ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2017)

Differentially private model selection with penalized and constrained likelihood

Jing Lei,

Carnegie Mellon University, Pittsburgh, USA

Anne-Sophie Charest,

Université Laval, Québec, Canada

Aleksandra Slavkovic and Adam Smith

Pennsylvania State University, State College, USA

and Stephen Fienberg

Carnegie Mellon University, Pittsburgh, USA

[Received April 2016. Revised August 2017]

Summary. In statistical disclosure control, the goal of data analysis is twofold: the information released must provide accurate and useful statistics about the underlying population of interest, while minimizing the potential for an individual record to be identified. In recent years, the notion of *differential privacy* has received much attention in theoretical computer science, machine learning and statistics. It provides a rigorous and strong notion of protection for individuals' sensitive information. A fundamental question is how to incorporate differential privacy in traditional statistical inference procedures. We study model selection in multivariate linear regression under the constraint of differential privacy. We show that model selection procedures based on penalized least squares or likelihood can be made differentially private by a combination of regularization and randomization, and we propose two algorithms to do so. We show that our privacy procedures are consistent under essentially the same conditions as the corresponding non-privacy procedures. We also find that, under differential privacy, the procedure becomes more sensitive to the tuning parameters. We illustrate and evaluate our method by using simulation studies and two real data examples.

Keywords: Consistency; Differential privacy; Information criteria; Model selection; Regression

1. Introduction

In data privacy research, the goal of data analysis is to provide accurate and useful statistical inference while preventing individual records from being identified. Such privacy protection is crucial in many statistical applications, namely for the analysis of census and survey data, medical and clinical studies, genetics data and Web user data collected on the Internet. In statistics, the treatment of confidential data has a long history under the name of 'statistical disclosure control' or 'statistical disclosure limitation'; see for example Dalenius (1977), Rubin (1993), Willenborg and De Waal (1996), Fienberg and Slavković (2010) and Hundepool *et al.*

Address for correspondence: Jing Lei, Department of Statistics, 132 Baker Hall, Carnegie Mellon University, Pittsburgh, PA 15213, USA.
E-mail: jinglei@andrew.cmu.edu

 0964–1998/17/181000

(2012). A long lasting challenge is to quantify rigorously the disclosure protection that is offered by privacy preserving data analysis procedures.

The notion of *differential privacy* has been introduced with the same objective in theoretical computer science by Dwork (2006) and Dwork *et al.* (2006). The general idea of differential privacy is to require for the outcome of a randomized data analysis procedure not to change much for small perturbations of the input data, so that one cannot infer from the output the presence or absence of any individual in the input data set, or infer some of the person's characteristics. This requirement can be rigorously quantified and does not depend on assumptions regarding the resources of the intruder, including any access to auxiliary information. Thus, differential privacy compares very favourably with measures of disclosure risk that are commonly used in the statistical disclosure control literature as it is more encompassing and a *worst-case definition*, but at the same time it has been criticized as too stringent from the perspective of achieving needed statistical data utility; for example, see Fienberg *et al.* (2010) and Karwa and Slavkovic (2012).

Over the last decade, there has been a rapid development of differentially private algorithms and procedures in both the computer science and the statistics literature. For example, the main focus in the computer science literature was on designing differentially private mechanisms and efficient algorithms for private data release, e.g. the Laplace noise perturbation mechanism (Dwork *et al.*, 2006), the exponential mechanism (McSherry and Talwar, 2007), releasing contingency tables (Barak *et al.*, 2007) and boosting (Hardt *et al.*, 2012). On the statistics side, research efforts include designing consistent and efficient differentially private point estimators (Dwork and Lei, 2009; Smith, 2011; Chaudhuri *et al.*, 2011; Lei, 2011; Bassily *et al.*, 2014; Karwa and Slavković, 2016), non-parametric density estimation (Wasserman and Zhou, 2010), hypothesis testing (Fienberg *et al.*, 2011; Johnson and Shmatikov, 2013; Uhler *et al.*, 2013; Yu *et al.*, 2014; Karwa *et al.*, 2014; Solea, 2014; Dwork *et al.*, 2015; Sheffet, 2015; Wang *et al.*, 2015; Gaboardi *et al.*, 2016) and statistical lower bounds (Chaudhuri and Hsu, 2011; Duchi *et al.*, 2013); there is also a large literature on private probably approximately correct learning, which echoes the concerns of statistical estimation in the context of classification (Kasiviswanathan *et al.*, 2011; Beimel *et al.*, 2010, 2014, 2015; Bun *et al.*, 2015; Karwa *et al.*, 2015).

In this paper we consider statistical model selection under the constraint of differential privacy. In the context of differential privacy, the data are held private and only the synthetic data or more commonly the inference results, such as point estimates, interval estimates and p-values, are released. Model selection is a particularly important and challenging task in privacy preserving data analysis and inference. First, model selection is a crucial step in the analysis of modern high dimensional complex data. In non-private statistical inferences, a good model balances between the modelling bias and potential of overfitting. Under privacy constraints, different models also have different levels of privacy risk. Generally speaking, fitting a more complicated model has a higher risk of privacy leakage as more information is revealed from the data. Therefore choosing a good model is not only necessary for accurate inference, but also for good privacy protection. Second, the model selection step itself reveals non-trivial information about the data, which may increase the risk of privacy leakage. For example, if the attacker knows all except one data entry, she can then make inference about the value of the unknown entry by reversing the model selection procedure. Therefore, to incorporate model selection in privacy preserving data analysis, the model selection step itself must be private.

Despite the fast development in combining statistical theory and methodology with differential privacy, model selection under privacy constraints has not been well understood. The problem of differentially private model selection is motivated by practical concerns: when private data analysis procedures are needed, it is rarely known which model is most appropriate

for the data. When releasing the whole data set is impractical because of privacy concerns and releasing point estimates for prespecified models has limited utility, an appropriate compromise is first to identify the best model and then to obtain and release consistent point estimates for this model. Both tasks need to be performed under privacy constraints, and thus we need methods for differentially private model selection.

We focus on classical linear regression model selection. In particular, we aim to provide insights for the following two questions:

(a) is it theoretically possible to do model selection with differential privacy under the classical conditions?, and
(b) what new practical concerns arise in model selection when differential privacy is required?

We first show that the answer to the first question is positive by proposing differentially private model selection procedures based on penalized least squares and likelihood which exhibit asymptotic utility guarantees. Here utility means that the procedure selects the correct model with high probability under appropriate regularity conditions. In other words, differentially private model selection is theoretically possible in the classical setting. More specifically, we propose a two-step differentially private model selection procedure: we first obtain a least square or maximum likelihood estimate under an l_1-constraint; then we use noisy optimization with regularization to obtain the best model. Interestingly, the l_1-constraint, which is usually employed in high dimensional problems, helps to achieve privacy even in low dimensions.

Second, from simulations and real data examples we observe that the finite sample behaviour of the method proposed depends crucially on the tuning parameter that is required by the procedure. For example, our algorithm imposes an l_1-norm constraint on the estimated regression coefficient. When the tuning parameter is conservatively chosen, i.e. the imposed upper bound of the l_1-norm is large, the utility is quite limited for the sample or moderate sample sizes. When the sample size is in the thousands, the dependence of utility on the tuning parameter is less significant, as predicted by the theory. More importantly, if auxiliary information is available, such as a good upper bound on the l_1-norm of the true regression coefficient, then we may be able to choose the tuning parameter more adaptively, with improved utility. This reveals a distinct feature of differentially private data analysis: some auxiliary information that does not affect classical inference may lead to a significant improvement in performance in differentially private analysis. Indeed, the l_1-bound on the regression coefficient which improves the utility of our differentially private procedure is useful in the lasso when the dimensionality is high, but not so much in the classical regime.

In relevant work, Chaudhuri and Vinterbo (2013) studied differentially private cross-validation, with primary examples being ridge regression and histogram estimation. The goal of cross-validation is to choose an estimation procedure that yields approximately the best predictive risk, which is substantially different from the goal of the current paper: selecting the correct model that generated the data. It is well known that, in general, cross-validation cannot produce consistent model selection even in the classical situation of low dimensional linear regression, unless the training sample ratio is chosen in a very unconventional way (Shao, 1993; Yang, 2007). Furthermore, the noisy minimization method in Chaudhuri and Vinterbo (2013) relies on global sensitivity of the estimate and validation procedures, and hence requires bounded data, whereas our method uses a novel high probability upper bound of the local sensitivity, which requires a less stringent boundedness condition.

The remainder of the paper is organized as follows. In Section 2, we review the definition and interpretation of differential privacy, as well as two general methods to create differentially private algorithms. Section 3 details the two proposed differentially private model selection

procedures, with proofs of their privacy guarantee, and notes on the choice of the tuning parameters. Statements and proofs of the utility guarantee of the two algorithms are given in Section 4. Empirical results, including a simulation study and two real data examples, are reported in Section 5. Section 6 provides a brief discussion. All proofs and technical details are collected in Appendix A.

The data that are analysed in the paper and the programs that were used to analyse them can be obtained from

`http://wileyonlinelibrary.com/journal/rss-datasets`

2. Differential privacy

Differential privacy requires that the output of a procedure is not drastically altered under small perturbations of the input data set, such that an attacker, regardless of his auxiliary information and computing power, can hardly recover the presence or absence of a particular individual in the data set. The notion of differential privacy is a property of that data analysis procedure, rather than of the output obtained.

2.1. Definition

To formalize, consider a data set $D = \{z_1, \ldots, z_n\} \in \mathcal{Z}^n$ consisting of n data points in sample space $\mathcal{Z}$. A data analysis procedure $\mathcal{T}$, possibly randomized, maps the data set D, together with a random input ω, to $\mathcal{T}(D) \equiv \mathcal{T}(D, \omega) \in S$, an output space. Here we assume that $(S, \mathcal{S})$ is a measurable space and $\mathcal{T}(D, \cdot) : \Omega \mapsto S$ is a measurable function. Whenever it is not confusing, we shall use $\mathcal{T}(D)$ to denote the random variable $\mathcal{T}(D, \omega)$.

For any two data sets D and D' of the same size, we use $\mathrm{Ham}(D, D')$ to denote their *Hamming distance*, which is the number of entries at which they differ, regardless of the order. For example, if $\mathcal{Z} = \mathbb{R}$ and $D = \{1, 2, 3, 4\}$ and $D' = \{2, 3, 4, 5\}$, then $\mathrm{Ham}(D, D') = 1$. In the rest of this paper, we always use D and D' to denote a pair of adjacent data sets that differ at the last entry: $D = \{z_1, \ldots, z_{n-1}, z_n\}$ and $D' = \{z_1, \ldots, z_{n-1}, z'_n\}$.

We can now state formally the property of differential privacy.

Definition 1 (ϵ differential privacy; Dwork *et al.* (2006)). Given a privacy parameter, $\epsilon > 0$, the procedure $\mathcal{T}$ satisfies ϵ *differential privacy* if

$$\sup_{\mathrm{Ham}(D,D')=1, A \in \mathcal{S}} \left| \log \left[\frac{P_\omega\{\mathcal{T}(D', \omega) \in A\}}{P_\omega\{\mathcal{T}(D, \omega) \in A\}} \right] \right| \leqslant \epsilon$$

where we define $\log(0/0) = 0$ for convenience.

