Efficient differentially private learning improves drug sensitivity prediction

Antti Honkela[1,*,†], Mrinal Das[2,*], Onur Dikmen[1] and Samuel Kaski[2,†]

[1] Helsinki Institute for Information Technology HIIT, Department of Computer Science, University of Helsinki, Finland
[2] Helsinki Institute for Information Technology HIIT, Department of Computer Science, Aalto University, Finland

Abstract

Users of a personalised recommendation system face a dilemma: recommendations can be improved by learning from data, but only if the other users are willing to share their private information. Good personalised predictions are vitally important in precision medicine, but genomic information on which the predictions are based is also particularly sensitive, as it directly identifies the patients and hence cannot easily be anonymised. Differential privacy [7, 8] has emerged as a potentially promising solution: privacy is considered sufficient if presence of individual patients cannot be distinguished. However, differentially private learning with current methods does not improve predictions with feasible data sizes and dimensionalities [10]. Here we show that useful predictors can be learned under powerful differential privacy guarantees, and even from moderately-sized data sets, by demonstrating significant improvements with a new robust private regression method in the accuracy of private drug sensitivity prediction [4]. The method combines two key properties not present even in recent proposals [26, 9], which can be generalised to other predictors: we prove it is asymptotically consistently and efficiently private, and demonstrate that it performs well on finite data. Good finite data performance is achieved by limiting the sharing of private information by decreasing the dimensionality and by projecting outliers to fit tighter bounds, therefore needing to add less noise for equal privacy. As already the simple-to-implement method shows promise on the challenging genomic data, we anticipate rapid progress towards practical applications in many fields, such as mobile sensing and social media, in addition to the badly needed precision medicine solutions.

*These authors contributed equally to this work.
†These authors jointly supervised the work.

1 Introduction

The widespread collection of private data, both by individuals and hospitals in the health domain, creates a major opportunity to develop new services by learning predictive models from the data. Privacy-preserving algorithms are required and have been proposed, but for instance anonymisation approaches [1, 18, 17] cannot guarantee privacy against adversaries with additional side information, and are poorly suited for genomic data where the entire data is identifying [13]. Guarantees of differential privacy [7, 8] remain valid even under these conditions [8], and differential privacy has arisen as the most popularly studied strong privacy mechanism for learning from data.

2 Efficient differentially private learning

Differential privacy [7, 8] is a formulation of reasonable privacy guarantees for privacy-preserving computation. It gives guarantees about the output of a computation and can be combined with complementary cryptographic approaches such as homomorphic encryption [12] if the computation process needs protection too. An algorithm $\mathcal{M}$ operating on a data set $\mathcal{D}$ is said to be *differentially private* if for any two data sets $\mathcal{D}$ and $\mathcal{D}'$, differing only by one sample, the ratio of probabilities of obtaining any specific result c is bounded as

$$\frac{p(\mathcal{M}(\mathcal{D}) = c)}{p(\mathcal{M}(\mathcal{D}') = c)} \leq \exp(\epsilon). \tag{1}$$

Because of symmetry between $\mathcal{D}$ and $\mathcal{D}'$ the probabilities need to be similar to satisfy the condition. Differential privacy is preserved in post-processing, which makes it flexible to use in complex algorithms. The ϵ is a privacy parameter interpretable as a privacy budget, with higher values corresponding to less privacy preservation. Differentially private learning algorithms are usually based on perturbing either the input [2, 7], output [7, 26] or the objective [3, 28].

Here we apply differential privacy to regression. The aim is to learn a model to predict the scalar target y_i from d-dimensional inputs x_i (Fig. 1a) as $y_i = f(x_i) + \eta_i$, where f is an unknown mapping and η_i represents noise and modelling error. We wish to design a suitable structure for f and a differentially private mechanism for efficiently learning an accurate private f from a data set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$.

We argue that a practical differentially private algorithm needs to combine two things: (i) it needs to provide *asymptotically efficiently private estimators* so that the excess loss incurred from preserving privacy will diminish as the number of samples n in the data set increases; (ii) it needs to *perform well on moderately-sized data.*

While the first requirement of asymptotic efficiency or consistency seems obvious, it is non-trivial to implement in practice and rules out some mechanisms published even quite recently [29]. The requirement was addressed in the Bayesian setting very recently [9], but the method failed to cover the second equally important criterion. Asymptotically consistently private methods always allow reaching stronger privacy with more samples.

It is difficult to prove optimality of a method on finite data so good performance needs to be demonstrated empirically. A design strategy for good methods controls the amount of shared private information. This has two components: (a) dimensionality needs to be reduced, to avoid the inherent incompatibility of privacy and high dimensionality which has been discussed previously [6], and (b) introducing robustness by bounding and transforming each variable (feature) to a tighter interval. Controlling the amount of shared information also introduces a trade-off: compared to the non-private setting, decreasing the dimensionality a lot may degrade the performance of the non-private approach, while a corresponding low-dimensional private algorithm may attain higher performance than a higher-dimensional one (see the results and Fig. 3a).

The essence of differential privacy is to inject a sufficient amount of noise to mask the differences between the computation results obtained from neighbouring data sets (differing by only one entry). The definition depends on the worst-case behaviour, which implies that suitably limiting the space of allowed results will reduce the amount of noise needed and potentially improve the results. In the output perturbation framework this can be achieved by bounding the possible outputs [26].

Here we propose a more powerful approach of bounding the data by projecting outliers to tighter bounds. The current standard practice in private learning is to linearly transform the data to desired bounds [28]. This is clearly sub-optimal as a few outliers can force a very small scale for other points. Significantly higher signal-to-privacy-noise ratio can be achieved by setting the bounds to cover the essential variation in the data and projecting the outliers separately inside these bounds. This approach also robustifies the analysis against outliers as the projection can be made independent of the outlier scale. In linear regression we call the resulting model *robust private linear regression*. It is illustrated in Fig. 1b, c.

3 Results

Genomics is an important domain for privacy-aware modelling, in particular for precision medicine. Many people wish to keep their and also their relatives' genomes private [20], and simple anonymisation is not sufficient to protect privacy since a genome is inherently identifiable [13]. Furthermore, individual genomes can be recovered from summary statistics [15] as well as phenotype data such as gene expression data [14]. On the other hand, previous research has shown that poorly implemented private models may put a patient to severe risk [10].

We apply the robust private linear regression model to predict drug sensitivity given gene expression data, in a setup where a small internal data set can be complemented by a larger set only available under privacy protection (Fig. 1a). We use data from the Genomics of Drug Sensitivity in Cancer (GDSC) project [27], and the setting and evaluation are similar as in the recent DREAM-NCI drug

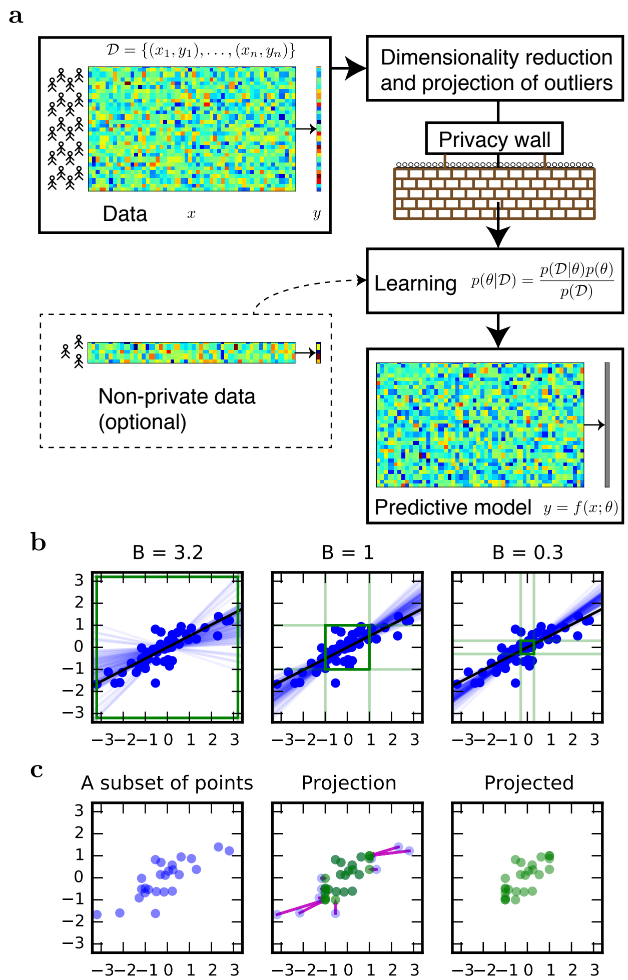

Figure 1: **Differentially private learning of a predictive model**. **a**, The modelling setup; most data (top) are available for learning only if their privacy can be protected. **b**, Bounding the data increasingly tightly (B; green square) brings 1D robust private linear regression models (blue lines illustrating the distribution of results of the randomised algorithm) closer to the non-private model (black line) as less noise needs to be injected. Blue points: data. **c**, The data are bounded in robust private linear regression by projecting outliers within the bounds (shown only for a subset of the points).

sensitivity prediction challenge [4]. The sensitivity of each drug is predicted with Bayesian linear regression based on expression of known cancer genes identified by the GDSC project [27] to limit the dimensionality. We achieve differential privacy by injecting noise to the sufficient statistics computed from the data, using the Wishart mechanism [16] to perturb the input covariance term and the Laplace mechanism [7] to perturb the target term. Full details are presented in Methods.

