2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing (HPSC), and IEEE Intl Conference on Intelligent Data and Security (IDS)

Privacy-Preserving Predictive Model Using Factor Analysis for Neuroscience Applications

Suprateek Kundu
Department of Biostatistics and Bioinformatics
Emory University, Atlanta, GA 30322
Email: suprateek.kundu@emory.edu

Shan Suthaharan
Department of Computer Science
UNC-Greensboro, Greensboro, NC 27407
Email: s_suthah@uncg.edu

***Abstract*—The purpose of this article is to present an algorithm which maximizes prediction accuracy under a linear regression model while preserving data privacy. This approach anonymizes the data such that the privacy of the original features is fully guaranteed, and the deterioration in predictive accuracy using the anonymized data is minimal. The proposed algorithm employs two stages: the first stage uses a probabilistic latent factor approach to anonymize the original features into a collection of lower dimensional latent factors, while the second stage uses an optimization algorithm to tune the anonymized data further, in a way which ensures a minimal loss in prediction accuracy under the predictive approach specified by the user. We demonstrate the advantages of our approach via numerical studies and apply our method to high-dimensional neuroimaging data where the goal is to predict the behavior of adolescents and teenagers based on functional magnetic resonance imaging (fMRI) measurements.**

I. Introduction

In neuroscience applications, linear regression models have been used as predictive models to address many problems using functional magnetic resonance imaging (fMRI) data from characterizing subjective pain intensity to predicting generalized anxiety disorders [1], [2]. As a motivating application, we consider predicting the generalized anxiety disorder (GAD) scores for children based on fMRI measurements in this paper [3]. In this application, the original fMRI measurements may encode information about the mental health status and clinical characteristics of a person, which may reveal privacy information of the person; hence, it needs to be protected.

Therefore, the fMRI measurements must be subject to a privacy preserving transformation, when shared with the users, in a manner that ensures that they can still be used to accurately predict the GAD scores for the test sample, while being protected. In addition to the transformed fMRI measurements, the user is also provided with the original GAD scores for the training dataset, but these scores are not released for subjects in the test data. This is because we want to protect the original GAD scores for the test subjects as they may reveal the cognitive or mental health status of the test subjects – this is the data sharing scenario that we considered in this paper.

Hence, our main goal is to present a computation approach that transforms the standard linear regression model into a privacy-preserving linear regression model. This transformation can benefit the application that involves brain imaging data in neuroscience such as fMRI and positron emission tomography, where data is collected on a large number of brain volumetric pixels or voxels. Note that the brain imaging data can be defined as a big data system due to its volume, complexity, scalability, and data heterogeneity [4], and contains information about the functional and structural characteristics of the brain [5], reflecting the cognitive abilities, mental health status, and clinical characteristics for an individual [6], [7].

In data privacy, four classes of privacy preserving methods have been studied in general [8]. The first class involves anonymization models, which suppress partial information about subjects in the data, with a view to ensure that the identity of the subject cannot be discovered [9], [10]. The second class involves perturbation models [11], [12] which add random noise or use signal transformations (e.g., Fourier or wavelet transformation) to produce perturbed versions of the original data. The third class involves matrix decomposition [13] which map the original data into a different vector subspace, and then transmit this transformed data to the user for analysis. The fourth class is distributed privacy preserving data mining [14], where the original data is distributed across multiple databases, with none of the database having access to the full data. Typically, local estimates are computed based on each database, and these are then aggregated to obtain the an estimate which is representative of the overall data.

In spite of a rich literature on privacy preserving algorithms, it may not be straightforward to apply the existing approaches to many emerging big data application, including neuroscience and medical science; thus, further research is needed to tackle these emerging issues. For example, the majority of the current perturbation approaches transmit data to the third party user which has the same number of features as the original data, which may result in computational and storage challenges on local machines having minimal system specifications for high dimensional data, while illuminating the privacy weaknesses. Also, since every variable in the data can contribute to a pattern that may reveal sensitive information, the anonymization must involve all variables. However, this is challenging since a greater level of perturbation needs to be applied to a large number of features to maintain the same privacy level [27]. This led [9] to suggest that the privacy for high-dimensional features may be better preserved by operating in a transformed basis instead of the vector space of the original features.

In a big data environment, often the information about

978-1-7281-0006-7/19/$31.00 ©2019 IEEE
DOI 10.1109/BigDataSecurity-HPSC-IDS.2019.00023

the high dimensional features can be captured via a lower dimensional set of independent latent factors, which represent different sources of variation. In the context of brain imaging data, these sources may correspond to brain signal patterns representing clinical characteristics (such as presence of a tumor), cognitive abilities, decision making instances and so on. It suggests the possibility of implementing a factor analysis (FA) based approach for analyzing such complex, high-dimensional, and heterogeneous data. There is a well established statistics literature for factor analysis models, which uses lower dimensional latent factors to model high dimensional features.

Factor models have been traditionally applied in behavioral and social sciences and psychometrics [15], where the latent factors have a natural interpretation as certain unobserved psychological traits. The FA models presented in [16] and [17], adopt the sparse characterization for dimensionality reduction in gene expression studies where p is large and n is small. In brain imaging, the independent component analysis (ICA) approach [18], which is a variant of the latent factor models, is successfully used for brain activation and network analysis. This points to the promise of FA based methods for privacy protection in neuroimaging applications, although no such advances have been made so far, to our knowledge.

