scientific reports

Check for updates

OPEN

Deconvoluting kernel density estimation and regression for locally differentially private data

Farhad Farokhi

Local differential privacy has become the gold-standard of privacy literature for gathering or releasing sensitive individual data points in a privacy-preserving manner. However, locally differential data can twist the probability density of the data because of the additive noise used to ensure privacy. In fact, the density of privacy-preserving data (no matter how many samples we gather) is always flatter in comparison with the density function of the original data points due to convolution with privacy-preserving noise density function. The effect is especially more pronounced when using slow-decaying privacy-preserving noises, such as the Laplace noise. This can result in under/over-estimation of the heavy-hitters. This is an important challenge facing social scientists due to the use of differential privacy in the 2020 Census in the United States. In this paper, we develop density estimation methods using smoothing kernels. We use the framework of deconvoluting kernel density estimators to remove the effect of privacy-preserving noise. This approach also allows us to adapt the results from non-parametric regression with errors-in-variables to develop regression models based on locally differentially private data. We demonstrate the performance of the developed methods on financial and demographic datasets.

Government regulations, such as the roll-out of the General Data Protection Regulation in the European Union (EU) (https://gdpr-info.eu), the California Consumer Privacy Act (https://oag.ca.gov/privacy/ccpa), and the development of the Data Sharing and Release Bill in Australia (https://www.pmc.gov.au/public-data/data-sharing-and-release-reforms.) increasingly prohibit sharing customer's data without explicit consent[1].

A strong candidate for ensuring privacy is differential privacy. Differential privacy intuitively uses randomization to provide plausible deniability for the data of an individual by ensuring that the statistics of privacy-preserving outputs do not change significantly by varying the data of an individual[2,3]. Companies like Apple (https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf), Google (https://developers.googleblog.com/2019/09/enabling-developers-and-organizations.html), Microsoft (https://www.microsoft.com/en-us/ai/ai-lab-differential-privacy), and LinkedIn (https://engineering.linkedin.com/blog/2019/04/privacy-preserving-analytics-and-reporting-at-linkedin) have rushed to develop projects and to integrate differential privacy into their products. Even, the US Census Bureau has decided to implement differential privacy in 2020 Census[4]. Of course, this has created much controversy pointing to "ripple effect on the many public and private organizations that conduct surveys based on census data"[5].

A variant of differential privacy is local differential privacy in which all data points are randomized before being used by the aggregator, who attempts to infer the data distribution or some of its properties[6–8]. This is in contrast with differential privacy in which the data is first processed and then obfuscated by noise. Local differential privacy ensures that the data is kept private from the aggregator by adding noise to the individual data entries before the aggregation process. This is a preferred choice when dealing with untrusted aggregators, e.g., third party service providers or commercial retailers with financial interests, or when it is desired to release an entire dataset publicly for research in a privacy-preserving manner[9]. Differential privacy is in spirit close to randomized response methods introduced originally in[10] to reduce potential bias due to non-response and social desirability when asking questions about sensitive topics. The randomized response can be used to conceal individual responses (i.e., protect individual privacy) so that the respondents are more inclined to answer truthfully[11–14]. In fact, for questions with binary answer, the randomized response method with forced response

Department of Electrical and Electronic Engineering, The University of Melbourne, Parkville, VIC 3010, Australia. email: farhad.farokhi@unimelb.edu.au

(i.e., a respondent determines whether to answer a sensitive question truthfully or with forced yes/no based on flipping a biased coin) is differentially private and the probability of the head for the coin determines the privacy budget in differential privacy[15]. However, differential is a more general and flexible methodology that can be used for categorical and non-categorical (i.e., continuous domain) questions[4,16,17]. This paper specifically consider the problem of analyzing privacy-preserving data on continuous domain, which is out of the scope of randomized response methodology.

Locally differential data can significantly distort our estimates of the probability density of the data because of the additive noise used to ensure privacy. The density of privacy-preserving data can become flatter in comparison with the density function of the original data points due to convolution of its density with privacy-preserving noise density. The situation can be even more troubling when using slow-decaying privacy-preserving noises, such as the Laplace noise. This concern is true irrespective of how many samples are gathered. This can result in under/over-estimation of the heavy-hitters, a common and worrying criticism of using differential privacy in the US Census[18].