Unlike with previous approaches, now prediction accuracy (ranking of new cell lines [4] to sensitive vs insensitive measured by Spearman's rank correlation; Fig. 2) improves when more privacy protected data is received. The proposed non-linear projection of the data to tighter bounds is the key to this success, as without it the method performs as poorly as the earlier ones.

To improve prediction performance in differentially private learning, trade-offs need to be made between dimensionality and amount of data (Fig. 3a), and between strength of privacy guarantees and amount of data (Fig. 3c), but the amount of optional non-private data matters significantly only when there is very little private data (Fig. 3b).

In Secs. A–B in the Supplementary Information we define asymptotic consistency and efficiency of private estimators relative to non-private ones and prove that the optimal convergence rate of differentially private Bayesian estimators to the corresponding non-private ones is $\mathcal{O}(1/n)$ for n samples, which is matched by our method. Unlike existing approaches [22, 24, 23], we compare the private estimators to the corresponding non-private ones, making the theory more easily accessible and more broadly applicable.

Robust private linear regression treats non-private and scrambled private data similarly in the model learning. An interesting next step for further improving the accuracy on very small private data would be to give a different weight to the clean and privacy-scrambled data by incorporating knowledge of the injected noise in the Bayesian inference, as has been proposed for generative models [25], but which is non-trivial in regression.

4 Methods

4.1 Linear regression model

The Bayesian linear regression model for scalar target y_i, with d-dimensional input x_i and fixed noise precision λ, is defined by

$$\begin{aligned} y_i|x_i &\sim N(x_i^T\beta, \lambda) \\ \beta &\sim N(0, \lambda_0 I), \end{aligned} \tag{2}$$

where β is the unknown parameter to be learnt. The λ and λ_0 are the precision parameters of the corresponding Gaussian distributions, and act as regularisers.

Given an observed data set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ with sufficient statistics $n\overline{xx} = \sum_{i=1}^n x_i x_i^T$ and $n\overline{xy} = \sum_{i=1}^n x_i y_i$, the posterior distribution of β is Gaussian,

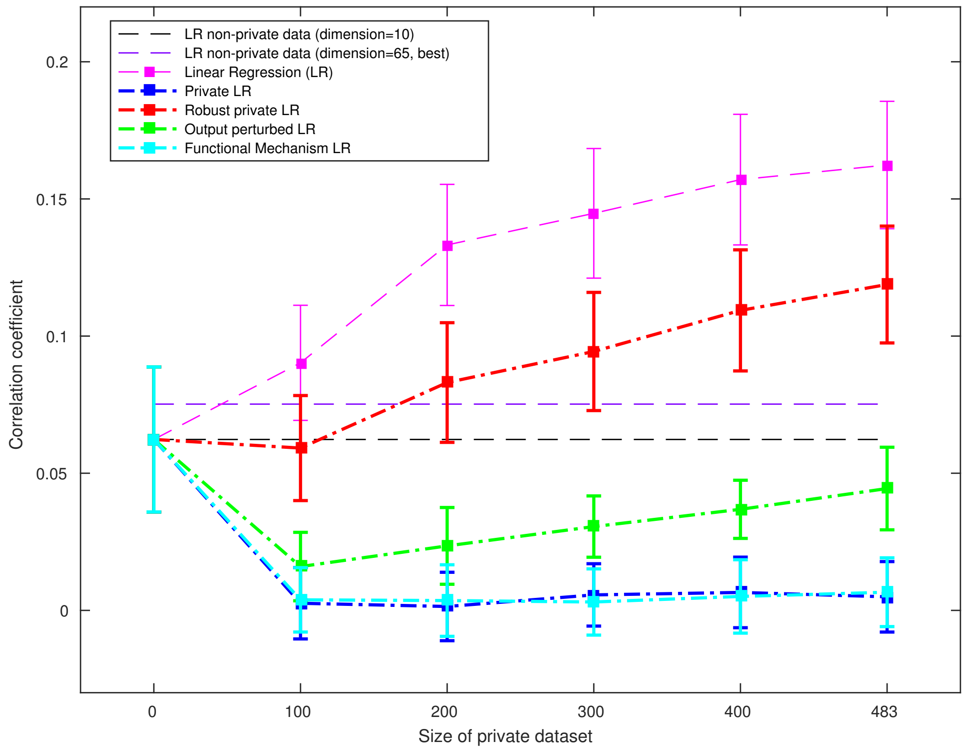

Figure 2: **Accuracy of drug sensitivity prediction in terms of Spearman's rank correlation coefficient over ranking cell lines by sensitivity to a drug (higher is better) increases with size of private data for the proposed robust private linear regression.** The state-of-the-art methods fail to utilise private data under strict privacy conditions. The baselines (horizontal dashed lines) are learned on 10 non-private data points; the private algorithms additionally have privacy-protected data (x-axis). The non-private algorithm (LR) has the same amount of additional non-privacy-protected data. All methods use 10-dimensional data except purple baseline showing the best performance with 10 non-private data points. Private methods use $\epsilon = 2$, corresponding results for $\epsilon = 1$ are in Fig. 5. The results are averaged over all drugs and 50-fold Monte Carlo cross-validation; error bars denote standard deviation over 50 Monte Carlo repeats. (See Methods for details.)

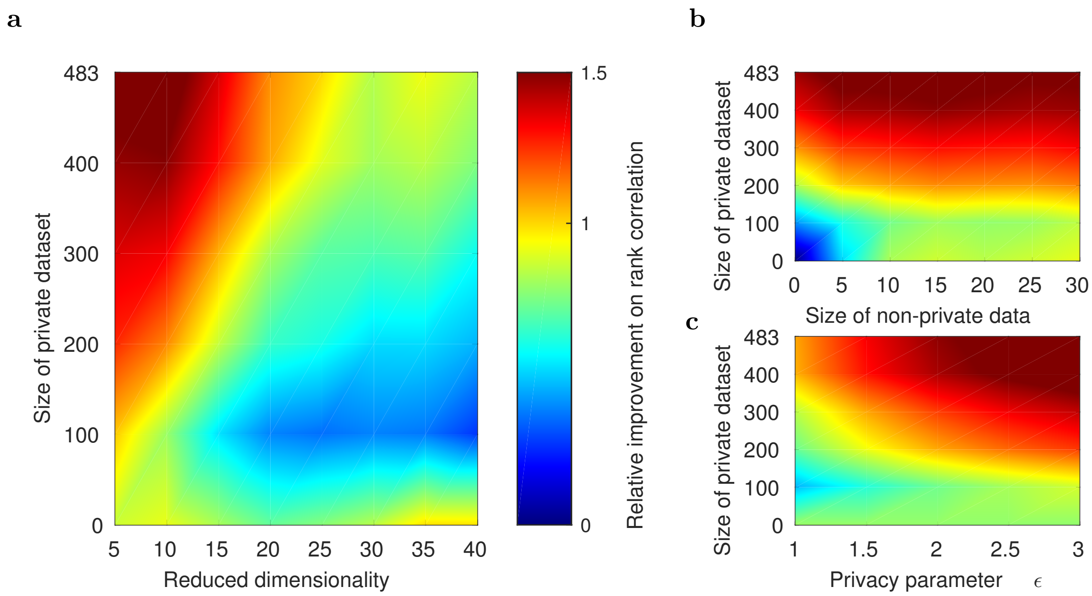

Figure 3: **Key trade-offs in differentially private learning.** Relative improvements over baseline (10 non-private data points). **a**, As the dimensionality increases, the models without private data improve whereas more data are needed to improve performance of the private methods. **b**, With enough private data, adding more non-private data does not significantly increase the performance. **c**, More data are needed if privacy guarantees are tighter (ϵ is smaller). Size of non-private data is 10 and $\epsilon = 2$ (except when otherwise noted).