Our main goal in this paper is to propose a FA approach that projects the high-dimensional feature space to a lower dimensional space of latent factors, in a manner which minimizes the loss in predictive accuracy. Transforming the features to latent factors effectively preserves the privacy of the data such that it is very difficult to recover the original features. This is due to the fact that the proposed approach projects the original features into a lower dimensional subspace, and the number of features in the original dataset cannot be recovered. Once the latent factor model has been fit to the data, we propose an optimization approach which re-calibrates these latent factors in order to minimize the loss in predictive accuracy under a linear regression model [19] based on the transformed data. We conduct extensive numerical studies to evaluate the prediction accuracy and density estimation accuracy under the proposed approach, and we also apply it to a brain imaging application. The highlights of the proposed approach are as follows:

- It results in dimensionality reduction via a probabilistic factor analysis, by projecting high dimensional features to a lower dimensional subspace. Hence, it compresses the volume of data and reduces storage needs and computational burden for third party users, who can simply store and analyze the transformed data in local machines.
- It provides a Bayesian factor analysis approach [20] for dimension reduction, which is used to provide credible intervals for the predicted values of the dependent variables. These intervals characterize measures of uncertainty and provide richer information than obtained under a point estimate, and is particularly relevant for big data environments containing inherent heterogeneity.
- It results in prediction accuracy which is higher than the competing privacy-preserving methods that involve an alternate matrix factorization.
- The proposed approach is computationally efficient, scalable to high dimensions, and allows convenient storage for of the transmitted data for third party users, all of which are key features for applicability to big data.

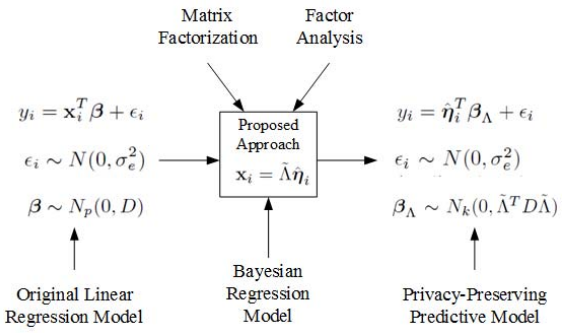

Fig. 1. An illustration of the proposed Bayesian approach that transforms a linear regression model into a privacy-preserving linear predictive model.

The rest of the paper is organized as follows: Section II presents the snapshot of our proposed approach, and introduces factor analysis models; Section III is dedicated to discussions regarding confidentiality protection of the proposed FA model; Section IV concerns the preservation of data utility under a linear regression model. An experiment analysis using simulated and real data sets is presented in Section V of this paper. Also note that, as a convention, we have used bold lowercase letters to represent vectors uppercase letters to represent matrices.

II. Proposed Approach

Fig. 1 provides a schematic diagram that illustrates how our approach transforms a linear regression model into a privacy-preserving linear regression model. This setup considers p covariates or features $\mathbf{x} = (x_1, \ldots, x_p)^T$, and a dependent variable y. The data for the i-th observation consists of the vector $(y_i, \mathbf{x}_i)$, where $\mathbf{x}_i \in \Re^p, i = 1, \ldots, n$, and each subject belongs to either the training data or the test dataset, as determined by the owner of the data or a trusted analyst having access to the original data. The objective here is to accurately predict the dependent variable for the test sample observations, based on transformed covariates designed to protect privacy.

Our main goal in this research work is to develop methods using parametrized linear regression models. In particular, we are interested in the following type of linear relationships between the dependent variable and the features:

$$y_i = \mathbf{x}_i^T\boldsymbol{\beta} + \epsilon_i,\ \epsilon_i \sim N(0, \sigma_e^2),\ \boldsymbol{\beta} \sim N(\mathbf{0}, \sigma_b^2 I_p), \tag{1}$$

where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$, ϵ_i are the error terms with residual variance σ_e^2, $\boldsymbol{\beta}$ are the $p \times 1$ dimensional covariate vector which captures the effect of the independent features on the outcome, and $N(\mathbf{0}, I_K)$ denotes a K dimensional Gaussian distribution with independently distributed components. Model (1) specifies a prior on the unknown regression coefficients $\boldsymbol{\beta}$ under a Bayesian approach, and uses the posterior distribution

to estimate $\boldsymbol{\beta}$. In this article, we treat σ_e^2 and σ_b^2 as tuning parameters which suffices for our privacy protection studies, although one could specify priors on them for a fully Bayesian treatment.

In the scope of this problem, the original features $\mathbf{x}$ can only be shared with a third party user after privacy preserving transformations, and data on the dependent variable can only be shared for the training data. Our goal is to transform the original features $\mathbf{x}$ into lower dimensional latent factors $\boldsymbol{\eta}$ under the approximation $\mathbf{x} = \Lambda\boldsymbol{\eta}$ using a Bayesian factor model. The factor loadings Λ, and the latent factors $\boldsymbol{\eta}$, are then re-calibrated to provide good prediction accuracy. The re-calibrated latent factors (but not the factor loadings) and the original values of the dependent variable are transmitted to the user as the training data, along with a test data containing the re-calibrated latent factors only. The user will also have a recommended model which can be fit to obtain an accurate prediction performance for the test dataset.

III. Confidentiality Protection via Matrix Decomposition

The key to the proposed algorithm is to project the high dimensional features onto a lower dimensional subspace of latent factors, with each such latent factor representing a potential source of variation. In particular, we propose the following exact equation of matrix decomposition:

$$\mathbf{x}_i = \tilde{\Lambda}\hat{\boldsymbol{\eta}}_i,\ i = 1, \ldots, n, \text{ or equivalently, } X^T = \tilde{\Lambda}\hat{E}, \tag{2}$$

where $\widehat{\boldsymbol{\eta}}_i$ represent the unknown K dimensional latent factors which encode the original dependent features with $\hat{E} = (\widehat{\boldsymbol{\eta}}_1, \ldots, \widehat{\boldsymbol{\eta}}_n)$, and $\tilde{\Lambda}$ represent factor loadings which relate the features to the latent factors. The above decomposition expresses the original vector of features as the weighted sum of confidential variables as $x_{ij} = \sum_{l=1}^{K} \tilde{\lambda}_{jl}\widehat{\eta_{li}}$. The number of latent factors K can either be pre-specified by the user or estimated in a data adaptive manner. The goal is to estimate the unknown quantities in (2) such that the loss in predictive accuracy is minimized when fitting the model using the transmitted data, which comprises of the latent factors $\widehat{\boldsymbol{\eta}}_1, \ldots, \widehat{\boldsymbol{\eta}}_n$. Another strong confidentiality feature is that it is not even possible to estimate the number of independent features in the original dataset from the latent factors.

