A Differentially Private Probabilistic Framework for Modeling the Variability Across Federated Datasets of Heterogeneous Multi-View Observations

Irene Balelli irene.balelli@inria.fr
Université Côte d'Azur, Inria Sophia Antipolis-Méditeranée, Epione Research Project, France
Santiago Silva santiago-smith.silva-rincon@inria.fr
Université Côte d'Azur, Inria Sophia Antipolis-Méditeranée, Epione Research Project, France
Marco Lorenzi marco.lorenzi@inria.fr
Université Côte d'Azur, Inria Sophia Antipolis-Méditeranée, Epione Research Project, France

Abstract

We propose a novel federated learning paradigm to model data variability among heterogeneous clients in multi-centric studies. Our method is expressed through a hierarchical Bayesian latent variable model, where client-specific parameters are assumed to be realization from a global distribution at the master level, which is in turn estimated to account for data bias and variability across clients. We show that our framework can be effectively optimized through expectation maximization (EM) over latent master's distribution and clients' parameters. We also introduce formal differential privacy guarantees compatibly with our EM optimization scheme. We tested our method on the analysis of multi-modal medical imaging data and clinical scores from distributed clinical datasets of patients affected by Alzheimers disease. We demonstrate that our method is robust when data is distributed either in iid and non-iid manners, even when local parameters perturbation is performed to reach suitable privacy levels. Our approach allows to quantify the variability of data, views and centers, while guaranteeing high-quality data reconstruction as compared to the state-of-the-art autoencoding models and federated learning schemes. Our code is available at `https://gitlab.inria.fr/epione/federated-multi-views-ppca`.

Keywords: Federated Learning, Hierarchical Generative Model, Heterogeneity, Differential Privacy

1. Introduction

The analysis of medical imaging datasets for the study of neurodegenerative diseases, requires the joint modeling of multiple *views* (or modalities), such as clinical scores and multi-modal medical imaging data. These views are generated through different processes for data acquisition, as for instance Magnetic Resonance Imaging (MRI) or Positron Emission Tomography (PET). Each view provides a specific information about the pathology, and the joint analysis of all views is necessary to improve diagnosis, for the discovery of pathological relationships or for predicting the disease evolution. Nevertheless, the integration of *multi-views* data, accounting for their mutual interactions and their joint variability, presents a number of challenges.

When dealing with high dimensional and noisy data it is crucial to be able to extract an informative lower dimensional representation to disentangle the relationships among observations, accounting for the intrinsic heterogeneity of the original complex data structure. From a statistical perspective, this implies the estimation of a model of the joint variability across views, or equiv-

alently the development of a joint *generative model*, assuming the existence of a common latent variable generating all views.

Several data assimilation methods based on dimensionality reduction have been developed (Cunningham and Ghahramani, 2015), and successfully applied to a variety of domains. The main goal of these methods is to identify a suitable lower dimensional latent space, where some key characteristics of the original dataset are preserved after projection. The most basic among such methods is Principal Component Analysis (PCA) (Jolliffe, 1986), where data are projected over the axes of maximal variability. More flexible approaches are Auto-Encoders (Wang et al., 2016), enabling to learn a low-dimensional representation minimizing the reconstruction error.

In some cases, Bayesian counterparts of the original dimensionality reduction methods have been developed, such as Probabilistic Principal Component Analysis (PPCA) (Tipping and Bishop, 1999), based on factor analysis, or, more recently, Variational Auto-Encoders (VAEs) (Kingma and Welling, 2014). VAEs are machine learning algorithms based on a generative function which allows probabilistic data reconstruction from the latent space. Encoder and decoder can be flexibly parametrized by neural networks (NNs), and efficiently optimized through Stochastic Gradient Descent (SGD). The added values of a Bayesian formulation is to provide a tool for sampling further observations from the estimated data distribution, and quantify the uncertainty of data and parameters. In addition, Bayesian model selection criteria, such as the Watanabe-Akaike Information Criteria (WAIC) (Gelman et al., 2014), allow to perform automatic model selection.

Multi-centric biomedical studies offer a great opportunity to significantly increase the quantity and quality of available data, hence to improve the statistical reliability of their analysis. Nevertheless, in this context, three main data-related challenges should be considered. 1) *Statistical heterogeneity* of *local datasets* (*i.e.* center-specific datasets): observations may be non-identically distributed across centers with respect to some characteristic affecting the output (*e.g.* diagnosis). Additional variability in local datasets can also come from data collection and acquisition bias (Kalter et al., 2019). 2) *Missing not at random views*: not all views are usually available for each center. 3) *Privacy* concerns: privacy-preserving laws are currently enforced to ensure the protection of personal data (*e.g.* the European General Data Protection Regulation - GDPR[1]), preventing the centralized analysis of data collected in multiple biomedical centers (Iyengar et al., 2018; Chassang, 2017). These limitations impose the need for extending data assimilation methods to handle decentralized heterogeneous data and missing views in local datasets.

Federated learning (FL) is an emerging paradigm specifically developed for the decentralized training of machine learning models. In order to guarantee data privacy, FL methods are conceived to avoid any sensitive data transfer among centers: raw data are processed within each center, which only shares local parameters with the master. Nevertheless, no formal privacy guarantees are provided on the shared statistics, which may still reveal sensitive information about individual data points used to train the model. Differential privacy (DP) is an established framework to provide theoretical guarantees about the anonymity of the shared statistics with respect to the training data points. Recent works (Abadi et al., 2016; Geyer et al., 2017; Triastcyn and Faltings, 2019) show the importance of combining FL and DP to prevent potential information leakage form the shared parameters while providing theoretical privacy guarantees both at the client and central server levels.

The standard aggregation method in FL is Federated Averaging (FedAvg) (McMahan et al., 2017a), which combines local models via weighted averaging. However, this aggregation scheme

1. `https://gdpr-info.eu/`

is sensitive to statistical heterogeneity, which naturally arises in federated datasets (Li et al., 2020), for example when dealing with multi-view data, or when data are not uniformly represented across data centers (e.g. non-iid distributed). In this case a faithful representation of the variability across centers is not guaranteed.

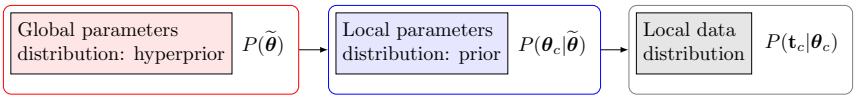


Figure 1: Hierarchical structure of Fed-mv-PPCA. Global parameters $\widetilde{\boldsymbol{\theta}}$ characterize the distribution of the local $\boldsymbol{\theta}_c$, which parametrize the local data distribution in each center.

We present here the Federated multi-view PPCA (Fed-mv-PPCA), a novel FL framework for data assimilation of heterogeneous multi-view datasets. Our framework is designed to account for the heterogeneity of federated datasets through a fully Bayesian formulation. Fed-mv-PPCA is based on a hierarchical dependency of the model's parameters to handle different sources of variability in the federated dataset (Figure 1). The method is based on a linear generative model, assuming Gaussian latent variables and noise, and allows to account for missing views and observations across datasets. In practice, we assume that there exists an ideal global distribution of each parameter, from which local parameters are generated to account for the local data distribution in each center. We show, in addition, that the privacy of the shared parameters of Fed-mv-PPCA can be explicitly quantified and guaranteed by means of DP. The code developed in Python is publicly available at `https://gitlab.inria.fr/epione/federated-multi-views-ppca`.

The paper is organized as follows: in Section 2 we provide a brief overview of the state-of-the-art and highlight the advancements provided by Fed-mv-PPCA. In Section 3 we describe the Fed-mv-PPCA framework: its extension to improve privacy preservation through DP is provided in Section 3.2. In Section 4 we show results with applications to synthetic data and to data from the Alzheimers Disease Neuroimaging Initiative dataset (ADNI). Section 5 concludes the paper with a brief discussion.

2. Related Works

The method presented in this paper falls within two main categories: Bayesian methods for data assimilation, and FL methods for heterogeneous datasets. Several methods for dimensionality reduction based on generative models have been developed in the past years, starting from the seminal work of PPCA by Tipping and Bishop (1999), to Bayesian Canonical Correlation Analysis (CCA) (Klami et al., 2013), which Matsuura et al. (2018) extended to include multiple views and missing modalities, up to more complex methods based on multi-variate association models (Shen and Thompson, 2019), developed, for example, to integrate multi-modal brain imaging data and high-throughput genomics data. More recent methods for the probabilistic analysis of multi-views datasests include the multi channel Variational Autoencoder (mc-VAE) by Antelmi et al. (2019) and Multi-Omics Factor Analysis (MOFA) by Argelaguet et al. (2018). MOFA generalizes PPCA for the analysis of multiple-omics data types, supporting different noise models to adapt to continu-

ous, binary and count data, while mc-VAE extends the classic VAE (Kingma and Welling, 2014) to jointly account for multiple-views data. Additionally, mc-VAE can handle sparse datasets: data reconstruction in testing can be inferred from available views, if some are missing.

Despite the possibility of performing data assimilation and integrate multiple views offered by the above methods, these approaches have not been conceived to handle federated datasets.

Statistical heterogeneity is a key challenge in FL and, more generally, in multi-centric studies (Li et al., 2020). To tackle this problem, Li et al. (2018) recently proposed the FedProx algorithm, which improves FedAvg by allowing for partial local work (*i.e.* adapting the number of local epochs) and by introducing a proximal term to the local objective function to avoid divergence due to data heterogeneity. Other methods have been developed under the Bayesian non-parametric formalism, such as probabilistic neural matching (Yurochkin et al., 2018), where the local parameters of NNs are federated depending on neurons similarities.

Despite significant improvements in the handling of statistical heterogeneity have been made since the development of FedAvg, state-of-the-art FL methods are currently essentially formulated for training schemes based on stochastic gradient descent, with principal applications to NNs based models. In addition, even if privacy is a major concern motivating the development of FL methods, classical works in FL does not provide strong theoretical guarantees for privacy preservation, leaving the door open to potential data leakage from malicious clients or the central server, such as through model inversion (Fredrikson et al., 2015). A solution to this problem can be provided by coupling FL schemes with DP (McMahan et al., 2017b). Nevertheless, beyond the specific application to NNs, we still lack of a consistent and privacy-compliant Bayesian framework for the estimation of local and global data variability, as part of a global optimization model, while accounting for data heterogeneity. This provides us motivation for the development of Fed-mv-PPCA and DP-Fed-mv-PPCA, a Bayesian framework for data assimilation from heterogeneous multi-views private federated datasets.

3. Methods

3.1 Federated multi-views PPCA

3.1.1 PROBLEM SETUP

We consider C independent centers. Each center $c \in \{1, \dots, C\}$ disposes of its private local dataset $T_c = \{\mathbf{t}_{c,n}\}_n$, with $|T_c| = N_c$. We assume that a total of K distinct views have been measured across all centers, and we allow missing views in some local dataset (*i.e.* some local dataset could be incomplete, including only measurements for $K_c < K$ views). For every $k \in \{1, \dots, K\}$, the dimension of the k^{th}-view (*i.e.* the number of features defining the k^{th}-view) is d_k, and we define $d := \sum_{k=1}^{K} d_k$. We denote by $\mathbf{t}_{c,n}^{(k)}$ the raw data of subject n in center c corresponding to the k^{th}-view, hence $\mathbf{t}_{c,n} = \left(\mathbf{t}_{c,n}^{(1)}, \dots, \mathbf{t}_{c,n}^{(K)}\right)$.

3.1.2 MODELING ASSUMPTIONS

The main assumption at the basis of Fed-mv-PPCA is the existence of a hierarchical structure underlying the data distribution. In particular, we suppose that there exist global parameters $\widetilde{\boldsymbol{\theta}}$, following a distribution $P(\widetilde{\boldsymbol{\theta}})$, able to describe the global data variability, *i.e.* the ensemble of local datasets. For each center, local parameters $\boldsymbol{\theta}_c$ are generated from $P(\boldsymbol{\theta}_c|\widetilde{\boldsymbol{\theta}})$, to account for the specific vari-

ability of the local dataset. Finally, local data $\mathbf{t}_c$ are obtained from their local distribution $P(\mathbf{t}_c|\boldsymbol{\theta}_c)$. Given the federated datasets, Fed-mv-PPCA provides a consistent Bayesian framework to solve the inverse problem and estimate the model's parameters across the entire hierarchy.