$p(\beta|\mathcal{D}) = N(\beta;\ \mu_*, \Lambda_*)$, with precision

$$\Lambda_* = \lambda_0 I + \lambda n\overline{xx} \tag{3}$$

and mean

$$\mu_* = \Lambda_*^{-1}(\lambda n\overline{xy}) \tag{4}$$

After learning with the training data set, the prediction of y_i using x_i is computed as follows:

$$\hat{y}_i = x_i^T \mu_*. \tag{5}$$

For evaluation we keep a part of the data set $\mathcal{D}$ aside (not used for training) and after predicting $\hat{y}_i$, we evaluate the error between the actual y_i and $\hat{y}_i$. In this paper, we do this using Spearman's rank correlation coefficient to evaluate how well the predictions separate sensitive and insensitive cell lines.

4.2 Differential privacy and efficiency

We apply differential privacy as defined in Eq. (1). We use *bounded differential privacy*, where two data sets are considered neighbouring if they contain the same number of elements n with $n-1$ equal elements. Compared to the other common alternative of *unbounded differential privacy*, in which two data sets are considered neighbouring if one is obtained from the other by adding or removing an element, bounded differential privacy makes it clear that the number of samples is not private which simplifies parameter tuning. The privacy parameter ϵ values are not directly comparable between the two formalisms, although an $\epsilon = k$ unbounded differentially private mechanism is always a $\epsilon = 2k$ bounded differentially private mechanism.

We define a private parameter estimation mechanism to be *asymptotically consistently private*, if the private estimate converges in probability to the corresponding non-private estimate as the number of samples increases. We show that the optimal rate of convergence of the private estimate to the corresponding non-private Bayesian estimate is $\mathcal{O}(1/n)$. Mechanisms reaching this convergence rate are called *asymptotically efficiently private*. A mechanism for estimating a model is called asymptotically consistently private with respect to a utility function if the utility of the private model converges in probability to the utility of the corresponding non-private model. For full detail of these definitions see Supplementary Information sections 1.1-1.2.

4.3 Robust private linear regression

The robust private linear regression is based on perturbing the sufficient statistics $n\overline{xx} = \sum_{i=1}^n x_i x_i^T$ and $n\overline{xy} = \sum_{i=1}^n x_i y_i$. We use independent $\frac{\epsilon}{2}$-differentially private mechanisms for perturbing both statistics. Together, they provide an ϵ-differentially private mechanism. The differentially private mechanism is based on a combination of the Wishart mechanism [16] and Laplace mechanism [7].

We project the outliers in the private data sets to fit the data in the interval $[-B_*, B_*]$ as

$$\begin{aligned} x_{ij} &= \max(-B_x, \min(x_{ij}, B_x)) \\ y_i &= \max(-B_y, \min(y_i, B_y)). \end{aligned} \tag{6}$$

After the projection, $\|x_i\|_\infty \leq B_x$ and $|y_i| \leq B_y$, and we add noise to $n\overline{xx}$ distributed as Wishart$(\frac{2dB_x^2}{\epsilon}\mathbf{I}_d, d+1)$ and to $n\overline{xy} = \sum_{i=1}^n x_i y_i$ distributed as Laplace$(0, b)$, where the scale parameter $b = \frac{4dB_xB_y}{\epsilon}$). This generalises earlier work on bounded variables [9] to the unbounded case by introducing the projection. Proof that this yields a valid asymptotically consistent and efficient differentially private mechanism is given in Supplementary Information section 2. We also show that a similar algorithm, applied to the estimation of a Gaussian mean, leads to an asymptotically consistent and efficient private estimate of the posterior mean, while the simpler input perturbation that perturbs the entire data set is not asymptotically consistently private.

The projection thresholds B_x and B_y are important parameters for good model performance. As illustrated in Fig. 4, they depend strongly on the size of the data set. We propose finding the optimal parameter values on an auxiliary synthetic data set of the same size, which was found to be effective in our case. We generate the auxiliary data set of n samples using a generative model similar to the one specified in Eq. 2:

$$\begin{aligned} x_i &\sim N(0, I_d) \\ y_i|x_i &\sim N(x_i^T\beta, \lambda) \\ \beta &\sim N(0, \lambda_0 I), \end{aligned} \tag{7}$$

where d is the dimension. We parameterise the projection thresholds as a function of the data standard deviation as

$$B_x = \omega_x\sigma_x, \quad B_y = \omega_y\sigma_y \tag{8}$$

$$\omega_x, \omega_y \in \{0.1\omega\}_{\omega=1}^{20}, \tag{9}$$

where the σ_x and σ_y are the standard deviations of x (considering all dimensions) and y, respectively. With all 400 pairs of (B_x, B_y) as specified above, we apply the outlier projection method of Eq. 6. We fit the model using the projected values and then compute the error with respect to the original values. The pair of (ω_x, ω_y) which gives the minimum error is used to define the (B_x, B_y) for the real data as in Eq. 8. In our case, $(\omega_x, \omega_y) = (0.3, 0.4)$. As the error we used Spearman's rank correlation between original $y_{1:n}$ and predicted $\bar{y}_{1:n}$ based on the model learnt with projected values.

4.4 Data and pre-processing

We used the gene expression and drug sensitivity data from the *Genomics of Drug Sensitivity in Cancer* (GDSC) project [27, 11] (release 5.0, June 2014, `http:`

//www.cancerrxgene.org) consisting of 124 drugs and a panel of 613 human cancer cell lines. Dimensionality of the gene expression data was reduced from $d = 13321$ down to 65 based on prior knowledge about genes that are frequently mutated in cancer, provided by the GDSC project at http://www.cancerrxgene.org/translation/Gene. We further ordered the genes based on their mutation counts as reported at http://cancer.sanger.ac.uk/cosmic/curation. Drug responses were quantified by log-transformed IC50 values (the drug concentration yielding 50% response) from the dose response data measured at 9 different concentrations. Each data point was normalised to have L2-norm $\|x_i\|_2 = 1$, which focuses the analysis on relative expression of the selected genes, and equalises the contribution of each data point. The mean was removed from drug sensitivities, $y_i := y_i - \text{mean}(y_{1:n})$. Data with missing drug responses were ignored, making the number of cell lines different across different drugs.

4.5 Experimental setup

We carried out a 50-fold Monte Carlo cross-validation process for different splits of the data set into train and test using different random seeds. For each repeat, we randomly split the 613 cell lines to 100 for testing and the rest for the training. We further randomly partitioned the training set to 30 non-private cell lines and used the rest as the private data set. In the experiments, we tested non-private data sizes from 0 to 30, and private data sizes from 100 to 483. The parameters λ, λ_0 in Eq. 2 are set to standard Gaussian distribution values $\lambda = \lambda_0 = 1$.

4.6 Alternative methods used in comparisons

We compared five models: (i) linear regression (LR) as defined in Eq. 2, (ii) robust private LR is the proposed method, and (iii) private LR is the proposed method without projection of the outliers, (iv) output perturbed LR [26], and (v) functional mechanism LR [28]. Output perturbed LR learns parameters β using the same LR model in Eq. 2, but instead of statistics the parameters are perturbed, in a data-independent manner. Our implementation of output perturbed LR makes use of minConf optimisation package [21]. For functional mechanism LR we used the code publicly available at https://sourceforge.net/projects/functionalmecha/.

4.7 Alternative interpretation: transformed linear regression

The outlier projection mechanism can also be interpreted to produce a transformed linear regression problem,

$$\phi_y(y_i)|x_i \sim N(\phi_x(x_i)^T \beta, \lambda), \tag{10}$$

where the functions $\phi_y()$ and $\phi_x()$ implementing the outlier projection can be defined as

$$\phi_y(y_i) = \max(-B_y, \min(B_y, y_i)) \tag{11}$$
$$\phi_x(x_i) = \max(-B_x, \min(B_x, x_i)). \tag{12}$$

The normalisation of data can also be included as a transformation. This interpretation makes explicit the flexibility in designing the transformations: the differential privacy guarantees will remain valid as long as the transformations obey the bounds

$$\phi_y(y_i) \in [-B_y, B_y], \quad \phi_x(x_i) \in [-B_x, B_x]. \tag{13}$$

Acknowledgements

We would like to thank Muhammad Ammad-ud-din for assistance in data processing and Otte Heinävaara for assistance in the theoretical analysis. This work was funded by the Academy of Finland (Centre of Excellence COIN; and grants 283193 (S.K. and M.D), 294238 and 292334 (S.K.), 278300 (A.H. and O.D.), 259440 and 283107 (A.H.)).

References

[1] R. Bayardo and R. Agrawal. Data privacy through optimal k-anonymization. In *Proc. 21st Int. Conf. Data Eng. (ICDE 2005)*, 2005.

[2] A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: the SuLQ framework. In *Proc. ACM Symp. Principles of Database Systems (PODS 2005)*, 2005.

[3] K. Chaudhuri and C. Monteleoni. Privacy-preserving logistic regression. In *Adv. Neural Inf. Process. Syst. 21*, 2008.

