Journal of Privacy and Confidentiality (2012) **4**, Number 1, 221–243

Privacy-Preserving Data Sharing in High Dimensional Regression and Classification Settings

Stephen E. Fienberg[*] and Jiashun Jin[†]

Abstract. We focus on the problem of multi-party data sharing in high dimensional data settings where the number of measured features (or the dimension) p is frequently much larger than the number of subjects (or the sample size) n, the so-called $p \gg n$ scenario that has been the focus of much recent statistical research. Here, we consider data sharing for two interconnected problems in high dimensional data analysis, namely the feature selection and classification. We characterize the notions of "cautious", "regular", and "generous" data sharing in terms of their privacy-preserving implications for the parties and their share of data, with focus on the "feature privacy" rather than the "sample privacy," though the violation of the former may lead to the latter. We evaluate the data sharing methods using a *phase diagram* from the statistical literature on multiplicity and Higher Criticism thresholding. In the two-dimensional phase space calibrated by the signal sparsity and signal strength, a phase diagram is a partition of the phase space and contains three distinguished regions, where we have no (feature) privacy violation, relatively rare privacy violations, and an overwhelming amount of privacy violation.

Keywords: Hamming Distance, Higher Criticism, LASSO, Marginal Regression, Noise addition, Phase Diagram, Variable Selection.

1 Introduction

There is now an extensive literature on privacy or confidentiality protection of statistical databases, including the notion of the risk-utility tradeoff, but most of the literature has tended to focus more on the protection against risk rather than on facilitating formal analyses in an accurate fashion. Additionally, existing research has more often focused on the release of a small number of summary statistics rather than on the more elaborated results and output from modern statistical methodology. This is especially the case for analyses based on high dimensional data. In this paper, we use a different approach and focus on high dimensional data. We consider the problem of data sharing among parties from the perspective of the utility of the results when useful features for classification and regression are both rare and weak. We then ask the reverse question

[*] Department of Statistics, Machine Learning Department, Living analytics Research Center, Cylab, Carnegie Mellon University, Pittsburgh, PA, `mailto:fienberg@stat.cmu.edu`

[†] Department of Statistics, Carnegie Mellon University, Pittsburgh, PA, `mailto:jiashun@stat.cmu.edu`

regarding the challenges to the privacy of the shared data, without adopting specific definitions of privacy—the approach found in the literature on differential privacy, for example. Because of this reverse perspective we use the descriptor: *privacy-preserving data sharing*. Because the sensitivity of "shared data" depends on the auxiliary information available to a party, the existence of such information can potentially compromise the acceptability of the data-sharing results we describe.

High dimensional data analysis, where the number of measured features (or the dimension) p is frequently much larger than the number of subjects (or the sample size) n, i.e., the so-called $p \gg n$ scenario, is a topic of major current interest in statistics. In many application problems, with genomics being an iconic example, large p usually means that the signals are sparse, and small n usually means that the individual signals are weak (e.g., see Donoho and Jin (2008)). For these reasons, it is challenging to make valid statistical inference (e.g., regression and prediction, classification, clustering), and the usual statistical challenges are magnified when different parties hold different parts of what we might conceptually consider as a single unified database. For example, multiple parties (hospitals or laboratories) may hold genomic data for the *same* type of cancer, and perhaps even for the same individuals, but assess it in different ways. When the data of each party are based on a very limited number of patients (which may due to the rarity of the cancer or to the difficulty in enrolling patients), the signal is so sparse and faint that it is impossible for each individual party to gain any valid insight with their share of data. Similarly, when the parties hold different genomic information on overlapping sets of patients, taking advantage of the full range of predictors is difficult at best. To overcome these difficulties, the parties can pool their data together for inference purposes. But the question is how to do this while at the same time offering some level of "privacy protection" for each party's data.

In this paper, we focus on privacy protection and data sharing in two separate but related contexts. Our goal is to link ideas on noise addition that are already in wide-spread use to models and methods for high dimensional regression and classification, where the statistician is interested in pursuing aspects of dimensionality reduction and/or variable selection and is faced with the problem of multiplicity in the choice of variables for inclusion, cf. an earlier proposal in Fienberg and Jin (2009). We utilize statistical results on Higher Criticism thresholding which address the problem, see e.g., Donoho and Jin (2009), Donoho and Jin (2008), and Jin (2009). At the core of our results is the so-called *Rare and Weak* signal model and the associated *phase diagram*. The latter prescribes (asymptotically) when it is possible to have successful classification/variable selection and when it is impossible to do so because the signals are too rare and weak. We then consider three different types of data sharing: Cautious Sharing, Regular Sharing, and Generous Sharing. We derive phase diagrams corresponding to different types of data sharing and use the results to elaborate the effect of different types of data sharing over the privacy-preservation.

Depending on the number of parties and the sample size of each party, data sharing may impact the statistical inference in three different ways: not helpful because the signals are too sparse and weak even with data sharing, substantially helpful, and non-substantially helpful as the signals are strong enough even without data sharing.

Cautious sharing is generally secure. Regular sharing may be secure if we are concerned only with the privacy of the Y variable (class label in the classification setting, response variable in the regression setting), but it is not secure if we are also concerned with the X variable. Generous sharing is usually not secure, and the attacker may be able to precisely identity the label Y which we wish to protect.

1.1 Overview of Related Literature

There is, at best, a limited literature dealing with privacy protection in high dimensional regression and classification settings, e.g., see Zhou et al. (2009). The differential privacy literature (e.g., see Dwork (2008), Dwork and Naor (2010), and Dwork and Smith (2009)) primarily focuses on noise addition, and some basic results can be adapted to the problem of multiple regression where, for example, one adds Laplacian noise to the regression coefficients or to the coefficients of a linear classifier. The difficulty with this approach is that to achieve a given level of privacy protection, one needs to add more noise as the number of coefficients to be protected increases; this ultimately compromises the utility of the released data. Indeed, Dinur and Nissim (2003) show that no "noisy database" can provide very accurate answers to too many queries, e.g., for coefficients in a regression model, for then one can use the queries to mount an attack on the database. Specific differential privacy approaches to logistic regression and support vector machines are given by Chaudhuri and Monteleoni (2008) and Sarwate et al. (2009), and Chauduri et al. (2011) suggest a unified approach to such problems. But in all of these settings there remain serious issues regarding the utility of the released data which result from the stringent differential privacy requirements.

In the statistical disclosure limitation literature, Ting et al. (2008) focus on data release of multivariate data and propose a multiplicative perturbation that preserves means and variance/covariances, and thus least squares regression coefficients and compares their results with approaches involving additive perturbation (Burridge, 2003; Muralidhar and Sarathy, 2006). O'Keefe and Good (2009) consider the release of regression results via remote analysis servers.