We assume that in each center c, the local data of subject n corresponding to the k^{th}-view, $\mathbf{t}_{c,n}^{(k)}$, follows the generative model:

$$\mathbf{t}_{c,n}^{(k)} = W_c^{(k)}\mathbf{x}_{c,n} + \boldsymbol{\mu}_c^{(k)} + \boldsymbol{\varepsilon}_c^{(k)}, \tag{1}$$

where $\mathbf{x}_{c,n} \sim \mathcal{N}(0, \mathbb{I}_q)$ is a q-dimensional latent variable, and $q < \min_k(d_k)$ is the dimension of the latent-space. $W_c^{(k)} \in \mathbb{R}^{d_k \times q}$ provides the linear mapping between latent space and observations for the k^{th}-view, $\boldsymbol{\mu}_c^{(k)} \in \mathbb{R}^{d_k}$ is the offset of the data corresponding to view k, and $\boldsymbol{\varepsilon}_c^{(k)} \sim \mathcal{N}\left(0, {\sigma_c^{(k)}}^2 \mathbb{I}_{d_k}\right)$ is Gaussian noise for the k^{th}-view. This formulation induces a Gaussian distribution over $\mathbf{t}_{c,n}^{(k)}$, implying:

$$\mathbf{t}_{c,n}^{(k)} \sim \mathcal{N}(\boldsymbol{\mu}_c^{(k)}, C_c^{(k)}), \tag{2}$$

where $C_c^{(k)} = W_c^{(k)} {W_c^{(k)}}^T + {\sigma_c^{(k)}}^2 \mathbb{I}_{d_k} \in \mathbb{R}^{d_k \times d_k}$. Finally, a compact formulation for $\mathbf{t}_{c,n}$ (*i.e.* considering all views concatenated) can be derived from Equation (1):

$$\mathbf{t}_{c,n} = W_c\mathbf{x}_{c,n} + \boldsymbol{\mu}_c + \Psi_c, \tag{3}$$

where $W_c, \boldsymbol{\mu}_c$ are obtained by concatenating all $W_c^{(k)}, \boldsymbol{\mu}_c^{(k)}$, and Ψ_c is a block diagonal matrix, where the k^{th}-block is given by $\boldsymbol{\varepsilon}_c^{(k)}$. The local parameters describing the center-specific dataset thus are $\boldsymbol{\theta}_c := \left\{\boldsymbol{\mu}_c^{(k)}, W_c^{(k)}, {\sigma_c^{(k)}}^2\right\}_k$. According to our hierarchical formulation, we assume that each local parameter in $\boldsymbol{\theta}_c$ is a realization of a common global prior distribution described by $\widetilde{\boldsymbol{\theta}} := \left\{\widetilde{\boldsymbol{\mu}}^{(k)}, \sigma_{\widetilde{\boldsymbol{\mu}}^{(k)}}, \widetilde{W}^{(k)}, \sigma_{\widetilde{W}^{(k)}}, \widetilde{\alpha}^{(k)}, \widetilde{\beta}^{(k)}\right\}_k$. In particular we assume that $\boldsymbol{\mu}_c^{(k)}$ and $W_c^{(k)}$ are normally distributed, while the variance of the Gaussian error, ${\sigma_c^{(k)}}^2$, follows an inverse-gamma distribution. Formally:

$$\begin{aligned}
\boldsymbol{\mu}_c^{(k)}|\widetilde{\boldsymbol{\mu}}^{(k)}, \sigma_{\widetilde{\boldsymbol{\mu}}^{(k)}} &\sim \mathcal{N}\left(\widetilde{\boldsymbol{\mu}}^{(k)}, \sigma^2_{\widetilde{\boldsymbol{\mu}}^{(k)}} \mathbb{I}_{d_k}\right), &(4)\\
W_c^{(k)}|\widetilde{W}^{(k)}, \sigma_{\widetilde{W}^{(k)}} &\sim \mathcal{MN}_{k,q}\left(\widetilde{W}^{(k)}, \mathbb{I}_{d_k}, \sigma^2_{\widetilde{W}^{(k)}} \mathbb{I}_q\right), &(5)\\
{\sigma_c^{(k)}}^2|\widetilde{\alpha}^{(k)}, \widetilde{\beta}^{(k)} &\sim \text{Inverse-Gamma}(\widetilde{\alpha}^{(k)}, \widetilde{\beta}^{(k)}), &(6)
\end{aligned}$$

where $\mathcal{MN}_{k,q}$ denotes the matrix normal distribution of dimension $d_k \times q$.

3.1.3 PROPOSED FRAMEWORK

The assumptions made in Section 3.1.2 allow to naturally define an optimization scheme based on Expectation Maximization (EM) locally, and on Maximum Likelihood estimation (ML) at the master level (Algorithm 1). Figure 2 shows the graphical model of Fed-mv-PPCA.

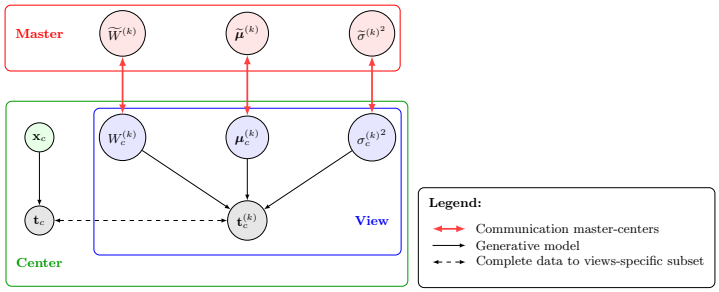


Figure 2: Graphical model of Fed-mv-PPCA. Thick double-sided red arrows relate nodes which are shared between center and master, while plain black arrows define the relations between the local dataset and the generative model parameters. Grey filled circles correspond to raw data: the dashed double-sided arrow simply highlights the complexity of the dataset, composed by multiple views.

Algorithm 1: Fed-mv-PPCA algorithm

Input : Rounds R; Iterations I; Latent space dimension q
Output: Global parameters $\widetilde{\boldsymbol{\theta}}$

for $r = 1, \ldots, R$ **do**
  for $c = 1, \ldots, C$ ***in parallel*** **do**
    Each center c initializes $\boldsymbol{\theta}_c$ using $P(\boldsymbol{\theta}_c|\widetilde{\boldsymbol{\theta}})$;
    I iterations of MAP estimation (EM + prior) to optimize $\boldsymbol{\theta}_c$ using $\widetilde{\boldsymbol{\theta}}$ as prior;
  end
  Each center c returns $\boldsymbol{\theta}_c$ to the master;
  The master collects $\boldsymbol{\theta}_c$, $c = 1, \ldots, C$ and estimates $\widetilde{\boldsymbol{\theta}}$ through ML;
  The master sends $\widetilde{\boldsymbol{\theta}}$ to all centers
end

With reference to Algorithm 1, the optimization of Fed-mv-PPCA is as follows:

Optimization. The master collects the local parameters $\boldsymbol{\theta}_c$ for $c \in \{1, \ldots, C\}$ and estimates the ML updated global parameters characterizing the prior distributions of Equations (4) to (6). Updated global parameters $\widetilde{\boldsymbol{\theta}}$ are returned to each center, and serve as priors to update the MAP estimation of the local parameters $\boldsymbol{\theta}_c$, through the M step on the functional $\mathbf{E}_{p(\mathbf{x}_{c,n}|\mathbf{t}_{c,n})} \ln \left(p(\mathbf{t}_{c,n}, \mathbf{x}_{c,n}|\boldsymbol{\theta}_c) p(\boldsymbol{\theta}_c|\widetilde{\boldsymbol{\theta}}) \right)$, where:

$$\mathbf{x}_{c,n}|\mathbf{t}_{c,n} \sim \mathcal{N}\left(\Sigma_c^{-1} {W_c}^T \Psi_c^{-1}(\mathbf{t}_{c,n} - \boldsymbol{\mu}_c), \Sigma_c^{-1}\right), \Sigma_c := (\mathbb{I}_q + W_c^T \Psi_c^{-1} W_c)$$

and

$$\begin{aligned}\langle \ln\left(p(\mathbf{t}_{c,n}, \mathbf{x}_{c,n}|\boldsymbol{\theta}_c)\right)\rangle \;=\; & -\sum_{n=1}^{N_c}\left\{\sum_{k=1}^{K}\left[\frac{d_k}{2}\ln\left({\sigma_c^{(k)}}^2\right)+\frac{1}{2{\sigma_c^{(k)}}^2}\|\mathbf{t}_{c,n}^{(k)}-\boldsymbol{\mu}_c^{(k)}\|^2+\right.\right.\\ & \frac{1}{2{\sigma_c^{(k)}}^2}tr\left({W_c^{(k)}}^T W_c^{(k)}\langle \mathbf{x}_{c,n}\mathbf{x}_{c,n}^T\rangle\right)\\ & \left.\left.-\frac{1}{{\sigma_c^{(k)}}^2}\langle \mathbf{x}_{c,n}\rangle^T {W_c^{(k)}}^T\left(\mathbf{t}_{c,i}^{(k),g}-\boldsymbol{\mu}_c^{(k)}\right)\right]+\frac{1}{2}tr\left(\langle \mathbf{x}_{c,n}\mathbf{x}_{c,n}^T\rangle\right)\right\},\end{aligned}$$

Initialization at round r=1. The latent-space dimension q, the number of local iterations I and the number of communication rounds R (*i.e.* number of complete cycles centers-master) are user-defined parameters. For the sake of simplicity, we set here the same number of local iterations for every center. Note that this constraint can be easily adapted to take into account systems heterogeneity among centers, as well as the size of each local dataset. At the first round, local parameters initialization, hence optimization, can be performed in two distinct ways: 1) each center can initialize randomly every local parameter, then perform EM through I iterations, maximizing the functional $\langle \ln\left(p(\mathbf{t}_{c,n}, \mathbf{x}_{c,n}|\boldsymbol{\theta}_c)\right)\rangle$; 2) the master can provide priors for at least some parameters, which will be optimized using MAP estimation as described above. In case of a random initialization of local parameters, the number of EM iterations for the first round can be increased: this can be seen as an exploratory phase.

The reader can refer to Appendix A for further details on the theoretical formulation of Fed-mv-PPCA and the corresponding optimization scheme.

3.2 Fed-mv-PPCA with Differential Privacy

Despite the Bayesian federated learning scheme deployed prevents any data transfer, it does not provide theoretical privacy guarantees on the shared statistics. Differential privacy (DP) (Dwork et al., 2014; Abadi et al., 2016) is considered as a privacy gold standard: it allows to quantify privacy protection properties of the considered algorithm, and to sanitize model parameters through output perturbation with the addition of a random noise mechanism. The noise strength has to be tuned to insure a good balance between privacy and utility of the outputs.

In Section 3.2.1 we recall the standard definition of differential privacy and related results which will be needed afterwards. A differentially private version of Fed-mv-PPCA (DP-Fed-mv-PPCA) is proposed in Section 3.2.2, allowing an improvement of the shared parameters' privacy.

3.2.1 DEFINITIONS AND RESULTS

We denote by D, D' two datasets: D and D' are said to be neighboring or adjacent datasets if they only differ by a datapoint $\mathbf{t}'$, $D = D' \cup \{\mathbf{t}'\}$. In this case we write $\|D - D'\| = 1$.

