Privacy and Regression Model Preserved Learning

Jinfeng Yi† **Jun Wang**† **Rong Jin***
† IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, USA
{jyi,wangjun}@us.ibm.com
* Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824, USA
rongjin@cse.msu.edu

Abstract

Sensitive data such as medical records and business reports usually contains valuable information that can be used to build prediction models. However, designing learning models by directly using sensitive data might result in severe privacy and copyright issues. In this paper, we propose a novel matrix completion based framework that aims to tackle two challenging issues simultaneously: i) handling missing and noisy sensitive data, and ii) preserving the privacy of the sensitive data during the learning process. In particular, the proposed framework is able to mask the sensitive data while ensuring that the transformed data are still usable for training regression models. We show that two key properties, namely *model preserving* and *privacy preserving*, are satisfied by the transformed data obtained from the proposed framework. In *model preserving*, we guarantee that the linear regression model built from the masked data approximates the regression model learned from the original data in a perfect way. In *privacy preserving*, we ensure that the original sensitive data cannot be recovered since the transformation procedure is irreversible. Given these two characteristics, the transformed data can be safely released to any learners for designing prediction models without revealing any private content. Our empirical studies with a synthesized dataset and multiple sensitive benchmark datasets verify our theoretical claim as well as the effectiveness of the proposed framework.

Introduction

Regression analysis is an important task that has found numerous applications in economics (Dielman 2001; Wu and Tseng 2002), politics (Black and Black 1973; Kousser 1973), health care (Bland and others 2000; Ryan and Farrar 2000), and social sciences (Ron 2002; Stevens 2009). Most of the studies under this topic focus on the learnability of the regression model but overlook the critical privacy protection issue. Note that many data, such as medical records, voting statistics and business reports, contains sensitive information, directly sharing them between the owner and learner may result in severe privacy concerns.

To address this issue, an initial thought is to make the data anonymous or perturb the sensitive information in the data. However, these methods cannot really protect the privacy. (Sweeney 1997) points out that more than 87% of American citizens can be uniquely identified by just observing their gender, ZIP code and birthdate. Also, in the well-known Netflix Prize, although the released Netflix movie rating data has been perturbed, (Narayanan and Shmatikov 2008) shows that it is possible to identify users of the Netflix data by linking them to IMDB dataset. An alternative approach to protect the privacy of sensitive information is to mask the sensitive data by either adding (Vaidya and Clifton 2004; Yu, Vaidya, and Jiang 2006; Gambs, Kégl, and Aïmeur 2007), or multiplying (Chen and Liu 2005; Liu, Kargupta, and Ryan 2006) a randomized matrix. However, there is always a trade-off between the utility and the privacy of the sensitive data. After introducing a sufficient amount of noises to the sensitive data, it has no guarantee that the regression model learned by the distorted data is the same as the regression model built using the original sensitive data. More importantly, information in sensitive data can be usually missing or noisy. For example, in health investigations, the reasons of missing data can be summarized as inappropriate study designs, equipment failure, side effects associated with the treatment, or even law issues (He 2010).

In this work, we aim to address these limitations by developing a novel learning-based framework that aims to learn a transformation from n sensitive data points to m ($m \neq n$) masked data points. Meanwhile, the proposed framework is able to address the following two problems: i) *missing data*, where a large number of features and responses in sensitive data are missing, and ii) *noisy data*, where sensitive data contains noises or outliers.

In our analysis, we show that the proposed framework ensures the following two key properties:

- *Model preserving*, with which we guarantee that the regression model learned from the masked data is the same as the regression model learned from the completed and denoised sensitive data;
- *Privacy preserving*, with which we ensure that sensitive data cannot be reliably recovered even when both the masked data and the learning algorithm are known.

Given these two characteristics, data owners can safely release the masked data to learners then directly applies the learned linear regression model to perform regression analysis on the original sensitive data.

Copyright © 2014, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

The core of the proposed framework is based on the theory of matrix completion (Candès and Tao 2010). In order to ensure that the regression model built by the masked data $\mathbf{Z}_m$ is the same as the regression model learned from the sensitive data $\mathbf{Z}_s$, we aim to ensure its sufficient condition, i.e., the columns of $\mathbf{Z}_m$ lie in the subspace spanned by the columns of sensitive data $\mathbf{Z}_s$. By assuming that the denoised sensitive data matrix is low-rank, a commonly used assumption in data analysis, we show that the combination of the sensitive data and the masked data should also be of low-rank. This enables us to cast the data masking problem into a problem of matrix completion. Moreover, to handle the missing values and noises in the sensitive data, we formulate a cost function in the form of squared loss. Through minimizing this loss under the nuclear norm regularization, we are able to simultaneously remove the noises from the sensitive data and learn optimal features/responses for the masked data. In addition, with the theory of matrix completion, we show that the original sensitive data cannot be recovered from the masked data. The proposed framework, to the best of our knowledge, is the first one that satisfies the properties of model preserving and privacy preserving simultaneously.

Related work

In this section, we review the existing work on privacy-preserving data mining and differential privacy.