Perhaps the closest relevant literature for the approach in this paper on privacy and regression comes in the domain of secure multiparty computation, e.g., see Amirbekyan and Estivill-Castro (2007), Du et al. (2004), Fienberg et al. (2008), Fienberg et al. (2012), Hall et al. (2011), and Sanil et al. (2004), where different parties hold parts of what we might think of as a combined database and are unwilling to share their data but are willing to have it used as part of a joint computation. This literature does not address the high dimensional problems except indirectly, since the goal there is to release the full vector of regression coefficients from "merged" databases, even though this might be done subject to differential privacy perturbations. Here we use the term "data sharing" to describe the willingness of the parties to have their data used in such computations and in several places in the paper we connect our approach to the secure multiparty computation literature.

1.2 Outline

The remainder of the paper is organized as follows. First, in Section 2 we discuss high dimensional classification and feature selection, and introduce the phase diagram. Second, in Section 3 we introduce data sharing and discuss their impact both in classification and feature selection, and in privacy-preservation. We characterize the notions of "cautious," "regular," and "generous" data sharing in terms of their privacy-preserving implications for the parties and their portions of the database. In Section 4, we turn to a high dimension regression-like database and consider protecting privacy by adding normal noise, cf. differential privacy which focuses on Laplacian noise addition. In Section 5, we cast the regression problem in the context of data sharing among two or more parties and, again, characterize the different notions of data sharing in terms of their privacy-preserving implications. Finally, in Section 6 we summarize the main results presented in this paper, and their relationship to the existing literature.

2 High Dimensional Feature Selection and Classification

In this section we consider data sharing in high dimensional classification. We begin by introducing the models and notation, and then turn to the phase diagram of high dimensional classification when useful features are both rare and weak. We conclude by discussing the implications of data sharing in a classification problem, with emphasis on their effect on the phase diagram.

We are primarily interested in the two-class classification problem, although extensions to multi-class classification are possible. Suppose we possess data for n different subjects (X_i, Y_i), $1 \leq i \leq n$, which we call the training dataset. Here, $Y_i \in \{-1, 1\}$ denotes the class label, and X_i is a p by 1 vector of measured features. The problem is how to use the data to train a classifier so that when we acquire a new subject, we can predict its label Y based on the feature X and the trained-classifier. We call X the test feature. The classification literature goes back at least to the classic 1936 paper by Fisher (Fisher, 1936), but recently attention has shifted from the setting where p is fixed and $n \to \infty$ to the high dimensional setting where $p \gg n$. The latter is of interest in many application areas, e.g., in cancer classification using genomic data, where the measured features are genes, proteins, etc., and $Y = \pm 1$ stands for cancer and normal.

We model the features as Gaussian so that $X_i \sim Normal(\mu_1, I_p)$ when $Y_i = 1$ and $X_i \sim Normal(\mu_2, I_p)$ when $Y_i = -1$. For simplicity, we assume the experiment is balanced so that the outcomes for Y are equally likely to be 1 or -1, and the data is centered so that $\mu_1 = \mu$ and $\mu_2 = -\mu$ for some contrast mean vector μ. The vector μ is unknown but is sparse in the sense that only a small proportion of its coordinates is nonzero. We say the j-th feature is useful if and only if $\mu(j) \neq 0$, $1 \leq j \leq p$.

2.1 Fisher's LDA and Feature Selection by Higher Criticism Thresholding

Let $w = (w(1), \ldots, w(p))'$ be a p by 1 weight vector (i.e., $w(j) \geq 0$ and $\sum_{j=1}^{p} w(j) = 1$), which may depend on the training data but not on the test data. Fisher's LDA (Fisher, 1936) takes a weighted average over different test features by

$$L(X) = \sum_{j=1}^{p} w(j)X(j).$$

Since the two classes are equally likely, LDA simply classifies the label $Y = \pm 1$ according to $L(X) >< 0$. When μ is known, the best choice of w should satisfy $w \propto \mu$ (Fisher, 1936). Unfortunately, μ is usually unknown, except for that it is sparse. Still, the sparsity of μ suggests that the optimal choice of w should also be sparse. This says that we should select only a small proportion of features and use them for classification. This is the problem of *feature selection*.

Let Z be the summarized z-scores for the training dataset:

$$Z = \frac{1}{\sqrt{n}} \sum_{i=1}^{p} Y_i X_i.$$

It is seen that $Z \sim Normal(\sqrt{n}\mu, I_p)$. A simple approach to feature selection is to select a threshold $t > 0$ and let

$$w(j) = \mathrm{sgn}(Z_j) \cdot 1_{\{|Z(j)| \geq t\}}.$$

In effect, only features with $Z(j)$ exceeding the threshold t in magnitude are selected for feature selection.

Seemingly, the key for feature selection is to set the threshold t. Donoho and Jin (2008) introduced a method that sets the threshold in a data-driven fashion which they refer to as the Higher Criticism Thresholding (HCT). To implement HCT, we follow three simple steps:

- For $1 \leq j \leq p$, let $\pi_j = P(|Normal(0,1)| \geq |Z(j)|)$ be the j-th p-value.
- Sort the p-values in the ascending order $\pi_{(1)} < \pi_{(2)} < \ldots < \pi_{(p)}$.
- let $\hat{j}$ be the index which maximizes the so-called Higher Criticism functional

$$\hat{j} = \hat{j}(Z, p) = \mathrm{argmax}_{\{1 \leq j \leq p/2\}} \left\{ \frac{(j/p) - \pi_{(j)}}{\sqrt{(j/p)(1 - j/p)}} \right\}.$$

The $\hat{j}$-th largest (in magnitude) z-score $|Z|_{(\hat{j})}$ is then the HCT which we denote by $\hat{t}_p^{HC} = \hat{t}_p^{HC}(Z)$.

Now, let

$$w^{HC}(j) = \mathrm{sgn}(Z(j))1_{\{|Z(j)| \geq \hat{t}_p^{HC}\}}, \qquad L^{HC}(X) = L^{HC}(X; w^{HC}) = \sum_{j=1}^{p} w^{HC}(j)X(j).$$

We reach the HCT classifier which classifies $Y = \pm 1$ according to $L^{HC}(X) >< 0$.

2.2 Phase Diagram for Classification and Feature Selection

For the contrast mean vector μ, we model its proportion of nonzero coordinates by ϵ,

$$\epsilon = \frac{1}{p}\#\{1 \leq j \leq p : \ \mu(j) \neq 0\}.$$

Recall that the vector of summarized z-scores is $Z \sim Normal(\sqrt{n}\mu, 1)$. We calibrate all nonzero coordinates of μ as having a common magnitude $\mu_0 = \tau/\sqrt{n}$ for some parameter $\tau > 0$. We are interested in the case where both ϵ is small and τ is moderately large, so that the useful features are both sparse and weak.

We use an asymptotic framework where $p \to \infty$, and the parameters (n, ϵ, τ) are linked to p through fixed parameters. In detail, fixing $(\vartheta, r) \in (0,1)^2$, we model

$$\epsilon = \epsilon_p = p^{-\vartheta}, \qquad \tau = \tau_p = \sqrt{2r \log p}.$$

When $r > 1$, the HCT literature (Donoho and Jin, 2008) demonstrates that almost all useful features can be accurately identified. In this case, the problem of feature selection is relatively easy; thus we focus on the case $0 < r < 1$. As (ϑ, r) ranges in $(0,1)^2$, the model covers the entire interesting range of sparsity level and signal strengths.