Definition 1 *A randomized algorithm $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ with domain $\mathcal{D}$ and range $\mathcal{R}$ is (ε, δ)-differentially private if for any $D, D' \in \mathcal{D}$ s.t. $\|D - D'\| = 1$ and for any $\mathcal{S} \in \mathcal{R}$:*

$$\mathbb{P}\left[\mathcal{M}(D) \in \mathcal{S}\right] \leq e^{\varepsilon}\left[\mathcal{M}(D') \in \mathcal{S}\right] + \delta$$

When $\delta = 0$, we simply say that the algorithm $\mathcal{M}$ is ε-differentially private.

A common mechanism to approximate a deterministic function or a query $f : \mathcal{D} \to \mathbb{R}^d$ with differential privacy is the addition of a random noise calibrated on the sensitivity of f.

Definition 2 *The l_p-sensitivity of a function $f : \mathcal{D} \to \mathbb{R}^d$ is defined as:*

$$\Delta_p f = \max_{\|D-D'\|=1} \|f(D) - f(D')\|_p$$

Classical mechanisms used for perturbation are the *Laplace mechanism* and the *Gaussian mechanism*. A Laplace (resp. Gaussian) mechanism is simply obtained by computing f, hence perturbing it with noise added from a Laplace (resp. Gaussian) distribution centered in the origin and with variance depending on the sensitivity of f:

$$\mathcal{M}(D) := f(D) + Noise,$$

where $Noise \sim \text{Laplace}\,(0, \text{std}_L(\Delta_p f))$ (resp. $Noise \sim \mathcal{N}\,(0, \text{var}_G(\Delta_p f))$).

Hereafter we detail the condition of a Laplace (resp. Gaussian) mechanism to preserve (ε, δ)-DP.

Theorem 3 *Given any function $f : \mathcal{D} \to \mathbb{R}^d$ and $\varepsilon > 0$, the Laplace mechanism defined as*

$$\mathcal{M}(D) := f(D) + (L_1, \ldots, L_d),$$

where L_i are iid drawn from Laplace$(0, \Delta_1 f/\varepsilon)$, preserves ε-DP.

Theorem 4 *Given any function $f : \mathcal{D} \to \mathbb{R}^d$ and $(\varepsilon, \delta) \in (0, 1)^2$, the Gaussian mechanism defined as*

$$\mathcal{M}(D) := f(D) + \mathcal{N}\left(\mathbf{0}, \left(\frac{\sqrt{2\ln(1.25\delta)}\Delta_2 f}{\varepsilon}\right)^2 \mathbb{I}_d\right),$$

preserves (ε, δ)-DP.

The formal proofs of Theorems 3-4 are provided *e.g.* by Dwork et al. (2014).

An improved Gaussian mechanism is further described by Zhao et al. (2019), with the advantages of 1) remaining valid for $\varepsilon > 1$ given $\delta \leq 0.5$ and 2) adding a smaller noise if compared to the result of Theorem 4 in the case $0 < \varepsilon \leq 1$.

Theorem 5 *Given any function $f : \mathcal{D} \to \mathbb{R}^d$, $\varepsilon > 0$, and $\delta \in (0, 0.5)$, the Gaussian mechanism defined as*

$$\mathcal{M}(D) := f(D) + \mathcal{N}\left(\mathbf{0}, \left(\frac{(c + \sqrt{c^2 + \varepsilon})\Delta_2 f}{\varepsilon\sqrt{2}}\right)^2 \mathbb{I}_d\right),$$

where $c = \sqrt{\ln\left(2/(\sqrt{16\delta + 1} - 1)\right)}$, preserves (ε, δ)-DP.

Theorems 4 and 5 can be extended to queries mapping to $\mathbb{R}^{d\times q}$ and matrix normal mechanisms:

Corollary 6 *Given any function $f : \mathcal{D} \to \mathbb{R}^{d\times q}$, $\varepsilon > 0$, and $\delta \in (0, 0.5)$, the matrix normal mechanism defined as*

$$\mathcal{M}(D) := f(D) + \mathcal{MN}_{d,q}\left(\mathbf{0}_{d,q}, \mathbb{I}_d, \left(\frac{(c+\sqrt{c^2+\varepsilon})\Delta_2 f}{\varepsilon\sqrt{2}}\right)^2 \mathbb{I}_q\right),$$

where $c = \sqrt{\ln\left(2/(\sqrt{16\delta+1}-1)\right)}$, preserves (ε, δ)-DP.

Proof The proof of Corollary 6 follows form Theorem 5 and the definition of matrix normal distribution, stating that $X \sim \mathcal{MN}_{d,q}(M, U, V)$ if and only if $\mathrm{vec}(X) \sim \mathcal{N}_{dq}(\mathrm{vec}(M), V \otimes U)$, where $\mathrm{vec}(M)$ denotes the vectorization of M and $\otimes$ denotes the Kronecker product. ■

We conclude this section by recalling the well known composition theorem (Dwork et al., 2014), which will be useful to quantify the global privacy budget for each center in the next sections.

Theorem 7 *For $i = 1, \ldots, k$, let $\mathcal{M}_i : \mathcal{D} \to \mathcal{R}_i$ be an $(\varepsilon_i, \delta_i)$-differentially private algorithm, and $\mathcal{M} : \mathcal{D} \to \prod_{i=1}^k \mathcal{R}_i$ defined as $\mathcal{M}(\mathcal{D}) := (\mathcal{M}_1(\mathcal{D}), \ldots, \mathcal{M}_k(\mathcal{D}))$. Then $\mathcal{M}$ is $\left(\sum_{i=1}^k \varepsilon_i, \sum_{i=1}^k \delta_i\right)$-differentially private.*

3.2.2 DIFFERENTIAL PRIVACY FOR LOCAL PARAMETERS

In this section we propose a reinforced federated learning scheme for Fed-mv-PPCA with DP to protect client-level privacy and avoid potential private information leakage from the shared local parameters.

We are interested in preserving the privacy of the shared local parameters $\boldsymbol{\theta}_c = \{\boldsymbol{\mu}_c^{(k)}, W_c^{(k)}, {\sigma_c^{(k)}}^2\}_k$, which can be done by the addition of some properly tuned random noise, as detailed in Section 3.2.1. Nevertheless, the client-level optimization scheme in Fed-mv-PPCA is based on an iterative algorithm: therefore we do not dispose of a closed formula to evaluate the sensitivity of each local parameter (*i.e.* the queries), nor an upper bound. To overcome this problem, we propose to perform difference clipping (Geyer et al., 2017; Zhang et al., 2021), one of the clipping strategies proposed for differentially private SGD models. Algorithm 2 outlines the optimization scheme for the DP-

Fed-mv-PPCA framework:

Algorithm 2: DP-Fed-mv-PPCA algorithm

Input : Rounds R; Iterations I; Latent space dimension q; Privacy parameters ε, δ
Output: Global parameters $\widetilde{\boldsymbol{\theta}}$

for $r = 1, \dots, R$ **do**
  for $c = 1, \dots, C$ ***in parallel*** **do**
    Initialize $\boldsymbol{\theta}_c$ using $P(\boldsymbol{\theta}_c|\widetilde{\boldsymbol{\theta}}[r-1])$;
    Update local parameters: I iterations of MAP estimation (EM + prior) to optimize $\boldsymbol{\theta}_c[r]$ using $\widetilde{\boldsymbol{\theta}}[r-1]$ as prior;
    Compute difference: $\Delta\boldsymbol{\theta}_c[r] := (\boldsymbol{\theta}_c[r] - \widetilde{\boldsymbol{\theta}}[r-1])$;
    Clip: $\overline{\Delta\boldsymbol{\theta}}_c[r] := \Delta\boldsymbol{\theta}_c[r] / \max\left(1, \|\Delta\boldsymbol{\theta}_c[r]\|_p / g(\sigma_{\widetilde{\boldsymbol{\theta}}}[r-1])\right)$;
    Perturb: $\mathcal{M}_{\boldsymbol{\theta}_c}[r] := \overline{\Delta\boldsymbol{\theta}}_c[r] + Noise(2g(\sigma_{\widetilde{\boldsymbol{\theta}}}[r-1]), \varepsilon, \delta)$;
    Return $\overline{\boldsymbol{\theta}}_c[r] := \mathcal{M}_{\boldsymbol{\theta}_c}[r] + \widetilde{\boldsymbol{\theta}}[r-1]$ to the master;
  end
  The master collects all $\overline{\boldsymbol{\theta}}_c[r]$ and estimates $\widetilde{\boldsymbol{\theta}}[r]$ through ML;
  The master sends $\widetilde{\boldsymbol{\theta}}[r]$ to all centers
end

Difference clipping and perturbation. With respect to Algorithm 1, difference clipping and perturbation are performed at the client level compatibly with the probabilistic formulation of the model:

1. The client computes the difference between the current local update and the initial prior (*i.e.* the corresponding global parameter obtained at the previous communication round, $r-1$):

$$\Delta\boldsymbol{\theta}_c[r] := (\boldsymbol{\theta}_c[r] - \widetilde{\boldsymbol{\theta}}[r-1])$$

2. The updated difference is clipped according to the standard deviation of the prior:

$$\overline{\Delta\boldsymbol{\theta}}_c[r] := \Delta\boldsymbol{\theta}_c[r] cdot \left(\max\left(1, \frac{\|\Delta\boldsymbol{\theta}_c[r]\|_p}{g(\sigma_{\widetilde{\boldsymbol{\theta}}}[r-1])}\right)\right)^{-1},$$

   where $g(\sigma_{\widetilde{\boldsymbol{\theta}}}[r-1]) := (\sigma_{\widetilde{\boldsymbol{\theta}}}[r-1]) \cdot \text{const}$, and the multiplicative constant is fixed by the user. This clipping mechanism forces the l_p norm of $\Delta\boldsymbol{\theta}_c[r]$ to be at most $g(\sigma_{\widetilde{\boldsymbol{\theta}}}[r-1])$. Consequently, the l_p sensitivity of $\overline{\Delta\boldsymbol{\theta}_c}[r]$ will be bounded by $2 \cdot g(\sigma_{\widetilde{\boldsymbol{\theta}}}[r-1])$.

3. The clipped difference is perturbed:

$$\mathcal{M}_{\boldsymbol{\theta}_c}[r] := \overline{\Delta\boldsymbol{\theta}}_c[r] + Noise(2g(\sigma_{\widetilde{\boldsymbol{\theta}}}[r-1]), \varepsilon, \delta)$$

   In particular for $\overline{\Delta\boldsymbol{\mu}^{(k)}}_c$ and $\overline{\Delta W^{(k)}}_c$ we propose to use a Gaussian (resp. matrix normal) mechanism (Theorem 5, resp. Corollary 6), in accordance to the supposed Gaussian prior distributions of these parameters, while a Laplace mechanism (Theorem 3) is used to perturb $\overline{\Delta{\sigma^{(k)}}^2}_c$.

4. The client adds again the prior and finally sends to the master $\overline{\boldsymbol{\theta}}_c[r] := \mathcal{M}_{\boldsymbol{\theta}_c}[r] + \widetilde{\boldsymbol{\theta}}[r-1]$.

Conversely to model clipping (Abadi et al., 2016; Wei et al., 2020), where the parameter update is directly clipped and perturbed, difference clipping has the advantage to allow reducing the magnitude of the perturbation: indeed, we expect the l_p norm of the difference $\Delta\boldsymbol{\theta}_c$ to be small compared to the l_p norm of $\boldsymbol{\theta}_c$. From a conceptual viewpoint, in our framework performing difference clipping with respect to the standard deviation of global parameters enforces local parameters to converge to the global parameters mean. On one hand, this implies a reduction of the ability of the framework in capturing the between-centers variability. On the other hand, this also allows to obfuscate the participation of the individual centers.