[4] J. C. Costello, L. M. Heiser, E. Georgii, M. Gönen, M. P. Menden, N. J. Wang, M. Bansal, M. Ammad-ud din, P. Hintsanen, S. A. Khan, J.-P. Mpindi, O. Kallioniemi, A. Honkela, T. Aittokallio, K. Wennerberg, NCI DREAM Community, J. J. Collins, D. Gallahan, D. Singer, J. Saez-Rodriguez, S. Kaski, J. W. Gray, and G. Stolovitzky. A community effort to assess and improve drug sensitivity prediction algorithms. *Nat. Biotechnol.*, 32(12):1202–1212, Dec 2014.

[5] P. Diaconis and D. Ylvisaker. Conjugate priors for exponential families. *Ann. Stat.*, 7(2):269–281, Mar 1979.

[6] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Privacy aware learning. *J. ACM*, 61(6):1–57, Dec 2014.

[7] C. Dwork. Differential privacy. In *Proc. 33rd Int. Colloq. on Automata, Languages and Prog. (ICALP 2006), Part II*, pages 1–12, 2006.

[8] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, Aug. 2014.

[9] J. Foulds, J. Geumlek, M. Welling, and K. Chaudhuri. On the theory and practice of privacy-preserving Bayesian data analysis. In *Proc. 32nd Conf. on Uncertainty in Artificial Intelligence (UAI 2016)*, Mar. 2016. arXiv:1603.07294 [cs.LG].

[10] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *Proc. 23rd USENIX Security Symp. (USENIX Security 2014)*, pages 17–32, 2014.

[11] M. J. Garnett, E. J. Edelman, S. J. Heidorn, C. D. Greenman, A. Dastur, K. W. Lau, P. Greninger, I. R. Thompson, X. Luo, J. Soares, Q. Liu, F. Iorio, D. Surdez, L. Chen, R. J. Milano, G. R. Bignell, A. T. Tam, H. Davies, J. A. Stevenson, S. Barthorpe, S. R. Lutz, F. Kogera, K. Lawrence, A. McLaren-Douglas, X. Mitropoulos, T. Mironenko, H. Thi, L. Richardson, W. Zhou, F. Jewitt, T. Zhang, P. O'Brien, J. L. Boisvert, S. Price, W. Hur, W. Yang, X. Deng, A. Butler, H. G. Choi, J. W. Chang, J. Baselga, I. Stamenkovic, J. A. Engelman, S. V. Sharma, O. Delattre, J. Saez-Rodriguez, N. S. Gray, J. Settleman, P. A. Futreal, D. A. Haber, M. R. Stratton, S. Ramaswamy, U. McDermott, and C. H. Benes. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, 483(7391):570–575, Mar 2012.

[12] C. Gentry. *A fully homomorphic encryption scheme*. PhD thesis, Stanford University, 2009.

[13] M. Gymrek, A. L. McGuire, D. Golan, E. Halperin, and Y. Erlich. Identifying personal genomes by surname inference. *Science*, 339(6117):321–324, Jan 2013.

[14] A. Harmanci and M. Gerstein. Quantification of private information leakage from phenotype-genotype data: linking attacks. *Nat. Methods*, 13(3):251–256, Mar 2016.

[15] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet.*, 4(8):e1000167, Aug 2008.

[16] W. Jiang, C. Xie, and Z. Zhang. Wishart mechanism for differentially private principal components analysis. In *Proc. Conf. AAAI Artif. Intell. 2016*, 2016. arXiv:1511.05680 [cs.CR].

[17] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Proc. 23rd Int. Conf. Data Eng. (ICDE 2007)*, 2007.

[18] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. *TKDD*, 1(1):3, Mar 2007.

[19] R. J. Muirhead. *Aspects of multivariate statistical theory.* Wiley, 1982.

[20] M. Naveed, E. Ayday, E. W. Clayton, J. Fellay, C. A. Gunter, J.-P. Hubaux, B. A. Malin, and X. Wang. Privacy in the genomic era. *ACM Comput. Surv.*, 48(1):1–44, Aug 2015.

[21] M. Schmidt, E. van den Berg, M. Friedlander, and K. Murphy. Optimizing costly functions with simple constraints: A limited-memory projected quasi-newton algorithm. In *Proc. 12th Int. Conf. on Artif. Intell. and Stat. (AISTATS 2009)*, 2009.

[22] A. Smith. Efficient, differentially private point estimators. Sept. 2008. arXiv:0809.4794 [cs.CR].

[23] Y.-X. Wang, J. Lei, and S. E. Fienberg. Learning with differential privacy: Stability, learnability and the sufficiency and necessity of ERM principle. Feb. 2015. arXiv: 1502.06309 [stat.ML].

[24] L. Wasserman and S. Zhou. A statistical framework for differential privacy. *J. Am. Stat. Assoc.*, 105(489):375–389, Mar 2010.

[25] O. Williams and F. McSherry. Probabilistic inference and differential privacy. In *Adv. Neural Inf. Process. Syst. 23*, 2010.

[26] X. Wu, M. Fredrikson, W. Wu, S. Jha, and J. F. Naughton. Revisiting differentially private regression: Lessons from learning theory and their consequences. Dec. 2015. arXiv:1512.06388 [cs.CR].

[27] W. Yang, J. Soares, P. Greninger, E. J. Edelman, H. Lightfoot, S. Forbes, N. Bindal, D. Beare, J. A. Smith, I. R. Thompson, S. Ramaswamy, P. A. Futreal, D. A. Haber, M. R. Stratton, C. Benes, U. McDermott, and M. J. Garnett. Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.*, 41(Database issue):D955–D961, Jan 2013.

[28] J. Zhang, Z. Zhang, X. Xiao, Y. Yang, and M. Winslett. Functional mechanism: Regression analysis under differential privacy. *PVLDB*, 5(11):1364–1375, 2012.

[29] Z. Zhang, B. Rubinstein, and C. Dimitrakakis. On the differential privacy of Bayesian inference. In *Proc. Conf. AAAI Artif. Intell. 2016*, 2016. arXiv:1512.06992 [cs.AI].

Supplementary Information

A Theoretical background

We argue that effective differentially private predictive modelling methods can be developed by a combination of:

i. An asymptotically efficiently private mechanism for which the effect of the noise added to guarantee privacy vanishes as the number of samples increases; and

ii. A way to limit the amount of private information to be shared. This yields better performance on finite data as less noise needs to be added for equivalent privacy. This can be achieved through a combination of two things:

   a. An approach to decrease the dimensionality of the data prior to the application of the private algorithm; and
   b. A method to focus the privacy guarantees to relevant variation in data.

Criterion i can be formally stated through additional loss in accuracy or utility of the estimates because of privacy. Our main asymptotic result is that the optimal convergence rate of a differentially private mechanism to a Bayesian estimate is $\mathcal{O}(1/n)$, which can be reached by our proposed mechanism.

Criterion ii is non-asymptotic and thus more difficult to address theoretically. It manifests itself in the constants in the convergence rates as well as empirical findings on the effect of dimensionality reduction and projecting outliers to tighter bounds as discussed in the main text and in Fig. 4.

A.1 Definition of asymptotic efficiency

We begin by formalisation of the theory behind Criterion i.

Definition 1. *A differentially private mechanism $\mathcal{M}$ is* asymptotically consistent with respect to an estimated parameter θ *if the private estimates $\hat{\theta}_{\mathcal{M}}$ given a data set $\mathcal{D}$ converge in probability to the corresponding non-private estimates $\hat{\theta}_{NP}$ as the number of samples, $n = |\mathcal{D}|$, grows without bound, i.e., if for any*[1] *$\alpha > 0$,*

$$\lim_{n\to\infty} \Pr\{\|\hat{\theta}_{\mathcal{M}} - \hat{\theta}_{NP}\| > \alpha\} = 0.$$

Definition 2. *A differentially private mechanism $\mathcal{M}$ is* asymptotically efficiently private with respect to an estimated parameter θ*, if the mechanism is asymptotically consistent and the private estimates $\hat{\theta}_{\mathcal{M}}$ converge to the corresponding non-private estimates $\hat{\theta}_{NP}$ at the rate $\mathcal{O}(1/n)$, i.e., if for any $\alpha > 0$ there exist constants C, N such that*

$$\Pr\{\|\hat{\theta}_{\mathcal{M}} - \hat{\theta}_{NP}\| > C/n\} < \alpha$$

[1] We use α in limit expressions instead of usual ϵ to avoid confusion with ϵ-differential privacy.

for all $n \geq N$*.*

The term asymptotically efficiently private in the above definition is justified by the following theorem, which shows that the rate $\mathcal{O}(1/n)$ is optimal for estimating expectation parameters of exponential family distributions. As it seems unlikely that better rates could be obtained for more difficult problems, we conjecture that this rate cannot be beaten for Bayesian estimates in general.

