Preserving Privacy of Continuous High-dimensional Data with Minimax Filters

Jihun Hamm
Department of Computer Science and Engineering, The Ohio State University, OH 43210, USA

Abstract

Preserving privacy of high-dimensional and continuous data such as images or biometric data is a challenging problem. This paper formulates this problem as a learning game between three parties: 1) data contributors using a filter to sanitize data samples, 2) a cooperative data aggregator learning a target task using the filtered samples, and 3) an adversary learning to identify contributors using the same filtered samples. *Minimax* filters that achieve the optimal privacy-utility trade-off from broad families of filters and loss/classifiers are defined, and algorithms for learning the filers in batch or distributed settings are presented. Experiments with several real-world tasks including facial expression recognition, speech emotion recognition, and activity recognition from motion, show that the minimax filter can simultaneously achieve similar or better target task accuracy *and* lower privacy risk, often significantly lower than previous methods.

1 Introduction

When databases of multiple subjects are released in public, the privacy of data contributors becomes an important issue. Several privacy preserving mechanisms for data publishing have been studied (see [8] for a review), including k-anonymity [21], secure multiparty computation [24], and differential privacy [6, 5, 4]. The majority of privacy-preserving mechanisms are designed for databases of categorical attributes, and may not be suitable for databases of continuous high-dimensional attributes such as photos, videos, audio clips, and biometric data. For example, it is argued that k-anonymity is ineffective for high-dimensional sparse databases [14], and differential privacy of high-dimensional continuous data is not as well understood as discrete data [19]. Furthermore, differential privacy guarantee deteriorates with the size of samples per subject [4], which is problematic if data consist of, e.g., a large number of video frames per subject. Differential privacy takes into account only the range of data values and not their distributions. While it provides a strong privacy guarantee, it costs unnecessarily high loss of utility if we are concerned with databases of a specific data distribution such as faces.

This paper takes a learning approach to preserving privacy of continuous high dimensional databases, with the aim of achieving anonymity for given and unseen data in terms of expected risks. Suppose a data aggregator collects data from multiple contributors[1], and sanitizes the data using a privacy mechanism before publishing. The sanitization mechanism considered in the paper is any differentiable family of non-invertible deterministic transformations of the original data samples [2], which will be referred to as 'filters'. An adversary's goal is to *identify* data contributors from the filtered data released in public using machine learning classifiers. The privacy problem can be viewed as a learning game between three parties: data contributors, a data aggregator, and an adversary. Data contributors choose a filter(s) to prevent adversaries from inferring contributor identities from the filtered data, while allowing the cooperative aggregator to learn a useful learning task(s) from the filtered data. In this context, the privacy breach and the utility of data can be measured by *classification risks*, of the subject-identification task and a target task(s), respectively[3]. A more formal description of the problem is as follows.

Appearing in Proceedings of the 18^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38.

[1] The terms 'contributor' and 'subject' will be used interchangeably in this paper.

[2] Randomized algorithms may also be used, but the main focus is on non-randomized algorithms.

[3] Other non-classification type tasks will be discussed in Section 2.2.

Problem: Suppose there are S data contributors. Let $x \in \mathbb{R}^D$ be a continuous high-dimensional data sample, $g(x;u) : \mathbb{R}^D \to \mathbb{R}^d$ a deterministic non-invertible filter parameterized by u, z a target task label, and y a subject identity label. For example, the goal is to allow recognition of 'smile' $z \in \{-1, 1\}$ without revealing the identity $y \in \{1, ..., S\}$ from a face image x after processing it with a linear or nonlinear filter $g(x;u)$. An adversary chooses the best model v from a family of loss/classification models to minimize (maximize) the risk (accuracy) of a subject classification task, which is called the *privacy risk*

$$f_{\text{priv}}(u, v) = -E[l(y^{\text{pred}}(g(x;u);v), y^{\text{true}})], \quad (1)$$

while the data aggregator also chooses the best model w from a (different) family of loss/classification models to minimize (maximize) the risk (accuracy) of a target task, which is called the *utility* or *target risk*

$$f_{\text{util}}(u, w) = E[l(z^{\text{pred}}(g(x;u);w), z^{\text{true}})]. \quad (2)$$

The privacy risk has a negative sign, since a gain for the adversary is a loss for data contributors, and vise versa. Finally, the goal for the data contributors is to choose a filter $g(x;u)$ from a family that attains the optimal utility-privacy risk trade-off.