Estimating probability distributions/densities under differential privacy is of extreme importance as it is often the first step in gaining more important insights into the data, such as regression analysis. However, most of the existing work on probability distributions estimation based on locally differential private data focuses on categorical data[19–23]. For categorical data (in contrast with numerical data), the privacy-preserving noise is no longer additive, e.g., the so-called exponential mechanism[24] or other boutique differential privacy mechanisms[25] are often employed that are not on the offer in the 2020 US Census. The density estimation results for categorical data are also related to de-noising results in randomized response methods[12]. The work on continuous domains is often done by binning or quantizing the domain. However, finding the optimal number of bins or quantization resolution depending on privacy parameters, data distribution, and number of data points is a challenging task.

In this paper, we take a different approach to density estimation by using kernels and thus eliminating the need to quantize the domain. Kernel density estimation is a non-parametric way to estimate the probability density function of a random variable using its samples proposed independently by Parzen[26] and Rosenblatt[27]. This methodology was extended to multi-variate variables in[28,29]. These estimators work based on batches of data; however, they can also be made recursive[30–32]. When the data samples are noisy because of measurement noise or, as in the case of this paper, privacy-preserving noise, we need to eliminate the effect of the additive noise kernel density estimation by deconvolution[33]. Therefore, we use the framework of deconvoluting kernel density estimators[33–36] to remove the effect of privacy-preserving noise, which is often in the form of Laplace noise[37]. This approach also allows us to adapt the results from non-parameteric regression with errors-in-variables[38–40] to develop regression models based on locally differentially private data. This is the first time that deconvoluting kernel density estimators have been used for analyze differentially-private data. This is an important challenge facing social science researchers and demographers following the changes administered in the 2020 Census in the United States[4].

Methods

Consider independently distributed data points $\{\mathbf{x}[i]\}_{i=1}^{n} \subset \mathbb{R}^q$, for some fixed dimension $q \geq 1$, from common probability density function $\phi_{\mathbf{x}}$. Each data point $\mathbf{x}[i] \in \mathbb{R}^q$ belongs to an individual. Under no privacy restrictions, the data points can be provided to the central aggregator to construct an estimate of the density $\phi_{\mathbf{x}}$ denoted by $\widehat{\phi}_{\mathbf{x}}$. We may use kernel K, which is a bounded even probability density function, to generate the density estimate $\widehat{\phi}_{\mathbf{x}}$. A widely recognized example of a kernel is the Gaussian kernel[41] in

$$K(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^q}} \exp\left(-\frac{1}{2}\mathbf{x}^\top \mathbf{x}\right). \tag{1}$$

In the big data regime $n \gg 1$, the choice of the kernel is not crucial to the accuracy of kernel density estimators so long as it meets the conditions in[34]. In this paper, we keep the kernel general. By using kernel K, we can construct the estimate

$$\widehat{\phi}_{\mathbf{x}}^{\mathrm{np}}(\mathbf{x}) = \frac{1}{nh^q} \sum_{i=1}^{n} K((\mathbf{x} - \mathbf{x}[i])/h), \tag{2}$$

where $h > 0$ is the bandwidth. The bandwidth is often selected such that $h \to 0$ as $n \to \infty$. The optimal rate of decay for the bandwidth has been established for families of distributions[33,34].

Remark 1 The problem formulation in this paper considers real-valued data as opposed as categorical data. This distinguishes the paper from the computer science literature on this topic, which primarily focuses on categorical data[19–23]. Real-valued data can arise in two situations. First, the posed question can be non-categorical, e.g., credit rating for loans or the interest rates of loans. We will consider this in one of our experimental results. However, aggregated categorical data can also be real-valued. For instance, the 2020 US Census reports the aggregate number of individuals from a race or ethnicity group within different counties. These numbers will be made differentially private as part of the US Census Bureau's privacy initiative[4]. Therefore, the methods developed in this paper are still relevant to categorical data, albeit in aggregated forms.

