulloch, Variable selection via Gibbs sampling, Journal of the American Statistical Association 88, 881 (1993).
6. Y. Ni, F. C. Stingo and V. Baladandayuthapani, Bayesian graphical regression, Journal of the American Statistical Association 114, 184 (2019).
7. Y. Wu, X. Jiang, J. Kim and L. Ohno-Machado, Grid binary logistic regression (glore): building shared models without sharing data, Journal of the American Medical Informatics Association 19, 758 (2012).
8. Y. Li, X. Jiang, S. Wang, H. Xiong and L. Ohno-Machado, Vertical grid logistic regression (vertigo), Journal of the American Medical Informatics Association 23, 570 (2016).
9. J. Xu, B. S. Glicksberg, C. Su, P. Walker, J. Bian and F. Wang, Federated learning for healthcare informatics, Journal of Healthcare Informatics Research 5, 1 (2021).
10. T. Li, A. K. Sahu, A. Talwalkar and V. Smith, Federated learning: Challenges, methods, and future directions, IEEE Signal Processing Magazine 37, 50 (2020).
11. Y. Laguel, K. Pillutla, J. Malick and Z. Harchaoui, A superquantile approach to federated learning with heterogeneous devices, in 2021 55th Annual Conference on Information Sciences and Systems (CISS), 2021.
12. X. Wang and D. B. Dunson, Parallelizing MCMC via Weierstrass sampler, arXiv preprint arXiv:1312.4605 (2013).
13. S. Srivastava, V. Cevher, Q. Dinh and D. Dunson, WASP: Scalable Bayes via barycenters of subset posteriors, in Artificial Intelligence and Statistics, 2015.
14. S. L. Scott, A. W. Blocker, F. V. Bonassi, H. A. Chipman, E. I. George and R. E. McCulloch, Bayes and big data: The consensus Monte Carlo algorithm, International Journal of Management Science and Engineering Management 11, 78 (2016).
15. C. Nemeth and C. Sherlock, Merging MCMC subposteriors through Gaussian-process approximations, Bayesian Analysis 13, 507 (2018).
16. Y. Ni, Y. Ji and P. Müller, Consensus Monte Carlo for random subsets using shared anchors, Journal of Computational and Graphical Statistics 29, 703 (2020).
17. Y. Ni, D. Jones and Z. Wang, Consensus variational and Monte Carlo algorithms for Bayesian nonparametric clustering, in 2020 IEEE International Conference on Big Data (Big Data), 2020.
18. Y. Ni, P. Müller, M. Diesendruck, S. Williamson, Y. Zhu and Y. Ji, Scalable Bayesian nonparametric clustering and classification, Journal of Computational and Graphical Statistics 29, 53 (2020).
19. L. J. Rendell, A. M. Johansen, A. Lee and N. Whiteley, Global consensus Monte Carlo, Journal of Computational and Graphical Statistics 30, 249 (2020).
20. D. Mesquita, P. Blomstedt and S. Kaski, Embarrassingly parallel MCMC using deep invertible transformations, in Uncertainty in Artificial Intelligence, 2020.
21. A. Chowdhury and C. Jermaine, Parallel and distributed MCMC via shepherding distributions, in International Conference on Artificial Intelligence and Statistics, 2018.
22. V. Plassier, M. Vono, A. Durmus and E. Moulines, DG-LMC: A turn-key and scalable synchronous distributed MCMC algorithm via Langevin Monte Carlo within Gibbs, in International Conference on Machine Learning, 2021.
23. S. Ahn, B. Shahbaba and M. Welling, Distributed stochastic gradient MCMC, in International Conference on Machine Learning, 2014.
24. K. El Mekkaoui, D. Mesquita, P. Blomstedt and S. Kaski, Federated stochastic gradient Langevin dynamics, in Uncertainty in Artificial Intelligence, 2021.
25. M. Vono, V. Plassier, A. Durmus, A. Dieuleveut and E. Moulines, QLSD: Quantised Langevin stochastic dynamics for Bayesian federated learning, in International Conference on Artificial Intelligence and Statistics, 2022.