Minimax filter: The optimal privacy risk value for data contributors is the minimum risk in face of the worst adversary:

$$\min_u \max_v f_{\text{priv}}(u, v), \quad (3)$$

and the optimal utility risk value is the minimum risk from a fully cooperative aggregator:

$$\min_u \min_w f_{\text{util}}(u, w). \quad (4)$$

Omitting the details for now, the minimax filter is defined as follows:

Definition 1. *Given a family of filters $\mathcal{G} = \{g : \mathbb{R}^D \to \mathbb{R}^d\}$, a family of loss/classifiers for the privacy risk $\{f_{\text{priv}} : \mathbb{R}^d \to \mathbb{R}\}$, a family of loss/classifiers for the target risk $\{f_{\text{util}} : \mathbb{R}^d \to \mathbb{R}\}$, a* **minimax filter** *$g(\cdot;u) \in \mathcal{G}$ parameterized by u, is the solution to the optimization problems (3) and (4).*

By definition, a minimax filter is an optimal filter for preserving privacy of data contributors for a given data distribution in terms of expected risks. After learning a filter, the data aggregrator can release the filtered data in public, with the assurance that no other filter from the family provides better privacy preservation and utility. Algorithms to find minimax filters are presented in Section 2, which builds on a classic method of minimax optimization (see [18] for a review). In this setting of a learning game among three parties, the data aggregator is assumed to act truthfully on behalf of the data contributors. However, the aggregator needs to store the original data from the contributors to find optimal filters, which leaves the aggreagtor vulnerable to privacy breaches by mistakes or by external attacks. Therefore, an additional protocol for learning optimal filters is proposed in the paper, where no party needs to access original data nor filter parameters of others during learning (see Figure 1). In addition, this protocol allows learning of **personalized** minimax filters to adapt to each contributor's data. For example, if medical data from multiple institutions are to be collected for aggregate analyses, individual filters for each institution (and for each subject from the institution) can be learned and implemented in a distributed way without storing all private data in a single location. The corresponding algorithms are presented in Section 3.

Contributions: This paper proposes a novel minimax formulation for optimal utility-privacy trade-off in high-dimensional continuous data; the paper provides practical algorithms that find minimax filters for a broad family of filters and losses/classifiers with assumptions of differentiability only; the paper also presents a distributed protocol for learning personalized minimax filters assuming untrusted aggregators.

Advantages of our approach compared to previous approaches are further explained in Related work and Experiments sections. In particular, experiments with several real-world tasks including facial expression recognition, speech emotion recognition, and activity recognition from motion, show that publicly available numerical databases are surprisingly susceptible to subject identification attacks, and that minimax filters can reduce the privacy risks to near chance levels without sacrificing utility by much.

2 Learning minimax filter

In this section, formulation and optimization methods of minimax filters are presented in detail.

2.1 Privacy risk

The privacy risk of data contributors was defined as the negative of the average subject classification loss (1). One can use either the multiclass classification loss (=identifying the subject out of S subjects)

$$f_{\text{priv}}(u, v) = \frac{1}{N} \sum_i -l(y(g(x_i;u);v), y_i), \quad (5)$$

where $y \in \{1, ..., S\}$, or the binary classification loss (=identifying if the owner of the sample x is s or not),

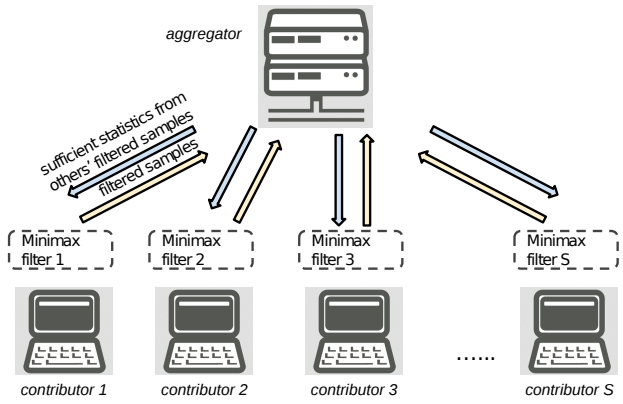

Figure 1: Distributed learning of individual minimax filters. Each contributor learns an individual minimax filter whose parameters are not revealed to the aggregator nor other contributors. All communications involve filtered samples (or sufficient statistics from them), and the aggregator only manages the concurrency of filtered samples.

where $y = -1$ means 'not the same subject' and $y = 1$ means 'the same'. With the binary loss, there are S such losses one for each subject, which can be averaged[4] as:

$$f_{\text{priv}}(u, \bar{v}) = \frac{1}{N} \sum_s \sum_{i \in I_s} -l(y(g(x_i; u); v_s), y_i), \quad (6)$$

where $\bar{v} = [v_1, ..., v_S]$ are the set of parameters for S subjects, and I_s is the sample index for subject s. The binary loss has one potential problem – there are $(S-1)$-times more negative samples from other subjects than positive samples from self, assuming the same number of sample per subject. Consequently, the high chance level accuracy $(S-1)/S$ gives a false sense of a high privacy risk. To address this, the paper uses a weighted risk that weighs the cost of true-positive $(S-1)$ times more than the cost of true-negatives, making the chance level accuracy 0.5 and not $(S-1)/S$. The modified binary loss is

$$\hat{l}(y(g(x_i)), y_i) = \begin{cases} (S-1) \cdot l(y(g(x_i)), y_i), & \text{if } y_i = 1 \\ l(y(g(x_i)), y_i), & \text{if } y_i = -1 \end{cases} \quad (7)$$

If each subject has different number of samples, the weights can be changed accordingly.