As discussed in the introduction, due to privacy restrictions, the exact data points $\{\mathbf{x}[i]\}_{i=1}^n$ might not be available to generate the density estimate in (2). The aggregator may only have access to noisy versions of these data points:

$$\mathbf{z}[i] = \mathbf{x}[i] + \mathbf{n}[i], \tag{3}$$

where $\mathbf{n}[i]$ is a privacy-preserving additive noise. To ensure differential privacy, Laplace additive noises is often used[37]. For any probability density ϕ, we use the notation $\text{supp}(\phi)$ to denote its support set, i.e., $\text{supp}(\phi) := \{\xi : \phi(\xi) > 0\}$.

Assumption 1 (*Bounded support*) $\text{supp}(\phi_{\mathbf{x}}) \subseteq \prod_{i=1}^q [\underline{x}_i, \overline{x}_i]$ for finite constants $\underline{x}_i \leq \overline{x}_i$.

Assumption 1 is without loss of generality as we are always dealing with bounded domains in social sciences with *a priori* known bounds on the data (e.g., the population of a region).

Definition 1 (*Local differential privacy*) The reporting mechanism in (3) is ϵ-(locally) differentially private for $\epsilon \geq 0$ if

$$\mathbb{P}\{\mathbf{x}[i] + \mathbf{n}[i] \in \mathcal{Z} | \mathbf{x}[i] = \mathbf{x}\} \leq \exp(\epsilon)\mathbb{P}\{\mathbf{x}[i] + \mathbf{n}[i] \in \mathcal{Z} | \mathbf{x}[i] = \mathbf{x}'\}, \quad \forall \mathbf{x}, \mathbf{x}' \in \text{supp}(\phi_{\mathbf{x}}),$$

for any Borel-measurable set $\mathcal{Z} \subseteq \mathbb{R}^q$.

Definition 1 ensures that the statistics of privacy-preserving output $\mathbf{x}[i] + \mathbf{n}[i]$, determined by its distribution, do not change "significantly" (the magnitude of change is bounded by the privacy parameter ϵ) if the data of individual $\mathbf{x}[i]$ changes. If $\epsilon \to 0$, the output becomes more noisy and a higher privacy guarantee is achieved. Laplace additive noise is generally used to ensure differential privacy. This is formalized in the following theorem, which is borrowed from[37].

Theorem 1 *Let $\{\mathbf{n}[i]\}_{i=1}^n$ be distributed according to the common multivariate Laplace density:*

$$\phi_{\mathbf{n}}(\mathbf{n}) = \frac{1}{2^q \prod_{j=1}^q b_j} \exp\left(-\sum_{j=1}^q \frac{|n_j|}{b_j}\right),$$

where n_j is the j-th component of $\mathbf{n} \in \mathbb{R}^q$. The reporting mechanism in (3) is ϵ-locally differentially private if $b_j = q(\overline{x}_j - \underline{x}_j)/\epsilon$ for $j \in \{1, \ldots, q\}$.

In what follows, we assume that the reporting policy in Theorem 1 is used to generate locally differentially private data points. Since $\{\mathbf{n}[i]\}_{i=1}^n$ are distributed according to the common density $\phi_{\mathbf{n}}(\mathbf{n})$, $\{\mathbf{z}[i]\}_{i=1}^q$ would also follow a common probability density, which is denoted by $\phi_{\mathbf{z}}$. Note that

$$\Phi_{\mathbf{z}}(\mathbf{t}) = \Phi_{\mathbf{x}}(\mathbf{t})\Phi_{\mathbf{n}}(\mathbf{t}), \tag{4}$$

where $\Phi_{\mathbf{z}}$, $\Phi_{\mathbf{x}}$, and $\Phi_{\mathbf{n}}$ are the characteristic functions of $\phi_{\mathbf{z}}$, $\phi_{\mathbf{x}}$, and $\phi_{\mathbf{n}}$. Using (4), we can use any approximation of $\Phi_{\mathbf{z}}$ to construct an approximation of $\Phi_{\mathbf{x}}$ and thus estimate $\phi_{\mathbf{x}}$. If we use kernel K for estimating density of $\mathbf{z}[i]$, $\forall i$, we get

$$\widehat{\phi}_{\mathbf{z}}(\mathbf{z}) = \frac{1}{nh^q} \sum_{i=1}^n K((\mathbf{z} - \mathbf{z}[i])/h).$$