Privacy budget

Theorem 8 *For sake of simplicity, let us choose the same ε, δ for all mechanisms considered above (a generalization to a parameter-specific choice of ε_i, δ_i is straightforward). The total privacy budget for the outputs of Algorithm 2 is $(3K\varepsilon, 2K\delta)$, where K is the total number of views.*

Proof The proof of Theorem 8 follows from Theorems 3-5 and Corollary 6, and by noting that data in each center are disjoint. In all centers, we are dealing with the mechanism $\mathcal{M} := (\mathcal{M}_{\boldsymbol{\mu}_c^{(k)}}, \mathcal{M}_{W_c^{(k)}}, \mathcal{M}_{\sigma_c^{(k)^2}})_k$, where for all k, $\mathcal{M}_{\boldsymbol{\mu}_c^{(k)}}$ and $\mathcal{M}_{W_c^{(k)}}$ are (ε, δ)-differentially private, while for all k, $M_{\sigma_c^{(k)^2}}$ is ε-differentially private. The result follows thanks to composition Theorem 7 and the invariance of differential privacy under post-processing. ■

Corollary 9 *If for local parameter $\theta_c \in \boldsymbol{\theta}_c$ the client-specific differential parameters are $(\varepsilon_c, \delta_c)$, then the total privacy budget for the corresponding global parameter $\widetilde{\theta}$ is bounded by $(\max(\varepsilon_c), \max(\delta_c))$.*

Proof The result directly follows from Theorem 8 and by considering Definition 1 and the monotonicity of the exponential function. ■

4. Applications

4.1 Materials

In the preparation of this article we used two datasets.

Synthetic dataset (SD): we generated 400 observations from (1), consisting of $k = 3$ views of dimension $d_1 = 15, d_2 = 8, d_3 = 10$. Each view was generated from a common 5-dimensional latent space. We randomly chose parameters $W^{(k)}, \boldsymbol{\mu}^{(k)}, \sigma^{(k)}$. Finally, to simulate heterogeneity, a randomly chosen sub-sample composed by 250 observations was shifted in the latent space by a randomly generated vector: this allowed to simulate the existence of two distinct groups in the population.

Alzheimer's Disease Neuroimaging Initiative dataset (ADNI)[2]: we consider 311 participants extracted from the ADNI dataset, among cognitively normal (NL) (104 subjects) and patients diagnosed with AD (207 subjects). All participants are associated with multiple data views: cognitive scores including MMSE, CDR-SB, ADAS-Cog-11 and RAVLT (CLINIC), Magnetic resonance imaging (MRI), Fluorodeoxyglucose-PET (FDG) and AV45-Amyloid PET (AV45) images. MRI morphometrical biomarkers were obtained as regional volumes using the cross-sectional pipeline of FreeSurfer v6.0 and the Desikan-Killiany parcellation. Measurements from AV45-PET and FDG-PET were estimated by co-registering each modality to their respective MRI space, normalizing by the cerebellum uptake and by computing regional amyloid load and glucose hypometabolism using PetSurfer pipeline and the same parcellation. Features were corrected beforehand with respect to intra-cranial volume, sex and age using a multivariate linear model. Data dimensions for each view are: $d_{\text{CLINIC}} = 7$, $d_{\text{MRI}} = 41$, $d_{\text{FDG}} = 41$ and $d_{\text{AV45}} = 41$. Further details on the demographics of the ADNI sample are provided in Appendix B, Table 3.

4.2 Benchmark

We compare our method to two state-of-the art data assimilation methods: Variational Autoencoder (VAE) (Kingma and Welling, 2014) and multi-channel VAE (mc-VAE) (Antelmi et al., 2019). To keep the modeling setup consistent across methods, both auto-encoders were tested by considering linear encoding and decoding mappings. In order to obtain the federated version of VAE and mc-VAE we use FedAvg (McMahan et al., 2017a), which is specifically conceived for stochastic gradient descent optimization. For both optimization methods and federation schemes we set to 100 the total number of communication rounds, of 15 epochs each, with the default learning rate.

For the sake of completeness, supplementary Table 7 and supplementary Figure 9 provide also results for both VAE and mc-VAE, and using FedProx (Li et al., 2018) as aggregation scheme with the proximal term λ varying from 0.01 to 0.5: this method is supposed to improve convergence in case of heterogeneous data distributions. Nevertheless no significant improvement as been observed comparing to the FedAvg scheme for the considered datasets and settings.

4.3 Results

We apply Fed-mv-PPCA to both SD and ADNI datasets, and quantify the quality of reconstruction and identification of the latent space with respect to the increasing number of centers, C, and the increasing data heterogeneity. We investigate also the ability of Fed-mv-PPCA in estimating the data variability and predicting the distribution of missing views. To this end, we consider 4 different scenarios of data distribution across multiple centers, detailed in Table 1.

For each experiment considered hereafter with Fed-mv-PPCA, we perform 3-fold Cross Validation (3CV) tests. For every test, local parameters are initialized randomly (*i.e.* no prior is provided by the master at the beginning), and the number of rounds is set to 100. Each round consists of 15 iterations for local MAP optimization, except the initialization round, which consists of 30 EM

2. The ADNI project was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI was to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessments can be combined to measure the progression of early Alzheimer's disease (AD) (see `www.adni-info.org` for up-to-date information).

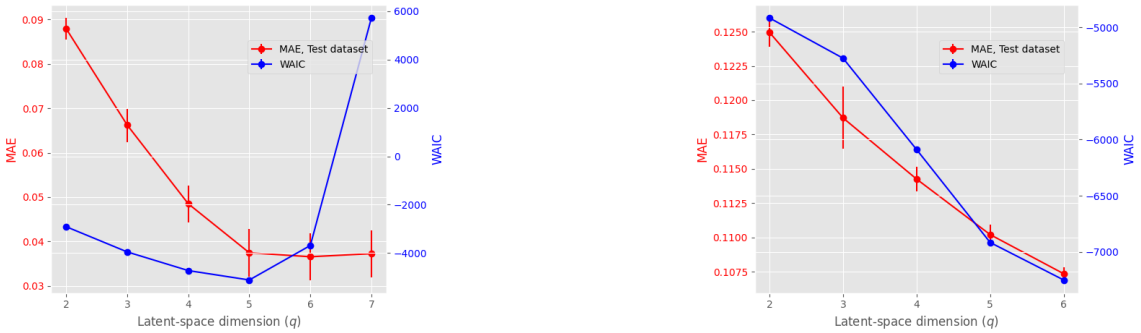
Table 1: Distribution of Datasets Across Centers.

Scenario	Description
IID	Data are iid distributed across C centers with respect to groups and all subjects dispose of a complete data raw
G	Data are non-iid distributed with respect to groups across C centers: $C/3$ centers includes subjects from both groups; $C/3$ centers only subjects from group 1 (AD in the ADNI case); $C/3$ centers only subjects from group 2 (NL for ADNI). All views have been measured in each center.
K	$C/3$ centers dispose of all observations; in $C/3$ centers the second view (MRI for ADNI) is missing; in $C/3$ centers the third view (FDG for ADNI) is missing. Data are iid distributed across C centers with respect to groups.
G/K	Data are non-iid distributed (scenario G) and there are missing views (scenario K).

iterations. Finally, when a centralized setting is tested, the number of rounds is set to 1 and the number of EM iterations to 800.

4.3.1 MODEL SELECTION

The latent space dimension q is an user defined parameter, with the only constraint $q < \min_k\{d_k\}$. To assess the optimal q, we consider the IID scenario and let q vary. We perform 10 times a 3-fold Cross Validation (3-CV), and split the train dataset across 3 centers. The resulting models are compared using the WAIC criterion. In addition, we consider the Mean Absolute reconstruction Error (MAE) in an hold-out test dataset: the MAE is obtained by evaluating the mean absolute distance between real data and data reconstructed using the global distribution. Figure 3 shows the evolution of WAIC and MAE with respect to the latent space dimension.



(a) SD

(b) ADNI

Figure 3: WAIC score and MAE for (a) the SD dataset and (b) the ADNI dataset. In both figures, the left y-axis scaling describes the MAE while the right y-axis scaling corresponds to the WAIC score.

Concerning the SD dataset, the WAIC strongly suggests $q = 5$ latent dimensions, hence demonstrating the ability of Fed-mv-PPCA to correctly recover the ground truth latent space dimension used to generate the data. Analogously, the MAE improves drastically up to the dimension $q = 5$, and subsequently stabilizes. For ADNI, the MAE improves for increasing latent space dimensions, and we obtain the best WAIC score for $q = 6$, suggesting that a high-capacity model is preferable to describe this larger dimensional dataset. Despite the agreement of MAE and WAIC for both datasets, the WAIC has the competitive advantage of providing a natural and automatic model selection measure in Bayesian models, which does not require testing data, conversely to MAE.

In the following experiments, we set the latent space dimension $q = 5$ for the SD dataset and $q = 6$ for the ADNI dataset.

Table 2: Results on ADNI dataset for all scenarios, and comparison with VAE and mc-VAE.

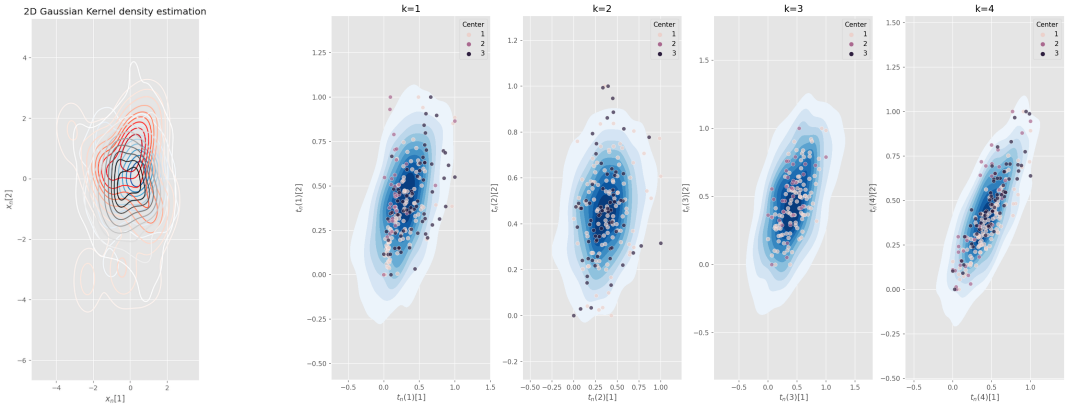
Scenario	Centers	Method	MAE Train	MAE Test	Accuracy in LS
IID	1 (centralized case)	**Fed-mv-PPCA**	**0.0805±0.0003**	**0.1110±0.0011**	0.8680±0.0379
		VAE	0.1055±0.0017	0.1344±0.0019	0.8003±0.0409
		mc-VAE	0.1382±0.0009	0.1669±0.0020	**0.8727±0.0319**
	3	**Fed-mv-PPCA**	**0.1027±0.0015**	**0.1073±0.0004**	0.8652±0.0270
		DP-Fed-mv-PPCA	0.1304±0.0047	0.1304±0.0041	0.8321±0.0388
		VAE	0.1172±0.0022	0.1192±0.0015	0.8289±0.0383
		mc-VAE	0.1602±0.0035	0.1567±0.0017	**0.8850±0.0262**
	6	**Fed-mv-PPCA**	**0.1203±0.0042**	**0.1074±0.0007**	0.8742±0.0267
		DP-Fed-mv-PPCA	0.1489±0.0051	0.1295±0.0029	0.8502±0.0347
		VAE	0.1357±0.0042	0.1191±0.0014	0.8224±0.0377
		mc-VAE	0.1840±0.0054	0.1563±0.0017	**0.8894±0.0230**
G	3	**Fed-mv-PPCA**	**0.1077±0.0090**	**0.1096±0.0011**	**0.8409±0.0293**
		DP-Fed-mv-PPCA	0.1362±0.0117	0.1340±0.0067	0.7977±0.0480
		VAE	0.1212±0.0077	0.1219±0.0015	0.7962±0.0440
		mc-VAE	0.1677±0.0156	0.1611±0.0025	0.8210±0.0464
	6	**Fed-mv-PPCA**	**0.1264±0.0126**	**0.10912±0.0011**	**0.8168±0.0324**
		DP-Fed-mv-PPCA	0.1585±0.0158	0.1340±0.0065	0.7898±0.0407
		VAE	0.1401±0.0114	0.1202±0.0016	0.7882±0.0534
		mc-VAE	0.1924±0.0219	0.1589±0.0018	0.8085±0.0464
K	3	Fed-mv-PPCA	0.0951±0.0086	0.1212±0.0109	0.8624±0.0303
		DP-Fed-mv-PPCA	0.1208±0.0081	0.1462±0.0092	0.8357±0.0329
	6	Fed-mv-PPCA	0.1107±0.0106	0.1293±0.0162	0.8720±0.0308
		DP-Fed-mv-PPCA	0.1434±0.0099	0.1604±0.0164	0.8515±0.0375
G/K	3	Fed-mv-PPCA	0.0995±0.0029	0.1271±0.0087	0.7338±0.0308
		DP-Fed-mv-PPCA	0.1287±0.0081	0.1547±0.0125	0.7164±0.0474
	6	Fed-mv-PPCA	0.1173±0.0061	0.1268±0.0088	0.7469±0.0202
		DP-Fed-mv-PPCA	0.1463±0.0088	0.1523±0.0104	0.7174±0.0387