One can potentially use existing matrix decomposition approaches to estimate $\tilde{\Lambda}$ and $\hat{\boldsymbol{\eta}}$ in (2). Using SVD, $X = V\Delta U$, where X is the $n \times p$ data matrix with rows as $\mathbf{x}_i^T, i = 1, \ldots, n,$, $V, U,$ are orthonormal matrices of dimension $n \times n$ and $p \times p$ respectively, and Δ is a rectangular diagonal matrix of dimension $n \times p$ containing elements as singular values. Using an approximation which involves K singular values, we have $\tilde{\mathbf{x}}_i^T = \mathbf{v}_i^T\Delta_K U_K$, where $\mathbf{v}_i^T$ denotes the i-th row of V, and $\Delta_K(n \times K)$ and $U_K(K \times p)$ are sub-matrices with the first K columns of Δ and U respectively. This translates to $\tilde{\mathbf{x}}_i = U_K^T\Delta_K^T\mathbf{v}_i = \Lambda\hat{\boldsymbol{\eta}}_i$, where $\Lambda = U_K^T$ and $\hat{\boldsymbol{\eta}}_i = \Delta_K^T\mathbf{v}_i$. Another possibility is to use a principal value decomposition (PCA), where the number of latent factors is fixed at $K = p$, and $\widehat{\boldsymbol{\eta}}_i = P\mathbf{x}_i, i = 1, \ldots, n$, with P denoting the matrix of eigen vectors corresponding to X^TX.

A. Factor Analysis Model

We propose a probabilistic factor analysis based approach which expresses the original features as a matrix decomposition in (2) along with an additive error term as follows

$$\mathbf{x}_i = \Lambda\boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i,\ \boldsymbol{\epsilon} \sim N(\mathbf{0}, \Sigma),\ \boldsymbol{\eta}_i \sim N(\mathbf{0}, I_K), \tag{3}$$

where $i = 1, \ldots, n$ and $\boldsymbol{\epsilon}_i$ is the vector of Gaussian residuals which add random perturbations to the matrix decomposition term in (3), $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$, and $\boldsymbol{\eta}$ are the K dimensional latent factors which are independently distributed Gaussian random variables. One can specify appropriate prior distributions on Λ and $\boldsymbol{\eta}$ under a Bayesian approach and obtain parameter estimates under these specifications. In particular, one can specify conjugate priors

$$\begin{aligned}\lambda_{jh} \mid (\phi_{jh}, \tau_h) &\sim N(0, \phi_{jh}^{-1}\tau_h^{-1}),\ \phi_{jh} \sim Ga(3/2, 3/2),\\ \tau_h &\sim Ga(a, 1),\ \sigma_j^{-2} \sim Ga(a_\sigma, b_\sigma),\ , j = 1, \ldots, p.\end{aligned} \tag{4}$$

which similar to the prior specification in [22], but without the provision for adaptive number of latent factors. We run a Markov Chain Monte Carlo (MCMC) [23], [24] to estimate model parameters $\hat{\Lambda}$ and $\widehat{\boldsymbol{\eta}}_1, \ldots, \widehat{\boldsymbol{\eta}}_n$, using posterior means.

One can potentially approximate the original features using $\hat{\mathbf{x}}_i^T = \hat{\Lambda}\hat{\boldsymbol{\eta}}_i, i = 1, \ldots, n$, where $\hat{\Lambda}$ denotes the posterior mean of Λ, and $\hat{\boldsymbol{\eta}}_i$ denotes the posterior mean of $\boldsymbol{\eta}_i, i = 1, \ldots, n$. However, instead of presenting the user with the perturbed features, we provide them with the latent factors $\hat{\boldsymbol{\eta}}_i, i = 1, \ldots, n$, after some recalibration which is described in the next section. Providing the third party users with the latent factors instead of the perturbed features $\hat{x}$ effectively suppresses information and makes it challenging for attackers to recover the original features. This is due to the fact that the original features can be expressed as a sum of a linear combination of the latent factors and a random error, i.e. $x_{im} = \sum_{l=1}^{K} \lambda_{ml}^*\eta_{li} + \epsilon_i, i = 1, \ldots, n, m = 1, \ldots, p$, and it is challenging to recover the true factor loadings λ^* without additional information, especially given the fact that the original number of features p is unknown. Moreover, incorporating an error term in (4) adds an additional random perturbation which ensures greater privacy of the original feature compared to the corresponding noise-free model.

It is also important to note that our transformed privacy-preserving linear regression model satisfies the theory of compressed sensing [25]; hence, as stated in [26], the signal $\mathbf{x}_i$ can be recovered only if it can be characterized by a sparse representation in the domain of a linear transform. Otherwise it is difficult by an attacker to recover the signal $\mathbf{x}_i$; hence, it also meets the privacy protection requirement.

B. Markov Chain Monte Carlo

In order to compute the parameter estimates under the Bayesian model (3-4), we adopt a MCMC approach [23], [24], which performs the parameter estimation processes by drawing random samples iteratively using posterior distributions.