Theorem 1. *The private estimates* $\hat{\theta}_{\mathcal{M}}$ *of an exponential family posterior expectation parameter* θ*, generated by a differentially private mechanism* $\mathcal{M}$ *that achieves* ϵ*-differential privacy for any* $\epsilon > 0$*, cannot converge to the corresponding non-private estimates* $\hat{\theta}_{NP}$ *at a rate faster than* $1/n$*. This is, assuming* $\mathcal{M}$ *is* ϵ*-differentially private, there exists no function* $f(n)$ *such that* $\limsup n f(n) = 0$ *and for all* $\alpha > 0$*, there exists a constant* N *such that*

$$\Pr\{\|\hat{\theta}_{\mathcal{M}} - \hat{\theta}_{NP}\| > f(n)\} < \alpha$$

for all $n \geq N$*.*

Proof. The non-private estimate of an expectation parameter of an exponential family is [5]

$$\hat{\theta}_{NP}|x_1, \ldots, x_n = \frac{n_0 x_0 + \sum_{i=1}^{n} x_i}{n_0 + n}. \tag{14}$$

The difference of the estimates from two neighbouring data sets differing by one element is

$$(\hat{\theta}_{NP}|\mathcal{D}) - (\hat{\theta}_{NP}|\mathcal{D}') = \frac{x - y}{n_0 + n}, \tag{15}$$

where x and y are the corresponding mismatched elements. Let $\Delta = \max(\|x - y\|)$, and let $\mathcal{D}$ and $\mathcal{D}'$ be neighbouring data sets including these maximally different elements.

Let us assume that there exists a function $f(n)$ such that $\limsup n f(n) = 0$ and for all $\alpha > 0$ there exists a constant N such that

$$\Pr\{\|\hat{\theta}_{\mathcal{M}} - \hat{\theta}_{NP}\| > f(n)\} < \alpha$$

for all $n \geq N$.

Fix $\alpha > 0$ and choose $M \geq \max(N, n_0)$ such that $f(n) \leq \Delta/4n$ for all $n \geq M$. This implies that

$$\|(\hat{\theta}_{NP}|\mathcal{D}) - (\hat{\theta}_{NP}|\mathcal{D}')\| = \frac{\Delta}{n_0 + n} \geq \frac{\Delta}{2n} \geq 2f(n). \tag{16}$$

Let us define the region $C_{\mathcal{D}} = \{t \mid \|(\hat{\theta}_{NP}|\mathcal{D}) - t\| < f(n)\}$. Based on our assumptions we have

$$\Pr(\hat{\theta}_{\mathcal{M}}|\mathcal{D} \in C_{\mathcal{D}}) > 1 - \alpha \tag{17}$$

$$\Pr(\hat{\theta}_{\mathcal{M}}|\mathcal{D}' \in C_{\mathcal{D}}) < \alpha \tag{18}$$

which implies that

$$\frac{\Pr(\hat{\theta}_{\mathcal{M}}|\mathcal{D} \in C_{\mathcal{D}})}{\Pr(\hat{\theta}_{\mathcal{M}}|\mathcal{D}' \in C_{\mathcal{D}})} > \frac{1-\alpha}{\alpha} \tag{19}$$

which means that $\mathcal{M}$ cannot be differentially private with $\epsilon < \log((1-\alpha)/\alpha) \to \infty$ as $\alpha \to 0$. □

A.2 Different utility functions

Definition 3. *Let $\mathcal{U}(\hat{\theta}_{NP}(\mathcal{D}))$ measure the utility of the non-private model $\hat{\theta}_{NP}$ estimated from data set $\mathcal{D}$ and let $\mathcal{U}(\hat{\theta}_{\mathcal{M}}(\mathcal{D}))$ measure the corresponding utility of the private model $\hat{\theta}_{\mathcal{M}}$ obtained using differentially private mechanism $\mathcal{M}$. The mechanism $\mathcal{M}$ is* asymptotically consistent with respect to a bounded utility $\mathcal{U}$ *if the random variables $\mathcal{U}(\hat{\theta}_{\mathcal{M}}(\mathcal{D}))$ converge in probability to $\mathcal{U}(\hat{\theta}_{NP}(\mathcal{D}))$ as the number of samples, $n = |\mathcal{D}|$, grows without bound, i.e., if for any $\alpha > 0$,*

$$\lim_{n\to\infty} \Pr\{|\mathcal{U}(\hat{\theta}_{\mathcal{M}}(\mathcal{D})) - \mathcal{U}(\hat{\theta}_{NP}(\mathcal{D}))| > \alpha\} = 0.$$

Theorem 2. *A differentially private mechanism $\mathcal{M}$ that is asymptotically consistent with respect to a set of parameters is asymptotically consistent with respect to any continuous utility that only depends on those parameters.*

Proof. If $\hat{\theta}_{\mathcal{M}}$ converges in probability to $\hat{\theta}_{NP}$ then by the continuous mapping theorem the value of $\mathcal{U}(\hat{\theta}_{\mathcal{M}})$ will converge in probability to $\mathcal{U}(\hat{\theta}_{NP})$. □

A.3 Example: Gaussian mean

Theorem 3. *Differentially private inference of the mean of a Gaussian variable, with Laplace mechanism to perturb the sufficient statistics, is asymptotically consistent with respect to the posterior mean.*

Proof. Let us consider the model

$$x_i \sim N(\mu, \Lambda)$$
$$\mu \sim N(\mu_0, \Lambda_0)$$

with μ as the unknown parameter and Λ and Λ_0 denoting the fixed prior precision matrices of the noise and the mean, respectively. We assume $||x_i||_1 \leq B$ and enforce this by projecting the larger elements to satisfy this bound.

Let the observed data set be $\mathcal{D} = \{x_i\}_{i=1}^n$ with sufficient statistic $n\bar{x} = \sum_{i=1}^n x_i$.

The non-private posterior mean is

$$\mu_{NP} = (\Lambda_0 + n\Lambda)^{-1}(\Lambda n\bar{x} + \Lambda_0\mu_0).$$

The corresponding private posterior mean is obtained by replacing $n\bar{x}$ with the perturbed version $n\bar{x}' = n\bar{x} + \delta$, where $\delta = (\delta_1, \ldots, \delta_d)^T \in \mathbb{R}^d$ with $\delta_j \sim \text{Laplace}(0, \frac{2Bd}{\epsilon})$ and $d = \dim(x_i)$, yielding

$$\mu_{DP} = (\Lambda_0 + n\Lambda)^{-1}(\Lambda(n\bar{x} + \delta) + \Lambda_0\mu_0).$$

The difference of the private and non-private means is

$$\begin{aligned}\|\mu_{DP} - \mu_{NP}\|_1 &= \|(\Lambda_0 + n\Lambda)^{-1}(\Lambda\delta)\|_1 \\ &= \|(\Lambda^{-1}\Lambda_0 + n \cdot I)^{-1}\delta\|_1 \le \frac{c}{n}\|\delta\|_1,\end{aligned}$$

which is valid for all $c > 1$ for large enough n. This implies that

$$\Pr\{\|\mu_{DP} - \mu_{NP}\|_1 \ge \alpha\} \le \Pr\left\{\frac{c}{n}\|\delta\|_1 \ge \alpha\right\} \to 0$$

as $n \to \infty$ for all $\alpha > 0$. □

Theorem 4. *Differentially private inference of the mean of a Gaussian variable with Laplace mechanism to perturb the input data set (naive input perturbation) is* not *asymptotically consistent with respect to the posterior mean.*

Proof. The mechanism is almost the same as in Theorem 3, but we now have $n\bar{x}' = n\bar{x} + \sum_{i=1}^{n}\delta_i)$ where $\delta_i = (\delta_{i1}, \ldots, \delta_{id})^T \in \mathbb{R}^d$ with $\delta_{ij} \sim \text{Laplace}(0, \frac{2Bd}{\epsilon})$. Similar computation as above yields

$$\begin{aligned}\|\mu_{DP} - \mu_{NP}\|_1 &= \left\|(\Lambda_0 + n\Lambda)^{-1}(\Lambda\sum_{i=1}^{n}\delta_i)\right\|_1 \\ &= \left\|(\frac{1}{n}\Lambda^{-1}\Lambda_0 + I)^{-1}\frac{1}{n}\sum_{i=1}^{n}\delta_i\right\|_1 \ge \frac{1}{2}\left\|\frac{1}{n}\sum_{i=1}^{n}\delta_i\right\|_1\end{aligned}$$

for sufficiently large n. By the central limit theorem the distribution of $\frac{1}{n}\sum_{i=1}^{n}\delta_i$ converges to a Gaussian with non-zero variance. Hence μ_{DP} does not converge to μ_{NP} for large n and the method is not asymptotically consistent. □

A.3.1 Asymptotic efficiency

Theorem 5. *ϵ-differentially private estimate of the mean of a d-dimensional Gaussian variable x bounded by $\|x_i\|_1 \le B$ in which the Laplace mechanism is used to perturb the sufficient statistics, is asymptotically efficiently private.*

Proof. In the proof of Theorem 3 we showed that

$$\|\mu_{DP} - \mu_{NP}\|_1 \le \frac{c}{n}\|\delta\|_1,$$

where $\delta = (\delta_1, \ldots, \delta_d)^T \in \mathbb{R}^D$ with $\delta_j \sim \text{Laplace}\left(0, \frac{2Bd}{\epsilon}\right)$.