Here, $\widehat{\phi}_{\mathbf{z}}$ is used to denote the approximation of $\phi_{\mathbf{z}}$. The characteristic function of $\widehat{\phi}_{\mathbf{z}}$ is given by

$$\widehat{\Phi}_{\mathbf{z}}(\mathbf{t}) = \Phi_K(h\mathbf{t})\widehat{\Phi}(\mathbf{t}),$$

where $\Phi_K(\mathbf{t})$ is the characteristic function of K and $\widehat{\Phi}(\mathbf{t})$ is the empirical characteristic function of measurements $\{\mathbf{z}[i]\}_{i=1}^n$, defined as

$$\widehat{\Phi}(\mathbf{t}) = \frac{1}{n} \sum_{i=1}^n \exp\left(i\mathbf{t}^\top \mathbf{z}[i]\right).$$

Therefore, the characteristic function of $\widehat{\phi}_{\mathbf{x}}$ is given by

$$\widehat{\Phi}_{\mathbf{x}}(\mathbf{t}) = \frac{\Phi_K(H\mathbf{t})\widehat{\Phi}(\mathbf{t})}{\Phi_{\mathbf{n}}(\mathbf{t})}$$

Further, note that

$$\begin{aligned}\Phi_{\mathbf{n}}(\mathbf{t}) =& \mathbb{E}\left\{\exp\left(i\mathbf{t}^\top \mathbf{n}\right)\right\}\\ =& \mathbb{E}\left\{\exp\left(it_1 n_1\right)\exp\left(it_2 n_2\right)\cdots\exp\left(it_q n_q\right)\right\}\\ =& \mathbb{E}\left\{\exp\left(it_1 n_1\right)\right\}\mathbb{E}\left\{\exp\left(it_2 n_2\right)\right\}\cdots\mathbb{E}\left\{\exp\left(it_q n_q\right)\right\}\\ =& \prod_{j=1}^{q}\frac{1}{1+b_j^2 t_j^2},\end{aligned}$$

where t_j is the j-th component of $\mathbf{t}\in\mathbb{R}^q$. We get

$$\widehat{\phi}_{\mathbf{x}}(\mathbf{x}) = \frac{1}{nh^q}\sum_{i=1}^{n}\widehat{K}_h((\mathbf{x}-\mathbf{z}[i])/h), \tag{5}$$

where

$$\begin{aligned}\widehat{K}_h(\mathbf{x}) &= \frac{1}{(2\pi)^q}\int_{\mathbb{R}^q}\exp(-i\mathbf{t}^\top\mathbf{x})\frac{\Phi_K(\mathbf{t})}{\Phi_{\mathbf{n}}(\mathbf{t}/h)}\mathrm{d}\mathbf{t}\\ &= \frac{1}{(2\pi)^q}\int_{\mathbb{R}^q}\exp(-i\mathbf{t}^\top\mathbf{x})\prod_{j=1}^{q}\left(1+\frac{b^2}{h^2}t_j^2\right)\Phi_K(\mathbf{t})\mathrm{d}\mathbf{t}\\ &= \prod_{j=1}^{q}\left(1-\frac{b_j^2}{h^2}\frac{\partial^2}{\partial x_j^2}\right)K(\mathbf{x}),\end{aligned}$$

where x_j is the j-th component of $\mathbf{x}\in\mathbb{R}^q$.

Under appropriate conditions on the kernel K[34], we can see that

$$\mathbb{E}\{\widehat{\phi}_{\mathbf{x}}(\mathbf{x})|\{\mathbf{x}_i\}_{i=1}^n\} = \widehat{\phi}_{\mathbf{x}}^{\mathrm{np}}(\mathbf{x}). \tag{6}$$