In the above notation, P_ω denotes the probability with respect to ω, which is the source of randomness in the data analysis procedure. Thus, the definition does not impose any conditions on the distribution of D—the privacy is required to hold for *all pairs of adjacent data sets*. This stringent condition means that procedures which satisfy definition 1 provide very strong privacy guarantees, even against adversaries who have considerable partial information about the data set (Ganta *et al.*, 2008). To satisfy this definition, any non-constant procedure must be randomized.

ϵ differential privacy is a strong requirement, as it takes the supremum over all possible neighbouring data sets of size n. A mild relaxation is (ϵ, δ) differential privacy.

Definition 2 ((ϵ, δ) differential privacy). Given $\epsilon > 0$ and $\delta \in (0, 1)$, a procedure satisfies (ϵ, δ) differential privacy if, for all measurable $A \subseteq \mathcal{S}$ and all neighbouring data sets D and D',

$$P_\omega\{\mathcal{T}(D) \in A\} \leqslant \mathrm{e}^\epsilon P_\omega\{\mathcal{T}(D') \in A\} + \delta.$$

Here the requirement of the original ϵ differential privacy is relaxed so that the distribution of $\mathcal{T}(D)$ only needs to be dominated by that of $\mathcal{T}(D')$ outside a set with probability no more than δ.

In our discussion of statistical applications we shall generally focus on data sets consisting of a sequence of n independent random samples from an underlying distribution, and the corresponding probability will be denoted P_D. Note that the differential privacy definition has no such assumption. We shall use $P_{D,\omega}$ to denote the overall randomness due to both the data and the randomness mechanism in the analysis.

2.2. Statistical interpretation of differential privacy

Privacy protection due to differential privacy can be interpreted from a Bayesian perspective, as in Abowd and Vilhuber (2008) and Kasiviswanathan and Smith (2008). Suppose that we have a prior distribution of the input data set D and then get to observe the random output $\mathcal{T}(D, \omega)$. Denote by P_i and Q_i the marginal prior and posterior of X_i, for the ith entry in D. Then, if the procedure $\mathcal{T}$ is differentially private and the prior P is a product measure on the entries of D, we have that $\mathrm{e}^{-\epsilon} \leqslant \mathrm{d}P_i/\mathrm{d}Q_i \leqslant \mathrm{e}^\epsilon$, thus limiting the information regarding X_i gained from the output $\mathcal{T}(D, \omega)$. A related hypothesis testing interpretation is given in theorem 2.4 of Wasserman and Zhou (2010).

We may also intuitively interpret differential privacy as a specific notion of robustness. It requires that the distribution of $\mathcal{T}(D)$ is not changed too much if D is perturbed in only one entry. However, the definition of differential privacy is also a *worst-case definition*, where the probability ratio needs to be uniformly bounded over all pairs of adjacent input data sets. A key ingredient to designing differentially private statistical procedures is to bridge the gap between the worst-case privacy guarantee and the average case statistical utility. It turns out that robustness and regularization are the most relevant structures to explore, as in Dwork and Lei (2009), Chaudhuri *et al.* (2011) and Smith and Thakurta (2013).

2.3. Designing differentially private algorithms

For any statistical task which we want to carry out, a specific randomized procedure $\mathcal{T}$ must be designed to take as input a database $D \in \mathcal{Z}^n$ and to return an element of the output space $\mathcal{S}$ while satisfying differential privacy. There are a few approaches which are sufficiently generic to be adaptable to various tasks and which are often used as building blocks for more complicated procedures. We present two of these methods, which we use later in the construction of our procedure. The first adds random noise to a non-private output and the second samples randomly from a set of candidate outputs.

2.3.1. Adding Laplace noise

The additive noise approach is applicable when the output space is a Euclidean space. For simplicity of presentation we consider here $\mathcal{S} = \mathbb{R}$. Let $T(D)$ be a deterministic (and hence non-private) mechanism. Define the *global sensitivity* G_T as

$$G_T = \sup_{\mathrm{Ham}(D,D')=1} |T(D) - T(D')|. \tag{1}$$

If $G_T < \infty$, then it is easy to check (Dwork *et al.*, 2006) that

$$\mathcal{T}(D) \equiv T(D) + \epsilon^{-1} G_T \zeta \tag{2}$$

satisfies ϵ differential privacy, where ζ is a standard double-exponential random variable with density function $0.5\exp(-|\zeta|)$ (which is also known as the Laplace distribution).

2.3.2. *The exponential mechanism and noisy optimization*

In the *exponential mechanism* (McSherry and Talwar, 2007), suppose that the output space S is equipped with a σ-algebra and a measure μ_0, and let $q : \mathcal{Z}^n \times S \mapsto \mathbb{R}$ be a score function that measures the quality of $s \in S$ in terms of its agreement with the input data set. Usually $q(D, s) = -|s - T(D)|$ for some deterministic procedure T. Denote $G_q = \sup_s G_{q(\cdot,s)}$, where $G_{q(\cdot,s)}$ is the global sensitivity of the mapping $q(\cdot, s)$. Let $\mathcal{T}(D)$ be the procedure that outputs a random sample from S with probability distribution Q whose density function satisfies

$$\frac{\mathrm{d}Q}{\mathrm{d}\mu_0}(s) \propto \exp\left\{\frac{\epsilon q(D, s)}{2G_q}\right\}.$$

Then $\mathcal{T}$ satisfies ϵ differential privacy.

When S is finite, we may also use additive noise to maximize $q(D, s)$ over s approximately. Let

$$\tilde{q}(D, s) = q(D, s) + 2\epsilon^{-1}\zeta G_{q(\cdot,s)}$$

be the privatized score, where ζ is an independent draw from the standard double-exponential distribution. Then

$$\mathcal{T}(D) = \arg\max_s \tilde{q}(D, s)$$

satisfies ϵ differential privacy and usually offers similar performance to that of the exponential mechanism.

2.3.3. *A generic scheme for (ϵ, δ) differential privacy*

In many statistical problems the sample space is not compact and hence $G_T = \infty$ for many statistics T such as the sample mean. A general strategy is to show that one can add much less noise for most average case data sets with (ϵ, δ) differential privacy (Dwork and Lei, 2009). These methods often involve the notion of *local sensitivity* of a deterministic procedure T:

$$G_T(D) = \sup_{D' : \mathrm{Ham}(D, D') = 1} |T(D) - T(D')|. \tag{3}$$

If $G_T(D)$ is finite and public, then one can show that adding noise to $T(D)$ as in result (2) with G_T replaced by $G_T(D)$ also gives ϵ differential privacy. Unfortunately, $G_T(D)$ depends on the data set and may contain sensitive information. However, there is a generic scheme based on this idea with valid privacy guarantee under the following two general conditions on the deterministic procedure T.

(a) For all $\epsilon > 0$ there is a real-valued function $G^*(D)$ and a randomized procedure $\mathcal{T}_\epsilon(D, g)$, which is ϵ differentially private if $g \geqslant G^*(D)$ and is assumed to be non-private.
(b) Given $\epsilon > 0$ and $\delta \in (0, 1)$, there is an ϵ differentially private mapping $G_{\epsilon,\delta}(D)$ satisfying $P_\omega\{G_{\epsilon,\delta}(D) \geqslant G^*(D)\} \geqslant 1 - \delta$ for all D.

An example of $G^*(D)$ is the local sensitivity of some procedure $T(D)$, and $\mathcal{T}$ is the noisy version as in result (2) calibrated to an upper bound of the local sensitivity.

Proposition 1. Under the above two assumptions, for any $\epsilon_1 + \epsilon_2 = \epsilon$, and $\delta \in (0, 1)$, $\mathcal{T}_{\epsilon_2}\{D, G_{\epsilon_1,\delta}(D)\}$ satisfies (ϵ, δ) differential privacy.

3. Differentially private model selection procedures

Popular model selection methods for linear regression include information criteria such as the Akaike information criterion (Akaike, 1974) and the Bayesian information criterion (BIC) (Schwarz *et al.*, 1978), cross-validation (Picard and Cook, 1984) and the more recent penalized least squares (e.g. Tibshirani (1996) and Fan and Li (2001)). In this paper we focus on penalized least squares and penalized profile likelihood estimators, both being variants of the classical approach based on information criteria.

3.1. Background: linear regression and model selection

In the linear regression model, the data points are independent, each consisting of a response $Y \in \mathbb{R}^1$, and a covariate $X \in \mathbb{R}^d$, which satisfies

$$Y = X^{\mathrm{T}}\beta_0 + W, \tag{4}$$

where $\beta_0 \in \mathbb{R}^d$ is the regression coefficient, and W is a Gaussian random variable, independent of X, with mean 0 and variance σ^2. The observed data set is $D = \{(X_i, Y_i) : 1 \leqslant i \leqslant n\}$, where $X_i = (X_{i1}, \ldots, X_{id})^{\mathrm{T}}$.

The model selection problem is to find the support of β_0: $M_0 \equiv \{j : \beta_0(j) \neq 0\}$. To give a precise formulation, consider a class of candidate models $\mathcal{M} \subseteq \{0,1\}^d$ that contains M_0. For each $M \in \mathcal{M}$, the corresponding hypothesis is $\beta_0 \in \Theta_M \equiv \{\beta \in \mathbb{R}^d : \beta_0(j) = 0,\ \forall\ j \notin M\}$.

Given a parameter (β, σ^2) and an observed data set, the log-likelihood is, ignoring constant terms,

$$l(\beta, \sigma^2; D) = \sum_{i=1}^{n} \left\{ -\frac{1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(Y_i - X_i^{\mathrm{T}}\beta)^2 \right\}.$$

We consider two cases separately.

3.1.1. Known variance

When σ^2 is known, we aim to find the β that maximizes the log-likelihood, which leads to model selection with least squares. Without loss of generality, we assume that $\sigma^2 = 1$. We then define

$$l(\beta; D) = -\frac{1}{2}\sum_{i=1}^{n}(Y_i - X_i^{\mathrm{T}}\beta)^2, \tag{5}$$

$$\hat{\beta}_M = \arg\min_{\beta \in \Theta_M} -2\,l(\beta; D), \tag{6}$$

$$l(M; D) = l(\hat{\beta}_M; D). \tag{7}$$

3.1.2. Unknown variance

In the more realistic setting that σ^2 is unknown, we can maximize over the nuisance parameter σ^2 to perform model selection with the profile likelihood. Ignoring constant terms, we obtain the profile log-likelihood for β:

$$l^*(\beta; D) = -\frac{n}{2}\log\left\{ \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i^{\mathrm{T}}\beta)^2 \right\}. \tag{8}$$

Maximizing this over all β in Θ_M, we obtain

$$l^*(M; D) \equiv \sup_{\beta \in \Theta_M} l^*(\beta; D). \tag{9}$$

3.1.3. *Information criteria*

We could simply maximize $l(M; D)$ or $l^*(M; D)$ as given above. However, more complex models tend to yield larger values of log-likelihood. The approach of information criteria thus minimizes the sum of negative log-likelihoods and a measure of model complexity. For example, in the case of unknown σ^2, we may choose the model by

$$\hat{M} = \arg\min_{M \in \mathcal{M}} -2 l^*(M; D) + \phi_n |M|, \tag{10}$$

where ϕ_n is the amount of penalty on the model complexity. The best-known and most widely used examples are the Akaike information criterion ($\phi_n = 2$) and BIC ($\phi_n = \log(n)$).