4.3.2 INCREASING HETEROGENEITY ACROSS DATASETS

To test the robustness of Fed-mv-PPCA's results, for each scenario of Table 1, we perform 10 times 3-CV to obtain train and test datasets, hence we split the train dataset across C centers. We compare our method to VAE and mc-VAE, using the same partition of train and test datasets for CV. For all methods we consider the MAE in both the train and test datasets, as well as the accuracy score in the Latent Space (LS) discriminating the groups (synthetically defined in SD or corresponding to the clinical diagnosis in ADNI). The classification was performed via Linear Discriminant Analysis (LDA) on the individual projection of test data in the latent space. In what follows we present a detailed description of results corresponding to the ADNI dataset. Results for the SD dataset are in line with what we observe for ADNI (see Supplementary Table 6 in Appendix B), and confirm that our method outperforms both VAE and mc-VAE in reconstruction in all scenarios. In addition, Fed-mv-PPCA outperforms in discrimination both methods in the non-iid setting, while mc-VAE shows slightly improved discriminating ability in the IID scenario.

IID distribution. We consider the IID scenario and split the train dataset across 1 to 6 centers. Table 2 shows that results from Fed-mv-PPCA are stable when passing from a centralized to a federated setting, and when considering an increasing number of centers C. We only observe a degradation of the MAE in the train dataset, but this does not affect the performance of Fed-mv-PPCA in reconstructing the test data. Moreover, irrespective of the number of training centers, Fed-mv-PPCA outperforms VAE and mc-VAE in reconstruction.

Heterogeneous distribution. We simulate an increasing degree of heterogeneity in 3 to 6 local datasets, to further challenge the models in properly recovering the global data. In particular, we consider both a non-iid distribution of subjects across centers, and missing not at random views in some local dataset. It is worth noting that scenarios implying datasets with missing views cannot be handled by VAE nor by mc-VAE, hence in these cases we reported only results obtained with our method.



(a) Latent space (b) Original space

Figure 4: G/K scenario. First two dimensions for (a) sampling from posterior distribution of latent variables $\mathbf{x}_{c,n}$, and (b) predicted distribution $\mathbf{t}_{c,n}^{(k)}$ against real data.

In Table 2 we report the average MAEs and Accuracy in the latent space for each scenario, obtained over 10 tests for the ADNI dataset. Fed-mv-PPCA is robust despite an increasing degree of heterogeneity in the local datasests. We observe a slight deterioration of the MAE in the test dataset in the more challenging non-iid cases (scenarios K and G/K), while we note a drop of the classification accuracy in the most heterogeneous setup (G/K). Nevertheless, Fed-mv-PPCA demostrates to be more stable and to perform better than VAE and mc-VAE when statistical heterogeneity is introduced.

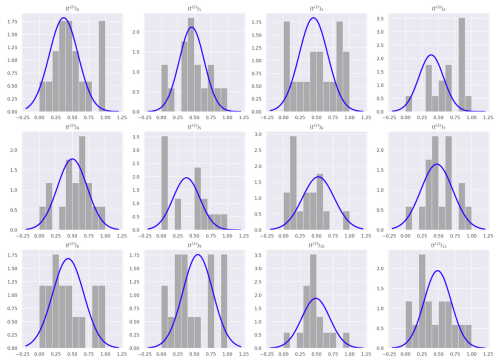


Figure 5: G/K scenario. Predicted testing distribution (blue curve) of sample features of the missing MRI view against real data (histogram).

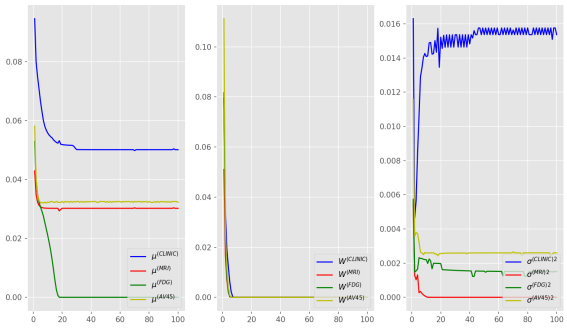
Figure 4 (a) shows the sampling posterior distribution of the latent variables, while in Figure 4 (b) we plot the predicted global distribution of the corresponding original space against observations, for the G/K scenario and considering 3 training centers. We notice that the variability of centers is well captured, in spite of the heterogeneity of the distribution in the latent space. In particular center 2 and center 3 have two clearly distinct means: this is due to the fact that subjects in these centers belong to two distinct groups (AD in center 2 and NL in center 3). Despite this, Fed-mv-PPCA is able to reconstruct correctly all views, even if 2 views are completely missing in some local datasets (MRI is missing in center 2 and FDG in center 3).

After convergence of Fed-mv-PPCA, each center is supplied with global distributions for each parameter: data corresponding to each view can therefore be simulated, even if some are missing in the local dataset. Considering the same simulation in the challenging G/K scenario as in Figure 4, in Figure 5 we plot the global distribution of some randomly selected features of a missing imaging view in the test center, against ground truth density histogram, from the original data. The global distribution provides an accurate description of the missing MRI view. Supplementary Figure 8 shows imputation for all features of the missing MRI and FDG views.

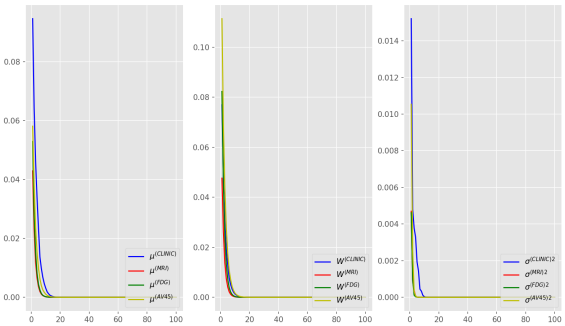
4.3.3 DIFFERENTIALLY PRIVATE FED-MV-PPCA

We repeated all experiments described in Section 4.3.2, using Fed-mv-PPCA with differential privacy. For each query $\theta_c \in \boldsymbol{\theta}_c$ we set $\varepsilon = 10$, $\delta = 0.01$. Finally, to perform difference clipping (Algorithm 2), we set the maximal l_p norm of the difference between the updated parameter at round r and the prior, $\overline{\Delta\boldsymbol{\theta}}_c[r]$, to be $\sigma_{\widetilde{\boldsymbol{\theta}}}[r-1]$.

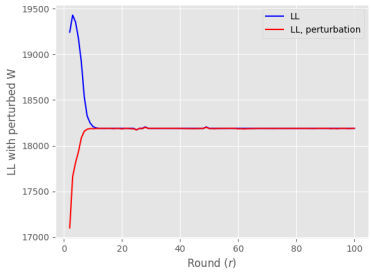
Evolution of the standard deviation of global parameters. To better understand the effect of performing difference clipping with respect to the priors, in Figure 6 (a-b) we plot the median evolution of the estimated standard deviation for each global parameter during training in the GK scenario, comparing Fed-mv-PPCA and DP-Fed-mv-PPCA. When DP is not introduced, we can see that standard deviations for most global parameters decrease, as expected, indicating harmonization of local parameters during training. This effect is enhanced when differential privacy is introduced: in this case all standard deviations drop towards 0 after approximately 20 communication rounds, meaning that the final global parameters distributions are strongly concentrated around their mean.



(a) Fed-mv-PPCA



(b) DP-Fed-mv-PPCA



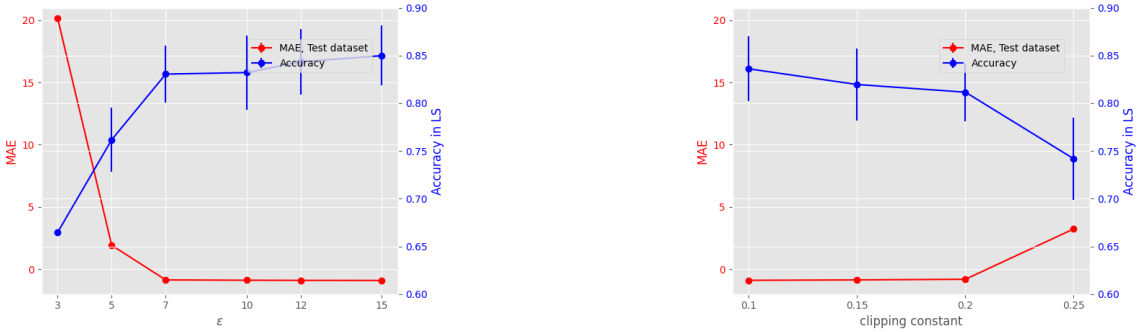
(c) Fed-mv-PPCA

Figure 6: (a-b)Evolution of the standard deviation of all global parameters for the GK scenario using 3 centers: comparison between Fed-mv-PPCA and DP-Fed-mv-PPCA. (c) Evolution of the expected log likelihood evaluated at each round using either $\boldsymbol{\theta}_c$ (blue curve) or $\overline{\boldsymbol{\theta}}_c$ (red curve).

DP parameters utility. We tested the utility of global parameters obtained with the differentially private Algorithm 2, to appreciate if data reconstruction and accuracy in the latent space are well preserved when the perturbation is performed at the client level (see Table 2, DP-Fed-mv-PPCA rows). As expected, we observe a deterioration of previous results, which increases with the number of training centers, due to the communication of a larger number of perturbed parameters. Nevertheless, results remain still coherent, and illustrate the utility of the differentially private global parameters.

In Figure 6 (c) we compare the median evolution across all centers and 3-CV tests of the total expected log-likelihood $\langle \ln \left(p(\mathbf{t}_{c,n}, \mathbf{x}_{c,n} | \boldsymbol{\theta}_c) \right) \rangle$ evaluated using the updated local parameters $\boldsymbol{\theta}_c$, and the ones obtained after difference clipping and perturbation, $\overline{\boldsymbol{\theta}}_c$. Both expected log-likelihoods converge starting from approximately 20 communication rounds, which is also due to the progressively more informative priors provided. Finally, Figure 7 shows the relationship between global parameters utility and both ε and the multiplicative constant used for difference clipping. We note that when (ε, δ) are fixed to $(1, 0.01)$, the clipping constant should be at most 0.2 to preserve a

reasonable utility of the model outputs. This further stresses the need of carefully tuning these DP parameters to ensure a good balance between privacy and utility.



(a) $\delta = 0.01$, clipping constant$= 1$ (b) $\varepsilon = 1, \delta = 0.01$

Figure 7: DP-Fed-mv-PPCA performance varying (a) the multiplicative constant for difference clipping and (b) ε. The MAE axes are in $\log_{10}$ scale.