As $p \to \infty$, the sample size $n = n_p$ may grow with p. We consider three types of such growths.

- *No growth*: $n_p = n_0$ is fixed as p grows to ∞.
- *Slow growth*: $1 \ll n_p \ll p^{\theta}$ for any $\theta > 0$ (e.g., $n = n_p = \log(p)$).
- *Regular growth*: $n_p = p^{\theta}$ for some parameter $\theta \in (0,1)$.

Introduce the *standard phase boundary curve* by

$$r = \rho(\vartheta) = \begin{cases} 0, & 0 < \vartheta \leq 1/2, \\ \vartheta - 1/2, & 1/2 < \vartheta \leq 3/4, \\ (1 - \sqrt{1-\vartheta})^2, & 3/4 < \vartheta < 1. \end{cases}$$

We define

$$\rho^{\star}(\vartheta) = \begin{cases} \frac{1}{n_0+1}\rho(\vartheta), & \star = \text{no growth}, \\ \rho(\vartheta), & \star = \text{slow growth}, \\ (1-\theta)\rho(\frac{\vartheta}{1-\theta}), & \star = \text{regular growth}. \end{cases}$$

The curve $r = \rho^\star(\vartheta)$ partitions the two-dimensional phase space $\{(\vartheta, r) : 0 < \vartheta < 1, 0 < r < 1\}$ into three sub-regions where we have very different results for classification and feature selection (because of the partition of phase space, we call it the phase diagram):

- *Region of Impossibility.* $\{(\vartheta, r) : 0 < \vartheta < 1, 0 < r < \rho^\star(\vartheta)\}$. In this region, asymptotically, it is impossible to classify well: the fraction of misclassified samples $\geq 1/2(1+o(1))$. In a sense, this is a very difficult situation where not much can be learned from the training data and random guessing is almost the best one can do for classification. Also, it is impossible to distinguish the useful features from the useless ones. Therefore, in this region, neither classification nor feature selection can be successful.
- *Region of Possibility.* $\{(\vartheta, r) : 0 < \vartheta < 1, \rho^\star(\vartheta) < r < \vartheta\}$. In this region, there are procedures that can successfully classify (and the HCT classifier is one of them). That is, the sum of Type I and Type II errors of such procedures tend to 0 as $p \to \infty$. See Donoho and Jin (2008) for details.

  In this region, however, it is still impossible to distinguish the useful features from the useless ones. For any feature selection method, we either have a very large false discovery rate (FDR) or a very large non-discovery rate (NDR). FDR is the ratio between the number of useless features that are misclassified as true features and the total number of estimated useful features, and NDR is the ratio between the number of true features that are misclassified as useless and the total number of useful features.
- *Region of Certainty.* $\{(\vartheta, r) : 0 < \vartheta < 1, \vartheta < r < 1\}$. In this region, there are procedures that can successfully classify (and the HCT classifier is one of them), but it is also possible to distinguish the useful features from the useless ones. In fact, if we use the FDR-controlling procedure of Benjamini and Hochberg (1995) where the FDR-controlling parameter is set to tend to 0 slowly enough, then the resulting procedure has a sum of FDR and NDR that tends to 0 as $p \to \infty$.

Figure 1 displays all three regions in the case where n has a slow growth. While the results are asymptotic, we note that finite n simulation can be found in Donoho and Jin (2009), where the convergence rate of the classification error is also studied.

3 Data Sharing in High Dimensional Feature Selection and Classification

An interesting problem in the area of confidentiality and privacy protection is data sharing without violation of privacy. We are concerned with the situation that two or more parties want to share data for better "joint" inference, but they do not want the "privacy" of their separate databases to be compromised. The context we envision is very broad. In this section, we discuss data sharing in the context of feature selection and classification. We continue the discussion in Section 5, in the context of high dimensional variable selection.

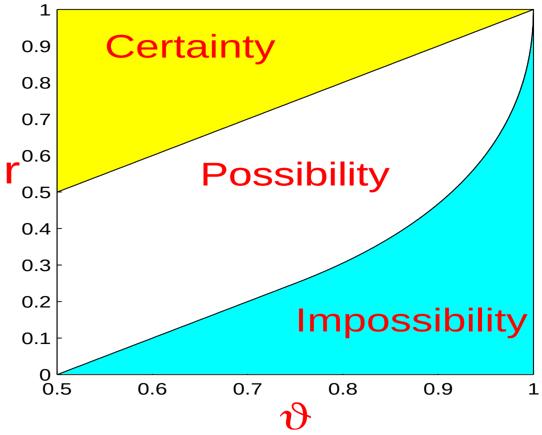

Figure 1: Partition of the phase space in high dimensional classification, where the x-axis and the y-axis calibrate the sparsity level and signal strength, respectively. In the cyan region, asymptotically, any trained classifier has a classification error $\gtrsim 1/2$ and thus fails completely. In the white region, successful classification is possible and there are trained classifiers (and HCT classifier is one of them) whose classification errors tend to 0. In the yellow region, it is not only possible to have successful classification, but also to identify almost all useful features.

Suppose now that we have N different parties, each of which has n_i, $1 \leq i \leq N$, copies of realizations of (X, Y). This is akin to the secure multi-party computation problem with horizontal partitioning. For simplicity, we focus on the case where

$$n_1 = n_2 = \ldots = n,$$

but the ideas can be generalized to much broader settings. These parties share a common goal: to build a trained classifier that can be used to predict the label when a new sample comes in.

The basic data of the parties take the form

$$(X_i^{(j)}, Y_i^{(j)}), \qquad 1 \leq i \leq n, \qquad 1 \leq j \leq N,$$

where i is the label for samples, and j is the label for parties. Each party may have a summarizing Z-vector: $Z^{(j)} = \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i X_i^{(j)}$, $1 \leq j \leq N$ where we take

$$Z^{(j)} \sim Normal(\sqrt{n}\mu, I_p). \tag{1}$$

Since useful features are both rare and weak, inference will be much easier if n is large. Unfortunately, n is typically small in many contemporary applications. Therefore, there are strong incentives for parties to share their data with one another. In fact, if all parties agree to share their data, they can use a common summarizing Z-vector

$$\tilde{Z} \equiv \frac{1}{\sqrt{N}} \sum_{j=1}^{N} Z^{(j)} \sim Normal(\sqrt{Nn}\mu, I_p), \tag{2}$$

where the signal is amplified by $\sqrt{N}$ times in strength. Still, they may want to do so in a manner that protects privacy of their individual databases.

There are many different scenarios we can envision regarding what data to focus on for privacy protection and what data to share. In many scenarios, it is more important to protect Y than X. For example, in a cancer study, knowing part (and sometimes all) of the coordinates of X does not necessarily lead to the identification of an individual's disease status, but knowing Y surely does. It is therefore of interest for us to protect Y, although we may also want to protect X.