In the case of known σ^2, the penalized minimization becomes

$$\hat{M} = \arg\min_{M \in \mathcal{M}} -2 l(M; D) + \phi_n |M|. \tag{11}$$

It is worth noting that the penalized least square estimator, with appropriate choice of ϕ_n, can also be used in the case of unknown σ^2.

3.2. *Towards differentially private model selection*

To perform model selection in a differentially private manner, we propose to apply the exponential mechanism or noisy minimization to the minimization problems (10) and (11), with $q(D, M)$ (here s in the general definition is replaced by M) given by the penalized likelihood

$$L^*(M; D) \equiv -2 l^*(M; D) + \phi_n |M|$$

in the case of penalized log-profile likelihood (10) or

$$L(M; D) \equiv -2 l(M; D) + \phi_n |M|$$

in the case of penalized least squares (11).

A key step in this approach is to evaluate and control the sensitivities of $L(M; D)$ and $L^*(M; D)$ as functions of D for any $M \in \mathcal{M}$.

To simplify the notation and to concentrate on the main idea, we shall assume in what follows that the data entries are bounded or standardized:

$$\begin{aligned} \max_{1 \leqslant i \leqslant n} |Y_i| &\leqslant r, \\ \max_{1 \leqslant i \leqslant n, 1 \leqslant j \leqslant d} |X_{ij}| &\leqslant 1, \end{aligned} \tag{12}$$

where r is a known number that can grow with n. Boundedness is typically required for differentially private data analysis. Standard methods for finding the range of the data set in a privacy preserving manner include those given in Dwork and Lei (2009) and Smith (2011).

3.2.1. *Sensitivity of least squares and profile log-likelihood*

As indicated in Section 3.1, we consider procedures based on two score functions for a model: the sum of squared residuals $l(M; D)$, and the profile likelihood $l^*(M; D)$.

To bound the sensitivity of either of them, we must bound the possible parameter vectors that we consider. We thus consider constrained versions of the two score functions. Given $R > 0$, we define

$$-2\,l_R(M;D) = \min_{\beta\in\Theta_M, \|\beta\|_1\leqslant R} -2\,l(\beta, D) \tag{13}$$

and

$$-2\,l_R^*(M;D) = \min_{\beta\in\Theta_M, \|\beta\|_1\leqslant R} -2\,l^*(\beta; D). \tag{14}$$

The sensitivity of the least squares loss is now easy to bound.

Lemma 1 (sensitivity of constrained least squares). When β has l_1-norm at most R and the data (X_i, Y_i) are restricted to $[-1, 1]^d \times [-r, r]$, the global sensitivity of the squared error loss functions $-2\,l(\beta;\cdot)$ and $-2\,l_R(M;\cdot)$ is at most $(r+R)^2$.

The proof of lemma 1 is both short and elementary, and hence has been omitted.

In the case of the constrained profile likelihood, we cannot bound the global sensitivity, but we can find a bound on the local sensitivity.

Lemma 2 (local sensitivity of constrained profile likelihood). Under the same conditions as in lemma 1, the local sensitivity of $-2\,l_R^*(M;\cdot)$ is no larger than

$$\frac{n(r+R)^2}{-2\,l_R(M;D) - (r+R)^2}.$$

According to the Laplace noise perturbation strategy that was introduced in Section 2.3, to achieve differential privacy we need to perturb the sample sums $-2l(\beta, D)$, $-2l_R(M;D)$ and $-2l_R^*(M;D)$ with Laplace noise scaled by their corresponding sensitivity or local sensitivity. Roughly speaking, these sample sums are proportional to n, and hence the upper bounds of their (local) sensitivities are of order $(r+R)^2$. For the additive Laplace noise to have negligible effect on the subsequent inference, we need $(r+R) \ll \sqrt{n}$. If $r+R$ is comparable with or larger than $\sqrt{n}$, the private estimates may be substantially different from the non-private estimates and hence have limited utility. For example, in the linear regression model that is considered in this paper, the ideal choice of parameter R is the l_1-norm of the true regression coefficient, and r will be upper bounded by the $R + \|W\|_\infty$, where $\|W\|_\infty$ is the maximum absolute value of the noise W. When the noise is Gaussian, we have $R + r \asymp 2R + \sqrt{\{2\log(n)\}}$ with high probability.

3.2.2. Algorithms

Lemma 1 implies that the penalized constrained least squares

$$L_R(M;D) = -2\,l_R(M;D) + |M|\phi_n$$

can easily be minimized in an ϵ differentially private manner. The complete algorithm is given in Table 1.

Next, lemma 2 gives an upper bound of the local sensitivity of $L_R^*(M;D)$. Now we apply the generic scheme of designing (ϵ, δ) differentially private algorithms described in Section 2.3.

First, let

$$G^*(D) \equiv \frac{n(r+R)^2}{\min_M -2\,l_R(M;D) - (r+R)^2}.$$

Then $G^*(D)$ is a uniform upper bound of the local sensitivity of $L_R(M;D)$ for all M. The only private part in $G^*(D)$ above is $\min_{M\in\mathcal{M}} -2l_R(M;D)$, which has global sensitivity $(r+R)^2$ according to lemma 1. Thus following the general procedure in Section 2.3, a valid choice of $G(D)$ for $-2l_R^*(M;D)$ is

Table 1. Algorithm 1: model selection via penalized constrained least squares

Input: data set $D = \{(X_i, Y_i) : i \in \{1, \ldots, n\}\}$, collection of models $\mathcal{M}$, parameters r, R, ϕ_n, ϵ
Output: estimated model $\hat{M} \in \mathcal{M}$
for each model M in $\mathcal{M}$ *do*

$$l_R(M; D) \leftarrow \max_{\beta \in \Theta_M, \|\beta\|_1 \leqslant R} -\frac{1}{2} \sum_{i=1}^{n} (Y_i - X_i^{\mathrm{T}} \beta)^2$$

$$L_R(M; D) \leftarrow -2 l_R(M; D) + \phi_n |M|$$

$$\tilde{L}_R(M; D) \leftarrow L_R(M; D) + \frac{2(r+R)^2}{\epsilon} Z_M, \text{ where } Z_M \text{ is a Laplace random variable}$$

return $\arg\min_{M \in \mathcal{M}} \tilde{L}_R(M; D)$

Table 2. Algorithm 2: model selection via penalized constrained profile likelihood

Input: data set $D = \{(X_i, Y_i) : i \in \{1, \ldots, n\}\}$, collection of models $\mathcal{M}$, parameters $r, R, \phi_n, \delta, \epsilon$
Output: estimated model $\hat{M} \in \mathcal{M}$
for each model M in $\mathcal{M}$ *do*

$$l_R^*(M; D) \leftarrow \max_{\beta \in \Theta_M, \|\beta\|_1 \leqslant R} -\frac{n}{2} \log \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i^{\mathrm{T}} \beta)^2 \right\}$$

$$L_R^*(M; D) \leftarrow l_R^*(M; D) + \phi_n |M|$$

$$\tilde{L}_R^*(M; D) \leftarrow L_R^*(M; D) + \frac{2G(D)}{\epsilon} Z_M \text{ where}$$

$$G(D) = \frac{n(r+R)^2}{\min_M -2 l_R(M; D) - (r+R)^2 + \epsilon^{-1}(r+R)^2 [Z_G - \log\{1/(2\delta)\}]}$$

$l_R(M; D)$ is as in algorithm 1, and Z_G and Z_M are Laplace random variables

return $\arg\min_{M \in \mathcal{M}} \tilde{L}_R^*(M; D)$

$$G(D) = \frac{n(r+R)^2}{\min_M -2 l_R(M; D) - (r+R)^2 + \epsilon^{-1}(r+R)^2 [Z_G - \log\{1/(2\delta)\}]}$$

where Z_G is a standard Laplace random variable.

Then we can construct the final estimators by using either exponential mechanism or the noisy minimization, both satisfying $(2\epsilon, \delta)$ differential privacy. The complete algorithm is given in algorithm 2 (Table 2).

3.3. Choosing the tuning parameters

Both algorithm 1 and algorithm 2 require two tuning parameters: an upper bound R of the l_1-norm of the regression coefficient, and the penalty parameter ϕ_n. In the context of privacy preserving data analysis, there are two requirements on the quality of any inference procedure: privacy and utility. Both proposed methods satisfy their corresponding ϵ differential privacy (algorithm 1) and $(2\epsilon, \delta)$ differential privacy (algorithm 2) for any choices of R and ϕ_n. Regarding utility, our theoretical analysis shows that both algorithms achieve consistent model selection for a wide range of R and ϕ_n.

However, our numerical experiments in Section 5 show that the performance of our proposed algorithms is sensitive to the choice of these tuning parameters. We acknowledge that fully data-driven and privacy preserving methods for choosing these tuning parameters remain a challenging and important open problem which is beyond the scope of this paper. Data-driven choice of penalty parameter is a difficult problem even without privacy constraints. Here we provide some heuristics on potential solutions.

Regarding choosing R, the ideal choice would be the l_1-norm of the true regression coefficient. In practice, a good choice of R should be close to $\|\beta_0\|_1$. This can be achieved by finding a differentially private estimate of the maximum l_1-norm of $\hat{\beta}_M$ over all $M \in \mathcal{M}$, which is feasible if only a single number is released.

Regarding choosing ϕ_n, a good choice of ϕ_n needs to be sufficiently large that it dominates any statistical sampling noise in the data and the additive noise due to privacy constraints. However, ϕ_n cannot be too large, because otherwise it will introduce substantial bias in the model selected. For the penalized constrained least squares algorithm, if we choose $\phi_n = \sigma^2 \log(n)$, then the procedure is almost the same as the BIC, except that the least square estimate comes with an (usually inactive) l_1-constraint. Therefore, if we replace σ^2 by a private estimate $\hat{\sigma}^2$, then we can mimic the BIC by choosing $\phi_n = \hat{\sigma}^2 \log(n)$. In our data examples, we illustrate that such a choice works well in practice and the results are stable under small changes of ϕ_n, provided that the sample size is moderately large.

3.4. Choosing the privacy parameter

The privacy property of a differentially private procedure is quantified by ϵ, the privacy parameter. According to the definition of ϵ differential privacy, a smaller value of ϵ implies stronger privacy protection. How to choose ϵ is usually a matter of policy and/or the users' preference. From a practical point of view, $\epsilon = 1$ means that the data release will not change the probability of any inferential outcome regarding an arbitrary individual by more than threefold, whereas $\epsilon = 0.1$ bounds such a change of probability by about 10%. The protection decreases exponentially in ϵ, so the practical protection becomes very weak for larger values of ϵ, such as $\epsilon \geqslant 10$.

In contrast, since most differentially private procedures add noise to statistics that is proportional to ϵ^{-1}, a smaller ϵ leads to larger noise, which will limit the utility of the statistic. Therefore, there is a privacy–utility trade-off in the choice of ϵ. In our simulation study, the choice of $\epsilon = 1$ gives reasonably good utility.