2.2 Joint utility-privacy risk

Trivial solutions to privacy risk minimization already exists – filters that output random or fixed numbers independent of actual data. However, such filters have no utility whatsoever for data aggregators. To avoid trivial solutions, it is necessary to include the secondary goal of minimizing the utility risk

$$f_{\text{util}}(u, w) = \frac{1}{N} \sum_i l(z(g(x_i; u); w), z_i), \quad (8)$$

by $\min_u \min_w f_{\text{util}}(u, w)$ which is equivalent to $\min_u [-\max_w (-f_{\text{util}}(u, w))]$. Consequently, a joint privacy and utility risk minimization can be defined using a constant ρ (say, 1) that weighs the relative importance of privacy and utility[5] :

$$\min_u \left[\max_v f_{\text{priv}}(u, v) - \rho \max_w (-f_{\text{util}}(u, w)) \right]. \quad (9)$$

Any task, including multiple tasks, may be used as a utility task(s) unless the target z and the identity y variables are totally dependent given data $g(x)$. Note that other non-classification type tasks can also be used as a target, such as the truthfulness of reconstructed data in terms of least-squares errors

$$f_{\text{util}}(u, w) = \frac{1}{N} \sum_i \|h(g(x_i; u); w) - x_i\|, \quad (10)$$

where $h(\cdot; w)$ is from a family of differentiable functions $\{\mathbb{R}^d \to \mathbb{R}^D\}$ which reconstruct the original sample using linear or nonlinear functions. Note that $h(g(x))$ cannot be an identity function if the filter is non-invertible such as dimensionality reduction.

2.3 Minimax optimization

The joint utility-privacy risk minimization (9) is an unconstrained continuous minimax problem (see [18] for a review). It is equivalent to the following problems:

$$\min_u [\max_v f_{\text{priv}}(u, v) - \rho \max_w (-f_{\text{util}}(u, w))] \quad (11)$$

$$= \min_u [\Phi_{\text{priv}}(u) - \rho \, \Phi_{\text{util}}(u)] = \min_u \Phi(u). \quad (12)$$

The minimax problem involving three variables u, v, w is a variation of a two-variable minimax problem $\min_u \max_v f(u, v) = \min_u \Phi(u)$. For simplicity, the two variable problem will be used for describing the minimax problem. Continuous minimax problems are more challenging than standard minimization problems for the following reasons. First, a closed-form max function $\Phi(u)$ usually does not exist; otherwise the problem is the same as the standard minimization. Second, $\arg\max_v f(u, v)$ may have multiple v's as solutions, depending on which $\Phi(u)$ becomes different functions. Third, it has two loops: an outer loop to

[4] Minimizing individual losses instead of the average loss will be considered in Section 3.

[5] The joint problem can also be formulated as minimizing the utility risk with a constrain on the privacy risk, which will be similar to solving (9) multiple times with an increasing ρ using interior-point methods.

minimize $\Phi(u)$ and an inner loop to maximize $f(u, v)$. However, there are several known classic algorithms for continuous minimax problems which require only weak assumptions on f. The f is assumed to be continuously differentiable with u and v, and $\nabla_u f$ continuously differentiable with v. A first-order optimization method was proposed by Panin [17] and refined later by Kiwiel [11], who uses a linear approximation of f at a fixed u along the direction q

$$f^l(q, v) = f(u, v) + \langle \nabla_u f(u, v),\ q \rangle, \tag{13}$$

and the approximate max function

$$\Phi^l(q) = \max_v f^l(q, v). \tag{14}$$

Using this approximation, a line search can be performed along the descent direction q that minimizes the max function Φ. In particular, with additional assumptions of Lipschitz continuity of $\nabla_u f$ and compactness of the domains for u, v, Kiwiel's algorithm monotonically decreases f for each iteration and converges to a stationary point u^*, i.e., a point u for which $\max_v \langle \nabla_u f(u^*, v),\ q \rangle \geq 0$ for all directions q. This paper uses Kiwiel's algorithm (Algorithm 1) to solve the joint utility-privacy optimization problem.

Algorithm 1 Kiwiel's algorithm

Main routine (see Supplementary Material for a full description)
Input: function pointers f, Φ, Φ^l_t, max iteration T
Init: Select u_1, v_1 randomly
Output: optimal parameter u_{T+1}
Begin:
  for $t = 1, ..., T$ **do**
    Solve $\Phi(u_t)$
    Find descent direction q_t by $\min_q \Phi^l(q)$ (14)
    Exit if solution converged
    Perform line search and update $u_{t+1} = u_t + \alpha_t \cdot q_t$
  end for