C. Choosing the Number of Factors

The number of factors in the latent factor model (3) can either be fixed by the user, or chosen adaptively using the data. In our numerical experiments, we usually fix the number of factors as that which minimizes the test error. However, one can alternatively use the approach in Bhattacharya and Dunson in [22], who propose adaptive shrinkage priors on the elements of Λ such that they become increasingly degenerate at zero as the number of factors in the model increase. This implies, that increasing the number of factors after a certain point becomes inconsequential and does not lead to improvement in model fitting, since the factor loadings of those additional factors are effectively very close to zero. The authors in that paper propose a threshold which is able to effectively determine the number of factors in an unsupervised manner. We refer the user to the original article for full details.

IV. Optimizing Data Utility

The matrix factorization step using latent factor models is designed to protect the data privacy so that the original features can not be recovered. However, neither the latent factors, nor the factor loadings, are calibrated to provide good predictive performance. In this section we re-calibrate these parameters with a goal to obtain improved prediction performance, as demonstrated in the numerical experiments section.

A. Calibration of Factor Loadings

In order to provide a closer approximation and reduce the influence of the residual noise on the perturbed features, we rescale the factor loadings as follows

$$\tilde{\Lambda} = X^T \hat{E}^T (\hat{E}\hat{E}^T)^{-1} \hat{E}^T, \tag{5}$$

where $\hat{E}$ is a $k \times n$ matrix with columns as $\hat{\boldsymbol{\eta}}_1, \ldots, \hat{\boldsymbol{\eta}}_n$. Equation (5) results in the original matrix of features being approximated as $\tilde{X}^T = \tilde{\Lambda}\hat{E}$, which provides improved data utility. Clearly, the L_1 error between the original and approximated features is given by

$$|X^T - \tilde{X}^T| = |X^T - \tilde{\Lambda}\hat{E}| = |X^T(I - \hat{E}^T(\hat{E}\hat{E}^T)^{-1}\hat{E}^T)|, \tag{6}$$

where $|\cdot|$ denotes the element-wise L_1 norm. These rescaled factor loadings are used to recalibrate the latent factors in a manner which results in improved prediction accuracy while preserving accuracy.

B. Recommended Model Fitting Algorithm

In order to protect confidentiality, we provide the user with a transformed dataset containing latent factors which will minimize the loss in predictive accuracy compared to the analysis under the original dataset. Note that one can express model (1) in terms of the latent factors, by substituting $\tilde{\mathbf{x}} = \tilde{\Lambda}\hat{\boldsymbol{\eta}}$ which yields

$$y_i = \hat{\boldsymbol{\eta}}_i^T \tilde{\Lambda}^T \boldsymbol{\beta} + \epsilon_i = \hat{\boldsymbol{\eta}}_i^T \boldsymbol{\beta}_\Lambda + \epsilon_i, \tag{7}$$

where $\epsilon_i \sim N(0, \sigma_e^2)$ and $\boldsymbol{\beta}_\Lambda = \tilde{\Lambda}^T\boldsymbol{\beta} \sim N_k(0, \sigma_b^2 \tilde{\Lambda}^T\tilde{\Lambda})$. We note that $\tilde{\mathbf{x}}$ is only an approximation of $\mathbf{x}$, hence there will be some loss in predictive accuracy under model (7), with the degree of the loss being dependent on the the error $|\mathbf{x} - \tilde{\mathbf{x}}|$. The third party user can obtain the estimate for $\boldsymbol{\beta}_\Lambda$ using a *maximum a-posteriori* or MAP estimator, which minimizes the negative log-posterior, or equivalently, minimizes the objective function below:

$$\frac{1}{\sigma_e^2}\sum (y_i - \hat{\boldsymbol{\eta}}_i^T \boldsymbol{\beta}_\Lambda)^2 + \alpha_1 \boldsymbol{\beta}_\Lambda^T (\sigma_b^2 \tilde{\Lambda}^T\tilde{\Lambda})^{-1} \boldsymbol{\beta}_\Lambda, \tag{8}$$

with respect to $\boldsymbol{\beta}_\Lambda$. This estimate can then be used to predict the dependent variable as $\hat{y}_i = \hat{\boldsymbol{\eta}}_i^T \hat{\boldsymbol{\beta}}_\Lambda, i = 1, \ldots, n$. The MAP estimator which minimizes (8) resembles a ridge regression type estimator, and can be obtained in a closed form as

$$\hat{\boldsymbol{\beta}}_\Lambda = (\frac{1}{\sigma_e^2}\hat{E}^T\hat{E} + \alpha_1 \frac{1}{\sigma_b^2}(\tilde{\Lambda}^T\tilde{\Lambda})^{-1})^{-1} (\frac{1}{\sigma_e^2}\sum_{i=1}^{n} y_i \boldsymbol{\eta}_i). \tag{9}$$

However, the estimator $\hat{\boldsymbol{\beta}}_\Lambda$ may not be adequate for providing optimal prediction performance, since there is nothing to guarantee that the predicted values under (7) based on the latent features will be close to the predicted values under model (1) using the original data. In order to obtain good prediction performance, the predicted values under the transformed model (7), and the original model (1) should be close. Suppose $\hat{\boldsymbol{\beta}}_\Lambda^*$ is the value of $\boldsymbol{\beta}_\Lambda$ in (7), which ensures the same predictive performance as in (1). In other words,

$$\begin{aligned} \hat{\boldsymbol{\eta}}_i^T \hat{\boldsymbol{\beta}}_\Lambda^* &= \mathbf{x}_i^T \hat{\boldsymbol{\beta}} \text{ for all } i = 1, \ldots, n, \\ \Rightarrow \hat{\boldsymbol{\beta}}_\Lambda^* &= (\hat{E}\hat{E}^T)^{-1}\hat{E}X\hat{\boldsymbol{\beta}}, \end{aligned} \tag{10}$$