4. Utility analysis

In privacy preserving data analysis, the privacy will be protected for *any* possible input data set. In other words, the privacy guarantee needs to cover the worst case and must be established with no distributional conditions on the data set. In the previous sections, our differentially private procedure requires only that the data are bounded, which can be verified or enforced easily in practice. In contrast, the statistical utility (e.g. consistency or rate of convergence) is usually based on common statistical assumptions on the data. To facilitate discussion, we first introduce some notation and assumptions.

4.1. Notation

Let $\mathbf{X}$ be the $n \times d$ design matrix, and $\mathbf{Y}$ the $n \times 1$ vector of Ys. For any $M \in \mathcal{M}$, let $\mathbf{X}_M$ be the $n \times |M|$ design matrix consisting of the columns in M. Then $\hat{\beta}_M$ is a $d \times 1$ vector that is

$(\mathbf{X}_M^{\mathrm{T}}\mathbf{X}_M)^{-1}\mathbf{X}_M^{\mathrm{T}}\mathbf{Y}$ on the entries in M and 0 elsewhere. It is the ordinary least square estimate under model M. The sample covariance is $\hat{\Sigma} = n^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}$.

4.2. Assumptions

We state below five assumptions required for our utility results and discuss their practical significance.

Our first assumption is a sparse linear model with Gaussian noise.

Assumption 1 (linear model with Gaussian noise). The data entries $(X_i, Y_i)_{i=1}^n$ are generated from the linear regression model (4) with a regression coefficient vector β_0 with $d_0 = \|\beta_0\|_0$ and $b_0 \equiv \min_{j:\beta_0(j)\neq 0} |\beta_0(j)|$. The noises W_i are independently and identically distributed Gaussian with mean 0 and variance σ^2.

Next we assume that the number of candidate models grows polynomially with sample size n. This is usually so when we search only over sparse models and the total number of variables grows polynomially in n.

Assumption 2 (candidate models). The set of candidate models $\mathcal{M}$ contains the true model $M_0 = \{j : \beta_0(j) \neq 0\}$ and has cardinality no more than n^α for some positive number α which is allowed to grow with n. The largest candidate model has $\bar{d}$ variables.

The worst-case sensitivity of the log-likelihood is difficult to control because the design matrix $\mathbf{X}$ may be poorly conditioned. Thus we need to add some singular value condition on the design matrix.

Assumption 3 (design matrix). The design matrix $\mathbf{X}$ is fixed, and the sample covariance satisfies the sparse eigenvalue condition

$$\kappa_0 \equiv \inf_{1 \leqslant \|\beta\|_0 \leqslant \bar{d}+d_0} \frac{\beta^{\mathrm{T}}\hat{\Sigma}\beta}{\|\beta\|_2^2} > 0.$$

Assumption 3 excludes the situation of linear dependence between columns of $\mathbf{X}_{M_0}$ and $\mathbf{X}_{M_0^{\mathrm{c}}}$. A similar condition has been considered in the literature on model selection consistency by using information criteria (Nishii, 1984). The condition that is stated here is for a fixed design matrix $\mathbf{X}$, but it holds with high probability for many random designs. Assumption 3 also implies an upper bound of the range of $\hat{\beta}_M$. Such a boundedness property will help to control the sensitivity of the log-likelihood.

Moreover, as we did in Section 3, we assume boundedness of the data entries, which is typically needed for developing differentially private procedures.

Assumption 4 (boundedness). Each entry of $\mathbf{Y}$ is bounded by r, and each entry of $\mathbf{X}$ is bounded by 1.

Finally, we assume that the sample size is sufficiently large, when compared with other quantities in the analysis such as the noise variance, and the inverse of privacy parameter ϵ.

Assumption 5 (sample size). The sample size n is sufficiently large that equations (18), (19) and (20) hold.

4.3. Utility of penalized constrained least squares

We first give a utility result for the noisy penalized constrained least squares estimator (algorithm 1).

Theorem 1. Assume that assumptions 1–5 hold. If ϕ_n satisfies

$$2A(1 \vee \sigma^2)\log(n) \leqslant \phi_n \leqslant \frac{1}{4 \vee (1+2d_0)}\kappa_0 b_0^2 \sigma^2 n,$$

where $A = 2(\alpha + c)$ for some $c > 0$, and

$$R \geqslant r\sqrt{\left(\frac{\bar{d}}{\kappa_0}\right)},$$

then the penalized constrained least squares estimator $\hat{M}$ given by algorithm 1 with privacy parameter ϵ and penalty parameter ϕ_n satisfies

$$P_{D,\omega}(\hat{M} \neq M) \leqslant n^{\alpha} \exp\left\{-\frac{\phi_n \epsilon}{4(R+r)^2}\right\} + \frac{1+2\bar{d}}{\sqrt{\{2\pi A \log(n)\}}} n^{-c}.$$

Recall that $P_{D,\omega}$ denotes taking probability over both the randomness in D and in the generation of Laplace random variables (denoted by ω) in the algorithm. Here we are assuming a fixed design matrix, so the randomness in D is equivalent to the randomness of the additive noise W.

4.4. Utility of penalized constrained profile likelihood

Now we provide a utility result for the noisy penalized profile likelihood estimator.

Theorem 2 (utility of penalized profile likelihood). Assume that assumptions 1–5 hold. If ϕ_n satisfies

$$4A\log(n) \leqslant \phi_n \leqslant \frac{2}{3|M_0|} n \log\left(1 + \frac{\kappa_0 b_0^2}{4\sigma^2}\right),$$

where $A = 2(\alpha + c)$ for some $c > 0$, and

$$R \geqslant r\sqrt{\left(\frac{\bar{d}}{\kappa_0}\right)}$$

then for any constant $c > 0$ and n sufficiently large as quantified in equations (18)–(20), the selected model $\hat{M}$ that is given by the noisy penalized constrained profile likelihood (algorithm 2) satisfies

$$P_{D,\omega}(\hat{M} \neq M_0) \leqslant n^{\alpha} \exp\left\{-\frac{\epsilon \sigma^2 \phi_n}{64(R+r)^2}\right\} + 3n^{-A/2} + 2\bar{d}n^{-c}.$$

5. Empirical results

We have shown in the previous section that the two proposed differentially private model selection procedures are consistent under similar conditions to those of the corresponding non-privacy algorithms. We now provide some empirical results for model selection via penalized constrained least squares, to illustrate both the utility of the algorithms and the influence of tuning parameter selection as discussed in Section 3.3.

5.1. Simulation study

We consider two different regression models, both of the form $Y = X^{\mathrm{T}}\beta_0 + W$, where X is an $n \times 6$ matrix with columns sampled independently from the uniform distribution on $[-1, 1]$ and W is standard normal, but the regression coefficients are different. The first model uses $\beta_0 = (1, 1, 1, 0, 0, 0)$ whereas the second model uses $\beta_0 = c(1.5, 1, 0.5, 0, 0, 0)$. Note that the

l_1-norm of β_0 is 3 in both cases, and thus the oracle value for R. However, the first model should be easier to recover from data.

In the simulations, we consider a small sample size ($n = 100$) and a moderately large sample size ($n = 1000$), as well as $R \in \{1, 2.5, 3.5, 10\}$, $\epsilon \in \{0.1, 1, 5, 10\}$ and $\phi \in [0, n/2]$. For each set of parameters, we sample 500 data sets from the true model, apply algorithm 1 to each of them and note the proportion of times that we correctly identify the correct model. We set r to be the maximum observed value of Y in the data set.

Fig. 1 shows the results for model 1 where $\beta_0 = (1, 1, 1, 0, 0, 0)$. Except with the smallest value of R, the procedure is very successful at choosing the correct model over a wide range of ϕ-values. The poor results for $R = 1$ are explained by the fact that, since R is much smaller than the true l_1-norm of β_0, the parameter cannot be estimated properly. When R is too large, the estimation of β_0 is unchanged, but a larger amount of noise is needed to satisfy differential privacy, which explains the small drop in utility. As expected, decreasing ϵ needed to provide stricter privacy also leads to decreased utility. Note, however, that for $n = 1000$ the method is very accurate for a wide range of ϕ even for ϵ as small as 1. The results also confirm our claim that the choice of ϕ is less crucial for larger sample size, as the procedure works well on an entire interval. Note also that the actual value of ϕ depends on n, with the optimal ϕ increasing roughly linearly with n.

Fig. 2 shows the results for where $\beta_0 = c(1.5, 1, 0.5, 0, 0, 0)$. The effects of n, R, ϕ and ϵ are very similar in this case, but the choice of R and ϕ is more important to achieve good utility. The sensitivity of the procedure to choices of R and ϕ thus depends on the structure of the true parameter β_0. Note, however, that for $n = 1000$ a proper choice for the parameters leads to completely accurate model selection with $\epsilon = 5$, and even very accurate for $\epsilon = 1$ with $R = 2.5$.

5.2. Application to real data sets

We first illustrate the results of model selection via penalized constrained least squares on a small data set of 97 observations, and then on a much larger data set with hundreds of thousands of observations.

5.2.1. Prostate cancer data set

The prostate cancer data set contains several clinical measures for 97 men with prostate cancer. In this paper, our goal is to predict the level of a prostate-specific antigen by using five continuous variables: the volume of the cancer, lcavol, the weight of the prostate, lweight, the age of the patient, age, the capsular penetration, lcp, and the benign prostatic hyperplasia amount, lbph. Except for age, all variables are taken on the log-scale. We also rescale all the variables to take values between -1 and 1. This does not violate privacy protection if we assume that the range of the input data is public. Where the range of data is sensitive, such a scaling can be done in a differentially private way by taking private estimators of upper and lower α-quantiles of the variable (Dwork and Lei, 2009), followed by truncation and scaling.

We consider all possible main effects models for the model selection procedure, for a total of 63 models to choose from. Model selection based on penalized maximum likelihood, without the constraint of differential privacy, selects the model with only two variables in addition to the intercept: the volume of the cancer and the weight of the prostate. Model selection is then performed by using algorithm 1 for various choices of R, ϕ and ϵ. We set r to the maximum value of Y.

Because of the lack of ground truth about the correct model, we compare the differentially privately selected and fitted model with the model selected and fitted by the BIC, which is a well-studied and commonly used non-privacy model selection procedure for such low dimensional regression problems. The non-privacy BIC model contains variables lcval and lweight, with an adjusted R^2-value 0.587.