We consider three types of data sharing:

- *Cautious Sharing*. All parties pool their data and run an algorithm in a black-box fashion. The black-box outputs some estimate of contrast mean vector μ, but does not allow any party to access the data of the others. We can accomplish this using secure multiparty computation methods as suggested above.
- *Regular sharing*. All parties release their summarizing Z-vector $Z^{(j)}$, but not $(X_i^{(j)}, Y_i^{(j)})$, $1 \leq i \leq n$, $1 \leq j \leq N$. In this approach, they release very limited information about their X variable and Y variable. In fact, if our goal is to protect the labels Y but not X, releasing $Z^{(j)}$ does not release any information on Y. Consider the case $n = 1$. Here $Z^{(j)} = Y \cdot X^{(j)}$, where $Y = \pm 1$. If an attacker only has access to the released vectors $Z^{(j)}$, but not to X or other side information, he/she could not tell whether $Y = 1$ or $Y = -1$. Note that situations where we have a large n are even more non-informative to the attacker. In such settings, the probability that the attacker can correctly identify the label Y is almost the same as that in the case where he/she does not have access to $Z^{(j)}$. In this sense, Regular Sharing is almost as secure as the Cautious Sharing.
- *Generous Sharing*. All parties release their X variables but not their Y-variables. At the same time, the black-box outputs the common summarizing Z-vector $\tilde{Z}$. Below we will see that this approach may significantly violate the privacy of individuals.

We now discuss the implications of data sharing on feature selection and classification. We discuss the cases of Cautious Sharing, Regular Sharing, and Generous Sharing separately.

3.1 Cautious Sharing

For the present purposes, we assume that the parties have two different goals associated with sharing their data: classification and feature selection.

Consider the case where the goal of the parties is merely classification, without much interest in the contrast mean vector μ or feature selection. Ideally, the black box can function as follows. First, it uses all data $(X_i^{(j)}, Y_i^{(j)})$, $1 \leq i \leq n$, $1 \leq j \leq N$, to calculate the vector $\tilde{Z}$ defined as in (2). Second, it uses $\tilde{Z}$ to estimate the HCT, denoted by $t_p^{HC}(\tilde{Z})$. Finally, it releases the weighted vector

$$\tilde{w}(j) = \mathrm{sgn}(\tilde{Z}(j)) \cdot 1_{\{|\tilde{Z}| \geq t_p^{HC}(\tilde{Z})\}}$$

to all parties. In the future, when a test feature X becomes available, each party can then construct

$$L(X) = \sum_{k=1}^{p} \tilde{w}(k) X(k),$$

and classify the corresponding Y label as ±1 according to $L(X) >< 0$. Consequently, the two-dimensional phase space $\{(\vartheta, r) : 0 < \vartheta < 1, 0 < r < 1\}$ can be partitioned as follows:

- *Region of Impossibility.* $\{(\vartheta, r) : 0 < \vartheta < 1, Nr < \rho^\star(\vartheta)\}$. In this region, even with data sharing, it is impossible for each party to have a successful classification.
- *Region of Substantially Helpful.* $\{(\vartheta, r) : 0 < \vartheta < 1, \rho^\star(\vartheta) < Nr < N\rho^\star(\vartheta)\}$. In this region, without data sharing, it is impossible for each party to have a successful classification; with data sharing, the black box algorithm yields a successful classification. Therefore, data sharing is substantially helpful.
- *Region of non-Substantially helpful.* $\{(\vartheta, r) : 0 < \vartheta < 1, r > \rho^\star(\vartheta)\}$. In this region, even without data sharing, each party is able to successfully classify (e.g., using the HCT classifier). For each party, data sharing is helpful in improving its classification results, but the improvement is not substantial.

In the above discussion, we assume that the contrast mean vector μ is not sensitive to the parties, and we have no intention to protect it. If this is not the case, then in the last step of the algorithm, the black box can choose not to release the weight vector w. If a party wishes to classify a test feature X, it has to input X to the black box, and the black-box outputs the predicted label only. Also, the black box should only allow a limited number of probes from each party, or there may be a security leak otherwise. To see the point, imagine that in the near future, a party has a continuous stream of unlabelled feature vectors rolling in. By probing the black box sufficiently many times, the party obtains labels for all of the components of its feature vector, and eventually it can use these labeled vectors to get a precise estimate of the vector μ. For example, consider the scenario where the party has probed the black box a total of m times, with vectors $X_1, X_2, \ldots, X_m$, and the black box predicts the corresponding labels as $\hat{Y}_1, \ldots, \hat{Y}_m$. Then the party can estimate μ by first obtaining the summarizing Z-vector, say, $Z = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \hat{Y}_i X_i$, and then estimate $\sqrt{m}\mu$ by hard thresholding,

$$\sqrt{m}\hat{\mu}(j) = Z(j) 1_{\{|Z(j)| \geq t\}}.$$

Jin (2009) shows that a good choice of t is $t = \sqrt{2\log(p)}$. Also, given that m is sufficiently large and that the prediction by the black box is reasonably accurate, the above thresholding scheme gives a good estimate of μ, and thus a security leak.

Next, we consider the case where the parties are not only interested in classification, but also feature selection. Suppose that the contrast mean vector μ is of common interest to all parties, and releasing it does not pose a security violation. In addition to the aforementioned classification rule, the black box can also perform feature selection as follows. Setting the FDR-controlling parameter q_p that tends to 0 slowly enough as $p \to \infty$ (e.g., $q_p = 1/\log(p)$), the black box applies the Benjamini and Hochberg procedures to $\tilde{Z}$. The procedure outputs a p by 1 vector of zeros and ones, $\mathcal{I}_p(\tilde{Z})$. The procedure determines the j-th feature as useful if and only if $\mathcal{I}_p(j) = 1$. Here again, the phase space partitions into three different regions as follows:

- *Region of Impossibility.* $\{(\vartheta, r):\ 0 < \vartheta < 1, Nr < \vartheta\}$. In this region, even with data sharing, it is impossible for each party to have a successful feature selection. For any procedures, we either have a large FDR or a large NDR.
- *Region of Substantially Helpful.* $\{(\vartheta, r):\ 0 < \vartheta < 1, \vartheta < Nr < N\vartheta\}$. In this region, without data sharing, it is impossible for each party to have a successful feature selection; with data sharing, the above algorithm yields a successful feature selection (i.e., the sum of FDR and NDR tends to 0 as $p \to \infty$).
- *Region of Non-Substantially Helpful.* $\{(\vartheta, r):\ 0 < \vartheta < 1, r > \vartheta\}$. In this region, even without data sharing, each party is able to successfully select features (e.g., using the HCT classifier). For each party, data sharing is helpful in improving the feature selection results, but the improvement is not substantial.

Additionally, Cautious Sharing is also secure, as long as the black box remains in the hands of a "trusted" third party or we use a more secure multi-party computation approach. Each party has very limited information about the other parties' X variables and Y variables, even when N is as small as 2.