Remarks: There is a special case of continuous minimax problems known as saddle point problems, in which $f(u, v)$ is convex in u and concave in v (consider $f(u, v) = u^2 - v^2$). Analogous to (strictly) convex problems, $f(u, v)$ has a global minimum (u^*, v^*) which satisfies $f(u^*, v) \leq f(u^*, v^*) \leq f(u, v^*)$, and its convergence rates are previously analyzed [15]. Unfortunately, the problem at hand is not a saddle-point problem. To see this, suppose one chooses a family of convex losses (e.g., least-squares, hinge, logistic, or exponential losses), and the family of linear classifiers. With a d-dimensional filter output $g(u) = [g_1(u), ..., g_d(u)]$, the Hessian of the loss $l(g(u); v)$ with respect u is (if it exists)

$$H_u[l] = (J_u[g])' H_g[l] (J_u[g]) + \sum_k \frac{dl}{dg_k} H_u[g_k], \tag{15}$$

where $H[\cdot]$ and $J[\cdot]$ are Hessian and Jacobian matrices. Since the first term $H_g[l]$ is positive semidefinite for a convex l, it will be difficult in general for the LHS $H_u[l]$ to be negative semidefinite. For example, with linear filters, the second term is zero ($H_u[g_k] = 0$), and consequently l is convex in both u and v, which is not a saddle-point problem. However, the presence of local minima does not prevent algorithms from finding a good solution in practice. In the experiments with real datasets, local minima did not pose a noticeable problem in achieving promising results.

Algorithm 2 Joint minimax algorithm

Input: data $\{(x_i, y_i, z_i)\}$, loss/classifier $l(\cdot\,; v, w)$, filter $g(\cdot\,; u)$, params for Kiwiel's algorithm
Output: minimiax filter params u
Subfunctions:
- f_{priv}: outputs the value and derivatives $\nabla_u df_{\text{priv}}, \nabla_v df_{\text{priv}}, \frac{d^2 f_{\text{priv}}}{dudv}$ given u, v
- f_{util}: similar to f_{priv}
- Φ_{priv}: outputs $\max_v f_{\text{priv}}(u, v)$ given u
- Φ_{util}: similar to Φ_{priv}
- Φ^l_{priv}: outputs $\max_v [f_{\text{priv}}(u, v) + \langle \nabla_u f_{\text{priv}}(u, v),\ q \rangle]$ and optimal v given u, q
- Φ^l_{util}: similar to Φ^l_{priv}

Main functions:
- $f(u, v, w)$: outputs $f = f_{\text{priv}}(u, v) + \rho\ f_{\text{util}}(u, w)$ and derivatives $\nabla_u f, \nabla_v f, \nabla_w f, \frac{d^2 f}{dudv}, \frac{d^2 f}{dudw}$ given u, v, w
- Φ: outputs $(\Phi_{\text{priv}}(u) - \rho\ \Phi_{\text{util}}(u))$ given u
- Φ^l: outputs $(\Phi^l_{\text{priv}}(u, q) - \rho\ \Phi^l_{\text{util}}(u, q))$, and optimal v, w given u and q

Begin:
  Call Algorithm 1 passing pointers to f, Φ, Φ^l and hyperparameters

2.4 Algorithm

The joint utility-privacy minimax algorithm is summarized in Algorithm 2. It is a meta-algorithm that wraps around the two-variable minimax solver (Algorithm 1). For given data and a given family of filters and a family of loss/classifiers, the algorithm defines subfunctions to evaluate f and its derivatives. It also defines sub-optimization routines to find Φ and Φ^l, which can use any appropriate optimizer for the given loss/classifier type. The final minimax solution is found by calling Kiwiel's algorithm (1) passing the

function pointers to f, Φ and Φ^l and other parameters as arguments.

3 Distributed learning protocol

Given a sufficient amount of data per subject, it is possible to learn *personalized* minimax filters adapted to each contributor instead of a common filter for all contributors. This will be especially useful if certain contributors have data distributions very different from the population distribution. Each contributor has its own filter parameters u_s and privacy task parameters v_s, and shares the common task parameters w with other contributors. Let $\bar{u} = \{u_1, ..., u_S\}$ and $\bar{v} = \{v_1, ..., v_S\}$ denote the sets of individual filter parameters and privacy risk parameters. The personalized minimax filters are found by solving

$$\min_{\bar{u}} \ \bar{\Phi}(\bar{u}) = \min_{\bar{u}} \ [\max_{\bar{v}} \bar{f}_{\text{priv}}(\bar{u}, \bar{v}) - \rho \max_{w}(-\bar{f}_{\text{util}}(\bar{u}, w))], \tag{16}$$

where the privacy and the utility risks are

$$\bar{f}_{\text{priv}}(\bar{u}, \bar{v}) \quad = \quad \frac{1}{N} \sum_{s} \sum_{i \in I_s} -l(y(g(x_i; u_s); v_s), y_i), \tag{17}$$

$$\bar{f}_{\text{util}}(\bar{u}, w) \quad = \quad \frac{1}{N} \sum_{s} \sum_{i \in I_s} l(z(g(x_i; u_s); w), z_i). \tag{18}$$

For the privacy risk, the weighted binary classification loss (7) is used, where the subject identity label y is -1 for the samples from other contributors and is 1 for the contributors' own samples. While this optimization can be solved in batch as in Section 2, there is an approach which provides additional privacy. In batch learning, an aggregator needs access to original data, which leaves the aggregator vulnerable to privacy breaches by mistakes or attacks. These can be prevented by a distributed learning protocol in which no party has access to original data nor filter parameters of others during learning. This is achieved by making the aggregator serve only as a coordinator of the distributed learning and delegating the actual optimization work to each contributor. Figure 1 illustrates this idea.