Because δ_j is Laplace, $|\delta_j|$ is exponential with

$$|\delta_j| \sim \text{Exponential}\left(\frac{\epsilon}{2Bd}\right)$$

and

$$\|\delta\|_1 = \sum_{j=1}^{d}|\delta_j| \sim \text{Gamma}\left(d, \frac{\epsilon}{2Bd}\right).$$

Given $\alpha > 0$ we can choose $C > cF^{-1}(1-\alpha; d, \epsilon/2Bd)$, where $F^{-1}(x; a, b)$ is the inverse cumulative distribution function of the Gamma distribution with shape a and rate b, to ensure that

$$\Pr\left\{\|\mu_{DP} - \mu_{NP}\|_1 > \frac{C}{n}\right\} \leq \Pr\left\{\frac{1}{n}\|\delta\|_1 > \frac{C}{n}\right\} = \Pr\{\|\delta\|_1 > C\} < \alpha. \tag{20}$$

□

A.3.2 Convergence rate

We can further study the probability of making an error of at least a given magnitude as

$$\begin{aligned}\Pr\{\|\mu_{DP} - \mu_{NP}\|_1 \geq \phi\} &\leq \Pr\left\{\frac{c}{n}\|\delta\|_1 \geq \phi\right\} \\ &= \Pr\left\{\mathrm{Gamma}\left(d, \frac{n\epsilon}{2Bcd}\right) \geq \phi\right\} \\ &= 1 - F\left(\phi; d, \frac{n\epsilon}{2Bcd}\right) = 1 - \frac{\gamma(d, \frac{n\phi\epsilon}{2Bcd})}{\Gamma(d)},\end{aligned} \tag{21}$$

where $F(x; a, b)$ is the cumulative distribution function of the Gamma distribution with shape a and rate b.

The formula in Eq. (21) unfortunately has no simple closed form expression. The result shows, however, that the n required to reach a certain level of performance is linear in B and $\frac{1}{\epsilon}$. The dependence on d is complicated, but it is in general super-linear as suggested by the mean of the gamma distribution in Eq. (21), $\frac{2Bd^2}{n\epsilon}$.

A.4 Example: Zhang et al., AAAI 2016, (arxiv:1512.06992)

In their paper Zhang et al. derive utility bounds for a number of mechanisms. The bounds are clearly insufficient to demonstrate the asymptotic efficiency of the corresponding methods. For Laplace mechanism applied to Bayesian network inference, their bound on excess KL-divergence as a function of the data set size n is

$$\mathcal{O}(mn \ln n)\left[1 - \exp\left(-\frac{n\epsilon}{2|\mathcal{I}|}\right)\right] + \sqrt{-\mathcal{O}(mn \ln n) \ln \delta}.$$

B Differentially private linear regression

Let us next consider the linear regression model with fixed noise Λ,

$$\begin{aligned} y_i | x_i &\sim N(x_i^T \beta, \Lambda) \\ \beta &\sim N(\beta_0, \Lambda_0), \end{aligned}$$

with β as the unknown parameter and Λ and Λ_0 denoting the precision matrices of the corresponding distributions.

Algorithm 1 Differentially private statistics release

function DiffPriSS(X, Y, ϵ, B_x, B_y)
  $n = |Y|$, $d = \dim(X)$
  $(C, D) = \textsc{project}(X, Y, B_x, B_y)$
  $P \sim \text{Wishart}(\frac{2dB_x^2}{\epsilon}\mathbf{I}_d, d+1)$
  for $i \in \mathbb{I}$ **do**
    $Q_i \sim \text{Laplace}(0, \frac{4dB_xB_y}{\epsilon})$
  end for
  $S_{xx} = CC' + P$
  $S_{xy} = CD + Q$
end function
function Project(X, Y, B_x, B_y)
  for $j = 1$ to n **do**
    for $i = 1$ to d **do**
      $C_{ij} = \max(-B_x, \min(B_x, X_{ij}))$
    end for
    $D_j = \max(-B_y, \min(B_y, Y_j))$
  end for
end function

Let the observed data set be $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ with sufficient statistics $n\overline{xx} = \sum_{i=1}^n x_i x_i^T$ and $n\overline{xy} = \sum_{i=1}^n x_i y_i$.

The non-private posterior precision of β is

$$\Lambda_{NP} = \Lambda_0 + \Lambda n\overline{xx}$$

and the corresponding posterior mean is

$$\mu_{NP} = \Lambda_{NP}^{-1}(\Lambda n\overline{xy} + \Lambda_0\beta_0). \tag{22}$$

The corresponding private posterior precision is obtained by replacing $n\overline{xx}$ with the perturbed version $n\overline{xx}' = n\overline{xx} + \Delta$, where Δ follows either Laplace or Wishart distribution according to the chosen mechanism, yielding

$$\Lambda_{DP} = \Lambda_0 + \Lambda(n\overline{xx} + \Delta).$$

Similarly using $n\overline{xy}' = n\overline{xy} + \delta$ with δ following the Laplace mechanism we obtain

$$\mu_{DP} = \Lambda_{DP}^{-1}(\Lambda(n\overline{xy} + \delta) + \Lambda_0\beta_0). \tag{23}$$

The mechanism is presented in detail in Algorithm 1.

B.1 The detailed mechanism

The function project in Algorithm 1 projects the data points into a useful space and computes the sufficient statistics.

Theorem 6. *Algorithm* DiffPriSS *in Algorithm 1 is ϵ-differentially private.*

Proof. (i) $S_{xx} = CC' + P$ is $\frac{\epsilon}{2}$-differentially private.

This follows from the proof of Jiang *et al.* [16] under slight modifications. After PROJECT we have

$$\|C\|_2^2 \le dB_x^2. \tag{24}$$

Following [16], we assume two data sets that are identical except replacing one vector so that projections of the original and modified data vectors are u and v. We denote the difference matrix of the two versions of CC' by $\Delta = vv^T - uu^T$. As [16] we have

$$\begin{aligned} \frac{p(g(D_1) = r)}{p(g(D_2) = r)} &= \exp\left[\frac{1}{2}\,\mathrm{trace}\left(\left(\frac{2dB_x^2}{\epsilon}\mathbf{I}_d\right)^{-1}\Delta\right)\right] \\ &\le \exp\left[\frac{1}{2}\left\|\left(\frac{2dB_x^2}{\epsilon}\mathbf{I}_d\right)^{-1}\right\|_2 \|\Delta\|_*\right], \end{aligned} \tag{25}$$

where $\|\Delta\|_* = \sum_i \sigma_i(\Delta)$ is the nuclear norm of Δ and $\sigma_i(\Delta)$ is the ith largest singular value of Δ.

We can obtain a stronger bound than [16] on the nuclear norm. As $\mathrm{rank}(\Delta) \le 2$,

$$\begin{aligned} \|\Delta\|_* \le \sqrt{2}\|\Delta\|_F &= \sqrt{2\,\mathrm{trace}(\Delta^T\Delta)} = \sqrt{2(\|u\|_2^4 + \|v\|_2^4 - 2(u^Tv)^2)} \\ &\le \sqrt{2(\|u\|_2^4 + \|v\|_2^4)} = \sqrt{4(dB_x^2)^2} = 2dB_x^2. \end{aligned} \tag{26}$$

Combining Eqs. (25) and (26) yields

$$\frac{p(g(D_1) = r)}{p(g(D_2) = r)} \le \exp\left[\frac{\epsilon}{4dB_x^2}\|\mathbf{I}_d^{-1}\|_2\|\Delta\|_*\right] \le \exp\left[\frac{2dB_x^2\epsilon}{4dB_x^2}\right] = \exp\left[\frac{\epsilon}{2}\right], \tag{27}$$

showing that $S_{xx} = CC' + P$ is $\frac{\epsilon}{2}$-differentially private.