One important topic to protect the sensitive information during the procedures of data analysis is privacy-preserving data mining (Aggarwal and Yu 2008). To do so, the owners of sensitive data usually provide modified or perturbed data entries to prevent the leaking of privacy information. In additive perturbation (Vaidya and Clifton 2004; Yu, Vaidya, and Jiang 2006; Gambs, Kégl, and Aïmeur 2007), a noise component drawn from some distributions is added to the data in order to mask the sensitive information. Usually, the added noise should be sufficiently large so that individual records cannot be recovered. Then some data mining techniques are developed to work with the aggregated distributions derived from the perturbed records. Additive perturbation is not suitable for high-dimensional data since the density estimation becomes increasingly inaccurate when the dimensionality of data is large than 10 (Aggarwal and Yu 2008). In contrast to additive perturbation, multiplicative perturbation (Chen and Liu 2005; Liu, Kargupta, and Ryan 2006) masks the sensitive information by multiplying a randomized matrix such that the distances between different data after perturbation are approximately preserved. Then a set of "transformation-invariant data mining models" can be applied to the perturbed data directly. However, multiplicative perturbations are not entirely safe from adversarial attacks because it was threatened by the two kinds of attacks "known input-output attack" and "known sample attack" (Aggarwal and Yu 2008). Other privacy-preserving data mining schemes (Verykios et al. 2004; Aggarwal and Yu 2008) include blocking the sensitive data entries, k-anonymity, combining several values into a coarser category, interchanging values of individual records, and releasing a sample of entire database. However, none of them is designed to preserve the regression models.

Differential privacy (Dwork 2006) is a new privacy standard that has attracted considerable attention in recent years. It provides formal privacy guarantees that do not depend on an adversary's background knowledge. As shown in (Hardt and Roth 2013), in order to preserve the data privacy, the added noise of any differentially private algorithms must scale polynomially with the dimensionality of the sensitive data. This indicates that the noisy terms can easily overwhelm the signals when data is relatively high-dimensional, thus significantly hurt the performance of regression.

Privacy and Regression Model Preserved Learning by Matrix Completion

In this section, we first present the problem statement and a general framework of privacy and regression model preserved learning. We then introduce the proposed algorithm for learning masked data when sensitive data is either fully-observed or only partially-observed. Finally, we present the analysis regarding the properties of *model preserving* and *privacy preserving*.

Problem Definition and Framework

Let $\mathcal{S} = \{S_1, \ldots, S_n\}$ be the set of n sensitive records, and let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathbb{R}^{(d+1)\times n}$ be a matrix where its first d rows are the features of the n sensitive records and its last row is a vector of all 1s. We introduce the last row in order to capture the bias term of the regression model. Following (Goldberg et al. 2010; Cabral et al. 2011), we assume that $\mathbf{X}$ is generated as follows: it starts from a $(d+1) \times n$ low-rank "pre"-feature matrix $\mathbf{X}_0$ with $\mathrm{rank}(\mathbf{X}_0) \ll \min(d+1, n)$, then the actual feature matrix $\mathbf{X}$ is generated by adding a Gaussian noise matrix $\mathbf{E}_X$ to $\mathbf{X}_0$ such that $|\mathbf{E}_X = \mathbf{X} - \mathbf{X}_0|_F$ is small. The logic behind the low-rank assumption is that, generally speaking, only a few factors can contribute to value changes of the sensitive data. Meanwhile, let $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n) \in \mathbb{R}^{t\times n}$ be the responses matrix with t multivariate measurements. We then define the soft responses $\mathbf{Y}_0 = (\mathbf{y}_1^0, \ldots, \mathbf{y}_n^0)$ as $\mathbf{Y}_0 = \mathbf{W}\mathbf{X}_0$, where $\mathbf{W} \in \mathbb{R}^{t\times(d+1)}$ is the underlying linear regression model. Both $\mathbf{X}_0$ and $\mathbf{Y}_0$ can be viewed as the denoised features and responses for the sensitive data. Then the problem of learning the regression model of sensitive data is essentially equivalent to learning the regression model of $\mathbf{X}_0$ and $\mathbf{Y}_0$.

To preserve both the privacy and regression model of sensitive data, we propose to compute a new feature matrix $\mathbf{A} \in \mathbb{R}^{(d+1)\times m}$ and a new response matrix $\mathbf{B} \in \mathbb{R}^{t\times m}$ such that (i) the linear regression model learned from $\mathbf{A}$ and $\mathbf{B}$ is the same as the model learned from $\mathbf{X}_0$ and $\mathbf{Y}_0$, and (ii) $\mathbf{X}_0$ and $\mathbf{Y}_0$ cannot be accurately recovered even if both $\mathbf{A}$ and $\mathbf{B}$ are known. To this end, we develop a matrix completion based framework to learn $\mathbf{A}$ and $\mathbf{B}$ without computing the regression model $\mathbf{W}$ explicitly. Then the owners of sensitive data can safely release the masked feature $\mathbf{A}$ and their associated responses $\mathbf{B}$ to a learner without taking privacy risks.

Privacy and Model Preserved Learning

In this subsection, we present the proposed matrix completion based framework of privacy and regression model preserved learning. Since the denoised feature $\mathbf{X}_0$ is a low-rank matrix and the soft response $\mathbf{Y}_0 = \mathbf{W}\mathbf{X}_0$ is a linear combination of $\mathbf{X}_0$, it is easy to verify that the matrix $[\mathbf{Y}_0; \mathbf{X}_0]$ is also of low-rank. When the target feature matrix $\mathbf{A}$ and response matrix $\mathbf{B}$ share the same regression model as $\mathbf{X}_0$ and $\mathbf{Y}_0$, we have $\mathbf{Y}_0 = \mathbf{W}\mathbf{X}_0$ and $\mathbf{B} = \mathbf{W}\mathbf{A}$. Thus, it directly follows that

$$[\mathbf{Y}_0,\ \mathbf{B}] = \mathbf{W}[\mathbf{X}_0,\ \mathbf{A}],$$

and we expect the following combined matrix

$$\mathbf{Z}_0 = \begin{bmatrix} \mathbf{Y}_0 & \mathbf{B} \\ \mathbf{X}_0 & \mathbf{A} \end{bmatrix}$$

is also of low-rank.