where $\hat{\beta}$ is the estimate for the regression coefficients in (1) under some model fitting algorithm (such as ordinary least squares, least absolute shrinkage and selection operator, elastic net, and so on), and $(\hat{E}\hat{E}^T)^{-1}$ is always computable as long as $K < n$, which is indeed the case for our applications involving a small to moderate number of factors. This relationship implies that $\hat{\boldsymbol{\beta}}_\Lambda$ should be constrained to be close to $\hat{\boldsymbol{\beta}}_\Lambda^*$ to achieve good prediction performance. The above can be achieved by augmenting the optimization function in (8) with a penalized term which enforces the relationship in (10) as

$$\begin{aligned} \text{arg min}_{\boldsymbol{\beta}_\Lambda} \sum_{i=1}^{n} (y_i - \hat{\boldsymbol{\eta}}_i^T \boldsymbol{\beta}_\Lambda)^2 + \alpha_1^* \boldsymbol{\beta}_\Lambda^T (\tilde{\Lambda}^T\tilde{\Lambda})^{-1} \boldsymbol{\beta}_\Lambda \\ + \alpha_2 |\boldsymbol{\beta}_\Lambda - \hat{\boldsymbol{\beta}}_\Lambda^*|, \end{aligned} \tag{11}$$

where $\alpha_1^* = \alpha_1 \sigma_e^2 / \sigma_b^2$ can be considered to be the ridge regression parameter. The first two terms in (11) are the same as in (8), however the third term imposes a L_1 penalty which forces $\boldsymbol{\beta}_\Lambda$ to be close to $\hat{\boldsymbol{\beta}}_\Lambda^*$. Equation (11) suggests that the third party user can obtain the predicted values as $\hat{\boldsymbol{\eta}}^T \hat{\boldsymbol{\beta}}_\Lambda$ using the transmitted dataset containing the latent factors, where $\hat{\boldsymbol{\beta}}_\Lambda$ is the solution to $\boldsymbol{\beta}_\Lambda$ in (11). Equation (11) can be written as

$$\begin{aligned} \text{arg min}_{\boldsymbol{\gamma}_\Lambda} \sum_{i=1}^{n} (y_i - \hat{\boldsymbol{\eta}}_i^T \hat{\boldsymbol{\beta}}_\Lambda^* - \hat{\boldsymbol{\eta}}_i^T \boldsymbol{\gamma}_\Lambda)^2 + \alpha_1^* \boldsymbol{\gamma}_\Lambda^T (\tilde{\Lambda}^T\tilde{\Lambda})^{-1} \boldsymbol{\gamma}_\Lambda \\ + \alpha_1^* \boldsymbol{\gamma}_\Lambda^T (\tilde{\Lambda}^T\tilde{\Lambda})^{-1} \hat{\boldsymbol{\beta}}_\Lambda^* + \alpha_2 |\boldsymbol{\gamma}_\Lambda|, \end{aligned} \tag{12}$$

where $\boldsymbol{\gamma}_\Lambda = \boldsymbol{\beta}_\Lambda - \hat{\boldsymbol{\beta}}_\Lambda^*$, and α_2, α_1^*, are the regularization parameters for ridge regression and lasso [19]. The form

of the objective function (12) appears to be non-standard, and it is not clear it can be solved using some existing algorithms. It turns out that (12) can be re-framed and solved using Lasso, which is detailed in the following Theorem. Let $\tilde{\Lambda}_{Ic}^T$ denote the $K \times K$ Cholesky factor for $(\tilde{\Lambda}^T\tilde{\Lambda})^{-1}$, $\mathcal{Y} = (\mathbf{y}^T, \tilde{\Lambda}_{Ic}^T\hat{\boldsymbol{\beta}}_{\Lambda}^*)^T$, $\hat{E}^*$ denote the $K \times n$ matrix with the i-th column as $\hat{\boldsymbol{\eta}}_i^* = (\tilde{\Lambda}_{Ic}^T)^{-1}\hat{\boldsymbol{\eta}}_i$, and denote $\mathcal{X}_{\alpha_1^*} = (\hat{E}^*, \sqrt{\alpha_1^*}I_K)^T$. Further denote by $||\cdot||^2$ the squared Euclidean norm.

Theorem I The solution in (12) can be obtained by equivalently minimizing the function $||\mathcal{Y} - \mathcal{X}_{\alpha_1^*}\boldsymbol{\gamma}_{\Lambda}^*||^2 + \alpha_2^*|\boldsymbol{\gamma}_{\Lambda}^*|$ with respect to $\boldsymbol{\gamma}_{\Lambda}^*$, where $\alpha_2^* = \alpha_2/det(\tilde{\Lambda}_{Ic}^T)$.

Theorem I suggests that the minimizer of (12) can be obtained via a Lasso type solution, for which many existing packages (such as *glmnet* in R) are available.

C. Transmitted Data Containing Rescaled Latent Factors

The user is supplied with the transformed dataset and asked to fit a lasso regression model corresponding to the covariate matrix $X_{\eta,\lambda_2}^* = \left(\hat{E}^*, \sqrt{\alpha_1}I_K\right)^T$ and the modified response vector $\mathcal{Y}$ for a series of positive α_1 values, so as to obtain series of optimal predicted values. In our experience, small values of α_2 provide the greatest reduction in prediction error. Note that none of the original features are transmitted and the third party user has no information about the dimension of the original feature space and hence it is very difficult to recover the original features from the transmitted data.