3.2 Regular Sharing

The discussion of Regular Sharing is similar to that of Cautious Sharing, and we keep it to the following observations:

- The vector $\tilde{Z}$ is the sufficient statistic of the vector μ, and knowing $\tilde{Z}$ means knowing all information related to estimating μ.
- Consequently, as far as our goal is classification and feature selection, knowing $\tilde{Z}$ is almost equivalent to knowing other parties' variables X and Y (but we don't so there is less chance of a security leak). Regular sharing plays a similar role to that of Cautious Sharing, and the phase space partitions in the same way as that of Cautious Sharing.

- If the contrast mean vector is sensitive, then in general we should not use Regular Sharing for security reasons.

On the privacy side, all we can learn about the j-th party is the summarizing Z-vector $\sum_{i=1}^{n} Y_i^{(j)} X_i^{(j)}$. Based on the Z-vector, unless n is very small (e.g., $n = 1$), nothing about $(X_i^{(j)}, Y_i^{(j)})$ can be learned. However, we must note that when n is small, something about $X_i^{(j)}$ can be learned through the knowledge of μ: given a nonzero coordinate of μ, the corresponding coordinate of $X_i^{(j)}$ must also tend to be large in the magnitude, though the signs can not be learned. Therefore, Regular Sharing does not pose security problems with respect to the variable Y, although it may release the information about the vector μ, and consequently some information on the variable X when n is small.

3.3 Generous Sharing

Generous Sharing has a role in classification and feature selection similar to that of Cautious Sharing or Regular Sharing, and the phase space partitions in the same way. What makes it very different from Cautious/Regular Sharing is on the security side. Note that we only consider Generous Sharing when we don't need to protect the X variable, so we focus the discussion on the security of Y.

In detail, in the Region of Impossibility, $Nr < \rho^\star(\vartheta)$. It is impossible to classify successfully. In this case, knowing another party's X variable does not help much in predicting the Y variable, so we don't have a serious privacy problem. In Regions of Substantially Help and Non-Substantially Help, $Nr > \rho^\star(\vartheta)$. Knowing the another party's X variables permits one to accurately predict the corresponding Y variables. Thus we have the possibility of a serious privacy violation.

To conclude this section, we note that data sharing is very helpful to the parties in both classification and feature selection when useful features are both rare and weak. If we are only concerned with protecting the Y variables, both Cautious Sharing and Regular Sharing preserve more privacy, but Generous Sharing does not always do so. If we are also concerned with protecting the contrast mean vector μ as well, Regular Sharing does not always preserve privacy, and thus we need to use Cautious Sharing. While the specific statements of the main results depend on our models and assumptions, the general claim and idea apply to much broader settings.

4 Protecting Privacy By Adding Noise and Variable Selection

In this section, we relate the approach of privacy-preserving noise addition to high dimensional regression analysis. Suppose we have a p by 1 vector β that contains sensitive or confidential data that we want to protect. For simplicity, we assume that the coordinates of β take values from $\{0, 1\}$. Such a situation may arise in applications where β is the vector of diagnostic results (e.g., HIV vs. non-HIV; cancer vs. normal).

We are interested in privacy-preserving data mining. To proceed, we save the data in a black box, which allows probing in the following fashion. Fix the number of probes $n \geq 1$ and $\sigma > 0$. For each $1 \leq i \leq n$, the database generates a p by 1 "weight" vector x_i (the coordinates of which don't have to be positive or have a sum of 1) and a sample $z_i \sim Normal(0, \sigma^2)$, and outputs to the querier the weight vector x_i as well as the weighted average of the coordinates of β masked by some Gaussian noise:

$$y_i = x_i'\beta + z_i.$$

We add Gaussian noise in order to make a connection to the main body of the literature on high dimensional regression, but extensions that allow for non-Gaussian noise are possible. These might allow a link to the literature on differential privacy but we have not explored such a possibility to date.

We would like to know how the results of statistical inference depend on the amount of added noise. Towards this end, we let

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{pmatrix}, \qquad X = \begin{pmatrix} x_1' \\ x_2' \\ \cdots \\ x_n' \end{pmatrix}, \qquad Z = \begin{pmatrix} z_1 \\ z_2 \\ \cdots \\ z_n \end{pmatrix}.$$

We can capture the probing sequence in the following linear model,

$$Y = X\beta + Z, \qquad Z \sim Normal(0, \sigma^2 I_n). \tag{3}$$

For simplicity, we assume the black box releases the variance parameter σ, and thus it is known to the querier. Dividing both sides of (3) by σ, we have

$$Y = X\beta + Z, \qquad Z \sim Normal(0, I_n), \tag{4}$$

where for simplicity we use the same notation but Y, β, Z are different from that in (3) by a factor $1/\sigma$ (especially, the nonzero coordinates of β now equal to $1/\sigma$). Fixing $0 < \epsilon < 1$, we model the nonzero coordinates of β as iid samples from a mixture of two point masses:

$$\beta_j \overset{iid}{\sim} (1-\epsilon)\nu_0 + \epsilon\nu_\tau, \qquad \tau \equiv \frac{1}{\sigma}, \tag{5}$$

where ν_a denotes the point mass at a. Additionally, we model x_i as independent and identically distributed samples from $Normal(0, 1/n \cdot I_p)$,

$$x_i \overset{iid}{\sim} Normal(0, \frac{1}{n} I_p), \tag{6}$$

where the factor $1/n$ is chosen so that the diagonal entries of $X'X$ are approximately 1. In the literature, this is called the Gaussian design (Genovese et al., 2012). Note that in a closely-related setting, Dinur and Nissim (2003) use a Bernoulli design. We note that the Gaussian assumption on X is non-essential, and the Gaussian distribution can be replaced by many other thin-tailed distributions. What is essential here is that the large coordinates of the Gram matrix $X'X$ are close to that of I_p, while the small coordinates of $X'X$ are uniformly small.

The model uses four parameters (p, n, ϵ, τ). Motivated by the need to protect "massive" datasets somehow, we suppose both p and n are large, but n is much smaller than p. We don't allow unlimited probing of the database merely for security reasons. Additionally, we suppose ϵ is small, which may arise in the case when β is the diagnostic record of a population that are not badly affected by HIV. Lastly, we suppose τ is small or at most moderately large. This is the case where the data base is well-protected so that the nonzero coordinates of β can not be easily identified by the querier.

4.1 Variable Selection, the LASSO, and Marginal Regression

In statistics, identifying nonzero coordinates of β is the well-known problem of variable selection. The goal for variable selection is almost the opposite of that of statistical disclosure limitation or privacy protection: in the former we want to study when it is possible to fully recover the vector β, whereas in the latter we want to study when it is possible to prohibit such a full recovery. Therefore, we argue that studying one problem leads to a better understanding of the other, and vice versa.

There are many approaches to variable selection in recent literature. Among them are the LASSO (Chen et al., 1998; Tibshirani, 1996) and marginal regression (Fan and Lv, 2008). Marginal regression is simple and convenient computationally, especially when p is large.