Given this observation, a direct thought for optimizing matrices $\mathbf{A}$ and $\mathbf{B}$ is to minimize the nuclear norm, which is the convex surrogate of the rank function, of matrix $\mathbf{Z}_0$. However, this simple approach is infeasible in practice since both of the denoised matrices $\mathbf{X}_0$ and $\mathbf{Y}_0$ are unknown. More severely, it could result in a trivial solution, e.g., both $\mathbf{A}$ and $\mathbf{B}$ are zero matrices.

In order to guarantee that the masked data $\mathbf{A}$ and $\mathbf{B}$ share the same regression model as sensitive data $\mathbf{X}_0$ and $\mathbf{Y}_0$, we aim to ensure its sufficient condition, i.e., the columns of masked data $\mathbf{Z}_m = [\mathbf{B}; \mathbf{A}]$ lie in the subspace spanned by the column space of sensitive data $\mathbf{Z}_s = [\mathbf{Y}_0; \mathbf{X}_0]$. In other words, this condition guarantees that there exist a $n \times m$ matrix $\mathbf{P}$ satisfying

$$\mathbf{Z}_m = \mathbf{Z}_s\mathbf{P}.$$

This observation inspires us to address the problem of trivial solution by generating a random feature matrix $\tilde{\mathbf{A}}$ that shares the same subspace as the noisy sensitive features $\mathbf{X}$. This enables us to learn a masked data $\mathbf{Z}_m = [\mathbf{B}; \mathbf{A}]$ by ensuring that i) the matrix $\mathbf{Z}_0 = [\mathbf{Z}_s, \mathbf{Z}_m]$ is of low-rank, and ii) the learned matrices $\mathbf{X}_0$, $\mathbf{Y}_0$ and $\mathbf{A}$ in $\mathbf{Z}_0$ are close to the observed matrices $\mathbf{X}, \mathbf{Y}$ and $\tilde{\mathbf{A}}$.

In order to estimate the matrix $\tilde{\mathbf{A}}$, we first generate a normalized $n \times m$ random matrix $\mathbf{P}$ whose entries take values $+1/\sqrt{m}$ and $-1/\sqrt{m}$ with equal probability $1/2$. We then transform the $(d+1) \times n$ matrix $\mathbf{X}$ to a $(d+1) \times m$ matrix $\tilde{\mathbf{A}}$, given by

$$\tilde{\mathbf{A}} = \mathbf{X}\mathbf{P}. \tag{1}$$

We normalize the projection matrix $\mathbf{P}$ since we want to ensure that the entries in matrices $\tilde{\mathbf{A}}$ and $\mathbf{X}$ have roughly the same magnitudes. The equation (1) guarantees that the columns in $\tilde{\mathbf{A}}$ lie in the subspace spanned by the columns of feature matrix $\mathbf{X}$.

Given the matrices $\mathbf{X}$, $\mathbf{Y}$ and $\tilde{\mathbf{A}}$, we rewrite $\mathbf{Z}_0$ as

$$\begin{aligned} \mathbf{Z}_0 &= \begin{bmatrix} \mathbf{Y} & \mathbf{B} \\ \mathbf{X} & \tilde{\mathbf{A}} \end{bmatrix} + \begin{bmatrix} \mathbf{E_Y} & \mathbf{0} \\ \mathbf{E_X} & \mathbf{E_A} \end{bmatrix} \\ &= \mathbf{Z} + \mathbf{E}. \end{aligned} \tag{2}$$

where $\mathbf{E_A} = \mathbf{A} - \tilde{\mathbf{A}}$, $\mathbf{E_X} = \mathbf{X} - \mathbf{X}_0$, and $\mathbf{E_Y} = \mathbf{Y} - \mathbf{Y}_0$ are the errors/noises in matrices $\mathbf{A}$, $\mathbf{X}$, and $\mathbf{Y}$, respectively. In order to capture such noises, we introduce a squared loss function, i.e. $\mathcal{L}(u, v) = \frac{1}{2}(u - v)^2$, to penalize large differences between the true features/responses and the observed ones.

Combining the above ideas, we propose a matrix completion based framework to simultaneously denoise the sensitive data and learn the masked data by solving the following optimization problem

$$\begin{aligned} \min_{\mathbf{Z}_0 \in \mathbb{R}^{(t+d+1)\times(n+m)}} \ & \mu||\mathbf{Z}_0||_* + \frac{1}{tn}\sum_{i=1}^{t}\sum_{j=1}^{n}\mathcal{L}(\mathbf{Z}_{ij}, [\mathbf{Z}_0]_{ij}) \\ & + \frac{C}{(d+1)(n+m)}\sum_{i=t+1}^{t+d+1}\sum_{j=1}^{n+m}\mathcal{L}(\mathbf{Z}_{ij}, [\mathbf{Z}_0]_{ij}), \end{aligned} \tag{3}$$

where $\|\cdot\|_*$ stands for the nuclear norm of matrix, and μ, C are positive trade-off parameters. In problem (3), we penalize features and responses separately since they may have different magnitudes. Once the optimal matrix $\mathbf{Z}_0$ is found, the masked feature matrix $\mathbf{A}$ and the corresponding response matrix $\mathbf{B}$ can be released to learners for training regression models.