D. Reporting Credible Intervals

The Bayesian factor analysis model (3)-(4) provides a collection of MCMC samples for $\hat{\boldsymbol{\eta}}_i, i = 1, \ldots, n$, as the output. For the mth MCMC sample with the value of the latent factor as $\boldsymbol{\eta}_{test}^{(m)}$ corresponding to the test sample observation $\mathbf{x}_{test}$, we can fit the above approach to obtain predicted value $\hat{y}_{test}^{(m)}$. One can then use the predicted values $\hat{y}_{test}^{(B+1)}, \ldots, \hat{y}_{test}^{(M)}$ over all the MCMC samples after burn-in B, to construct prediction credible intervals. In addition to the predicted value $\hat{y}_{test}$, the 95% credible intervals provide a measure of uncertainty that is often important in applications such as neuroimaging analysis.

V. Experimental Analysis

We performed a variety of numerical studies - using simulated and real data - to evaluate the performance of our approach, and compared our approach to a few other competing methods. In particular, we look at the percentage loss in predictive accuracy under the privacy preservation, compared to a predictive model involving ridge regression using the original dataset. This is computed by expressing the difference in the mean squared error (MSE) under a ridge regression model using the original data and the privacy preserving approach, as a percentage of the MSE under the former. The MSE is a commonly used metric for measuring prediction accuracy – it measures the squared L_2 norm between the predicted values and the true samples. It is computed as

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2,$$

where $\hat{\mu}_i$ is the fitted value of the observations for the i-th sample that is represented by y_i.

A. Experiments with Simulated Data

In this experiment we studied the predictive accuracies for different sample sizes n, varying dimensions p, and different number of factors K. We used 20,000 MCMC iterations with a burn in of 5000, for fitting the Bayesian factor model (3).

We compared the predictive performance of the proposed approach with that of two other privacy preserving methods. The first method is based on factor analysis, where the Bayesian latent factor model (3) is fit to the features x to obtain posterior estimates of factor loadings ($\hat{\Lambda}$) and latent factors ($\hat{\boldsymbol{\eta}}$), using MCMC samples; then a perturbed version of the original independent features is computed as $\hat{\mathbf{x}}_i = \hat{\Lambda}\hat{\boldsymbol{\eta}}_i, i = 1, \ldots, n$. The second competing approach involves a singular value decomposition, where the perturbed independent features are obtained using the first k singular values, using the method in [9]. The number of singular values is chosen to be the same as the number of factors used in the proposed approach, to ensure a fair comparison. Once these perturbed independent features ($\hat{x}$) are obtained under these two competing approaches, we then report the predictive accuracy under a ridge regression model using the perturbed dataset $(y_i, \hat{\mathbf{x}}_i), i = 1, \ldots, n$.

For the simulations that we performed, we generated data using the following linear regression model:

$$y_i = \mathbf{x}_i^T\boldsymbol{\beta}_0 + \epsilon_i,\ \mathbf{x}_i \sim N_p[0, \Lambda_0\Lambda_0^T + \text{diag}(\sigma_{1*}^2, \ldots, \sigma_{p*}^2)],$$

where Λ_0 has dimension $p \times k_0$, and $\boldsymbol{\beta}_0 = (1, 1, 1, 1, 1, \mathbf{0})$ is a $p \times 1$ vector with the first five elements as 1 and the rest as 0. For our simulations, we choose $p = 10,000$ features. The above represents a linear regression model, where the independent features are generated from a latent factor model having k_0 factors, which was fixed to be equal to 10. When reporting the results under the proposed approach, the number of latent factors was adaptively chosen using the method in multiplicative Gamma method in [19] as highlighted in Section III. The results under the alternate approaches was also reported corresponding to the number of factors chosen under the proposed method using the multiplicative Gamma prior approach. For reporting the predictive test MSE corresponding, we split the data into a training and test set using a 70-30 ratio. We note that the goal of our analysis is not to estimate the latent factors, or recover the true regression coefficients $\boldsymbol{\beta}_0$, but rather to predict the dependent variable with good accuracy.

Fig. (2) shows the decrease in predictive accuracy loss for the proposed approach and the other two competing approaches, as the sample size n is increased, but the number of independent features is fixed ($p = 10,000$). The number of factors (k) chosen adaptively as described in the preceding paragraph. From the Figure, it is clear that the proposed

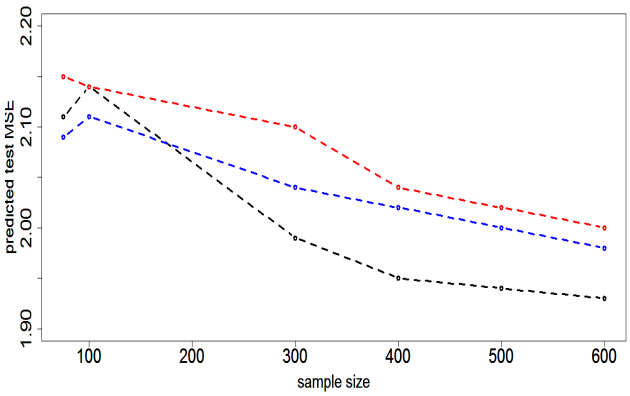

Fig. 2. Decrease of predictive MSE when sample size increases. The black, blue and red lines correspond to the proposed method, the SVD-based method and the Bayesian 2-stage approach.