3.1 Algorithm

The proposed solution for optimizing the personalized risk (16) is to use the block coordinate descent. The objective is minimized by finding optimal parameters for each contributor s while keeping other contributors' parameters fixed

$$\min_{u_s} \ [\max_{v_s} f_{\text{priv}}(u_s, v_s) - \rho \ \max_{w} -f_{\text{util}}(u_s, w)], \tag{19}$$

and the process is repeated until convergence (see Sec. 2.7 [2] for a general discussion on convergence). The protocol is described in Algorithm 3. Initially, contributor s has its own data X_s and target class labels Z_s. Before the distributed process begins, a common initial minimax filter g learned from public domain data in batch, is distributed to all contributors. A contributor s sees its own data, current filtered samples $g_{s'}(X_{s'})$ of all others $s' \neq s$, and the target labels Z for all contributors. The contributor then updates its minimax filter by a few iterations of joint utility-privacy algorithm (Algorithm 2). Intuitively, what each contributor does in its iteration, is updating its filter u_s so that its filtered data g_s is seemingly *indistinguishable* from others' data.

Algorithm 3 Distributed learning protocol

Main routine on aggregator
Input: T, S, initial filter $g(\cdot; u)$
Output: final filtered data $g(X)$
Init: Send initial params u and receive filtered samples $g_s(X_s)$ and labels Z_s, $s = 1, ..., S$. Re-distribute $\{Z_s\}$.
Begin:
for $t = 1, ..., T$ **do**
  for $s = 1, ..., S$ **do**
    Call Contributor s' routine with updated values $\{g_{s'}(X'_s)\}$
    Receive and update $g_s(X_s)$
  end for
end for

Contributor s' routine
Input: data X_s and labels Z_s, loss/classifier $l(\cdot\,; v, w)$, filter $g(\cdot\,; u)$, and hyperparams
Output: individual minimax filter params u_s
Init: Receive initial u_s from the aggregator, and send back $g_s(X_s)$ and Z_s
Begin:
for $t = 1, ..., T_s$ **do**
  Use Algorithm 2 to solve (19) and update u_s
  Send $g_s(X_s)$ to the aggregator
end for

Comments: Multiple contributors routines may be called in parallel for asynchronous updates of $g_s(X_s)$. Additionally, to reduce communication loads, it is possible to communicate only the sufficient statistics required for computing the loss and its derivatives instead of actual filtered data $g_s(X_s)$.

4 Related work

Utility-privacy trade-offs using the notion of differential privacy have been studied analytically, in particular in the context of the statistical estimation [20, 1, 3] and learnability [10]. For a non-differentially private approach, Krause et al. [12] studied the NP-

hardness of optimal utility-privacy trade-off in discrete attribute selection, and demonstrated near-optimality of greedy selection when the attributes are conditionally independent, which, however, is unlikely for high-dimensional continuous data. Also, previous work used mutual information to quantify utility and privacy risks, which are difficult to estimate in practice for high-dimensional continuous data, while this paper uses maximum empirical loss which is readily evaluated. Preserving privacy of high-dimensional face images using deterministic algorithms has been proposed in [16, 9, 7, 23]: [16] applies k-anonymity to images; [7] learns a linear filter using Partial Least Squares to reduce the covariance between filtered data and private labels; [23] also learns a linear filter using the log-ratio of the Fisher's Linear Discriminant Analysis metrics. Our approach differs from these in several aspects: It is not limited to linear filters and is applicable to arbitrary differentiable nonlinear filters such as multilayer neural networks; It directly optimizes the utility-privacy risk instead of optimizing heuristic criteria such as covariance differences or LDA log-ratios.

5 Experiments

The proposed algorithms are evaluated using three real-world datasets: face data for gender/expression recognition, speech data for emotion recognition, and accelerometry data for activity recognition.

Filters: Linear dimensionality reduction is mainly used as filters ($x \mapsto G^T x$), with two-layer neural networks as nonlinear filters when applicable. The following methods of choosing the optimal filters are compared:

- Linear, non-private: random subspace projection (Rand). G is a random full rank $D \times d$ matrix.
- Linear, non-private: PCA. G is the eigenvectors corresponding to d largest eigenvalues of $\mathrm{Cov}(x)$.
- Linear, private: Private Partial Least Squares (PPLS), using Algorithm 1 from [7].
- Linear, private: Discriminately Decreasing Discriminability (DDD) [23] with a mask-type filter from the code[6].
- Linear and nonlinear, private: Minimax filter (2).

Remarks: DDD requires analytical solutions to eigenvalue problems which are unavailable in multiclass problems, and is used only in the binary problem with the face database. Also, DDD code uses a mask-type filter, and the dimension d is same as the image size. The dimension d is also irrelevant to nonlinear Minimax since it does not use linear dimensionality reduction.

[6] `http://mplab.ucsd.edu/~jake`

Classifier/loss: Logistic regression is used a classifier for both utility and privacy risks, where the loss $l(y(g(x_i;u),y_i;v)$ is the negative log-likelihood. For optimization subroutines Φ and Φ^l in Algorithm 2, L-BFGS is used with the 1st and 2nd-order derivatives $\nabla_u l, \nabla_v l, \frac{d^2 l}{dudv}$ of the log likelihood.