Note that in many cases, the sensitive features or responses can be corrupted, we then discuss how to handle the issue of missing entries in the sensitive data. In the following, we only focus on the problem of recovering the denoised sensitive data from the corrupted one. This is due to the reason that, as long as the sensitive data is recovered, we can follow the same procedure as discussed before to generate masked data.

Suppose the entries in $\mathbf{X}$ and $\mathbf{Y}$ are missing at random. We denote by $\Omega_{\mathbf{X}}$ and $\Omega_{\mathbf{Y}}$ to be the index sets of observed entries in sensitive features and responses, respectively. We define a family of matrix projection operators $\mathbf{P}_\Omega$ that takes a $p \times q$ matrix $\mathbf{E}$ as the input and outputs a new matrix $\mathbf{P}_\Omega(\mathbf{E}) \in \mathbb{R}^{p\times q}$ as

$$[\mathbf{P}_\Omega(\mathbf{E})]_{ij} = \begin{cases} \mathbf{E}_{ij} & (i,j) \in \Omega \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

This projection operator guarantees that only the observed entries in the matrix can be projected into the space where we apply matrix completion. Note that (i) the denoised sensitive data matrix $\mathbf{Z}_s = [\mathbf{Y}_0; \mathbf{X}_0]$ is of low-rank, and (ii) both $||\mathbf{P}_{\Omega_{\mathbf{X}}}(\mathbf{X} - \mathbf{X}_0)||_F$ and $||\mathbf{P}_{\Omega_{\mathbf{Y}}}(\mathbf{Y} - \mathbf{Y}_0)||_F$ should be relatively small, we can cast the problem of recovering the sensitive data matrix $\mathbf{Z}_s$ into the following convex optimization problem

$$\begin{aligned} \min_{\mathbf{Z}_s \in \mathbb{R}^{(t+d+1)\times n}} \ & \mu||\mathbf{Z}_s||_* + \frac{1}{|\Omega_{\mathbf{Y}}|}\sum_{ij\in\Omega_{\mathbf{Y}}}\mathcal{L}(\mathbf{Y}_{ij}, \mathbf{Y}_{0_{ij}}) \\ & + \frac{C}{|\Omega_{\mathbf{X}}|}\sum_{ij\in\Omega_{\mathbf{X}}}\mathcal{L}(\mathbf{X}_{ij}, \mathbf{X}_{0_{ij}}). \end{aligned} \tag{5}$$

It is evident that the optimization problem (5) is a variant of the optimization problem (3). In more detail, problem (5) and problem (3) are equivalent when $m = 0$, $|\Omega_{\mathbf{X}}| = n(d+1)$ and $|\Omega_{\mathbf{Y}}| = nt$. We use the efficient Fixed Point Continuation method (Ma, Goldfarb, and Chen 2011) to optimize the problems (3) and (5).

Analysis

In this subsection, we analyze the properties of the proposed framework. Specifically, we show that the proposed framework satisfies both the properties of *model preserving* as well as *privacy preserving*.

i) Model Preserving Let $\mathbf{Z}_s = [\mathbf{Y}_0; \mathbf{X}_0]$ be the denoised responses and features of n sensitive data, and let $\mathbf{Z}_e = [\mathbf{B}; \mathbf{A}]$ be the responses and features of m masked data. The following theorem shows the property of model preserving:

Theorem 1. *Let $\mathbf{X} = \mathbf{X}_0 + \mathbf{E}$, where $\mathbf{E}_{ij} \sim N(0, \sigma^2)$. Let $\tilde{\mathbf{A}} = \mathbf{XP}$ and $\hat{\mathbf{B}}$ be the optimal solution of the optimization problem*

$$\min_{\mathbf{B} \in \mathbb{R}^{t \times m}} \left|\left| \begin{bmatrix} \mathbf{Y}_0 & \mathbf{B} \\ \mathbf{X}_0 & \tilde{\mathbf{A}} \end{bmatrix} \right|\right|_*. \tag{6}$$

Then the columns of the matrix $\mathbf{Z}_{\mathbb{E}} = \mathbb{E}[\hat{\mathbf{B}}; \tilde{\mathbf{A}}]$ lie in the subspace spanned by the columns of sensitive data $\mathbf{Z}_s = [\mathbf{Y}_0; \mathbf{X}_0]$, where $\mathbb{E}[\mathbf{M}]$ is the expectation of matrix $\mathbf{M}$.

Note that $\mathbf{Y}_0 = \mathbf{WX}_0$. Using Theorem 1, it is easy to verify that the learned masked data also satisfying $\mathbb{E}(\hat{\mathbf{B}}) = \mathbf{WA}$. Although Theorem 1 only shows that *the expectation of* the masked data share the same regression model as the original sensitive data, our empirical studies further verify that the masked data approximates the regression model learned from the original data in an almost perfect way.

Based on Theorem 1, we have the following corollary showing that the nuclear norm minimization problem (6) can be approximately reduced to the problem of finding a $n \times m$ transformation matrix $\mathbf{T}$.