The LASSO is a shrinkage and selection method for linear regression that minimizes the usual sum of squared errors, usually with a bound on the sum of the absolute values of the coefficients. The LASSO solution for the linear model problem in equation (4), $\hat{\beta}^{LASSO}$, is the vector that minimizes the following quantity:

$$\frac{1}{2}\|Y - X\mu\|^2 + \lambda\|\mu\|_1,$$

where $\|\cdot\|_1$ denotes the ℓ^1-norm and $\lambda > 0$ is a tuning parameter.

Marginal regression involves fixing a threshold t. If we multiply both sides of equation (4) by X' and denote the result as $\tilde{Y}$,

$$\tilde{Y} = X'X\beta + X'Z. \tag{7}$$

Marginal regression estimates β by

$$\hat{\beta}_j^{MR} = \begin{cases} \tilde{Y}_j, & |\tilde{Y}_j| \geq t, \\ 0, & |\tilde{Y}_j| < t, \end{cases} \qquad 1 \leq j \leq p;$$

in the statistical literature, this is called *hard thresholding*.

4.2 Phase Diagram for Linear Regression

Similar to the discussion in Section 2.2, we approach the variable selection problem with an asymptotic framework where p tends to ∞, and (ϵ, τ, n) are linked to p through fixed parameters. In detail, fixing $(\vartheta, \theta) \in (0,1)^2$ and $r > 0$, we model

$$\epsilon = \epsilon_p = p^{-\vartheta}, \qquad n = n_p = p^{\theta}, \qquad \tau = \tau_p = \sqrt{2r \log p}.$$

We assume

$$(1 - \vartheta) < \theta,$$

so that the number of signals (nonzero coordinates of β) is much smaller than the sample size n. This condition is almost necessary for successful variable selection (Donoho, 2006).

For a variable selection procedure that yields the quantity $\hat{\beta}$, we measure the errors by the Hamming distance:

$$\mathrm{Hamm}_p(\hat{\beta}) = \mathrm{Hamm}_p(\hat{\beta}; \epsilon_p, \tau_p, n_p) = E_{\epsilon_p, \tau_p}\left[E\biggl(\sum_{j=1}^{p} 1\{\mathrm{sgn}(\beta_j) \neq \mathrm{sgn}(\hat{\beta}_j)\} \biggr) \right].$$

In other words, we make an error if and only if either a signal is classified as a noise or a noise is classified as a signal. In terms of the Hamming distance, the variable selection problem also displays a watershed phenomenon similar to that we observed in classification, and the two-dimensional phase space $\{(\vartheta, r) : \ 0 < \vartheta < 1, \, r > 0\}$ similarly partitions into three regions, as illustrated in Figure 2, as follows:

- *Region of Exact Recovery.* $\{(\vartheta, r) : \ 0 < \vartheta < 1, \, r > (1 + \sqrt{1 - \vartheta})^2\}$. In this region, if we set the tuning parameter of the LASSO by $\lambda = \lambda_p = [(\vartheta + r)/(2r)]\tau_p$ and the threshold of marginal regression by $t = t_p = [(\vartheta + r)/(2r)]\tau_p$, then as $p \to \infty$,

  $$\mathrm{Hamm}_p(\hat{\beta}^{LASSO}) \to 0, \qquad \mathrm{Hamm}_p(\hat{\beta}^{MR}) \to 0.$$

  As a result, the probability that the LASSO or the marginal regression makes one or more errors in variable selection tends to 0. Therefore, both the LASSO and marginal regression yield exact recovery with overwhelming probability.

- *Region of Almost Full Recovery.* $\{(\vartheta, r) : \ 0 < \vartheta < 1, \, \vartheta < r < (1 + \sqrt{1 - \vartheta})^2\}$. In this region, on one hand, for any variable selection procedure $\hat{\beta}$,

  $$\frac{\mathrm{Hamm}(\hat{\beta})}{p\epsilon_p} \geq L_p p^{-(\vartheta - r)^2/(4r)}.$$

  On the other hand, for the LASSO and the marginal regression (where λ and t are set as above, respectively),

  $$\frac{\mathrm{Hamm}(\hat{\beta}^{LASSO})}{p\epsilon_p} \leq L_p p^{-(\vartheta - r)^2/(4r)}, \qquad \frac{\mathrm{Hamm}(\hat{\beta}^{MR})}{p\epsilon_p} \leq L_p p^{-(\vartheta - r)^2/(4r)}.$$

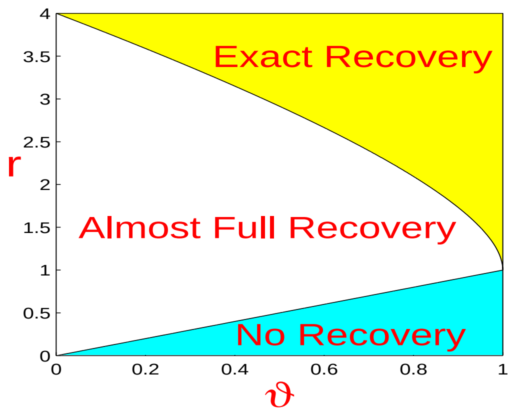

Figure 2: Phase diagram for variable selection, where the x-axis and y-axis calibrate the sparsity level and the signal strength, respectively. In the cyan region, asymptotically, successful recovery is impossible, and the Hamming error of any variable selection procedure $\gtrsim p\epsilon_p$. In the white region, it is possible to recover most of the signals (e.g., by the LASSO or marginal regression), but it is impossible to recover all signals. In the yellow region, it is possible to recover all signals (e.g., by the LASSO or marginal regression) with overwhelming probability.

Here, $L_p > 0$ denotes a multi-log(p) term that satisfies (a) $\lim_{p\to\infty} L_p p^{\delta} = \infty$ and (b) $\lim_{p\to\infty} L_p p^{-\delta} = 0$ for any $\delta > 0$. Note that the exponent of $p\epsilon_p p^{-(\vartheta-r)^2/(4r)}$ is $(1-\vartheta) - (\vartheta - r)^2/(4r) > 0$. Therefore, in this region, it is impossible to have exact recovery, but it is possible to have procedures (e.g., the LASSO or marginal regression) that yield almost full recovery .

- *Region of no recovery.* $\{(\vartheta, r) : 0 < r < \vartheta\}$. In this region, for all variable selection procedure $\hat{\beta}$,

$$\frac{\text{Hamm}(\hat{\beta})}{p\epsilon_p} \geq (1 + o(1)).$$

  In this region, variable selection is very difficult and all procedures fail completely (asymptotically).

See Genovese et al. (2012) for proofs and details.

Thus far, we have focused on privacy protection by adding noise. We now move to a different but closely related setting, where we discuss data sharing.

5 Privacy-Preserving Data Sharing in Variable Selection

We now continue the discussion from Section 3 on data sharing, but in the context of high dimensional variable selection.