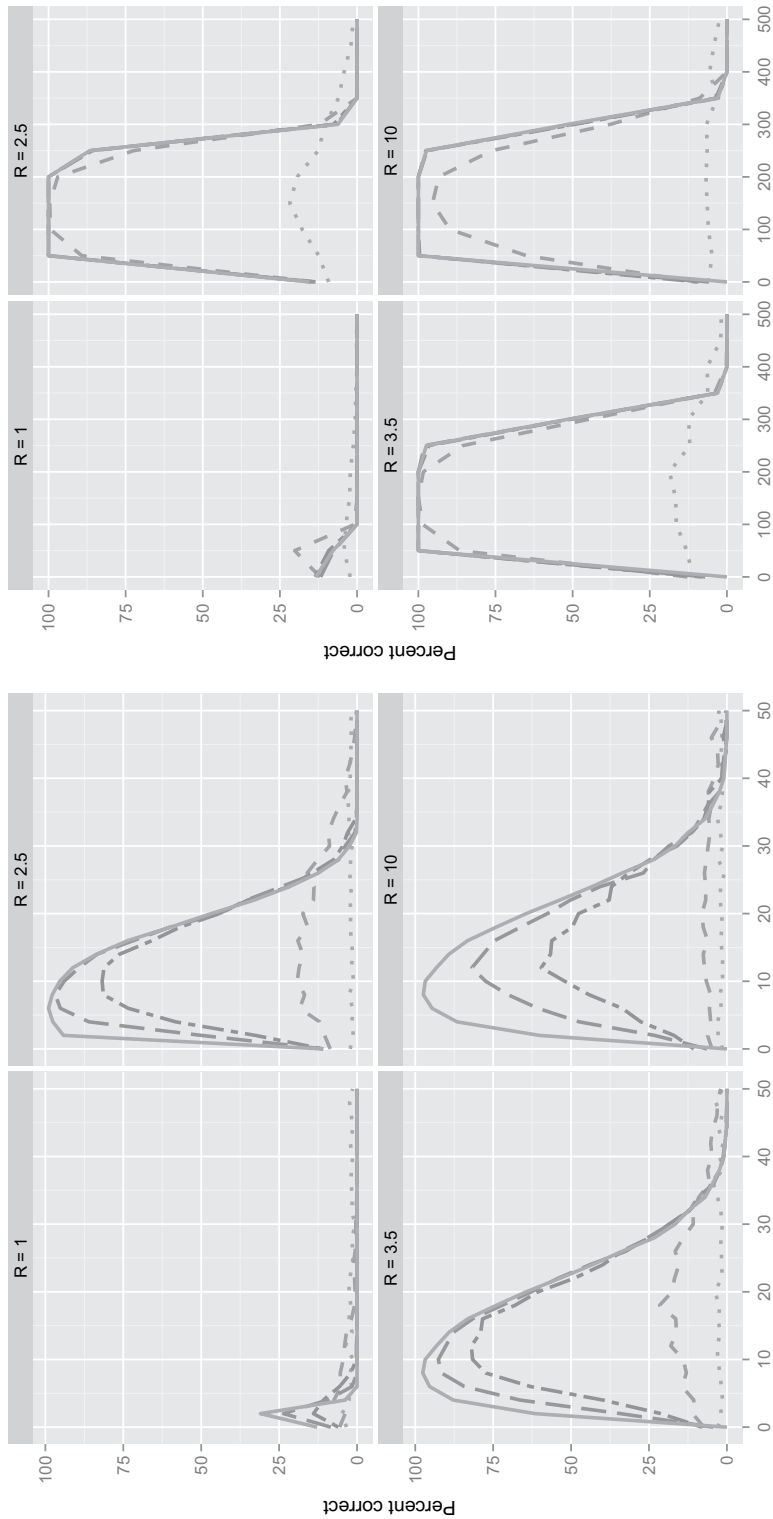

Fig. 1 Proportion of correct model selection from 500 independent replications of algorithm 1 selecting between the 63 possible models when the true model has $\beta = (1, 1, 1, 0, 0, 0)$ (various values of n, R, ϕ and ϵ are illustrated, and r is set to the maximum value of Y; there is a large range of ϕ for which the privacy procedure does as well as without privacy, and in fact picks the correct model; as n increases, the task becomes easier; note that the proper value of ϕ increases with n as well) (· · · · ·, $\epsilon = 0.1$; – – –, $\epsilon = 1$; - — - —, $\epsilon = 5$; — —, $\epsilon = 10$; ———, no differential privacy): (a) $n = 100$; (b) $n = 1000$

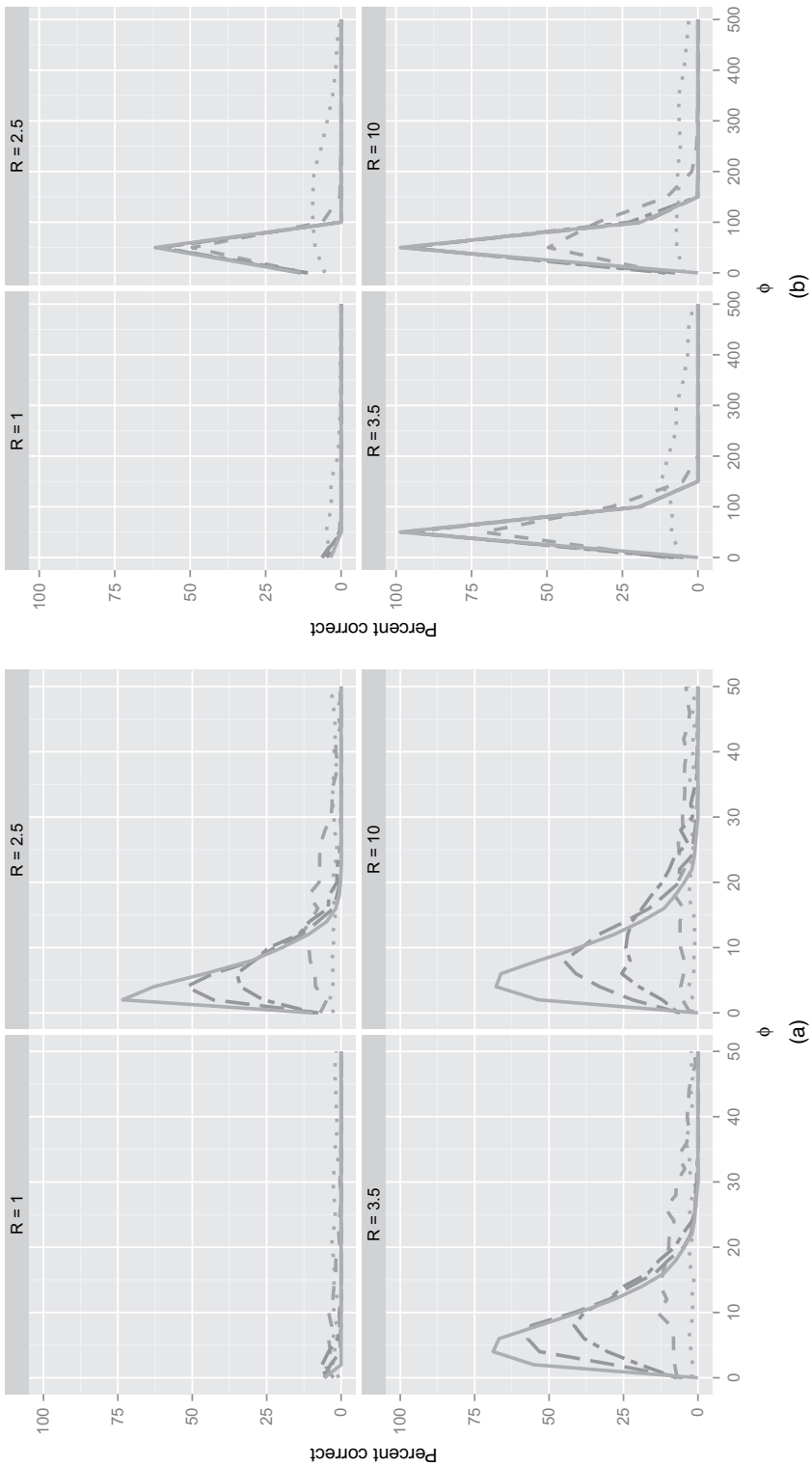

Fig. 2 Proportion of correct model selection from 500 independent replications of algorithm 1 selecting between the 63 possible models when the true model has $\beta = (1.5, 1, 0.5, 0, 0, 0)$ (various values of n, R, ϕ and ϵ are illustrated, and r is set to the maximum value of Y; the utility of the procedure depends crucially on the choices for R and ϕ; as n increases, the task becomes easier; note that the proper value of ϕ increases with n as well) (· · · · · ·, $\epsilon = 0.1$; – – –, $\epsilon = 1$; - — -, $\epsilon = 5$; — — —, $\epsilon = 10$; ————, no differential privacy): (a) $n = 100$; (b) $n = 1000$

Table 3. Average relative adjusted R^2-value over 1000 repetitions on the prostate data†

R	*Results for* $\epsilon = 1$				*Results for* $\epsilon = 5$			
	$\phi = 1$	$\phi = 2$	$\phi = 4$	$\phi = 8$	$\phi = 1$	$\phi = 2$	$\phi = 4$	$\phi = 8$
4	0.80	0.79	0.79	0.79	0.86	0.86	0.86	0.86
6	0.79	0.79	0.78	0.78	0.85	0.85	0.86	0.86
8	0.78	0.78	0.77	0.77	0.85	0.85	0.85	0.86
10	0.77	0.77	0.75	0.75	0.85	0.85	0.86	0.86

†The baseline R^2-value is 0.587.

Table 4. Frequency of inclusion of each variable in the 1000 repeated applications of the privacy model selection

ϕ	*lcavol*	*lweight*	*age*	*lbph*	*lcp*
1	0.85	0.51	0.48	0.51	0.54
2	0.83	0.53	0.47	0.47	0.58
4	0.83	0.49	0.45	0.45	0.49
8	0.83	0.44	0.39	0.41	0.43

Table 3 reports the average relative adjusted R^2-value of the privacy fit and the BIC fit for a variety of combinations of the tuning parameter (ϵ, ϕ, R). The average is taken over 1000 repeated applications of the (random) differentially private model selection procedure. We see that, when $\epsilon = 1$, all values of R and ϕ lead to a relative adjusted R^2 between 0.75 and 0.8. When $\epsilon = 5$, the values increase to above 0.85. In such a small sample, the performance is stable regarding the choices of R and ϕ, but it decreases when stronger privacy is required (smaller ϵ). All standard errors are of the order of 10^{-3} and have been omitted. The value of ϕ corresponding to the BIC with known noise variance is about 2.42.

Table 4 reports the frequencies of each variables being included over 1000 repeated applications of the privacy model selection procedure, with $R = 4$, $\epsilon = 1$ and $\phi \in (1, 2, 4, 8)$. Again, the performance is stable across different choices of ϕ. Similar results for other values of ϵ and R have been omitted. It is clear that the variable lcavol (the volume of cancer) is by far the most stably chosen variable. In addition to variable lcavol, the non-privacy BIC method also selects lweight. The selection frequencies of lweight are comparable with those of lcp. In fact, the model with the three variables lcavol, lweight and lcp has the second-best BIC score among all candidate models.

5.2.2. Housing data set

Since the constrained optimization can be implemented on the basis of only sufficient statistics, the privacy procedure scales very well for much larger data sets. The housing data set contains several variables measured on 348 189 houses sold in the San Francisco Bay area between 2003 and 2006. In addition to the price of the sale, the data include the year of the transaction (an ordinal variable with four levels), the latitude and longitude of the house, the county in which it is located (a categorical variable with nine levels) and a continuous measure of its size. We

preprocess the data to remove houses with price outside the range \$105 000–905 000 and size larger than 3000 ft^2. We also combine some of the small counties into a new indicator variable. All predictors are also scaled so that they take values in $[-1, 1]$. The resulting data set contains 235 760 sample points with 12 variables without intercept, including base square footage, bsqft, lot square footage, lsqft, time of transaction, time, age of house, age, latitude, lat, longitude, long, number of bedrooms, nbr, Alameda County, ala, Contra Costa County, cc, Marin plus San Francisco plus San Mateo, mss, Napa plus Sonoma, ns, and Santa Clara, sc.