Corollary 1. *The convex optimization problem (6) can be approximately relaxed to the following optimization problem:*

$$\min_{\mathbf{T} \in \mathbb{R}^{n \times m}} ||\mathbf{T}||_F^2 + \lambda ||\tilde{\mathbf{A}} - \mathbf{X}_0 \mathbf{T}||_F^2, \tag{7}$$

where $||\mathbf{T}||_F$ is the Frobenius norm of matrix $\mathbf{T}$.

We skip the proof of Theorem 1 and Corollary 1 due to space limitation. Corollary 1 basically shows that the proposed nuclear norm minimization problem is close to the problem of learning a $n \times m$ transformation matrix which maps the sensitive data to the masked data. However, the proposed optimization problem (3) is more desirable than the optimization problem (7) since: (i) the denoised feature matrix $\mathbf{X}_0$ is unknown and we cannot optimize problem (7) without knowing it, and (ii) problem (3) is robust to noisy information in sensitive data while problem (7) is not.

ii) Privacy Preserving We then discuss the privacy preserving property of the proposed framework. Specifically, the proposed framework can preserve privacy of sensitive data in the following aspects:

- **Number of sensitive records** Since learners can only observe the features and responses of m masked records and m can be significantly different from n, they cannot identify the number of sensitive records.
- **Privacy of $\mathbf{X}$ and $\mathbf{Y}$** Even if the number of sensitive records n is leaked, learners still have no chance to recover the sensitive data features $\mathbf{X}$ and responses $\mathbf{Y}$. This is because that the proposed framework approximately learns a $n \times m$ transformation matrix $\mathbf{T}$ and this matrix is completely blind to the learners. Indeed, directly optimizing the problem (3) using the masked data $\mathbf{A}$ and $\mathbf{B}$ will lead to a trivial solution, e.g., both the recovered matrices $\mathbf{X}$ and $\mathbf{Y}$ are all zero matrices.
- **Privacy of individual record.** In addition, the proposed framework enjoys an even higher privacy standard that is similar to the property of differential privacy (Dwork 2006). In more detail, even though the matrices $\mathbf{A}$ and $\mathbf{B}$, as well as the $n-1$ sensitive records are known, the learner still cannot recover the last column of the sensitive record. Based on the theory of matrix completion (Candès and Tao 2010), there is no chance to recover a single column if no entry in that column is observed. Specifically, even when all of the $n+m-1$ columns of the matrix $\mathbf{Z}_0$ are known, the learner still cannot complete the last unknown column, thus cannot recover the last sensitive record.

Note that by implicitly identifying the linear regression model (i.e., a hyperplane in a $(d+1)$-dimensional space) using the sensitive data, the proposed framework simultaneously erases all the sensitive information and randomly generate m new data points (masked data) that are lying on this hyperplane. This is the key reason that why the proposed framework satisfies both properties of *model preserving* and *privacy preserving*.

Experiments

In this section, we first use simulated data to verify our theoretical claim, i.e., the columns of masked data $\mathbf{Z}_m = [\mathbf{B}; \mathbf{A}]$ approximately lie in the subspace spanned by the columns of sensitive data $\mathbf{Z}_s = [\mathbf{Y}_0; \mathbf{X}_0]$. We then use two benchmark datasets for regression to verify the effectiveness of the proposed data masking framework.

Experiment with Synthesized Data

We first conduct experiments with simulated data to verify that the learned masked data $\mathbf{Z}_m$ is spanned by the same subspace as the sensitive data $\mathbf{Z}_s$, the key condition to ensure the property of model preserving. To this end, we first generate two random matrices $\mathbf{U}$ and $\mathbf{V}$ with sizes equaling to 500×20 and $1,000 \times 20$. We then construct a low-rank denoised feature matrix as $\mathbf{X}_0 = \mathbf{UV}^\top$. By adding some Gaussian noises to the entries of $\mathbf{X}_0$, we generate a noisy feature matrix $\mathbf{X}$, i.e., $\mathbf{X} = \mathbf{X}_0 + \mathbf{E}_X$ with $\mathbf{E}_{X_{i,j}} \sim N(0, \sigma^2)$. Also, we set the regression model $\mathbf{W}$ as a randomly generated 500×10 matrix, indicating that there are 10 multivariate measurements in total. This enables us to compute the soft label matrix via $\mathbf{Y}_0 = \mathbf{W}^\top \mathbf{X}_0$. Finally, we also add some random Gaussian noises to matrix $\mathbf{Y}_0$ for simulating noisy response as $\mathbf{Y} = \mathbf{Y}_0 + \mathbf{E}_Y$ with $\mathbf{E}_{\mathbf{Y}_{i,j}} \sim N(0, \sigma^2)$. We use noisy features $\mathbf{X}$ and responses $\mathbf{Y}$ as the input of the proposed framework then analyze whether the learned masked data $\mathbf{Z}_m = [\mathbf{B}; \mathbf{A}]$ lies

Table 1: Projection errors with different levels of the added noise

Variances σ^2	0.05	0.1	0.2	0.3	0.4	0.5
Errors $\mathcal{E}$	0	0	0	0.01	0.01	0.02

Table 2: Projection errors with different levels of the added noise when 80% of entries in sensitive data are missing