5. Conclusions

In spite of the large amount of currently available multi-site biomedical data, we still lack of reliable analysis methods to be applied in multi-centric applications, in compliance with privacy. To tackle this challenge, Fed-mv-PPCA proposes a hierarchical generative model to perform data assimilation of federated heterogeneous multi-views data. The Bayesian approach allows to naturally handle statistical heterogeneity across centers and missing views in local datasets, while providing an interpretable model of data variability. We show that Fed-mv-PPCA can be further coupled with differential privacy. Compatibly with our Bayesian formulation, we provide formal privacy guarantees of the proposed federated learning scheme against potential private information leakage from the shared statistics. Our applications demonstrate that Fed-mv-PPCA is robust with respect to an increasing degree of heterogeneity across training centers, and provides high-quality data reconstruction, outperforming competitive methods in all scenarios. Moreover, when differential privacy is introduced, we provide an investigation of the method performance according to difference privacy budget scenarios. It is worth noting that three DP hyperparameters play a key role, and could affect the performance of DP-Fed-mv-PPCA: the privacy budget parameters (ϵ, δ), and the clipping constant multiplying $\sigma_{\widetilde{\boldsymbol{\theta}}}$. These parameters are tightly related and all contribute to determine the magnitude of the noise used for perturbing the updated difference $\overline{\Delta\boldsymbol{\theta}}$. Indeed, increasing either ε or δ, or reducing the multiplicative constant in the clipping mechanism, implies the addition of a smaller noise, hence the improvement of the overall utility of the global model. Nevertheless, ε and δ should be kept small enough to ensure a reasonable privacy preservation.

Further extensions of this work will focus on introducing sparsity on the reconstruction weights, to improve the robustness of the approach to non-informative dimensions and modalities. Moreover, in order to improve the robustness of DP-Fed-mv-PPCA, non-Gaussian data likelihood and priors could be introduced in the future, to better account for heavy-tailed distributions defined by outliers datasets and centers.

Acknowledgments

This work received financial support by the French government, through the 3IA Côte dAzur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002, and by the ANR JCJC project Fed-BioMed, ref. num. 19-CE45-0006-01. The authors are grateful to the OPAL infrastructure from Université Côte d'Azur for providing resources and support.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimers Association; Alzheimers Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; ElanPharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; andTransition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (`www.fnih.org`). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimers Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for NeuroImaging at the University of SouthernCalifornia.

Ethical Standards

The work follows appropriate ethical standards in conducting research and writing the manuscript, following all applicable laws and regulations regarding treatment of animals or human subjects.

Conflicts of Interest

The authors declare that they have no conflict of interests.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

Luigi Antelmi, Nicholas Ayache, Philippe Robert, and Marco Lorenzi. Sparse multi-channel variational autoencoder for the joint analysis of heterogeneous data. In Kamalika Chaud-

huri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 302–311. PMLR, 2019. URL http://proceedings.mlr.press/v97/antelmi19a.html.

R. Argelaguet, B. Velten, D. Arnol, S. Dietrich, T. Zenz, J. C. Marioni, F. Buettner, W. Huber, and O. Stegle. Multi-omics factor analysis-a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol*, 14(6):e8124, 2018.

Gauthier Chassang. The impact of the eu general data protection regulation on scientific research. *ecancermedicalscience*, 11, 2017.

John P Cunningham and Zoubin Ghahramani. Linear dimensionality reduction: Survey, insights, and generalizations. *The Journal of Machine Learning Research*, 16(1):2859–2900, 2015.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1322–1333, 2015.

Andrew Gelman, Jessica Hwang, and Aki Vehtari. Understanding predictive information criteria for bayesian models. *Statistics and computing*, 24(6):997–1016, 2014.

Robin C Geyer, Tassilo Klein, and Moin Nabi. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557*, 2017.

Arun Iyengar, Ashish Kundu, and George Pallis. Healthcare informatics and privacy. *IEEE Internet Computing*, 22(2):29–31, 2018.

Ian T Jolliffe. Principal components in regression analysis. In *Principal component analysis*, pages 129–155. Springer, 1986.

Joeri Kalter, Maike G Sweegers, Irma M Verdonck-de Leeuw, Johannes Brug, and Laurien M Buffart. Development and use of a flexible data harmonization platform to facilitate the harmonization of individual patient data for meta-analyses. *BMC research notes*, 12(1):164, 2019.

Diederik P Kingma and Max Welling. Stochastic gradient vb and the variational auto-encoder. In *Second International Conference on Learning Representations, ICLR*, volume 19, 2014.

Arto Klami, Seppo Virtanen, and Samuel Kaski. Bayesian canonical correlation analysis. *Journal of Machine Learning Research*, 14(Apr):965–1003, 2013.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127*, 2018.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020.

A Llera and CF Beckmann. Estimating an inverse gamma distribution. *arXiv preprint arXiv:1605.01019*, 2016.

Toshihiko Matsuura, Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. Generalized bayesian canonical correlation analysis with missing modalities. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017a.

H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. *arXiv preprint arXiv:1710.06963*, 2017b.

Li Shen and Paul M Thompson. Brain imaging genomics: Integrated analysis and machine learning. *Proceedings of the IEEE*, 108(1):125–162, 2019.

Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

Aleksei Triastcyn and Boi Faltings. Federated learning with bayesian differential privacy. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 2587–2596. IEEE, 2019.

Yasi Wang, Hongxun Yao, and Sicheng Zhao. Auto-encoder based dimensionality reduction. *Neurocomputing*, 184:232–242, 2016.

Kang Wei, Jun Li, Ming Ding, Chuan Ma, Howard H Yang, Farhad Farokhi, Shi Jin, Tony QS Quek, and H Vincent Poor. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security*, 15:3454–3469, 2020.

Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Nghia Hoang, and Yasaman Khazaeni. Probabilistic federated neural matching. 2018.

Xinwei Zhang, Xiangyi Chen, Mingyi Hong, Zhiwei Steven Wu, and Jinfeng Yi. Understanding clipping for federated learning: Convergence and client-level differential privacy. *arXiv preprint arXiv:2106.13673*, 2021.

Jun Zhao, Teng Wang, Tao Bai, Kwok-Yan Lam, Zhiying Xu, Shuyu Shi, Xuebin Ren, Xinyu Yang, Yang Liu, and Han Yu. Reviewing and improving the gaussian mechanism for differential privacy. *arXiv preprint arXiv:1911.12060*, 2019.

Appendix A. Theoretical derivation of Fed-mv-PPC method

Problem setting

We consider C centers, each center $c \in \{1, \dots, C\}$ disposing of data from N_c subjects, each consisting of $K_c \leq K$ views of dimension d_k, $\sum_{k=1}^{K} d_k = d$.

For each k, c and each $n \in \{1, \dots, N_c\}$, the generative model is:

$$\mathbf{t}_{c,n}^{(k)} = W_c^{(k)} \mathbf{x}_{c,n} + \boldsymbol{\mu}_c^{(k)} + \boldsymbol{\varepsilon}_c^{(k)}, \tag{7}$$

where:

- $\mathbf{t}_{c,n}^{(k)} \in \mathbb{R}^{d_k}$ denotes the raw data of the k^{th}-view of the sample indexed by n in center c, which belongs to group g.
- $\mathbf{x}_{c,n} \sim \mathcal{N}(0, \mathbb{I}_q)$ is a q-dimensional latent variable, $q \leq \min_k(d_k)$ being a suitable user-defined latent-space dimension.
- $W_c^{(k)} \in \mathbb{R}^{d_k \times q}$ provides the linear mapping between the two sets of variables for the k^{th}-view.
- $\boldsymbol{\mu}_c^{(k)} \in \mathbb{R}^{d_k}$ allows data corresponding to view k to have a non-zero mean.
- $\boldsymbol{\varepsilon}_c^{(k)} \sim \mathcal{N}\left(0, {\sigma_c^{(k)}}^2 \mathbb{I}_{d_k}\right)$ is a Gaussian noise for the k^{th}-view.

A compact formulation for $\mathbf{t}_{c,n}$ (*i.e.* considering all views concatenated) can be easily derived from Equation (7):

$$\mathbf{t}_{c,n} = W_c \mathbf{x}_{c,n} + \boldsymbol{\mu}_c + \boldsymbol{\varepsilon}_c, \tag{8}$$

where:

- $\mathbf{t}_{c,n} = \left[{\mathbf{t}_{c,n}^{(1)}}^T, \dots, {\mathbf{t}_{c,n}^{(K)}}^T \right]^T \in \mathbb{R}^d$
- $W_c = \left[{W_c^{(1)}}^T, \dots, {W_c^{(K)}}^T \right]^T \in \mathbb{R}^{d \times q}$
- $\boldsymbol{\mu}_c = \left[{\boldsymbol{\mu}_c^{(1)}}^T, \dots, {\boldsymbol{\mu}_c^{(K)}}^T \right]^T \in \mathbb{R}^d$
- $\boldsymbol{\varepsilon}_c = \left[{\boldsymbol{\varepsilon}_c^{(1)}}^T, \dots, {\boldsymbol{\varepsilon}_c^{(K)}}^T \right]^T \sim \mathcal{N}(0, \Psi_c)$,

  where Ψ_c is a diagonal block-matrix, $\Psi_c = diag\left({\sigma_c^{(1)}}^2 \mathbb{I}_{d_1}, \dots, {\sigma_c^{(K)}}^2 \mathbb{I}_{d_K} \right)$

Note that for the sake of simplicity we represented all K views. If in center c the k^{th}-view is absent, than it will be simply removed, *e.g.* one would have:
$\mathbf{t}_{c,n} = \left[{\mathbf{t}_{c,n}^{(1)}}^T, \dots, {\mathbf{t}_{c,n}^{(k-1)}}^T, {\mathbf{t}_{c,n}^{(k+1)}}^T, \dots, {\mathbf{t}_{c,n}^{(K)}}^T \right]^T$, $\mathbf{t}_{c,n} \in \mathbb{R}^{d-d_k}$.

For each center c and each k we want to estimate $\boldsymbol{\theta}_c := \left\{\boldsymbol{\mu}_c^{(k)}, W_c^{(k)}, {\sigma_c^{(k)}}^2\right\}$ assuming that all local parameters are a realization of a common global distribution, to be estimated as well. The latter, provide a global model, which should be able to describe data across all centers.

Parameter $\boldsymbol{\mu}$

We assume that $\forall c, k$:

$$\boldsymbol{\mu}_c^{(k)} | \widetilde{\boldsymbol{\mu}}^{(k)}, \sigma^2_{\widetilde{\boldsymbol{\mu}}^{(k)}} \sim \mathcal{N}\left(\widetilde{\boldsymbol{\mu}}^{(k)}, \sigma^2_{\widetilde{\boldsymbol{\mu}}^{(k)}} \mathbb{I}_{d_k}\right) \tag{9}$$

Step 1. *(In each center)*: Estimate $\boldsymbol{\mu}_c^{(k)}[s+1]$ given $\left(\widetilde{\boldsymbol{\mu}}^{(k)}, \sigma^2_{\widetilde{\boldsymbol{\mu}}^{(k)}}\right)[s]$ (iteration s is denoted by $[s]$).

From Equation (7), the marginal distribution of $\mathbf{t}_{c,n}^{(k),g}$ is:

$$\mathbf{t}_{c,n}^{(k)} \sim \mathcal{N}(\boldsymbol{\mu}_c^{(k)}, C_c^{(k)}),$$

where $C_c^{(k)} = W_c^{(k)} {W_c^{(k)}}^T + {\sigma_c^{(k)}}^2 \mathbb{I}_{d_k}$, $C_c^{(k)} \in \mathbb{R}^{d_k \times d_k}$.