Hall and Fienberg (2010) discuss the problem of privacy-preserving record linkage to form the database, and Hall et al. (2011) and Fienberg et al. (2012) describe the release of regression and logistic regression results for a given model from linkable databases. For the present purposes, we characterize the merged database using the linear model introduced in (3):

$$Y = X\beta + z, \qquad X = X_{n,p}, \qquad z \sim Normal(0, I_n).$$

There are many types of data combination from the parties. For example, the rows of X may come from different parties (horizontal partitioning of X), or the columns of X may come from different parties (vertical partitioning of X). In other words, different parties may hold data from different samples of individuals or data on different attributes for the same individuals. We focus our discussion on the horizontal partitioning case. Despite the similarity to the setting in the preceding section, the results here are different.

We suppose (Y, X) has the following partition,

$$Y = \begin{pmatrix} Y^{(1)} \\ Y^{(2)} \\ \dots \\ Y^{(N)} \end{pmatrix}, \qquad X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \\ \dots \\ X^{(N)} \end{pmatrix}, \tag{8}$$

where $Y^{(j)}$ is an $n_j \times 1$ vector and $X^{(j)}$ is an $n_j \times p$ matrix, $1 \leq j \leq N$. For simplicity, we assume $n_1 = n_2 = \dots n_N = n$, and rows of X are iid samples from

$$Normal(0, \frac{1}{nN} I_p).$$

Additionally, fixing $\vartheta \in (0, 1)$ and $r > 0$, we model β as

$$\beta_j \overset{iid}{\sim} (1 - \epsilon_p)\nu_0 + \epsilon_p \nu_{\tau_p}, \qquad \epsilon_p = p^{-\vartheta}, \qquad \tau_p = \sqrt{2r \log p}.$$

The primary interest is on identifying the nonzero coordinates of β. We are concerned with protecting Y, and not too much on protecting X, although the latter may also be of interest.

For simplicity, we assume the added noise as independently Gaussian. Extension to settings where the added noise has more complicated structures is possible (for other options of added noise, see Blum et al. (2008), Hardt and Rothblum (2010), and Hardt et al. (2010) , for example). Our focus is on the privacy associated with the sparse signal vector β, which differs somewhat from the focus in these papers. We leave the study along this line to the future.

Similar to our discussion in Section 3, we focus on Cautious Sharing and Generous Sharing and omit the discussion of Regular Sharing because of its similarly to that of Cautious Sharing.

- *Cautious Sharing.* All parties pool the data into a black box, which outputs some estimate, say, $\hat{\beta}$. For example, $\hat{\beta}$ can be the labels of coordinates of β that are estimated as nonzero. An individual party does not have access to the X variables of other parties. The vector $\hat{\beta}$ is all that is shared by all parties.
- *Generous Sharing.* Similar to Cautious Sharing, each party has access to the vector $\hat{\beta}$ which is computed in a black-box. The difference is that each party has access to the X variables of others, but not the Y variable.

Note that the major difference of the two types of sharing is not the performance of variable selection, but the privacy-preservation.

In the two-dimensional phase space $\{(\vartheta, r) :\ 0 < \vartheta < 1, r > 0\}$, two types of sharing yield the same partition of regions as follows:

- *Region of No Recovery.* $\{(\vartheta, r) :\ 0 < \vartheta < 1,\ 0 < Nr < \vartheta\}$. In this region, for any variable selection procedures $\hat{\beta}$, the Hamming distance satisfies $\mathrm{Hamm}_p(\hat{\beta}) \gtrsim p\epsilon_p$, even with data sharing. Therefore, successful variable selection is impossible.
- *Region of Substantially Helpful.* $\{(\vartheta, r) :\ 0 < \vartheta < 1, \vartheta < Nr < N\vartheta\}$. In this region, without data sharing, successful variable selection is impossible; with data sharing, both the LASSO and marginal regression yield almost full recovery. Therefore, data sharing is very helpful.
- *Region of Non-Substantially Helpful.* $\{(\vartheta, r) :\ 0 < \vartheta < 1,\ r > \vartheta\}$. In this region, even without data sharing, both the LASSO and marginal regression yield almost full recovery. Of course, the performance of these procedures are even better with data sharing, but the improvement is not substantial.

Not surprisingly, data sharing has a very positive impact on variable selection.

With regard to privacy, we recall that our primary interest is in protecting Y. Here we see that Cautious Sharing offers a form of security: one party has no access to the X variable of other parties, so it does not directly raise an issue associated with the privacy of them. In contrast, Generous Sharing is not secure: one party has access to both $\hat{\beta}$ and the X-variables of the other parties, and can conveniently use this information to predict the corresponding Y variables.

The result is closely related to work on phase-transition by Dwork et al. (2007), but is different in important ways. There are four key parameters in the regression model: the dimension p, the sample size n, the sparsity level ϵ_p, and the signal strength τ_p. The focus of Dwork et al. (2007) has been on the interaction of p and n, while that of the current paper is on the interaction of ϵ_p and τ_p. In fact, the signal vector β is assumed to be sparse here but not necessarily sparse in Dwork et al. (2007). For these reasons, our work focuses on very different settings from that in Dwork et al. (2007), and the phase-transition phenomena found in this paper is very different from that found in Dwork et al. (2007). Also, Dwork et al. (2007) uses square error loss to measure the risk and we use the Hamming distance (note that the focus of this paper is on the privacy of the nonzero coordinates of β, not the whole vector). Last, the lower bound argument in Dwork et al. (2007) is very different from that in our paper. In Dwork et al. (2007), by "failure", they mean the failure of a specific algorithm, namely the linear programming (LP) algorithm mentioned there. In our setting, by "failure", we mean the failure of any methods, computationally feasible or not. Therefore, our claim on the lower bound is much stronger.

6 Discussion

Much of the literature on privacy and confidentiality protection of statistical databases has focused on the release of summaries extracted from the data, e.g., noisy versions of sufficient statistics under a model, and has used a limited set of ideas from statistical theory. But with the growing size and complexity of databases and the desire to extract information from them, there is the need to think about aspects of privacy protection in different contexts and with more varied methodology. Tools for high dimensional data analysis are now a central part of the statistical literature, and methods often focus on situations where the signals tend to be both sparse and weak, a direct result of the so-called "large dimension and small sample size" scenario. One way to amplify the signal is to increase the sample size through the sharing of different but related datasets across parties.

We have investigated aspects of privacy protection for two related problems associated with high dimensional statistical analysis: feature selection and classification, and variable selection in a linear regression model. We characterize both problems using the phase diagram that comes from the modern statistical literature on the topic. For the classification problem, the phase diagram consists of three different phases, in which correspondingly, successful classification is impossible, successful classification is possible but identifying all useful features is impossible, and both successful classification and identifying all useful features is possible. For the variable selection problem, the phase diagram also consists of three different phases in a similar fashion, in which correspondingly, we have no recovery, almost full recovery, and exact recovery. We view the phase diagram as a benchmark for evaluating the inferential performance of different procedures and for evaluating the privacy protection impact of data sharing. In particular, we investigate the impact of three different types of data sharing that we label as Cautious Sharing, Regular Sharing, and Generous Sharing. Depending on the number of parties and the sample size of each party, data sharing may impact the statistical inference in three different ways: not helpful because the signals are too sparse and weak even with data sharing, substantially helpful, and non-substantially helpful as the signals are strong enough even without data sharing. Cautious sharing is generally secure. Regular sharing may be secure if we are concerned only with the privacy of the

Y variable (class label in the classification setting, response variable in the regression setting), but is not if we are also concerned with the X variable. Generous sharing is usually not secure, and the attacker may be able to precisely identity the label Y which we wish to protect.