Variances σ^2	0.05	0.1	0.2	0.3	0.4	0.5
Errors $\mathcal{E}$	0	0	0	0.01	0.01	0.01

in the subspace spanned by the columns of sensitive data $\mathbf{Z}_s = [\mathbf{Y}_0; \mathbf{X}_0]$. Following (Yi et al. 2013), we define a projection operator $\mathbf{P}_k$ as $\mathbf{P}_k = \mathbf{U}_r\mathbf{U}_r^\top$, where $\mathbf{U}_r$ are the top r left singular vectors of matrix $\mathbf{Z}_s$ and r is the rank of $\mathbf{Z}_s$. Then the projection error is expressed as

$$\mathcal{E} = \max_{1\leq i\leq m} \frac{1}{m}\|(\mathbf{Z}_m)_i - \mathbf{P}_k(\mathbf{Z}_m)_i\|_2. \tag{8}$$

To verify the robustness of the proposed framework, we vary the variance σ^2 in the range $\{0.05, 0.1, 0.2, 0.3, 0.4, 0.5\}$. For each σ^2, we optimize the masked data $\mathbf{Z}_m = [\mathbf{B}; \mathbf{A}]$ using the proposed framework, then compute the projection error $\mathcal{E}$ using Eq. (8). Table 1 shows the average projection errors over 10 runs. We observe that the projection errors equal to 0 when σ^2 is no greater than 0.2, indicating that the masked data truly lies in the subspace spanned by the column space of the sensitive data when the added noise is not too large. Even when the variance is as large as 0.5, the projection error is still merely 0.02, verifying that the proposed framework is also robust to high levels of noises.

To further verify our theoretical claim in the case of missing data, we randomly mark 80% entries of the feature matrix $\mathbf{X}$ and the response matrix $\mathbf{Y}$ as unobserved. We then use the partially observed data to generate masked data. The average projection errors over 10 runs are summarized in Table 2. It is expected that, with a large portion of entries missing, the task of identifying the underlying column space of sensitive data would become more challenging. However, by comparing the Table 1 and Table 2, we observe that the proposed method can yield the same projection errors as the complete data even when a large portion of entries in sensitive data are missing. We conjecture that by recovering missing entries through the matrix completion problem (5), the recovered sensitive data tends to be close to the denoised data matrix $\mathbf{Z}_s = [\mathbf{Y}_0; \mathbf{X}_0]$, making it a good starting point to learn masked data by optimizing the problem (3). To sum up, this experiment verifies that the proposed method can accurately recover the column space of sensitive data even when the sensitive data is both noisy and highly corrupted (i.e., a majority of sensitive data is missing).

Experiment with Benchmark Datasets

We then evaluate the proposed framework on two sensitive benchmark datasets. They are

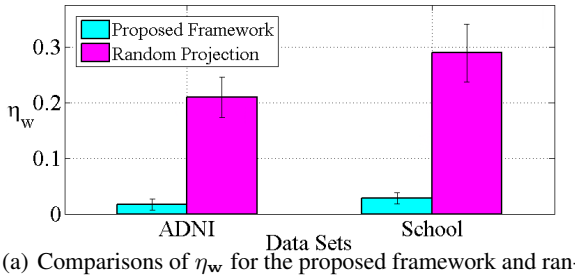
1. *ADNI* dataset (Zhou et al. 2013) that is from the Alzheimer's Disease Neuroimaging Initiative database. This dataset contains 328 Magnetic Resonance Imaging (MRI) features of 675 patients. The responses of this data are the Mini-Mental State Exam (MMSE) scores at six time points M06, M12, M18, M24, M36, and M48.
2. *School* data set (Gong, Ye, and Zhang 2012) that is from the Inner London Education Authority (ILEA). This dataset is consists of examination records of $15,362$ students from 139 secondary schools in years 1985, 1986 and 1987. Each sample is represented by 27 binary attributes which include year, gender, examination score, and so on. The responses of this data are the examination scores and we have a total of 139 tasks with each task corresponding to one school.

Since the proposed framework is the first method that simultaneously satisfies the properties of model preserving and privacy preserving, there is no existing strong baseline algorithm that can be compared. Note that the proposed framework can be approximately reduced to a model preserved multiplicative approach, we compare it to the random projection method (Liu, Kargupta, and Ryan 2006) that multiplies a random matrix to sensitive data in order to preserve its privacy. To this end, we randomly generate a $\tilde{d} \times (d+1)$ projection matrix $\mathbf{P}$ satisfying that the expectation of $\mathbf{P}^\top\mathbf{P}$ equals to an identity matrix. We then project the original sensitive features to a $\tilde{d}$-dimensional space using the projection matrix $\mathbf{P}$, i.e., $\tilde{\mathbf{X}} = \mathbf{P}\mathbf{X}$. After obtaining the linear regression model $\tilde{\mathbf{W}}$ learned using $\tilde{\mathbf{X}}$ and $\mathbf{Y}$, we can project it back to a regression model $\mathbf{W}$ in the original $(d+1)$-dimensional space by $\mathbf{W} = \tilde{\mathbf{W}}\mathbf{P}^\top$. In our experiments, the dimensionality $\tilde{d}$ of random projection matrix $\mathbf{P}$ is set to be $[d/2]$.