Parameters: A small regularization factor ($\lambda = 10^{-6}$) is used for logistic regression, and $\rho = 1$ for joint utility-privacy risks. The number of iterations for Minimax subroutines is set to $T_{\mathrm{aux}} = 5$ for auxiliary routines (see Supplementary Material) and $T_s = 1$ for distributed routines (Algorithm 3.) The main iteration for Minimax is stopped manually when the progress is slow ($T = 20 - 200$). Other hyperparameters for Algorithm 1 are in Supplementary Material.

5.1 Gender/expression recognition from face

The Genki database [23] consists of face images with varying poses and facial expressions. The original data is used unchanged, which have $N = 1740$ training images (50% male and 50% female; 50% smile and 50% non-smile) downscaled to 1616 pixels. The test set has 100 images (50 males and 50 females;50 smiling and 50 non-smiling) non-overlapping with the training set. The dimensionality of the original data is $D = 256$, and the filter is tested with $d = 50, 100, 150, 200$. The dataset has gender and expression labels but no subject label. Consequently, gender classification is used as the private task and expression classification is used as the target task. Six methods are compared: Rand, PCA, PPLS, DDD, Minimax 1 (linear), Minimax 2 (nonlinear). For the nonlinear Minimax 2, a two-layer sigmoid neural network is used with of a small number of hidden nodes ($D \times 10 \times 10$). The nonlinear network is pretrained as a stacked denoising autoencoders [22] followed by supervised backpropagation with the target task.

Figure 2 shows the test accuracy. The dotted lines are level sets of utility-privacy trade-off (=target task accuracy - private task accuracy) shown for a reference. Our methods (Minimax 1,2) achieve the highest accuracy for the target task (expression) and nearly the lowest accuracy for the private task (gender) at the same time. Compared to the linear Minimax 1, the nonlinear Minimax 2 is slightly better in the target task accuracy and is slightly worse in the privacy risks. For all dimensions d, Minimax 1,2 achieve the best compromise (=closer to the top-left corner) of all methods. In terms of privacy risk, Minimax 1,2 and DDD achieve almost the chance level accuracy (=0.5), which implies a strong privacy preservation. DDD comes close to Minimax, while another private method PPLS is not very successful in reducing the privacy risk. As expected, non-private methods (Rand, PCA)

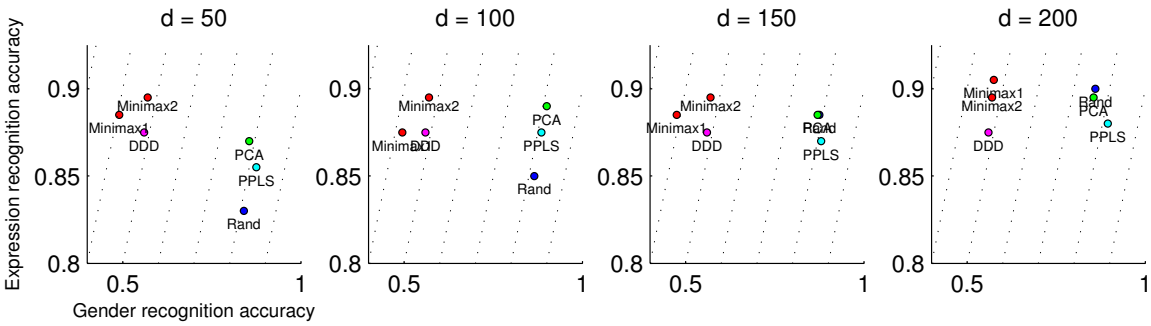

Figure 2: Genki database: Expression recognition vs and gender recognition from faces.

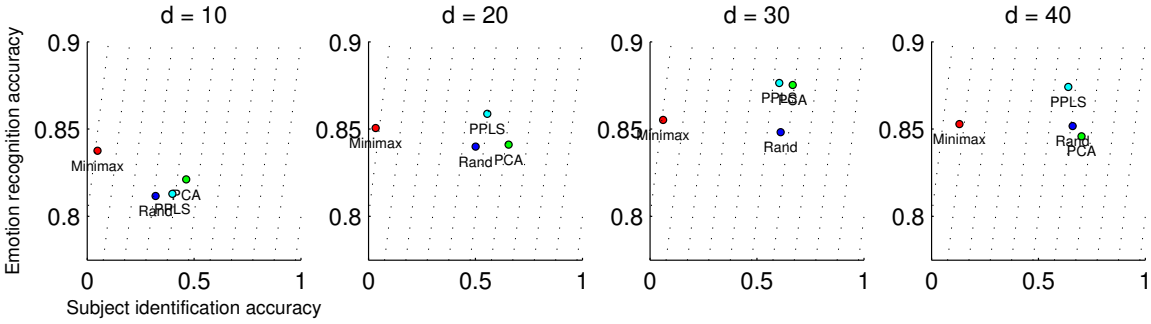

Figure 3: Enterface: Emotion recognition vs and subject identification from speech.

do not reduce the privacy risk. As dimension d increases from 50 to 200, the accuracy of both the target and the private tasks increases (=toward top-right) for all methods, but the value of utility-privacy trade-off (=accuracy of target task − accuracy of private task) remains relatively the same regardless of d. (Note that d is irrelevant to Minimax 2 and DDD.)