Our work is loosely intertwined with several topics in privacy and confidentiality protection, including secure multi-party computation, privacy protection through noise addition, and differential privacy. Our methodology and approach differ from those in the current literature in important ways precisely because our focus is on high dimensional data.

Acknowledgments

The research was partially supported by Army contract DAAD19-02-1-3-0389 to Cylab, by NSF Awards BCS0941518 and DMS0908613 to the Department of Statistics at Carnegie Mellon University, and by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office through a grant for the joint Carnegie Mellon/Singapore Management University Living Analytics Research Centre.

References

Amirbekyan, A. and Estivill-Castro, V. (2007). Privacy-preserving regression algorithms. In *Proceedings of the 7th WSEAS International Conference on Simulation, Modelling and Optimization (SMO'07)*. Stevens Point, Wisconsin, USA: World Scientific and Engineering Academy and Society (WSEAS). 37–45.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 29(4):1165–1188.

Blum, A., Ligett, K., and Roth, A. (2008). A learning theory approach to non-interactive database privacy. In *Proceedings of the 40^{th} Annual ACM Symposium on Theory of Computing (STOC'08)*. ACM Press. 609–618.

Burridge, J. (2003). Information preserving statistical obfuscation. *Statistics and Computing*, 13(4):321–327.

Chaudhuri, K. and Monteleoni, C. (2008). Privacy-preserving logistic regression. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou (eds.), *Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems (NIPS'08)*. MIT Press. 289–296.

Chauduri, K. C., Monteleoni, C., and Sarwate, A. D. (2011). Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12:1069–1109.

Chen, S., Donoho, D., and Saunders, M. (1998). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61.

Dinur, I. and Nissim, K. (2003). Revealing information while preserving privacy. In *Proceedings of the ACM SIGMOD Principles of Database Systems Conference (PODS'03)*. ACM Press. 202–210.

Donoho, D. (2006). For most large underdetermined systems of linear equations the minimal ℓ^1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(7):907–934.

Donoho, D. and Jin, J. (2008). Higher criticism thresholding: Optimal feature selection when useful features and rare and weak. *Proceedings of the National Academy of Sciences*, 105(39):14790–14795.

— (2009). Feature selection by higher criticism thresholding: Optimal phase diagram. *Philosophical Transcations of the Royal Society, Series A*, 367:4449–4470.

Du, W., Han, Y.-S., and Chen, S. (2004). Privacy preserving multivariate statistical analysis: Linear regression and classification. In M. W. Berry, U. Dayal, C. Kamath, and D. Skillicorn (eds.), *2004 SIAM International Conference on Data Mining*. Lake Buena Vista, Florida: ACM.

Dwork, C. (2008). Differential privacy: A survey of results. In *Proceedings of the 5th International Conference on Theory and Applications of Models of Computation (TAMC'08)*. Berlin, Heidelberg: Springer-Verlag. 1–19.

Dwork, C., McSherry, F., and Talwar, K. (2007). The price of privacy and the limits of LP decoding. In *Proceedings of the 39th Annual ACM Symposium on Theory of Computing (STOC'07)*. ACM Press. 85–94.

Dwork, C. and Naor, M. (2010). On the difficulties of disclosure prevention in statistical databases or the case for differential privacy. *Journal of Privacy and Confidentiality*, 2(1):93–107.

Dwork, C. and Smith, A. (2009). Differential privacy for statistics: What we know and what we want to learn. *Journal of Privacy and Confidentiality*, 1(2):135–154.

Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space (with discussion). *Journal of the Royal Statistical Society, Series B*, 70:849–911.

Fienberg, S. E., Hall, R., and Nardi, Y. (2012). Achieving both valid and secure logistic regression analysis on aggregated data from different private sources. *Journal of Privacy and Confidentiality*, 4(1):189–220.

Fienberg, S. E. and Jin, J. (2009). Statistical disclosure limitation for data access. In L. Liu and M. T. Özsu (eds.), *Encyclopedia of Database Systems*. Springer. 2783–2789.

Fienberg, S. E., Nardi, Y., and Slavkovic, A. B. (2008). Valid statistical analysis for logistic regression with multiple sources. In C. S. Gal, P. B. Kantor, and M. E. Lesk (eds.), *Proceedings of the Workshop on Priavacy and Security, Interdisciplinary Studies in Information Privacy and Security (ISIPS)*, vol. 5661 of *LNCS*. Springer. 82–94.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188.

Genovese, C., Jin, J., Wasserman, L., and Yao, Z. (2012). A comparison of the LASSO and marginal regression. *Journal of Machine Learning Research*, 13:2069–2105.

Hall, R. and Fienberg, S. E. (2010). Privacy-preserving record linkage. In J. Domingo-Ferrer and E. Magkos (eds.), *Privacy in Statistical Databases 2010 (PSD 2010)*, vol. 6344 of *LNCS*. Berlin: Springer. 269–283.

Hall, R., Fienberg, S. E., and Nardi, Y. (2011). Secure multiple linear regression based on homomorphic encryption. *Journal of Official Statistics*, 27(4):669–691.

Hardt, M., Ligett, K., and McSherry, F. (2010). A simple and practical algorithm for differentially-private data release. arXiv:1012.4763.

Hardt, M. and Rothblum, G. (2010). A multiplicative weights mechanism for interactive privacy-preserving data analysis. In *Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science (FOCS 2010)*. IEEE. 61–70.

Jin, J. (2009). Impossibility of successful classification when useful features are rare and weak. *Proceedings of the National Academy of Sciences*, 106(22):8859–8864.

Muralidhar, K. and Sarathy, R. (2006). Data shuffling: A new masking approach for numerical data. *Management Science*, 52(5):658–670.

O'Keefe, C. M. and Good, N. M. (2009). Regression output from a remote analysis server. *Data & Knowledge Engineering*, 68(11):1175–1186.

Sanil, A. P., Karr, A. F., Lin, X., and Reiter, J. P. (2004). Privacy preserving regression modelling via distributed computation. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04)*. New York: ACM. 677–682.

Sarwate, A. D., Chaudhuri, K., and Monteleoni, C. (2009). Differentially private support vector machines. *CoRR*, abs/0912.0071.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288.

Ting, D., Fienberg, S. E., and Trottini, M. (2008). Random orthogonal matrix masking methodology for microdata release. *International Journal of Information and Computer Security*, 2(1):86–105.

Zhou, S., Lafferty, J., and Wasserman, L. (2009). Compressed and privacy-sensitive sparse regression. *IEEE Transactions on Information Theory*, 55(2):846–866.