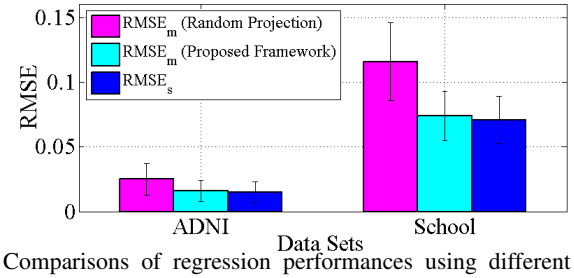
We first conduct experiments when both features and responses are fully observed. For both of these two sensitive datasets, we randomly sample 70% of the records as training data to generate the masked data for training regression models. We treat the remaining 30% of records as testing data for evaluating the regression performance. To evaluate the effect of model preserving property, we apply support vector regression (SVR) (Smola and Schölkopf 2004) with a linear kernel to both learned masked data and the original training (sensitive) data. We denote the linear regression models learned from the masked and sensitive data as $\mathbf{W}_m$ and $\mathbf{W}_s$, respectively. To obtain optimal models, we apply 5-fold cross validations on both masked and sensitive data with the regularization parameter of SVR ranging from 2^{-5} to 2^5. The number of masked data m is set to be $[n/3]$. To measure the difference between two regression models $\mathbf{W}_e$ and $\mathbf{W}_s$, we use

$$\eta_{\mathbf{W}} = \frac{|\mathbf{W}_m - \mathbf{W}_s|_F}{\max(|\mathbf{W}_m|_F, |\mathbf{W}_s|_F)}$$

as an evaluation metric. In addition, we also employ the Root Mean Squared Error (RMSE) (Gunst and Mason 1977) to evaluate the regression performance. We denote $\mathrm{RMSE}_{\mathbf{W}_m}$ as the performance of applying $\mathbf{W}_m$ to the testing data, and $\mathrm{RMSE}_{\mathbf{W}_s}$ as the performance of applying $\mathbf{W}_s$ to the testing data. Each experiment is repeated 10 times and the average performance are reported.

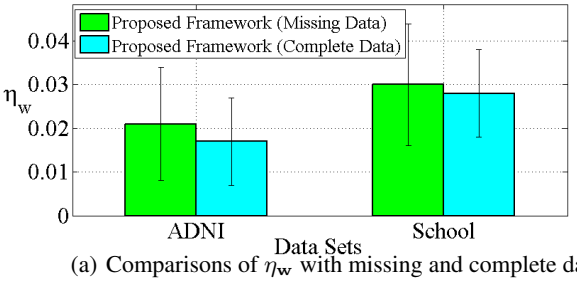


(a) Comparisons of $\eta_{\mathbf{w}}$ for the proposed framework and random projection

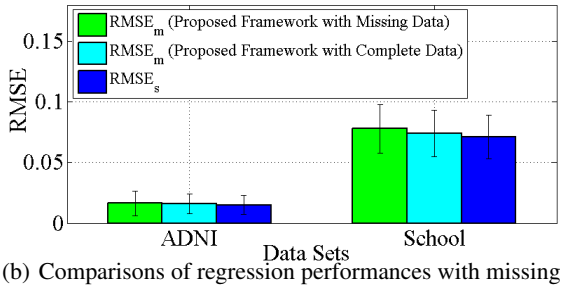


(b) Comparisons of regression performances using different methods

Figure 1: Comparisons of regression models trained directly by sensitive data, by masked data generated from the proposed framework, and by perturbed data generated from random projections.



(a) Comparisons of $\eta_{\mathbf{w}}$ with missing and complete data



(b) Comparisons of regression performances with missing and complete data

Figure 2: Comparison of regression models trained directly by sensitive data, by masked data generated from the complete sensitive data, and by masked data generated from the partially-observed sensitive data.

Figure 1 compares the regression models learned using the masked data generated by the proposed framework and the regression models trained using the perturbed data generated by the random projection method. Figure 1(a) shows that for the proposed method, the regression models learned by the masked data and the regression models learned by the original training (sensitive) data are very close to each other. In contrast, random projection method cannot preserve the regression model since the normalized model differences $\eta_{\mathbf{w}}$ are significantly larger than 0 in both datasets. In Figure 1(b), we observed that by directly applying the regression models learned from the masked data to the testing data, we can achieve almost the same performance as using the regression models built from the original sensitive data. As a comparison, random projection approach suffers from a large increase of regression errors in terms of RMSE, verifying that it cannot preserve the regression models.

We further conduct experiments to evaluate how the proposed framework performs when a considerable portion of entries in sensitive data are missing. We randomly mark 80% entries of the sensitive data features and responses as unobserved. We then use the partially-observed sensitive data to generate masked data by the proposed framework. Since random projection cannot be applied to missing data, we compare the performance of the proposed framework with missing data to the proposed framework with complete data, and the experimental results averaged over 10 random trials are shown in Figure 2.

From Figure 2(a), we observed that even when 80% of the sensitive features and responses are missing, the normalized model difference $\eta_{\mathbf{w}}$ are still very small (less than 0.04), indicating that the learned masked data can still preserve the regression models of the original fully-observed data. Figure 2(b) shows that when applying the trained regression models to the testing data, the models learned using partially-observed sensitive data can achieve very similar RMSEs as the models directly trained using the complete sensitive data. This is not surprising since the difference between two regression models is very small, as indicated by Figure 2 (a). In summary, the experiments verify that the proposed framework can satisfy the property of model preserving, and it is also robust to missing and noisy entries in both sensitive features and responses.