(ii) CD is a $d \times 1$ vector where d is the cardinality of $\mathbb{I}$ and each element of CD is computed as follows:

$$\forall i \in \mathbb{I}, \ \ CD_i = \sum_{j=1}^{n} C_{ij}D_j, \tag{28}$$

where $|C_{ij}| \le B_x$ and $|D_j| \le B_y$, and thus the sensitivity of CD is $2dB_xB_y$. Thus, $S_{xy} = CD + Q$ is $\frac{\epsilon}{2}$-differentially private.

Therefore, releasing S_{xx} and S_{xy} together by DiffPriSS is ϵ-differentially private. ☐

B.2 Asymptotic consistency and efficiency

Theorem 7. *Differentially private inference of the posterior mean of the weights of linear regression with Laplace or Wishart mechanism to perturb the sufficient statistics is asymptotically consistent with respect to the posterior mean.*

Proof. Using Eqs. (22)–(23) we can evaluate

$$\begin{aligned}
\|\mu_{DP} - \mu_{NP}\|_1 &= \left\|\Lambda_{DP}^{-1}(\Lambda(n\overline{xy} + \delta) + \Lambda_0\beta_0) - \Lambda_{NP}^{-1}(\Lambda n\overline{xy} + \Lambda_0\beta_0)\right\|_1 \\
&\leq \left\|\Lambda_{DP}^{-1}(\Lambda(n\overline{xy} + \delta) + \Lambda_0\beta_0) - \Lambda_{DP}^{-1}(\Lambda n\overline{xy} + \Lambda_0\beta_0)\right\|_1 \\
&\quad + \left\|\Lambda_{DP}^{-1}(\Lambda n\overline{xy} + \Lambda_0\beta_0) - \Lambda_{NP}^{-1}(\Lambda n\overline{xy} + \Lambda_0\beta_0)\right\|_1 \\
&= \left\|\Lambda_{DP}^{-1}\Lambda\delta\right\|_1 + \left\|(\Lambda_{DP}^{-1} - \Lambda_{NP}^{-1})(\Lambda n\overline{xy} + \Lambda_0\beta_0)\right\|_1 \\
&= \left\|(\Lambda_0 + \Lambda(n\overline{xx} + \Delta))^{-1}\Lambda\delta\right\|_1 \\
&\quad + \left\|\left[(\Lambda_0 + \Lambda(n\overline{xx} + \Delta))^{-1}\right.\right. \\
&\qquad \left.\left. - (\Lambda_0 + \Lambda(n\overline{xx}))^{-1}\right](\Lambda n\overline{xy} + \Lambda_0\beta_0)\right\|_1 \\
&= \left\|(\Lambda_0 + \Lambda(n\overline{xx} + \Delta))^{-1}\Lambda\delta\right\|_1 \\
&\quad + \left\|\left[\left(\frac{1}{n}\Lambda_0 + \Lambda\left(\overline{xx} + \frac{1}{n}\Delta\right)\right)^{-1}\right.\right. \\
&\qquad \left.\left. - \left(\frac{1}{n}\Lambda_0 + \Lambda\overline{xx}\right)^{-1}\right]\left(\Lambda\overline{xy} + \frac{1}{n}\Lambda_0\beta_0\right)\right\|_1 .
\end{aligned}$$

Assuming $\overline{xx} > 0$, the first term clearly approaches 0 as $n \to \infty$. For the second term, as $n \to \infty$, $(\frac{1}{n}\Lambda_0 + \Lambda(\overline{xx} + \frac{1}{n}\Delta))^{-1} \to (\frac{1}{n}\Lambda_0 + \Lambda\overline{xx})^{-1}$ and as $(\Lambda\overline{xy} + \frac{1}{n}\Lambda_0\beta_0)$ is bounded, the second term also approaches 0 as $n \to \infty$. This shows that μ_{DP} converges in probability to μ_{NP}. □

Theorem 8. *ϵ-differentially private inference of the posterior mean of the weights of linear regression with the Wishart mechanism of Algorithm 1 to perturb the sufficient statistics is asymptotically efficiently private.*

Proof. From the proof of Theorem 7 we have

$$\begin{aligned}
&\|\mu_{DP} - \mu_{NP}\|_1 \leq \left\|(\Lambda_0 + \Lambda(n\overline{xx} + \Delta))^{-1}\Lambda\delta\right\|_1 \\
&+ \left\|\left[\left(\frac{1}{n}\Lambda_0 + \Lambda\left(\overline{xx} + \frac{1}{n}\Delta\right)\right)^{-1} - \left(\frac{1}{n}\Lambda_0 + \Lambda\overline{xx}\right)^{-1}\right]\left(\Lambda\overline{xy} + \frac{1}{n}\Lambda_0\beta_0\right)\right\|_1 .
\end{aligned} \tag{29}$$

The first term can be bounded easily as

$$\begin{aligned}
\left\|(\Lambda_0 + \Lambda(n\overline{xx} + \Delta))^{-1}\Lambda\delta\right\|_1 &= \left\|(\Lambda^{-1}\Lambda_0 + \Delta + n\overline{xx})^{-1}\delta\right\|_1 \\
&\leq \left\|(\Lambda^{-1}\Lambda_0 + \Delta + n\overline{xx})^{-1}\right\|_1 \|\delta\|_1 \\
&\leq \frac{c_1}{n}\left\|(\overline{xx})^{-1}\right\|_1 \|\delta\|_1
\end{aligned} \tag{30}$$

where $c_1 > 1$. The bound is valid for any $c_1 > 1$ as n gets large enough.

Similarly as in the proof of Theorem 5,

$$\|\delta\|_1 \sim \text{Gamma}\left(d, \frac{\epsilon}{4dB_xB_y}\right). \tag{31}$$

Given $\alpha > 0$ we can choose similarly as in the proof of Theorem 5

$$C_1 > c_1 F^{-1}(1-\alpha/2; d, \epsilon/(4dB_xB_y)) \left\|(\overline{xx})^{-1}\right\|_1,$$

where $F^{-1}(x;\alpha,\beta)$ is the inverse distribution function of the Gamma distribution with shape α and rate β, to ensure that

$$\Pr\left\{\|(\Lambda_0 + \Lambda(n\overline{xx}+\Delta))^{-1}\Lambda\delta\|_1 > \frac{C_1}{n}\right\} < \frac{\alpha}{2}. \tag{32}$$

The second term can be bounded as

$$\begin{aligned}
&\left\|\left[\left(\frac{1}{n}\Lambda_0 + \Lambda\left(\overline{xx}+\frac{1}{n}\Delta\right)\right)^{-1} - \left(\frac{1}{n}\Lambda_0 + \Lambda\overline{xx}\right)^{-1}\right]\left(\Lambda\overline{xy}+\frac{1}{n}\Lambda_0\beta_0\right)\right\|_1 \\
&= \left\|\left[\left(\frac{1}{n}\Lambda^{-1}\Lambda_0 + \overline{xx}+\frac{1}{n}\Delta\right)^{-1} - \left(\frac{1}{n}\Lambda^{-1}\Lambda_0 + \overline{xx}\right)^{-1}\right]\left(\overline{xy}+\frac{1}{n}\Lambda^{-1}\Lambda_0\beta_0\right)\right\|_1 \\
&= \frac{1}{n}\left\|\left(\frac{1}{n}\Lambda^{-1}\Lambda_0 + \overline{xx}+\frac{1}{n}\Delta\right)^{-1}\Delta\left(\frac{1}{n}\Lambda^{-1}\Lambda_0 + \overline{xx}\right)^{-1}\left(\overline{xy}+\frac{1}{n}\Lambda^{-1}\Lambda_0\beta_0\right)\right\|_1 \\
&\le \frac{1}{n}\left\|\left(\frac{1}{n}\Lambda^{-1}\Lambda_0 + \overline{xx}+\frac{1}{n}\Delta\right)^{-1}\Delta\left(\frac{1}{n}\Lambda^{-1}\Lambda_0 + \overline{xx}\right)^{-1}\right\|_1 \left\|\overline{xy}+\frac{1}{n}\Lambda^{-1}\Lambda_0\beta_0\right\|_1 \\
&\le \frac{1}{n}\left\|\left(\frac{1}{n}\Lambda^{-1}\Lambda_0 + \overline{xx}+\frac{1}{n}\Delta\right)^{-1}\right\|_1 \|\Delta\|_1 \\
&\quad \left\|\left(\frac{1}{n}\Lambda^{-1}\Lambda_0 + \overline{xx}\right)^{-1}\right\|_1 \left\|\overline{xy}+\frac{1}{n}\Lambda^{-1}\Lambda_0\beta_0\right\|_1 \\
&\le \frac{c_2}{n}\left\|(\overline{xx})^{-1}\right\|_1 \|\Delta\|_1 \left\|(\overline{xx})^{-1}\right\|_1 \|\overline{xy}\|_1,
\end{aligned}$$

where similarly as in Eq. (30), the bound is valid for any $c_2 > 1$ as n gets large enough. Here $\|\Delta\|_1$ is the l_1-norm of the matrix Δ that follows the Wishart distribution $\Delta \sim \text{Wishart}(\frac{2dB_x^2}{\epsilon}\mathbf{I}_d, d+1)$. We can bound it as

$$\|\Delta\|_1 \le \sqrt{d}\lambda_1(\Delta),$$

where $\lambda_1(\Delta)$ is the largest eigenvalue of the matrix. The exact distribution of $\lambda_1(\Delta)$ is given in [19, Corollary 9.7.2], but as it is very complicated we do not reproduce it here. Given $\alpha > 0$ we can nevertheless use that distribution to select C_2 such that

$$\Pr\left\{\lambda_1(\Delta) > \frac{C_2}{c_2\sqrt{d}\left\|(\overline{xx})^{-1}\right\|_1^2 \|\overline{xy}\|_1}\right\} < \frac{\alpha}{2}. \tag{33}$$