We consider again models with main effects only, for a total 8191 models to choose the best model from. As above, we apply the model selection procedures with various values of R, ϕ and ϵ, and use the maximum value of Y as r. Similarly, the resulting models are compared with the model given by the BIC method. The adjusted R^2-value of the non-privacy BIC model is 0.282.

Table 5 reports the average relative adjusted R^2-value of the privacy fit and the BIC fit for a variety of combinations of the tuning parameter (ϵ, ϕ, R), each over 1000 repeated applications. We see that, except the clearly underfitting case $R = 10$, all other combinations of tuning parameters give very similar and good performance, with relative adjusted R^2 very close to 1. All standard errors are very small, of the order of 10^{-5}, and have been omitted. This suggests that, given a moderately large sample size, the performance of the method proposed is insensitive to ϕ, and also insensitive to R as long as R is sufficiently large to avoid underfitting. The value of ϕ corresponding to the BIC with known noise variance is about 8.8. Results for other values of ϵ and ϕ are available and quite similar.

Table 6 reports the frequencies of each variable being included over 1000 repeated applications, with $R = 35$, $\epsilon = 1$ and $\phi \in (4, 8, 16, 32)$. Again, the performance is stable across different choices of ϕ. Similar results for other values of ϵ and R have been omitted. It is clear that the variables lsqft, ala and sc are significantly less important than other variables, with frequency of inclusion between 0.4 and 0.6. The variables bsqft and age are quite important but are not always

Table 5. Average relative adjusted R^2-value over 1000 repetitions on the housing data†

R	*Results for* $\epsilon = 1$				*Results for* $\epsilon = 5$			
	$\phi = 4$	$\phi = 8$	$\phi = 16$	$\phi = 32$	$\phi = 4$	$\phi = 8$	$\phi = 16$	$\phi = 32$
10	0.623	0.623	0.623	0.623	0.624	0.624	0.624	0.623
25	0.995	0.995	0.995	0.995	0.998	0.998	0.998	0.998
35	0.997	0.997	0.997	0.996	1	1	1	0.999
100	0.994	0.993	0.993	0.993	0.999	0.999	0.999	0.999

†The baseline R^2 value is 0.282.

Table 6. Frequency of inclusion of each variable in the 1000 repeated applications of the private model selection

ϕ	*bsqft*	*lsqft*	*time*	*lat*	*long*	*age*	*nbr*	*ala*	*cc*	*mss*	*ns*	*sc*
4	0.85	0.47	1	1	1	0.84	1	0.60	0.99	1	0.92	0.58
8	0.88	0.49	1	1	1	0.83	1	0.60	0.98	1	0.91	0.60
16	0.85	0.48	1	1	1	0.86	1	0.58	0.97	1	0.91	0.58
32	0.83	0.45	1	1	1	0.80	1	0.56	0.97	1	0.90	0.54

included in the privately selected model. The other variables are clearly important, with inclusion frequency close to 1. The BIC method excludes the variable sc, which has the second-lowest frequency of inclusion.

6. Discussion

With modern data acquisition and storage techniques enabling the collection and analysis of huge amounts of personal information in a multitude of formats, protecting individual privacy inevitably becomes of crucial concern for modern data analysis. Although differential privacy offers a mathematically strong and elegant privacy guarantee, the nature of such a conservative constraint in the context of everyday statistical analysis remains unclear. Previous work on statistical analysis with differential privacy mainly focused on simple statistical queries, such as location and scale statistics, regression coefficients and simple hypothesis testing, and network analysis. In this paper we considered the more challenging problem of model selection in the classical setting. We showed that standard techniques for differentially private data analysis can be combined with known statistical tools such as penalized least square or information criteria to construct privacy preserving model selection procedures with strong utility. We proposed two algorithms for this task and proved privacy and utility results for each of them.

Our procedures feature a double regularization, as they include both constrained estimation in the fitting step and penalization in the model comparison step. Thus the method involves two tuning parameters. A key observation, which was illustrated in Section 5, is that, although we have proven good large sample properties for a wide range of penalty parameter ϕ_n and R, the practical performance is sensitive to the particular choices for these parameters. In other words, these tuning parameters play a very important and unique role in designing differentially private procedures with good practical performance. Although, in low dimensional settings, the regularization parameter is not needed for statistical inference without privacy constraints, when privacy is a concern, a carefully chosen amount of regularization can lead to stable, low sensitivity estimators even in the worst case. Appropriate choice of both tuning parameters thus reflects the need for some additional information in the data that will be useful for differentially private procedures, but not necessary in traditional inference methods. It will be an interesting future topic to give a general characterization of such privacy-related quantities, and to develop differentially private methods to estimate these quantities.

Appendix A: Proofs and technical details

A.1. Proofs

We first provide proofs of the theorems.

A.1.1. Proof of proposition 1

$$
\begin{aligned}
P_\omega[\mathcal{T}\{D, G(D)\} \in A] &\leqslant P_\omega[\mathcal{T}\{D, G(D)\} \in A, G(D) \geqslant G^*(D)] + \mathbb{P}\{G(D) < G^*(D)\} \\
&\leqslant \int_{u \geqslant G^*(D)} P_\omega\{\mathcal{T}(D, u) \in A\}\, \mathrm{d}P_{G(D)}(u) + \delta \\
&\leqslant \int_{u \geqslant G^*(D)} \mathrm{e}^{\epsilon_2} P_\omega\{\mathcal{T}(D', u) \in A\}\, \mathrm{d}P_{G(D)}(u) + \delta \\
&\leqslant \int_{u \geqslant G^*(D)} \mathrm{e}^{\epsilon_2} P_\omega\{\mathcal{T}(D', u) \in A\} \mathrm{e}^{\epsilon_1}\, \mathrm{d}P_{G(D')}(u) + \delta \\
&\leqslant \mathrm{e}^{\epsilon} P_\omega\{\mathcal{T}(D', G(D')) \in A] + \delta.
\end{aligned}
$$

A.1.2. Proof of lemma 2

Given a candidate model M, the log-likelihood is, ignoring constant terms,

$$-2l^*(M;D) = n\log\left\{n^{-1}\sum_{i=1}^{n}(Y_i - X_i^{\mathrm{T}}\hat{\beta}_M)^2\right\}.$$

Let $\hat{\beta}'_{R,M}$ be the constrained least square estimator with input data set D'. By boundedness of Y_i, X_i and $\hat{\beta}_{R,M}$ we have $(Y_i - X_i^{\mathrm{T}}\beta)^2 \leqslant (r+R)^2$ for all β such that $\|\beta\|_1 \leqslant R$. Then, using the fact that $\log(x/y) \leqslant (x-y)/y$ for $x \geqslant y$, we have

$$\begin{aligned}
|2l_R^*(M;D') - 2l_R^*(M;D)| &= n\left|\log\left\{\frac{l_R(M;D)}{l_R(M;D')}\right\}\right| \\
&\leqslant \max\left\{\frac{(Y'_n - X_n'^{\mathrm{T}}\hat{\beta}_{R,M})^2}{n^{-1}\sum_{i=1}^{n}(Y_i - X_i^{\mathrm{T}}\hat{\beta}_{R,M})^2}, \frac{(Y_n - X_n^{\mathrm{T}}\hat{\beta}'_{R,M})^2}{n^{-1}\sum_{i=1}^{n}(Y'_i - X_i'^{\mathrm{T}}\hat{\beta}'_{R,M})^2}\right\} \\
&\leqslant \frac{n(r+R)^2}{-2l_R(M;D) - (r+R)^2}.
\end{aligned}$$

Next we prove our main utility theorems. The proofs of theorem 1 and theorem 2 rely on the following result, which is a consequence of simple linear algebra.

Proposition 2. We have $\|\hat{\beta}_M\|_1 \leqslant \sqrt{(\bar{d}/\kappa_0)}r$ for all $M \in \mathcal{M}$. As a result, if $R \geqslant \sqrt{(\bar{d}/\kappa_0)}r$ then $\hat{\beta}_{R,M} = \hat{\beta}_M$, $l(M;D) = l_R(M;D)$ and $l_R^*(M;D) = l^*(M;D)$ for all $M \in \mathcal{M}$.

Proposition 2 enables us to remove the l_1-constraint in our analysis. Although proposition 2 implies that the l_1-constraint is inactive, such a constraint cannot be removed from the algorithms because it is used to bound the worst-case sensitivity of the estimators.

A.1.3. Proof of theorem 1

We prove theorem 1 for noisy minimization. The argument can be adapted simply to cover the exponential mechanism.

Define

$$\Delta(M) = [-2l(M;D) + 2l(M_0;D)].$$

Then, according to proposition 2, we can work directly with the unconstrained version:

$$\begin{aligned}
P_{D,\omega}(\hat{M} \neq M) &\leqslant \sum_{M' \in \mathcal{M}} P_{D,\omega}\left\{-2l(M';D) + \phi_n|M'| + \frac{2(R+r)^2}{\epsilon}Z_{M'} \leqslant -2l(M_0;D) + \phi_n|M_0| + \frac{2(R+r)^2}{\epsilon}Z_{M_0}\right\} \\
&\leqslant n^{\alpha} P_{D,\omega}\left[\frac{2(R+r)^2}{\epsilon}(Z_1 - Z_2) \geqslant \sup_{M' \neq M_0}\{\Delta_{M'} + (|M'| - |M_0|)\phi_n\}\right].
\end{aligned}$$

For $M' \supset M_0$, using claim 2 in Appendix A.2 we have that $\sup_{M':M' \supset M_0} \Delta_{M'} \geqslant -A\sigma^2(|M'| - |M_0|)\log(n)$ with probability at least

$$1 - \frac{\bar{d} - d_0}{\sqrt{\{2\pi A \log(n)\}}}n^{-c}.$$

Thus with the same probability we have

$$\sup_{M' \supset M_0} \Delta_{M'} + (|M'| - |M_0|)\phi_n \geqslant (|M'| - |M_0|)\{\phi_n - A\sigma^2\log(n)\} \geqslant \phi_n/2. \tag{15}$$

For $M' \not\subseteq M_0$, we have by claim 3 in Appendix A.2, with probability at least

$$1 - \frac{1 + \bar{d}}{\sqrt{\{2\pi A \log(n)\}}}n^{-c},$$

$$\sup_{M' \not\subseteq M} \Delta_{M'} + (|M'| - |M_0|)\phi_n \geqslant \tfrac{1}{2}\kappa_0 b_0^2 \sigma^2 n - |M_0|\phi_n \geqslant \phi_n/2. \tag{16}$$

Thus conditioning on $E_2^c \cap E_3^c$, which has probability at least

$$1 - \frac{1+2\bar{d}}{\sqrt{\{2\pi A \log(n)\}}} n^{-c},$$

we have

$$P_\omega(\hat{M} \neq M_0) \leqslant n^\alpha \exp\left\{ -\frac{\phi_n}{2} \frac{\epsilon}{2(R+r)^2} \right\}.$$

A.1.4. Proof of theorem 2

Consider events E_1–E_5 as defined in Appendix A.2. We focus on the event $(\cup_{k=0}^5 E_k)^c$, which has probability at least $1 - 3n^{-A/2} - 2\bar{d}n^{-c}$.