Conclusions and Future Work

In this paper, we propose a framework for privacy and regression model preserved learning. The key idea is to cast the problem of data masking into a problem of matrix completion. The masked data are generated through filling unknown entries that tend to maintain the low-rank matrix structure. We show that the masked data using the proposed framework satisfies both properties of *model preserving* and *privacy preserving*. This ensures that data owners can safely release sensitive data to learners for training regression models and the learned model can be directly applied to the original sensitive data. In addition, by exploiting the strengths of loss functions as well as matrix completion techniques, the proposed framework is robust to both *noises* and *missing entries* that often occur in the sensitive data. Our empirical studies with a synthesized dataset and two real-world sensitive datasets verify our theoretical claims, and also show promising performance of the proposed algorithm. Our future directions include the extensions to privacy preserving for the classification problem, as well as preserving nonlinear regression models.

References

Aggarwal, C. C., and Yu, P. S. 2008. *Privacy-Preserving Data Mining - Models and Algorithms*, volume 34 of *Advances in Database Systems*. Springer.

Black, E., and Black, M. 1973. The wallace vote in alabama: A multiple regression analysis. *The Journal of Politics* 35(03):730–736.

Bland, M., et al. 2000. *An introduction to medical statistics.* Number Ed. 3. Oxford University Press.

Cabral, R. S.; De la Torre, F.; Costeira, J. P.; and Bernardino, A. 2011. Matrix completion for multi-label image classification. In *NIPS*.

Candès, E. J., and Tao, T. 2010. The power of convex relaxation: near-optimal matrix completion. *IEEE Transactions on Information Theory* 56(5):2053–2080.

Chen, K., and Liu, L. 2005. Privacy preserving data classification with rotation perturbation. In *ICDM*, 589–592.

Dielman, T. E. 2001. *Applied regression analysis for business and economics*. Duxbury/Thomson Learning.

Dwork, C. 2006. Differential privacy. In *ICALP (2)*, 1–12.

Gambs, S.; Kégl, B.; and Aïmeur, E. 2007. Privacy-preserving boosting. *Data Min. Knowl. Discov.* 14(1):131–170.

Goldberg, A. B.; Zhu, X.; Recht, B.; Xu, J.-M.; and Nowak, R. D. 2010. Transduction with matrix completion: Three birds with one stone. In *NIPS*, 757–765.

Gong, P.; Ye, J.; and Zhang, C. 2012. Robust multi-task feature learning. In *KDD*, 895–903. ACM.

Gunst, R. F., and Mason, R. L. 1977. Biased estimation in regression: an evaluation using mean squared error. *Journal of the American Statistical Association* 72(359):616–628.

Hardt, M., and Roth, A. 2013. Beyond worst-case analysis in private singular vector computation. In *STOC*, 331–340.

He, Y. L. 2010. Missing data analysis using multiple imputation getting to the heart of the matter. *Circulation-cardiovascular Quality and Outcomes* 3(1):98–U145.

Kousser, J. M. 1973. Ecological regression and the analysis of past politics. *The Journal of Interdisciplinary History* 4(2):237–262.

Liu, K.; Kargupta, H.; and Ryan, J. 2006. Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. *IEEE Trans. Knowl. Data Eng.* 18(1):92–106.

Ma, S.; Goldfarb, D.; and Chen, L. 2011. Fixed point and bregman iterative methods for matrix rank minimization. *Math. Program.* 128(1-2):321–353.

Narayanan, A., and Shmatikov, V. 2008. Robust de-anonymization of large sparse datasets. In *IEEE Symposium on Security and Privacy*, 111–125.

Ron, A. 2002. Regression analysis and the philosophy of social science: A critical realist view. *Journal of Critical Realism* 1(1):119–142.

Ryan, M., and Farrar, S. 2000. Using conjoint analysis to elicit preferences for health care. *BMJ: British Medical Journal* 320(7248):1530.

Smola, A. J., and Schölkopf, B. 2004. A tutorial on support vector regression. *Statistics and computing* 14(3):199–222.

Stevens, J. 2009. *Applied multivariate statistics for the social sciences*. Taylor & Francis US.

Sweeney, L. 1997. Weaving technology and policy together to maintain confidentiality. *The Journal of Law, Medicine & Ethics* 25(2-3):98–110.

Vaidya, J., and Clifton, C. 2004. Privacy preserving naive bayes classifier for vertically partitioned data. In *SDM*.

Verykios, V. S.; Bertino, E.; Fovino, I. N.; Provenza, L. P.; Saygin, Y.; and Theodoridis, Y. 2004. State-of-the-art in privacy preserving data mining. *SIGMOD Record* 33(1):50–57.

Wu, B., and Tseng, N.-F. 2002. A new approach to fuzzy regression models with application to business cycle analysis. *Fuzzy Sets and Systems* 130(1):33–42.

Yi, J.; Zhang, L.; Jin, R.; Qian, Q.; and Jain, A. K. 2013. Semi-supervised clustering by input pattern assisted pairwise similarity matrix completion. In *ICML (3)*, 1400–1408.

Yu, H.; Vaidya, J.; and Jiang, X. 2006. Privacy-preserving svm classification on vertically partitioned data. In *PAKDD*, 647–656.

Zhou, J.; Liu, J.; Narayan, V. A.; and Ye, J. 2013. Modeling disease progression via multi-task learning. *NeuroImage* 78(0):233 – 248.