The corresponding log-likelihood gives:

$$\mathcal{L}_c^{(k)} = -\frac{1}{2}\left\{N_c d_k \ln(2\pi) + N_c \ln|C_c^{(k)}| + \sum_{n=1}^{N_c}\left(\mathbf{t}_{c,n}^{(k)} - \boldsymbol{\mu}_c^{(k)}\right)^T \left(C_c^{(k)}\right)^{-1}\left(\mathbf{t}_{c,n}^{(k)} - \boldsymbol{\mu}_c^{(k)}\right)\right\} \tag{10}$$

Therefore, for each center c and for all $k \in \{1, \ldots, K\}$, the following optimization problem should be considered:

$$\max_{\boldsymbol{\mu}_c^{(k)}} \mathcal{L}_c^{(k)} + \ln p\left(\boldsymbol{\mu}_c^{(k)}\right),$$

where:

$$\ln p\left(\boldsymbol{\mu}_c^{(k)}\right) = -\frac{1}{2\sigma^2_{\widetilde{\boldsymbol{\mu}}^{(k)}}}\left(\boldsymbol{\mu}_c^{(k)} - \widetilde{\boldsymbol{\mu}}^{(k)}\right)^T \left(\boldsymbol{\mu}_c^{(k)} - \widetilde{\boldsymbol{\mu}}^{(k)}\right) + const,$$

where $const$ collects terms which are independents from $\boldsymbol{\mu}_c^{(k)}$. We obtain:

$$\mu_c^{(k)}[s+1] = \left[N_c \mathbb{I}_{d_k} + \frac{1}{\sigma^2_{\widetilde{\boldsymbol{\mu}}^{(k)}}[s]} C_c^{(k)}\right]^{-1}\left[\sum_{n=1}^{N_c} \mathbf{t}_{c,n}^{(k)} + \frac{1}{\sigma^2_{\widetilde{\boldsymbol{\mu}}^{(k)}}[s]} C_c^{(k)} \widetilde{\boldsymbol{\mu}}^{(k)}[s]\right] \tag{11}$$

Step 2. *(In the master)*: Estimate $\left(\widetilde{\boldsymbol{\mu}}^{(k)}[s+1], \sigma^2_{\widetilde{\boldsymbol{\mu}}^{(k)}}\right)[s+1]$ given $\boldsymbol{\mu}_c^{(k)}[s+1]$ for all c.

Using (9), we obtain the following log-likelihood:

$$\mathcal{L} = \sum_{c=1}^{C} \ln p(\boldsymbol{\mu}_c^{(k)}) = \sum_{c=1}^{C}\left\{const - \frac{1}{\sigma^2_{\widetilde{\boldsymbol{\mu}}^{(k)}}} \|\boldsymbol{\mu}_c^{(k)} - \widetilde{\boldsymbol{\mu}}^{(k)}\|^2\right\} \tag{12}$$

By imposing $\partial_{\left(\widetilde{\boldsymbol{\mu}}^{(k)},\sigma^2_{\widetilde{\boldsymbol{\mu}}^{(k)}}\right)}((12)) = 0$ we obtain:

$$\widetilde{\boldsymbol{\mu}}^{(k)}[s+1] = \frac{1}{C}\sum_{c=1}^{C}\boldsymbol{\mu}_c^{(k)}[s+1] \tag{13}$$

and

$$\sigma^2_{\widetilde{\boldsymbol{\mu}}^{(k)}}[s+1] = \frac{1}{Cd_k}\sum_{c=1}^{C}\left\|\boldsymbol{\mu}_c^{(k)}[s+1] - \widetilde{\boldsymbol{\mu}}^{(k)}[s+1]\right\|^2 \tag{14}$$

Complete-data log-likelihood

From Equations (7)-(8) one can derive the following marginal distributions:

$$\mathbf{t}_{c,n}^{(k)}|\mathbf{x}_{c,n} \sim \mathcal{N}\left(W_c^{(k)}\mathbf{x}_{c,n} + \boldsymbol{\mu}, {\sigma_c^{(k)}}^2\mathbb{I}_{d_k}\right)$$

and

$$\mathbf{x}_{c,n}|\mathbf{t}_{c,n} \sim \mathcal{N}\left(\Sigma_c^{-1}B_c(\mathbf{t}_{c,n} - \boldsymbol{\mu}_c), \Sigma_c^{-1}\right),$$

where:

- $\Sigma_c := (\mathbb{I}_q + W_c^T\Psi_c^{-1}W_c) = \left(\mathbb{I}_q + \sum_{k=1}^{K}\frac{1}{\left(\sigma_c^{(k)}\right)^2}{W_c^{(k)}}^T W_c^{(k)}\right) \in \mathbb{R}^{q\times q}$
- $B_c := {W_c}^T\Psi_c^{-1} = \left[\frac{{W_c^{(1)}}^T}{\left(\sigma_c^{(1)}\right)^2}\ldots,\frac{{W_c^{(K)}}^T}{\left(\sigma_c^{(K)}\right)^2}\right] \in \mathbb{R}^{q\times d}$

Hence:

- $\langle\mathbf{x}_{c,n}\rangle = \Sigma_c^{-1}B_c(\mathbf{t}_{c,n} - \boldsymbol{\mu}_c)$
- $\langle\mathbf{x}_{c,n}\mathbf{x}_{c,n}^T\rangle = \Sigma_c^{-1} + \langle\mathbf{x}_{c,n}\rangle\langle\mathbf{x}_{c,n}\rangle^T$

The joint distribution of $\mathbf{t}_{c,n}$ and $\mathbf{x}_{c,n}$ follows ($p(\mathbf{t}_{c,n},\mathbf{x}_{c,n}) = p(\mathbf{t}_{c,n}|\mathbf{x}_{c,n})p(\mathbf{x}_{c,n})$), hence the expectation of the complete-data log-likelihood for each center c with respect to $p(\mathbf{x}_{c,n}|\mathbf{t}_{c,n})$:

$$\begin{aligned}\langle\mathcal{L}_{Cc}\rangle \;=\; -\sum_{n=1}^{N_c}\Bigg\{\sum_{k=1}^{K}\Bigg[&\frac{d_k}{2}\ln\left({\sigma_c^{(k)}}^2\right) + \frac{1}{2{\sigma_c^{(k)}}^2}\|\mathbf{t}_{c,n}^{(k)} - \boldsymbol{\mu}_c^{(k)}\|^2 + \frac{1}{2{\sigma_c^{(k)}}^2}tr\left({W_c^{(k)}}^T W_c^{(k)}\langle\mathbf{x}_{c,n}\mathbf{x}_{c,n}^T\rangle\right)\\ &-\frac{1}{{\sigma_c^{(k)}}^2}\langle\mathbf{x}_{c,n}\rangle^T{W_c^{(k)}}^T\left(\mathbf{t}_{c,i}^{(k),g} - \boldsymbol{\mu}_c^{(k)}\right)\Bigg] + \frac{1}{2}tr\left(\langle\mathbf{x}_{c,n}\mathbf{x}_{c,n}^T\rangle\right)\Bigg\},\end{aligned} \tag{15}$$

Parameter W

We assume that $\forall c, k$:

$$W_c^{(k)}|\widetilde{W}^{(k)}, \sigma^2_{\widetilde{W}^{(k)}} \sim \mathcal{MN}_{d_k,q}\left(\widetilde{W}^{(k)}, \mathbb{I}_{d_k}, \sigma^2_{\widetilde{W}^{(k)}}\mathbb{I}_q\right) \tag{16}$$

Step 1. ***(In each center)*****:** Estimate $W_c^{(k)}[s+1]$ given $\left(\widetilde{W}^{(k)}, \sigma^2_{\widetilde{W}^{(k)}}\right)[s]$.

For each center c, we consider the following optimization problem:

$$\max_{W_c^{(k)}} \langle \mathcal{L}_{Cc} \rangle + \ln p\left(W_c^{(k)}\right),$$

where $\ln p\left(W_c^{(k)}\right) = -\frac{1}{2\sigma^2_{\widetilde{W}^{(k)}}} tr\left(\|W_c^{(k)} - \widetilde{W}^{(k)}\|_2^2\right) + const.$

It follows:

$$W_c^{(k)}[s+1] = \left[\sum_{n=1}^{N_c} (\mathbf{t}_{c,n}^{(k)} - \boldsymbol{\mu}_c^{(k)}) \langle \mathbf{x}_{c,n} \rangle^T + \frac{{\sigma_c^{(k)}}^2}{\sigma^2_{\widetilde{W}^{(k)}}[s]} \widetilde{W}^{(k)}[s]\right] \left[\sum_{n=1}^{N_c} \langle \mathbf{x}_{c,n} \mathbf{x}_{c,n}^T \rangle + \frac{{\sigma_c^{(k)}}^2}{\sigma^2_{\widetilde{W}^{(k)}}[s]} \mathbb{I}_q\right]^{-1}$$

Step 2. ***(In the master)*****:** Estimate $\left(\widetilde{W}^{(k)}, \sigma^2_{\widetilde{W}^{(k)}}\right)[s+1]$ given $W_c^{(k)}[s+1]$ for all c. Proceeding as for parameter $\boldsymbol{\mu}$ and using (16):

$$\widetilde{W}^{(k)}[s+1] = \frac{1}{C} \sum_{c=1}^{C} W_c^{(k)}[s+1] \tag{17}$$

and

$$\sigma^2_{\widetilde{W}^{(k)}}[s+1] = \frac{1}{Cd_kq} \sum_{c=1}^{C} tr\left[\left(W_c^{(k)}[s+1] - \widetilde{W}^{(k)}[s+1]\right)^T \left(W_c^{(k)}[s+1] - \widetilde{W}^{(k)}[s+1]\right)\right] \tag{18}$$

Parameter σ^2

We assume that $\forall c, k$:

$${\sigma_c^{(k)}}^2 | \widetilde{\sigma}^{(k)^2} \sim \text{Inverse-Gamma}(\alpha^{(k)}, \beta^{(k)}), \tag{19}$$

so that:

$$Var\left({\sigma_c^{(k)}}^2\right) = \frac{{\beta^{(k)}}^2}{(\alpha^{(k)} - 1)^2(\alpha^{(k)} - 2)} := \widetilde{\sigma}^{(k)^2} \tag{20}$$

Step 1. ***(In each center)*****:** Estimate ${\sigma_c^{(k)}}^2[s+1]$ given $\left(\alpha^{(k)}, \beta^{(k)}\right)[s]$.

For each center c, we consider the following optimization problem:

$$\max_{{\sigma_c^{(k)}}^2} \langle \mathcal{L}_{Cc} \rangle + \ln p\left({\sigma_c^{(k)}}^2\right),$$

where $\ln p\left({\sigma_c^{(k)}}^2\right) = -(\alpha^{(k)} + 1) \ln\left({\sigma_c^{(k)}}^2\right) - \frac{\beta^{(k)}}{{\sigma_c^{(k)}}^2} + const$:

It follows:

$$\left(\sigma_c^{(k)}\right)^2[s+1] = \frac{1}{N_c d_k + 2(\alpha^{(k)}[s]+1)}\left\{\sum_{n=1}^{N_c}\left[\|t_{c,n}^{(k)} - \mu_c^{(k)}\|^2 + tr\left(W_c^{(k)^T} W_c^{(k)} \langle x_{c,n} x_{c,n}^T\rangle\right)\right.\right.$$
$$\left.\left. -2\langle x_{c,n}\rangle^T W_c^{(k)^T}\left(t_{c,n}^{(k)} - \mu_c^{(k)}\right)\right] + 2\beta^{(k)}[s]\right\} \quad (21)$$

Step 2. ***(In the master)*:** Estimate $\left(\alpha^{(k)}, \beta^{(k)}\right)[s+1]$ given ${\sigma_c^{(k)}}^2[s+1]$ for all c.

In order to estimate the parameters of the inverse-gamma distribution, we use the (ML1) method described by Llera and Beckmann Llera and Beckmann (2016).

Appendix B. Supplementary Tables and Figures

Table 3: Demographics of the clinical sample from the Alzheimer's Disease Neuroimaging Initiative (ADNI).

Group	Sex	Count	Age	Range
AD	Female	94	71.58 (7.59)	55.10 - 90.30
	Male	113	74.37 (7.19)	55.90 - 89.30
NL	Female	58	73.76 (4.61)	65.10 - 84.70
	Male	46	75.39 (6.58)	59.90 - 85.60

Table 4: Data Types.

View	Dim.	Description
CLINIC	7	Cognitive assessments
MRI	41	Magnetic resonance imaging
FDG	41	Fluorodeoxyglucose-Positron Emission Tomography (PET)
AV45	41	AV45-Amyloid PET

Table 5: Latent Space Dimension Assessment.