Because $R \geqslant r\sqrt{(\bar{d}/\kappa_0)}$ by proposition 2 we have $-2l^*(M; D) = -2l_R^*(M; D)$ for all M under consideration.

Next we bound the difference between $-2l^*(M; D)$ and $-2l^*(M_0; D)$. Denote

$$\Delta^*(M) = -2l^*(M; D) + 2l^*(M_0; D).$$

In the case $M \supset M_0$, applying the fact that $-\log(1-x) \leqslant x/(1-x)$ for $x \in (0, 1)$ to

$$x = 1 - \frac{\sum_{i=1}^n (Y_i - X_i^T \hat{\beta}_M)^2}{\sum_{i=1}^n (Y_i - X_i^T \hat{\beta}_0)^2}$$

we have

$$\begin{aligned} 0 \leqslant -\Delta^*(M) &= n \log\left\{ \frac{\sum_{i=1}^n (Y_i - X_i \hat{\beta}_0)^2}{\sum_{i=1}^n (Y_i - X_i \hat{\beta}_M)^2} \right\} \leqslant n \frac{-\Delta(M)}{-2l(M; D)} \\ &\leqslant 2A(|M| - |M_0|)\log(n), \end{aligned} \tag{17}$$

where the last step follows from the fact that we are not in the event $E_0 \cup E_1 \cup E_2$ defined in Appendix A.2.

In the case $M \not\subseteq M_0$,

$$\Delta^*(M) = n \log\left\{ 1 + \frac{\Delta(M)}{-2l(M_0; D)} \right\} \geqslant n \log\left(1 + \frac{\kappa_0 b_0^2}{4\sigma^2} \right)$$

because we are in E_0^c (which implies that $-2l(M_0; D) \leqslant 2n\sigma^2$ by claim 1) and E_3^c (which implies that $\Delta(M) \geqslant n\kappa_0 b_0^2 \sigma^2/2$ by claim 3).

Therefore, when

$$4A \log(n) \leqslant \phi_n \leqslant \frac{2}{3|M_0|} n \log\left(1 + \frac{\kappa_0 b_0^2}{4\sigma^2} \right),$$

we have

$$L_R^*(M; D) - L_R^*(M_0; D) \geqslant \phi_n/2, \qquad \forall\, M \neq M_0.$$

On the event considered, we also have $G(D) \leqslant 4(R+r)^2\sigma^{-2}$ by claim 5. Then we can bound the error probability by

$$\begin{aligned} P_\omega\{\hat{M} \neq M_0 | D, G(D)\} &= \sum_{M \in \mathcal{M}, M \neq M_0} P_\omega(\hat{M} = M) \\ &\leqslant \sum_{M \in \mathcal{M}, M \neq M_0} \exp\left[-\frac{\epsilon}{4G(D)}\{L_R^*(M; D) - L_R^*(M_0; D)\}\right] \\ &\leqslant n^\alpha \exp\left\{-\frac{\epsilon\phi_n\sigma^2}{64(R+r)^2}\right\}. \end{aligned}$$

A.2. Further proof details

Here we give details and summarize the multiple 'with high probability' statements in the utility analysis. Recall that

$$\Delta(M) = [-2l(M; D) + 2l(M_0; D)].$$

Let $A = 2(\alpha + c)$; assume that

$$\frac{n}{\log(n)} \geqslant \max\left\{\frac{64A\sigma^2}{\kappa_0 b_0^2}, \frac{4\bar{d}\sigma^2}{\kappa_0 b_0^2}, 8A\bar{d}\right\}, \tag{18}$$

$$d_0 \leqslant n/8, \tag{19}$$

$$\epsilon \geqslant \frac{(R+r)^2[(A/2)\log(n) + \log\{1/(2\delta)\}]}{n\sigma^2/4 - (R+r)^2}. \tag{20}$$

We define the following events and give the corresponding probability bounds.

(a) Define

$$E_0 := \{D : -2l(M_0; D)\sigma^{-2} \geqslant n - d_0 + \sqrt{\{2(n-d_0)A\log(n)\}} + A\log(n)\}, \tag{21}$$

$$E_1 := \{D : -2l(M_0; D)\sigma^{-2} \leqslant n - d_0 - \sqrt{\{2(n-d_0)A\log(n)\}}\}. \tag{22}$$

Claim 1. $P_D(E_0) \leqslant n^{-A/2}$ and $P_D(E_1) \leqslant n^{-A/2}$. Also, using equation (18), we have $-2l(M; D) \leqslant 2n\sigma^2$ on E_0^c.

Proof. Without loss of generality, assume that $\sigma^2 = 1$. Then $-2l(M_0; D) = \|\mathbf{Y} - P_{\Pi_0}\mathbf{Y}\|_2^2 = \|P_{\Pi_0^\perp}\mathbf{W}\|_2^2$ is a χ^2 random variable with $n - d_0$ degrees of freedom. The first part of claim 1 follows from lemma 1 of Laurent and Massart (2000). The second part of claim 1 can be verified directly.

(b) Define

$$E_2 := \left\{\inf_{M \in \mathcal{M}, M \supset M_0} \frac{\Delta(M)}{\sigma^2(|M| - |M_0|)} < -A\log(n)\right\}. \tag{23}$$

Claim 2.

$$P_D(E_2) \leqslant \frac{\bar{d} - d_0}{\sqrt{\{2\pi A\log(n)\}}} n^{-c}.$$

Proof. Because $M \supset M_0$,

$$\begin{aligned} 0 \leqslant -\Delta(M) &= \|\mathbf{Y} - \mathbf{X}\hat{\beta}_{M_0}\|_2^2 - \|\mathbf{Y} - \mathbf{X}\hat{\beta}_M\|_2^2 \\ &= \|\mathbf{X}\beta_0 + \mathbf{W} - P_{\Pi_0}(\mathbf{X}\beta_0 + \mathbf{W})\|_2^2 - \|\mathbf{X}\beta_0 + \mathbf{W} - P_{\Pi_M}(\mathbf{X}\beta_0 + \mathbf{W})\|_2^2 \\ &= \|\mathbf{W} - P_{\Pi_0}(\mathbf{W})\|_2^2 - \|\mathbf{W} - P_{\Pi_M}(\mathbf{W})\|_2^2 = \|(P_{\Pi_M} - P_{\Pi_0})(\mathbf{W})\|_2^2. \end{aligned}$$

Because $M_0 \subset M$, $P_{\Pi_M} - P_{\Pi_0}$ is a projection operator of dimension $|M| - |M_0|$. Using a tail probability bound for Gaussian random variables, we have, for all $A = 2(\alpha + c) > 0$,

$$\mathbb{P}\{\Delta(M)\sigma^{-2} \leqslant -A(|M|-|M_0|)\log(n)\} \leqslant \frac{|M|-|M_0|}{\sqrt{\{2\pi A\log(n)\}}} n^{-A/2} \leqslant \frac{\bar{d}-d_0}{\sqrt{\{2\pi A\log(n)\}}} n^{-A/2}.$$

The desired result follows from the union bound.

(c) For $M \not\supseteq M_0$, let $M_1 = M_0 \backslash M$, $M_2 = M_0 \cap M$ and $J_M^* = \sqrt{(n\kappa_0 b_0^2 |M_1|)}$. Define

$$E_3 := \left\{ \inf_{M\in\mathcal{M}, M\not\supseteq M_0} \Delta(M) \leqslant J_M^{*2} - 2\sigma J_M^* \sqrt{\{A\log(n)\}} - (|M|-|M_2|)A\log(n) \right\}. \tag{24}$$

Claim 3.

$$P_D(E_3) \leqslant \frac{1+\bar{d}}{\sqrt{\{2\pi A\log(n)\}}} n^{-c}.$$

and under E_3^c and equation (18)

$$\Delta(M) \geqslant \tfrac{1}{2}\kappa_0 b_0^2 \sigma^2 n, \qquad \forall M \in \mathcal{M}, \quad M \not\supseteq M_0.$$

Proof. Denote $\beta_{0,M}$ the vector that agrees with β_0 on M and 0 elsewhere:

$$\begin{aligned}
\Delta(M) &= \|\mathbf{Y} - P_{\Pi_M}(\mathbf{Y})\|_2^2 - \|P_{\Pi_0^\perp}(\mathbf{W})\|_2^2 \\
&= \|P_{\Pi_M^\perp}(\mathbf{X}\beta_0) + P_{\Pi_M^\perp}(\mathbf{W})\|_2^2 - \|P_{\Pi_0^\perp}(\mathbf{W})\|_2^2 \\
&= \|P_{\Pi_M^\perp}(\mathbf{X}_{M_1}\beta_{0,M_1})\|_2^2 + 2\langle P_{\Pi_M^\perp}(\mathbf{X}_{M_1}\beta_{0,M_1}), P_{\Pi_M^\perp}(\mathbf{W})\rangle + \|P_{\Pi_M^\perp}(\mathbf{W})\|_2^2 \\
&\quad - \|P_{\Pi_0^\perp}(\mathbf{W})\|_2^2 \\
&\geqslant \|P_{\Pi_M^\perp}(\mathbf{X}_{M_1}\beta_{0,M_1})\|_2^2 + 2\langle P_{\Pi_M^\perp}(\mathbf{X}_{M_1}\beta_{0,M_1}), \mathbf{W}\rangle - \|P_{\Pi_{0,M}\cap\Pi_0^\perp}(\mathbf{W})\|_2^2,
\end{aligned}$$

where $\Pi_{0,M}$ is the linear subspace spanned by $X_{M_0\cup M}$.

Let $J = \|P_{\Pi_M^\perp}(\mathbf{X}_{M_1}\beta_{0,M_1})\|_2$. Then the second term in the above equation is distributed as $N(0,\sigma^2 J^2)$. By the eigenvalue condition, we have $J^2 \geqslant \kappa_0 n b_0^2 |M_1|$, where κ_0 is the minimum sparse eigenvalue and b_0 is a lower bound of the signal level. To see this, observe that $M \cap M_1 = \emptyset$ and

$$\begin{aligned}
J &= \|P_{\Pi_M^\perp}(\mathbf{X}_{M_1}\beta_{0,M_1})\| \\
&= \left\| (\mathbf{X}_{M_1}, \mathbf{X}_M) \begin{pmatrix} \beta_{0,M_1} \\ -(\mathbf{X}_M^{\mathrm{T}}\mathbf{X}_M)^{-1}\mathbf{X}_M^{\mathrm{T}}\mathbf{X}_{M_1}\beta_{0,M_1} \end{pmatrix} \right\| \\
&\geqslant \sqrt{(n\kappa_0)} \left\| \begin{pmatrix} \beta_{M_1} \\ -(X_M^{\mathrm{T}}X_M)^{-1}X_M^{\mathrm{T}}X_{M_1}\beta_{M_1} \end{pmatrix} \right\| \\
&\geqslant \sqrt{(n\kappa_0)}\|\beta_{M_1}\| \geqslant \sqrt{(n\kappa_0|M_1|)}b_0,
\end{aligned}$$

where the first inequality uses the sparse eigenvalue condition.