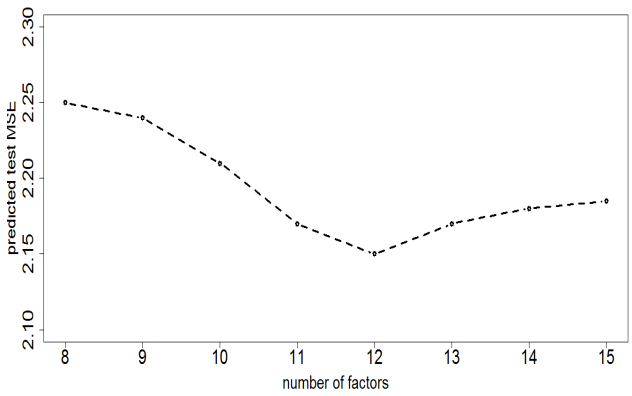

Fig. 3. Predictive test MSE over varying number of factors for the proposed approach.

approach maintains a lower test MSE and hence has a higher predictive accuracy consistently over all sample sizes. Fig. (3) shows how the predictive accuracy changes as the number of factors k is changed for a given sample size ($n = 100$) and feature dimension ($p = 10,000$). In particular, we manually pre-specify the number of latent factors and fit the proposed model and the alternative approaches corresponding to a varying number of latent factors. For the dimensions that we consider, we find that the predictive MSE decreases (i.e. predictive accuracy increases) as the number of factors is increased, but eventually the predictive MSE plateaus off when the number of factors is increased beyond a certain threshold. In addition, we also observe that the lowest MSE across varying number of factor is achieved when $k = 12$ which is different than the true number of factors (10) used for generating the data. The simulation study clearly illustrates the advantages of the proposed approach in terms of predictive accuracy over varying sample sizes and number of factors.

B. Experiments with fMRI Data

We analyze data from the Philadelphia Neurodevelopment Cohort (PNC) study, a large-scale, NIMH funded initiative to understand the developmental trajectory of the brain from childhood to adolescence (Satterthwaite et al., 2014). The PNC study contains psychiatric and cognitive phenotype information on a sample of $n = 9428$ participants ages 8-21; a subsample (n = 1445) of these participants received multimodal neuroimaging and they have functional magnetic resonance imaging (fMRI) scans as well as Diffusion Tensor Imaging (DTI) scans. We work with a subset of individuals (49 subjects with 24 boys and 25 girls between ages 8-21 years) on whom we have fMRI data and their phenotypic behavioral data. We are interested in predicting the cognitive scores (generalized anxiety disorder or GAD) for these individuals based on their gender, age, and high-dimensional fMRI measurements over 100,000 voxels in the brain, using a privacy preserving transformation on the fMRI data in order to protect the confidentiality of the fMRI measurements. The fMRI data was pre-processed using standard pipelines as in Higgins et. al (2018) in order to remove unrelated noise to the fMRI signal.

We split the total sample size into a training and a test set using a 70-30 split, and fit the proposed model and the two competing approaches in the simulation section and compare the MSE values. The MSE under the proposed approach was 3.63 (3.56,3.67), while the MSE under the two-step Bayesian model was 3.90 (3.82,3.94) and the SVD-based approach was 3.82. The numbers in parenthesis for the proposed approach and the two-step Bayesian approach correspond to the lower and upper credible intervals for the predicted test sample MSE that were compured as follows. We compute the test MSE corresponding to each MCMC sample value for the model parameters, and use this collection of test sample MSE values over all the MCMC samples to construct a 2-σ or 95% credible intervals for the predicted MSE value.

In order to evaluate the MSE in terms of the best and worst case scenarios, we compute the MSE under a linear model using gender, age, and the original fMRI measurements as covariates (best case scenario) and also compute the MSE for the test sample using only the intercept, age, and gender without any other covariates (worst case scenario). The MSE for the best and worst case scenario was given by 3.41 and 3.99 respectively. This illustrates that the best case MSE is closer to the 95% credible interval for the predicted MSE under the proposed method, and that the worst case MSE is closer to the 95% credible interval under the two-step Bayesian model. Clearly, the proposed factor analysis based approach shows significantly predictive advantages compared to the other two competing approaches, while highlights the utility of the proposed privacy-preserving method for high-dimensional neuroimaging studies.

The number of factors which was used for our analysis was 18 which yielded the lowest MSE in the test sample. We ran our approach by fixing the number of factors between

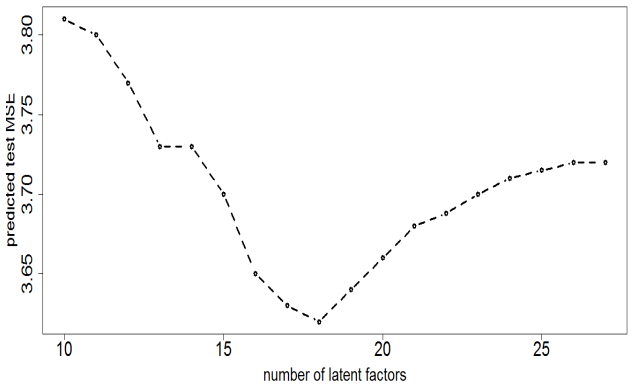

Fig. 4. Predictive test MSE for PNC study over varying number of factors.

8 and 25, and obtained the lowest MSE when the number of factors was set to 18. We did not see a significant change in the predicted MSE when the number of factors was increased beyond 25. A plot of the predictive MSE over varying number of factors for the PNC application is shown in Fig. 4. The low number of factors used in our analysis clearly demonstrates the drastic dimension reduction under our model, with a goal to preserve the privacy of the high-dimensional fMRI data.

VI. Conclusion

We have proposed a novel approach which preserve the privacy of the feature vectors via dimension reduction using factor analysis; it is designed to maximize the predictive accuracy with respect to the response variable. Our goal is to reduce the feature space into lower dimensional latent factors, such that the predictive accuracy of a linear regression model is minimally compromised. We have used a factor analysis based model for dimension reduction; however, our approach can be extended to more general dimension reduction approaches such as manifold learning and compressed sensing. We will explore these approaches for privacy-preservation with a goal to maximize the predictive or classification accuracy in future work. A bonus key feature of the proposed approach is that it can instantaneously update the transformed dataset as soon as a new data point arises, and send this to the user who may be interested in re-running the analysis using the updated data.