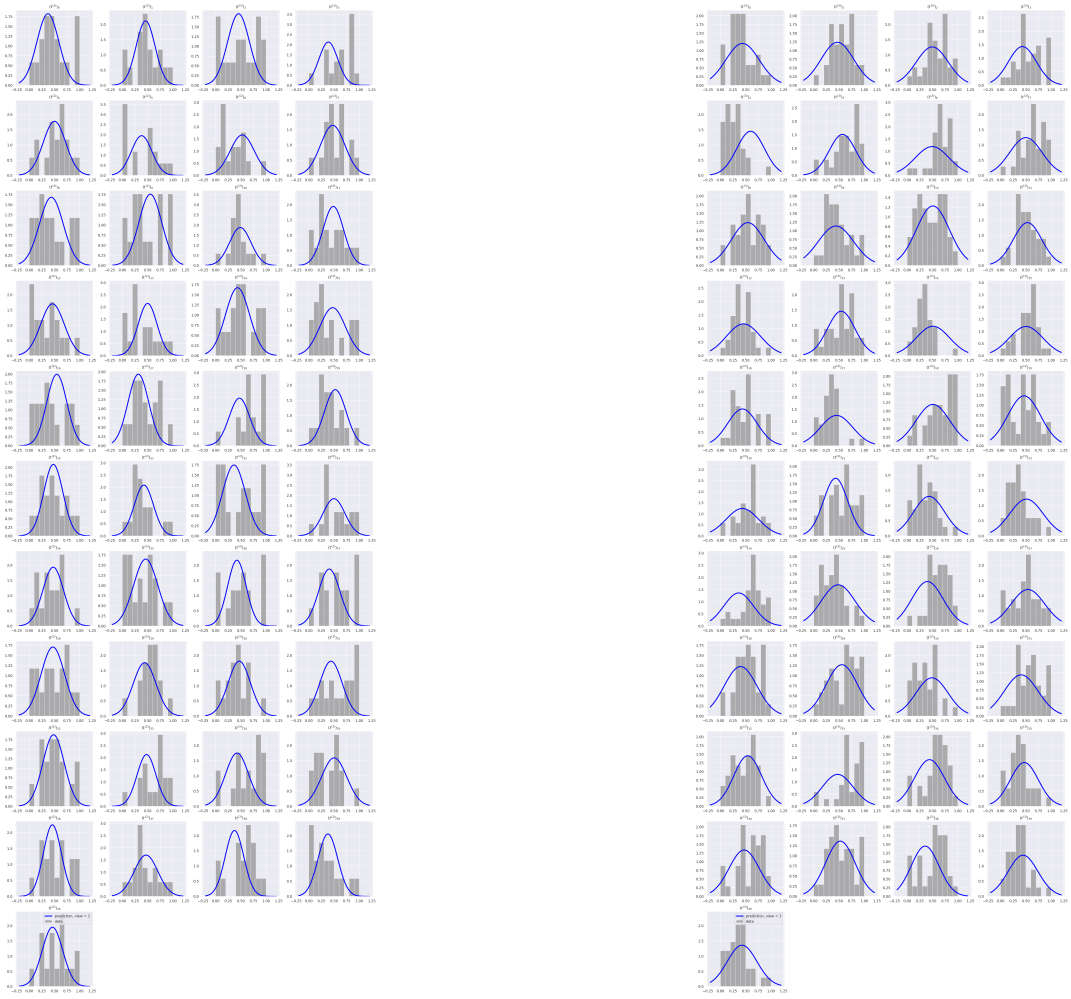
	SD			ADNI		
q	**WAIC**	**MAE Train**	**MAE Test**	**WAIC**	**MAE Train**	**MAE Test**
2	-2911	0.0886±0.0031	0.0879±0.0024	-4916	0.1240±0.0017	0.1249±0.0011
3	-3954	0.0640±0.0029	0.0662±0.0038	-5275	0.1170±0.0028	0.1187±0.0023
4	-4725	0.0450±0.0036	0.0485±0.0042	-6088	0.1113±0.0016	0.1142±0.0009
5	**−5114**	0.0327±0.0038	0.0375±0.0054	-6915	0.1064±0.0017	0.1102±0.0007
6	-3688	0.0313±0.0032	0.0366±0.0052	**−7250**	0.1028±0.0014	0.1073±0.0005
7	5722	0.0320±0.0036	0.0373±0.0053	-	-	-

Table 6: Results on SD dataset for all scenarios, and comparison with VAE and mc-VAE.

Scenario	Centers	Method	MAE Train	MAE Test	Accuracy in LS
IID	1 (centralized case)	**Fed-mv-PPCA**	**0.0124±3.e^{-5}**	**0.0405±0.0037**	1±0
		VAE	0.0851±0.0039	0.1011±0.0048	1±0
		mc-VAE	0.1236±0.0099	0.1382±0.0087	1±0
	3	**Fed-mv-PPCA**	**0.0320±0.0024**	**0.0373±0.0035**	1±0
		DP-Fed-mv-PPCA	0.0858±0.0111	0.0848±0.0099	1±0
		VAE	0.0683±0.0073	0.0702±0.0073	1±0
		mc-VAE	0.1172±0.0030	0.1146±0.0046	1±0
	6	**Fed-mv-PPCA**	**0.0422±0.0052**	**0.0371±0.0039**	1±0
		DP-Fed-mv-PPCA	0.0843±0.0093	0.0738±0.0076	1±0
		VAE	0.0769±0.0093	0.0680±0.0080	1±0
		mc-VAE	0.1295±0.0055	0.1134±0.0030	1±0
G	3	**Fed-mv-PPCA**	**0.0432±0.0074**	**0.0433±0.0026**	**0.9930±0.0093**
		DP-Fed-mv-PPCA	0.0960±0.0151	0.0951±0.0144	0.9873±0.0176
		VAE	0.0787±0.0135	0.0698±0.0082	0.9835±0.0272
		mc-VAE	0.1562±0.0086	0.1497±0.0076	0.9732±0.0512
	6	**Fed-mv-PPCA**	**0.0538±0.0101**	**0.0420±0.0048**	**0.9995±0.0019**
		DP-Fed-mv-PPCA	0.0945±0.0129	0.0813±0.0114	1±0
		VAE	0.0891±0.0148	0.0685±0.0063	0.9918±0.0428
		mc-VAE	0.1758±0.0154	0.1495±0.0112	0.9607±0.0398
K	3	Fed-mv-PPCA	0.0320±0.0052	0.0455±0.0069	1±0
		DP-Fed-mv-PPCA	0.0922±0.0137	0.1048±0.0151	1±0
	6	Fed-mv-PPCA	0.0402±0.0065	0.0448±0.0088	1±0
		DP-Fed-mv-PPCA	0.0959±0.0105	0.1014±0.0119	1±0
G/K	3	Fed-mv-PPCA	0.0395±0.0068	0.0567±0.0108	0.7812±0.02179
		DP-Fed-mv-PPCA	0.1144±0.0215	0.1343±0.0235	0.7852±0.0526
	6	Fed-mv-PPCA	0.0499±0.0104	0.0575±0.0128	0.7785±0.0222
		DP-Fed-mv-PPCA	0.1070±0.0139	0.1119±0.0144	0.7887±0.0449

Table 7: Results on ADNI dataset for all scenario G using VAE (resp. mc-VAE), and FedProx as aggregation scheme with the proximal term λ varying from 0.01 to 0.5.

Centers	Method	λ	MAE Train	MAE Test	Accuracy in LS
3	VAE	0 (FedAvg)	0.1172±0.0022	0.1192±0.0015	0.8289±0.0383
		0.01	0.1209±0.0074	0.1215±0.0013	0.7962±0.0438
		0.05	0.1215±0.0076	0.1218±0.0015	0.8009±0.0425
		0.1	0.1214±0.0075	0.1220±0.0018	0.8067±0.0399
		0.2	0.1218±0.0077	0.1221±0.0016	0.7977±0.0469
		0.3	0.1212±0.0075	0.1216±0.0017	0.7865±0.0443
		0.4	0.1212±0.0074	0.1217±0.0014	0.8033±0.0355
		0.5	0.1214±0.0077	0.1218±0.0020	0.7878±0.0420
	mc-VAE	0 (FedAvg)	0.1602±0.0035	0.1567±0.0017	0.8850±0.0262
		0.01	0.1674±0.0155	0.1605±0.0028	0.8185±0.0494
		0.05	0.1667±0.0153	0.1604±0.0028	0.8156±0.0444
		0.1	0.1674±0.0154	0.1609±0.0022	0.8249±0.0399
		0.2	0.1676±0.0156	0.1610±0.0025	0.8217±0.0431
		0.3	0.1676±0.0157	0.1610±0.0029	0.8184±0.0511
		0.4	0.1673±0.0155	0.1607±0.0021	0.8275±0.0426
		0.5	0.1679±0.0157	0.1613±0.0025	0.8229±0.0408
6	VAE	0 (FedAvg)	0.1357±0.0042	0.1191±0.0014	0.8224±0.0377
		0.01	0.1400±0.0114	0.1198±0.0022	0.7804±0.0470
		0.05	0.1403±0.0115	0.1203±0.0021	0.7827±0.0411
		0.1	0.1406±0.0116	0.1205±0.0019	0.7847±0.0531
		0.2	0.1407±0.0117	0.1207±0.0018	0.7837±0.0433
		0.3	0.1404±0.0115	0.1207±0.0018	0.7837±0.0569
		0.4	0.1405±0.0116	0.1203±0.0020	0.7753±0.0546
		0.5	0.1406±0.0113	0.1205±0.0023	0.7776±0.0501
	mc-VAE	0 (FedAvg)	0.1840±0.0054	0.1563±0.0017	0.8894±0.0230
		0.01	0.1932±0.0220	0.1596±0.0019	0.8140±0.0420
		0.05	0.1927±0.0219	0.1592±0.0016	0.8101±0.0484
		0.1	0.1932±0.0221	0.1595±0.0022	0.8043±0.0399
		0.2	0.1930±0.0219	0.1596±0.0020	0.8066±0.0441
		0.3	0.1931±0.0221	0.1595±0.0019	0.8217±0.0453
		0.4	0.1931±0.0220	0.1594±0.0018	0.8111±0.0419
		0.5	0.1934±0.0221	0.1596±0.0022	0.8021±0.0581



(a) MRI (b) FDG

Figure 8: Global distribution of all features of missing views in the Test dataset, for the G/K scenario. In this scenario, 1/3 of all subjects in the Test dataset do not dispose of MRI data and 1/3 do not dispose of FDG data. In both figure, the blue curve denotes the predicted global distribution of all features of the (a) MRI view and (b) the FDG view. Gray histograms correspond to real data in the Test dataset.

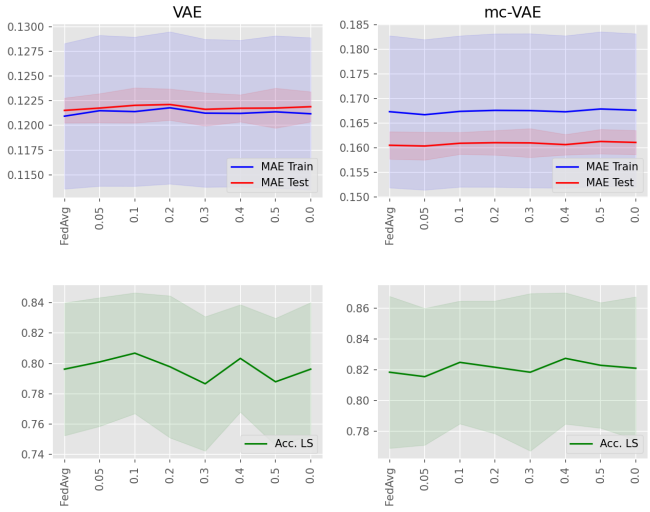


Figure 9: ADNI data, scenario G with 3 centers. MAE and accuracy in the latent space varying the FedProx proximal parameter λ.