Note that the dimension of $\Pi_M \cap \Pi_0^\perp$ is at most $|M|-|M_2|$. Let $J^* := \sqrt{(\kappa_0 n b_0^2)}$. Using equation (18) we know that $J \geqslant J^* \geqslant 4\sqrt{\{A\log(n)\}}$. Then with probability at least

$$1 - (1+|M|-|M_2|)n^{-A/2}/\sqrt{\{2\pi A\log(n)\}} \leqslant 1 - \frac{1+\bar{d}}{\sqrt{\{2\pi A\log(n)\}}} n^{A/2},$$

we have

$$\begin{aligned}
\Delta(M)\sigma^{-2} &\geqslant J^2 - 2\sigma J\sqrt{\{A\log(n)\}} - (|M|-|M_2|)\sigma^2 A\log(n) && (25)\\
&\geqslant J^{*2} - 2\sigma J^*\sqrt{\{A\log(n)\}} - \sigma^2\bar{d}A\log(n) && (26)\\
&\geqslant \tfrac{1}{2}\kappa_0 b_0^2 n && (27)
\end{aligned}$$

where the last inequality uses equation (18). The claim follows from the union bound.

(d) Define

$$E_4 := \{D : \inf_{M\in\mathcal{M}} -2l(M;D) \leqslant \sigma^2 n/2\}. \tag{28}$$

Claim 4. $E_4 \subseteq E_1 \cup E_2 \cup E_3$.

To see this, first note that, on $(E_1 \cup E_2 \cup E_3)^c$, we have $\min_{M \in \mathcal{M}} -2l(M; D) \geqslant \sigma^2[n - d_0 - \sqrt{\{2(n-d_0)A\log(n)\}} - A(\bar{d} - d_0)\log(n)] \geqslant n\sigma^2/2$ by using equations (18) and (19).

(e) Define

$$E_5 := \{z_G < -A\log(n)/2\}. \tag{29}$$

Claim 5. $P_\omega(E_5) = \frac{1}{2}n^{-A/2} \leqslant n^{-A/2}$. On $E_4^c \cap E_5^c$, we have, using equations (18)–(20),

$$G(D) \leqslant \frac{4(R+r)^2}{\sigma^2}.$$

The proof is elementary and so has been omitted.

References

Abowd, J. M. and Vilhuber, L. (2008) How protective are synthetic data? In *Privacy in Statistical Databases* (eds J. Domingo-Ferrer and Y. Saygin), pp. 239–246. New York: Springer.

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Trans. Autom. Control*, **19**, 716–723.

Barak, B., Chaudhuri, K., Dwork, C., Kale, S., McSherry, F. and Talwar, K. (2007) Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In *Proc. 26th Symp. Principles of Database Systems*, pp. 273–282.

Bassily, R., Smith, A. and Thakurta, A. (2014) Private empirical risk minimization: efficient algorithms and tight error bounds. In *Proc. 55th A. Symp. Foundations of Computer Science*, pp. 464–473. New York: Institute of Electrical and Electronics Engineers.

Beimel, A., Brenner, H., Kasiviswanathan, S. P. and Nissim, K. (2014) Bounds on the sample complexity for private learning and private data release. *Mach. Learn.*, **94**, 401–437.

Beimel, A., Kasiviswanathan, S. P. and Nissim, K. (2010) Bounds on the sample complexity for private learning and private data release. In *Proc. Theory of Cryptography Conf.*, pp. 437–454.

Beimel, A., Nissim, K. and Stemmer, U. (2015) Learning privately with labeled and unlabeled examples. In *Proc. 26th A. Symp. Discrete Algorithms, San Diego, Jan. 4th–6th*, pp. 461–477.

Bun, M., Nissim, K., Stemmer, U. and Vadhan, S. P. (2015) Differentially private release and learning of threshold functions. In *Proc. 56th A. Symp. Foundations of Computer Science, Berkeley, Oct. 17th–20th*, pp. 634–649.

Chaudhuri, K. and Hsu, D. (2011) Sample complexity bounds for differentially private learning. In *Proc. Conf. Learning Theory*, pp. 155–186.

Chaudhuri, K., Monteleoni, C. and Sarwate, A. D. (2011) Differentially private empirical risk minimization. *J. Mach. Learn. Res.*, **12**, 1069–1109.

Chaudhuri, K. and Vinterbo, S. A. (2013) A stability-based validation procedure for differentially private machine learning. In *Advances in Neural Information Processing Systems*, pp. 2652–2660. Red Hook: Curran Associates.

Dalenius, T. (1977) Towards a methodology for statistical disclosure control. *Statist. Tidskr.*, **15**, 429–444.

Duchi, J. C., Jordan, M. I. and Wainwright, M. J. (2013) Local privacy and statistical minimax rates. In *Proc. 54th A. Symp. Foundations of Computer Science*, pp. 429–438. New York: Institute of Electrical and Electronics Engineers.

Dwork, C. (2006) Differential privacy. In *Proc. 33rd Int. Colloq. Automata, Languages and Programming*, pp. 1–12.

Dwork, C. and Lei, J. (2009) Differential privacy and robust statistics. In *Proc. 41st A. Symp. Theory of Computing*. New York: Association for Computing Machinery.

Dwork, C., McSherry, F., Nissim, K. and Smith, A. (2006) Calibrating noise to sensitivity in private data analysis. In *Proc. 3rd Theory of Cryptography Conf.*, pp. 265–284.

Dwork, C., Su, W. and Zhang, L. (2015) Private false discovery rate control. *Preprint arXiv:1511.03803*. Microsoft Research, Mountain View.

Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Ass.*, **96**, 1348–1360.

Fienberg, S. E., Rinaldo, A. and Yang, X. (2010) Differential privacy and the risk-utility tradeoff for multi-dimensional contingency tables. In *Proc. Int. Conf. Privacy in Statistical Databases*, pp. 187–199. New York: Springer.

Fienberg, S. and Slavković, A. (2010) Data privacy and confidentiality. In *International Encyclopedia of Statistical Science* (ed. M. Lovric), pp. 342–345. New York: Springer.

Fienberg, S. E., Slavkovic, A. and Uhler, C. (2011) Privacy preserving gwas data sharing. In *Proc. 11th Int. Conf. Data Mining*, pp. 628–635. New York: Institute of Electrical and Electronics Engineers.

Gaboardi, M., Lim, H., Rogers, R. and Vadhan, S. (2016) Differentially private chi-squared hypothesis testing: goodness of fit and independence testing. In *Proc. 33rd Int. Conf. Machine Learning*.

Ganta, S. R., Kasiviswanathan, S. P. and Smith, A. (2008) Composition attacks and auxiliary information in data privacy. In *Proc. 14th Int. Conf. Knowledge Discovery and Data Mining*, pp. 265–273. New York: Association for Computing Machinery.

Hardt, M., Ligett, K. and McSherry, F. (2012) A simple and practical algorithm for differentially private data release. In *Advances in Neural Information Processing Systems*, pp. 2339–2347. Red Hook: Curran Associates.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K. and de Wolf, P.-P. (2012) *Statistical Disclosure Control*. Hoboken: Wiley.

Johnson, A. and Shmatikov, V. (2013) Privacy-preserving data exploration in genome-wide association studies. In *Proc. 19th Int. Conf. Knowledge Discovery and Data Mining, New York*, pp. 1079–1087. New York: Association for Computing Machinery.

Karwa, V., Kifer, D. and Slavkovic, A. (2015) Private posterior distributions from variational approximations. *Preprint arXiv:1511.07896*. Pennsylvania State University, State College.

Karwa, V. and Slavkovic, A. (2012) Differentially private graphical degree sequences and synthetic graphs. In *Privacy in Statistical Databases* (eds J. Domingo-Ferrer and I. Tinnirello), pp. 273–285. Berlin: Springer.

Karwa, V. and Slavković, A. (2016) Inference using noisy degrees: differentially private beta-model and synthetic graphs. *Ann. Statist.*, **44**, 87–112.

Karwa, V., Slavkovic, A. and Krivitsky, P. (2014) Differentially private exponential random graphs. In *Privacy in Statistical Databases* (ed. J. Domingo-Ferrer), pp. 143–155. New York: Springer.

Kasiviswanathan, S. P., Lee, H. K., Nissim, K., Raskhodnikova, S. and Smith, A. (2011) What can we learn privately? *SIAM J. Comput.*, **40**, 793–826.

Kasiviswanathan, S. P. and Smith, A. (2008) On the semantics of differential privacy: a bayesian formulation. *Preprint arXiv:0803.3946*. General Electric Global Research.

Laurent, B. and Massart, P. (2000) Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, **28**, 1302–1338.

Lei, J. (2011) Differentially private m-estimators. In *Advances in Neural Information Processing Systems*, pp. 361–369.

McSherry, F. and Talwar, K. (2007) Mechanism design via differential privacy. In *Proc. 48th A. Symp. Foundations of Computer Science*, pp. 94–103.

Nishii, R. (1984) Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.*, **12**, 758–765.

Picard, R. R. and Cook, R. D. (1984) Cross-validation of regression models. *J. Am. Statist. Ass.*, **79**, 575–583.

Rubin, D. B. (1993) Statistical disclosure limitation. *J. Off. Statist.*, **9**, 461–468.

Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.

Shao, J. (1993) Linear model selection by cross-validation. *J. Am. Statist. Ass.*, **88**, 486–494.

Sheffet, O. (2015) Differentially private least squares: estimation, confidence and rejecting the null hypothesis. *Preprint arXiv:1507.02482*. Harvard University, Cambridge.

Smith, A. (2011) Privacy-preserving statistical estimation with optimal convergence rates. In *Proc. 43rd A. Symp. Theory of Computing*. New York: Association for Computing Machinery.

Smith, A. and Thakurta, A. (2013) Differentially private feature selection via stability arguments, and the robustness of the lasso. In *Proc. 26th A. Conf. Learning Theory, Princeton, June 12th–14th*, pp. 819–850.

Solea, E. (2014) Differentially private hypothesis testing for normal random variables. *Master's Thesis*. Pennsylvania State University, State College.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B, **58**, 267–288.

Uhler, C., Slavkovic, A. and Fienberg, S. E. (2013) Privacy-preserving data sharing for genome-wide association studies. *J. Priv. Confidentlty*, **5**, 137–166.

Wang, Y., Lee, J. and Kifer, D. (2015) Differentially private hypothesis testing, revisited. *Preprint arXiv:1511.03376*. Carnegie Mellon University, Pittsburgh.

Wasserman, L. and Zhou, S. (2010) A statistical framework for differential privacy. *J. Am. Statist. Ass.*, **105**, 375–389.

Willenborg, L. and De Waal, T. (1996) *Statistical Disclosure Control in Practice*. New York: Springer Science and Business Media.

Yang, Y. (2007) Consistency of cross validation for comparing regression procedures. *Ann. Statist.*, **35**, 2450–2473.

Yu, F., Fienberg, S. E., Slavkovic, A. B. and Uhler, C. (2014) Scalable privacy-preserving data sharing methodology for genome-wide association studies. *J. Biomed. Inform.*, **50**, 133–141.