Acknowledgment

This research is partially supported by the Institute for Quantitative Methods and Theory at Emory University through the Visiting Faculty Fellowship awarded to Shan Suthaharan.

References

[1] A. Marquand, M. Howard, M. Brammer, C. Chu, S. Coen and J. Mouro-Miranda. “Quantitative prediction of subjective pain intensity from whole-brain fMRI data using Gaussian processes.” Neuroimage, 49(3), 2178-2189, 2010.

[2] G. Valente, A. L. Castellanos, G. Vanacore and E. Formisano. “Multivariate linear regression of high-dimensional fMRI data with multiple target variables.” Human brain mapping, 35(5), 2163-2177, 2014.

[3] J. R. Strawn, A. M. Wehry, M. P. DelBello, M. A. Rynn and S. Strakowski. “Establishing the neurobiologic basis of treatment in children and adolescents with generalized anxiety disorder.” Depression and Anxiety, 29(4), 328-339, 2012.

[4] L. Pessoa. ”Understanding brain networks and brain organization.” Physics of life reviews 11, no. 3 (2014): 400-435.

[5] T. Xu, K. R. Cullen, B. Mueller, M. W. Schreiner, K. O. Lim, C. S. Schulz, K. K. Parhi. Network analysis of functional brain connectivity in borderline personality disorder using resting-state fMRI, NeuroImage: Clinical (2016), doi: 10.1016/j.nicl.2016.02.006

[6] B. Rypma, J. S. Berger, V. Prabhakaran, B. M. Bly, D. Y. Kimberg, B. B. Biswal and M. D’esposito. Neural correlates of cognitive efficiency. Neuroimage, 33(3), 969-979, 2006.

[7] F. Collette, M. Hogge, E. Salmon and M. Van der Linden. Exploration of the neural substrates of executive functioning by functional neuroimaging. Neuroscience, 139(1), 209-221, 2006.

[8] A. Machanavajjhala and J. P. Reiter. ”Big privacy: Protecting confidentiality in big data.” XRDS: Crossroads, The ACM Magazine for Students, vol. 19, no. 1, pp. 20-23, 2012.

[9] T. A. Lasko and S. A. Vinterbo. “Spectral anonymization of data.” IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 3, pp. 437-446, 2010.

[10] M. J. Silva, P. Rijo and A. Francisco. “Evaluating the Impact of Anonymization on Large Interaction Network Datasets.” In Proceedings of the First International Workshop on Privacy and Security of Big Data, pp. 3-10. ACM, 2014.

[11] K. Muralidhar, and R. Sarathy. “A theoretical basis for perturbation methods.” Statistics and Computing, 13(4), pp. 329-335, 2003.

[12] S. Suthaharan. ”A correlation-based subspace analysis for data confidentiality and classification as utility in CPS.” In Communications and Network Security (CNS), 2016 IEEE Conference on, pp. 426-431. IEEE, 2016.

[13] S. Xu, J. Zhang, D. Han and J. Wang. Singular value decomposition based data distortion strategy for privacy protection. Knowledge and Information Systems, 10(3), 383-397, 2006.

[14] T. Fukasawa, J. Wang, T. Takata and M. Miyazaki. An effective distributed privacy-preserving data mining algorithm. In International Conference on Intelligent Data Engineering and Automated Learning. pp. 320-325). Springer, Berlin, Heidelberg, 2004.

[15] B. D. Zumbo and E. K. Chan. Reflections on validation practices in the social, behavioral, and health sciences. In Validity and validation in social, behavioral, and health sciences (pp. 321-327). Springer, Cham, 2014.

[16] M. West. Bayesian factor regression models in the large p, small n paradigm. Bayesian Statist. 7:72332, 2003.

[17] C. M. Carvalho, J. Chang, J. E. Lucas, J. R. Nevins, Q. Wang and M. West. High-dimensional sparse factor modeling: applications in gene expression genomics. Journal of the American Statistical Association, 103(484), 1438-1456, 2008.

[18] A. Hyvrinen and E. Oja. Independent component analysis: algorithms and applications. Neural networks, 13(4-5), 411-430, 2000.

[19] S. Suthaharan. Machine learning models and algorithms for big data classification - Thinking with examples for effective learning. Integrated Series in Information Systems, vol. 36, pp. 1-359, Springer US. 2016

[20] C. Chang, S. Kundu and Q. Long. Scalable Bayesian Variable Selection for High Dimensional Structured Data. https://arxiv.org/abs/1604.07264, 2016.

[21] S. Kundu, and D. Dunson. “Bayesian Variable Selection in Semiparametric Linear Models.” Journal of the American Statistical Association, Theory and Methods, 109, pp. 437-447, 2014.

[22] A. Bhattacharya and D. B. Dunson. ”Sparse Bayesian infinite factor models.” Biometrika, vol. 98, no. 2, pp. 291-306, 2011.

[23] C. Robert, and G. Casella. Introducing Monte Carlo Methods with R. Springer Science & Business Media, 2009.

[24] G. C. McDonald, and D. I. Galarneau. ”A Monte Carlo evaluation of some ridge-type estimators.” Journal of the American Statistical Association, vol. 70, no. 350, pp. 407-416, 1975.

[25] D. L. Donoho. Compressed sensing. IEEE Transactions on information theory, 52(4), 1289-1306, 2006.

[26] A. Adler, M. Elad and M. Zibulevsky. Compressed Learning: A Deep Neural Network Approach. arXiv preprint arXiv:1610.09615, 2016.

[27] N. Spruill. Proceedings of the Section on Survey Research Methods. American StatisticalAssociation; 1983. The confidentiality and analytic usefulness of masked business microdata; p.602-607.