Combining Eqs. (32) and (33) shows that

$$\Pr\left\{\|\mu_{DP} - \mu_{NP}\|_1 > \frac{C_1 + C_2}{n}\right\} < \alpha. \tag{34}$$

□

B.3 Convergence rate

Using Chebysev's inequality together with Eq. (31) we can show that with high probability

$$\|\delta\|_1 = \mathcal{O}\left(\frac{d^2 B_x B_y}{\epsilon}\right)$$

and thus

$$\left\|(\Lambda_0 + \Lambda(n\overline{xx} + \Delta))^{-1}\Lambda\delta\right\|_1 = \mathcal{O}\left(\frac{d^2 B_x B_y \left\|(\overline{xx})^{-1}\right\|_1}{n\epsilon}\right). \tag{35}$$

The distribution of $\lambda_1(\Delta)$ in [19, Corollary 9.7.2] is not practical. The bound of [16], on the other hand, implies that with high probability

$$\begin{aligned}\lambda_1(W) &= \mathcal{O}\left(d\log d\,\lambda_1\left(\frac{2dB_x^2}{\epsilon}\mathbf{I}_d\right)\right)\\ &= \mathcal{O}\left(\frac{d^2 B_x^2 \log d}{\epsilon}\right)\end{aligned}$$

which implies

$$\|\Delta\|_1 = \mathcal{O}\left(\frac{d^{5/2} B_x^2 \log d}{\epsilon}\right). \tag{36}$$

Combining Eqs. (29)–(36) yields

$$\|\mu_{DP} - \mu_{NP}\|_1 = O\left(\frac{d^2 B_x B_y \|\overline{xx}^{-1}\|_1 + d^{5/2} B_x^2 \log d \left\|(\overline{xx})^{-1}\right\|_1^2 \|\overline{xy}\|_1}{n\epsilon}\right)$$

with high probability.

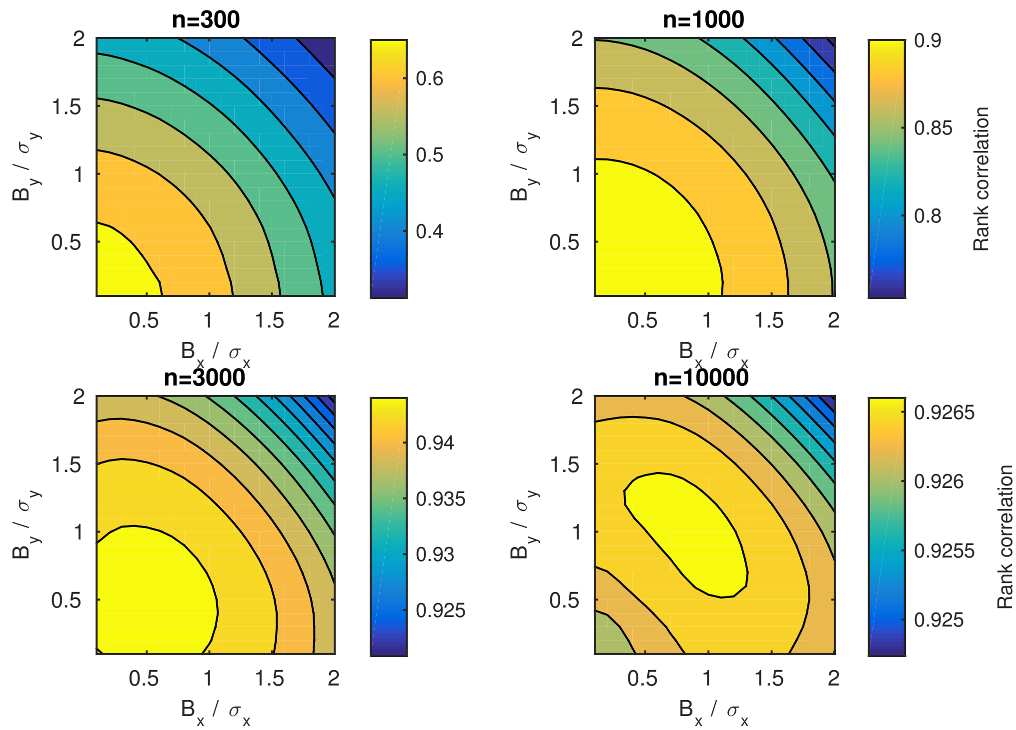

Figure 4: Illustration of the effect of projecting the outliers in linear regression, for different sample sizes n with 10-dimensional synthetic data, evaluated by Spearman's rank correlation between the predicted and true values. The x and y axes denote the projection thresholds as a function of standard deviations of data. Top right corner illustrates projection threshold at 2 standard deviations, no outlier projection would be further to top right. Higher values (yellow) are better. The result illustrates a clear benefit from the projection for moderate sample sizes, but the benefit decreases for really large sample sizes.

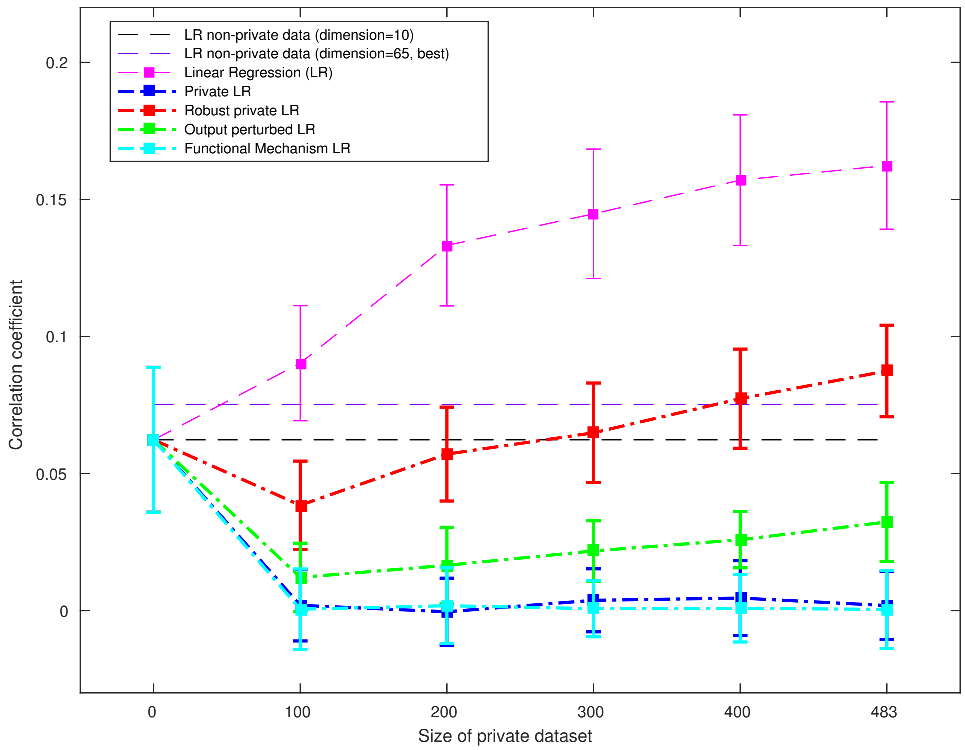

Figure 5: This is a complement to Figure 2 with more stringent privacy. Here we show Spearman's rank correlation coefficients (ρ) between the measured ranking of the cell lines and the ranking predicted by the models using $\epsilon = 1$. The baselines (horizontal dashed lines) are learned on 10 non-private data points; the private algorithms additionally have privacy-protected data (x-axis). The non-private algorithm (LR) has the same amount of additional non-privacy-protected data. All methods use 10-dimensional data except purple baseline showing the best performance with 10 non-private data points. The results are averaged over all drugs and 50-fold Monte Carlo cross-validation; error bars denote standard deviation over 50 Monte Carlo repeats. The result shows that more data are needed for good prediction performance under more stringent privacy.