26. H.-Y. Chen and W.-L. Chao, FedBE: Making Bayesian model ensemble applicable to federated

learning, arXiv preprint arXiv:2009.01974 (2020).

27. M. Yurochkin, M. Agarwal, S. Ghosh, K. Greenewald, N. Hoang and Y. Khazaeni, Bayesian nonparametric federated learning of neural networks, in International Conference on Machine Learning, 2019.
28. C. M. Carvalho, N. G. Polson and J. G. Scott, Handling sparsity via the horseshoe, in Artificial Intelligence and Statistics, 2009.
29. A. Bhadra, J. Datta, N. G. Polson and B. Willard, Lasso meets horseshoe: A survey, Statistical Science 34, 405 (2019).
30. E. Makalic and D. F. Schmidt, A simple sampler for the horseshoe estimator, IEEE Signal Processing Letters 23, 179 (2015).
31. N. Meinshausen and P. Bühlmann, High-dimensional graphs and variable selection with the lasso, The Annals of Statistics 34, 1436 (2006).
32. P. Spirtes, C. N. Glymour, R. Scheines and D. Heckerman, Causation, Prediction, and Search (MIT press, 2000).
33. J. Pearl, Causality: Models, Reasoning and Inference, 2nd edn. (Cambridge University Press, USA, 2009).
34. P. Spirtes, Directed cyclic graphical representations of feedback models, in Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, UAI'95 (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1995).
35. T. Richardson, A discovery algorithm for directed cyclic graph, in Proceedings of the Twelfth International Conference on Uncertainty in Artificial Intelligence, UAI'961996.
36. G. Lacerda, P. L. Spirtes, J. Ramsey and P. O. Hoyer, Discovering cyclic causal models by independent components analysis, in Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence, 2008.
37. J. Li, Y. Lu, R. Akbani, Z. Ju, P. L. Roebuck, W. Liu, J.-Y. Yang, B. M. Broom, R. G. Verhaak, D. W. Kane et al., TCPA: a resource for cancer functional proteomics data, Nature methods 10, 1046 (2013).
38. M. Kim, J. Park, M. Bouhaddou, K. Kim, A. Rojc, M. Modak, M. Soucheray, M. J. McGregor, P. O'Leary, D. Wolf et al., A protein interaction landscape of breast cancer, Science 374, p. eabf3066 (2021).
39. R. Esteve-Puig, R. Gil, E. Gonzalez-Sanchez, J. J. Bech-Serra, J. Grueso, J. Hernandez-Losa, T. Moline, F. Canals, B. Ferrer, J. Cortes et al., A mouse model uncovers lkb1 as an uvb-induced dna damage sensor mediating cdkn1a (p21waf1/cip1) degradation, PLoS Genetics 10, p. e1004721 (2014).
40. R. L. Grossman, A. P. Heath, V. Ferretti, H. E. Varmus, D. R. Lowy, W. A. Kibbe and L. M. Staudt, Toward a shared vision for cancer genomic data, New England Journal of Medicine 375, 1109 (2016).
41. Y. Feng, M. Spezia, S. Huang, C. Yuan, Z. Zeng, L. Zhang, X. Ji, W. Liu, B. Huang, W. Luo et al., Breast cancer development and progression: Risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis, Genes & Diseases 5, 77 (2018).
42. P. Müller, G. Parmigiani and K. Rice, FDR and Bayesian multiple comparisons rules, in Proceedings of the 8th Valencia World Meeting on Bayesian Statistics, (Oxford University Press, 2006).
43. M. Kanehisa, M. Furumichi, Y. Sato, M. Ishiguro-Watanabe and M. Tanabe, KEGG: integrating viruses and cellular organisms, Nucleic acids research 49, D545 (2021).
44. Y. Ni, F. C. Stingo and V. Baladandayuthapani, Bayesian nonlinear model selection for gene regulatory networks, Biometrics 71, 585 (2015).