5.2 Emotion recognition from speech

The Enterface database [13] is an audiovisual emotion databases of 43 speakers from 14 nations reading pre-defined English sentences in six induced emotions. From the raw speech signals sampled in 48 KHz, MFCC coefficients are computed using 20 ms windows with 50% overlap and 13 Mel-frequency bands. The mean, max, min, and standard deviation of the MFCC coefficients over the duration of each sentence are computed, resulting in $N = 427$ samples of $D = 52$ dimensional feature vectors from $S = 43$ subjects. Each subject's samples are randomly split to generate training (80%) and test (20%) sets. Average test accuracy over 10 such trials is reported. Linear filters with dimensions $d = 10, 20, 30, 40$ are used. The target task is a binary classification of 'happy' and 'non-happy' emotions from speech, and the privacy risks is the multiclass ($S = 43$) subject classification using the multiclass loss (5).

Figure 3 shows the test accuracy. The target task accuracy of Minimax is comparable to PPLS, PCA and Rand, but the privacy risk is significantly lower than other methods, near the chance level ($1/S = 0.02$) compared to $0.4 - 0.6$ of non-private methods. This also shows that seemingly harmless statistics (mean, max, min, s.d. of MFCC) are quite susceptible to identification attacks without privacy mechanisms. Similar to experiment 1, the accuracy of both the target and the private tasks increases with the dimension d for all methods, and the value of utility-privacy trade-off remains similar regardless of d.

5.3 Activity recognition from accelerometry

The UCI Human Activity Recognition Dataset [13] is a collection of motion sensor data on a smartphone by 30 subjects performing six activities (*walking*, *walking upstairs*, *walking downstairs*, *sitting*, *standing*, *laying*). Various time and frequency domain variables are extracted from the signal, resulting in $N = 10299$ samples of $D = 561$ dimensional feature vectors from 30 subjects, which are used in the experiment unchanged. Out of 30 subjects, 15 subjects are chosen randomly (call it domain 1) and the other 15 subjects (call it domain 2) are left for the subsequent experiment on distributed learning. For each domain, each subject's samples are randomly split to generate training (50%)

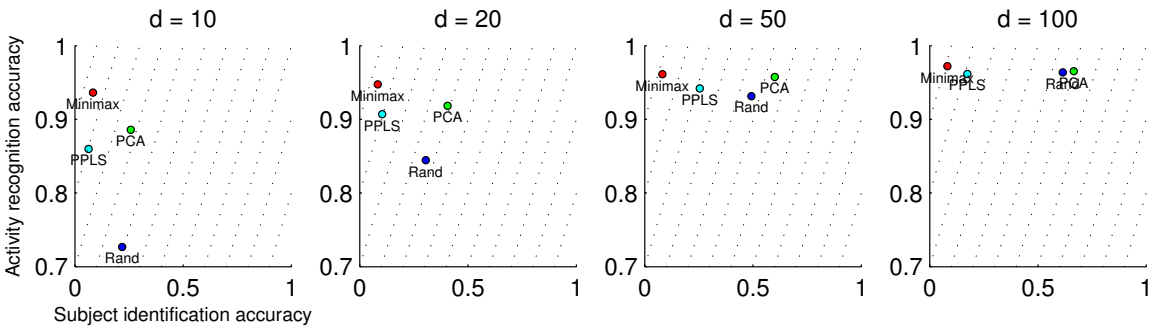

Figure 4: UCI: Activity recognition vs subject identification from accelerometry.

and test (50%) sets. In this experiment only domain 1 subjects are used. At each trial, the subjects and the training/test sets are randomized, and the average test accuracy over 10 such trials is reported. Linear filters with dimensions $d = 10, 20, 50, 100$ are used. The target task is a multiclass ($C = 6$) classification of activity, and the privacy risks is the multiclass ($S = 15$) subject classification risk (5).

Figure 4 shows the test accuracy. The target task accuracy of Minimax is much higher than others at $d = 10, 20$ and becomes comparable to others at $d = 50, 100$. However the private task accuracy of Minimax is lower than PPLS and significantly lower than PCA and Rand, close to the chance level ($1/S = 0.067$). The figure also shows that accelerometry data are susceptible ($0.2 - 0.7$) to identification attacks without privacy mechanisms. For all dimensions d, Minimax achieves the best compromise (=closer to the top-left corner) of all methods similar to previous experiments. Also, similar to previous experiments, the accuracy of both the target and the private tasks roughly increases with d for all methods, but the value of utility-privacy trade-off remains similar.

5.4 Distributed learning of individual filters

In this experiment, the protocol for distributed learning of individual minimax filters (Algorithm 3) is evaluated using the same accelerometry data. For this purpose, samples from the $S = 15$ subjects in domain 2 (who are not in domain 1) are used. The individual filters are initialized with a common minimax filter learned from domain 1 (with $d = 20$). The accuracy of target and privacy tasks over the iteration in Algorithm 3 is computed. The target task is activity recognition as before, and the privacy task is weighted binary subject identification (7), as the goal is for each contributor to learn its own filter.

Figure 5 shows the test accuracy averaged over 10 trials. The target task accuracy increases slightly from 0.936 ± 0.013 (before learning) to 0.946 ± 0.010 (after 20 iterations), and the privacy risk decreases from 0.674 ± 0.011 (before learning) to 0.538 ± 0.017 (after 20 iterations) where the chance level is 0.5. These improvements demonstrate that individual minimax filters can be learned successfully in a distributed setting using the proposed protocol.

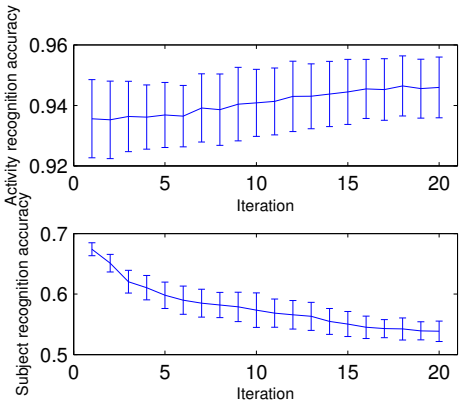

Figure 5: Distributed learning of individual minimax filters from accelerometry dataset.

6 Conclusion

This work presents a practical privacy-preserving mechanism for publishing continuous high-dimensional datasets. Minimax filters are defined to achieve the optimal utility-privacy trade-off given a family of filters and a family of loss/classifiers. Algorithms for finding minimax filters in batch and distributed settings are presented and demonstrated on real datasets. Experiments show that publicly available multisubject datasets per se are surprisingly susceptible to subject identification attacks, and that even linear minimax filters can reduce the privacy risks close to chance level without sacrificing target task accuracy by much.

References

[1] M. S. Alvim, M. E. Andrés, K. Chatzikokolakis, P. Degano, and C. Palamidessi. Differential privacy: on the trade-off between utility and information leakage. In *Formal Aspects of Security and Trust*, pages 39–54. Springer, 2012.

[2] D. P. Bertsekas. *Nonlinear programming*. Athena Scientific, 1999.

[3] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Local privacy and statistical minimax rates. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pages 429–438. IEEE, 2013.

[4] C. Dwork. Differential privacy. In *Automata, languages and programming*, pages 1–12. Springer, 2006.

[5] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, pages 265–284. Springer, 2006.

[6] C. Dwork and K. Nissim. Privacy-Preserving Data Mining on Vertically Partitioned Databases. In *Proc. CRYPTO*. Springer, 2004.

[7] M. Enev, J. Jung, L. Bo, X. Ren, and T. Kohno. Sensorsift: balancing sensor data privacy and utility in automated face understanding. In *Proceedings of the 28th Annual Computer Security Applications Conference*, pages 149–158. ACM, 2012.

[8] B. Fung, K. Wang, R. Chen, and P. S. Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Comp. Surveys (CSUR)*, 42(4):14, 2010.

[9] R. Gross, L. Sweeney, F. De La Torre, and S. Baker. Semi-supervised learning of multi-factor models for face de-identification. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[10] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

[11] K. Kiwiel. A direct method of linearization for continuous minimax problems. *Journal of optimization theory and applications*, 55(2):271–287, 1987.

[12] A. Krause and E. Horvitz. A utility-theoretic approach to privacy and personalization. In *AAAI*, volume 8, pages 1181–1188, 2008.

[13] O. Martin, I. Kotsia, B. Macq, and I. Pitas. The enterface´05 audio-visual emotion database. In *Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on*, pages 8–8. IEEE, 2006.

[14] A. Narayanan and V. Shmatikov. Robust deanonymization of large sparse datasets. In *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, pages 111–125. IEEE, 2008.

[15] A. Nedić and A. Ozdaglar. Subgradient methods for saddle-point problems. *Journal of optimization theory and applications*, 142(1):205–228, 2009.

[16] E. M. Newton, L. Sweeney, and B. Malin. Preserving privacy by de-identifying face images. *Knowledge and Data Engineering, IEEE Transactions on*, 17(2):232–243, 2005.

[17] V. Panin. Linearization method for continuous minmax problem. *Cybernetics and Systems Analysis*, 17(2):239–243, 1981.

[18] B. Rustem and M. Howe. *Algorithms for worst-case design and applications to risk management*. Princeton University Press, 2009.

[19] A. Sarwate and K. Chaudhuri. Signal processing and machine learning with differential privacy: Algorithms and challenges for continuous data. *Signal Processing Magazine, IEEE*, 30(5):86–94, 2013.

[20] A. Smith. Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 813–822. ACM, 2011.

[21] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[22] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of ICML*, pages 1096–1103. ACM, 2008.

[23] J. Whitehill and J. Movellan. Discriminately decreasing discriminability with learned image filters. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2488–2495. IEEE, 2012.

[24] A. C. Yao. Protocols for secure computations. In *2013 IEEE Symp. Found. Comp. Sci.*, pages 160–164. IEEE, 1982.