Therefore, $\widehat{\phi}_{\mathbf{x}}(\mathbf{x})$ in (5) is effectively an unbiased estimate of $\widehat{\phi}_{\mathbf{x}}^{\mathrm{np}}(\mathbf{x})$ in (2). In average, we are canceling the effect of the differential privacy noise. Selecting bandwidth (or smoothing parameter) h is an important aspect of kernel estimation. In[26], it was shown that $\lim_{n\to\infty} h = 0$ guarantees asymptotic ubiasedness (i.e., point-wise convergence of the kernel density estimate to the true density function) while $\lim_{n\to\infty} nh = +\infty$ is required to ensure asymptotic consistency. Many studies have focused on finding optimal bandwidth[29,42–44]. Numerical methods based on cross validation for setting the bandwidth are proposed in[45,46]. Often, it is recommended to compare the results from different bandwidth selection algorithms to avoid misleading conclusions caused by over-smoothing or under-smoothing of the density estimate[47]. These results have been also extended to noisy measurements with deconvoluting kernel density estimation[34,48]. If h scales according to $n^{-1/5}$, $\widehat{\phi}_{\mathbf{x}}(\mathbf{x})$ is a consistent estimator of $\phi_{\mathbf{x}}$ as $n\to\infty$, i.e., $\widehat{\phi}_{\mathbf{x}}(\mathbf{x})$ converges $\phi_{\mathbf{x}}$ point-wise for all $\mathbf{x}\in\mathrm{supp}(\phi_{\mathbf{x}})$[34]. Note that by selecting $h=\mathcal{O}(n^{-1/5})$, we get

$$\int\mathbb{E}\{\widehat{\phi}_{\mathbf{x}}(\mathbf{x})-\phi_{\mathbf{x}}(\mathbf{x})\}^2 = \mathcal{O}(n^{-4/5}),$$

where $\mathcal{O}$ denotes the Bachmann–Landau notation. Therefore, $\int\mathbb{E}\{\widehat{\phi}_{\mathbf{x}}(\mathbf{x})-\phi_{\mathbf{x}}(\mathbf{x})\}^2\to 0$ as $n\to\infty$. This means that the effect of the differential-privacy noise is effectively negligible on large datasets.

For regression analysis, we consider independently distributed data points $\{(\mathbf{x}[i],\mathbf{y}[i])\}_{i=1}^n$ from common probability density function. We would like to understand the relationship between inputs $\mathbf{x}[i]$ and outputs $\mathbf{y}[i]$ for all i. Similarly, we assume that we can only access noisy privacy-preserving inputs $\{\mathbf{z}[i]\}_{i=1}^n$ instead of accurate inputs $\{\mathbf{x}[i]\}_{i=1}^n$. Following the argument above, we can also construct the Nadaraya–Watson kernel regression (see, e.g.[49]) as

$$\widehat{m}(\mathbf{x}) := \frac{\sum_{i=1}^n\widehat{K}_h((\mathbf{x}-\mathbf{z}[i])/h)\mathbf{y}[i]}{\sum_{i=1}^n\widehat{K}_h((\mathbf{x}-\mathbf{z}[i])/h)}. \tag{7}$$

Under appropriate conditions on the kernel K and the bandwidth h[40], $\widehat{m}(\mathbf{x})$ converges to $\mathbb{E}\{\mathbf{y}|\mathbf{x}\}$ almost surely. In practice the bandwidth can be computed by minimizing the cross-validation cost, i.e., the error of estimating each $\mathbf{y}[j]$ using the Nadaraya–Watson kernel regression constructed from $\{(\mathbf{z}[i],\mathbf{y}[i])\}_{i\in\{1,\ldots,n\}\setminus\{j\}}$ averaged over all choices of ℓ. The optimal bandwidth is given by

$$\arg\min_h\sum_{j=1}^{n}\ell(\mathbf{y}[j],\widehat{m}_{-j}(\mathbf{x}[j])), \tag{8}$$

where ℓ is a fitness function, e.g., $\ell(\mathbf{y},\mathbf{y}') = \|\mathbf{y}-\mathbf{y}'\|_2^2$, and $\widehat{m}_{-j}(\mathbf{x})$ is the Nadaraya–Watson kernel regression constructed from $\{(\mathbf{z}[i],\mathbf{y}[i])\}_{i\in\{1,\ldots,n\}\setminus\{j\}}$:

$$\widehat{m}_{-j}(\mathbf{x}) := \frac{\sum_{i\in\{1,\ldots,n\}\setminus\{j\}}\widehat{K}_h((\mathbf{x}-\mathbf{z}[i])/h)\mathbf{y}[i]}{\sum_{i\in\{1,\ldots,n\}\setminus\{j\}}\widehat{K}_h((\mathbf{x}-\mathbf{z}[i])/h)}.$$

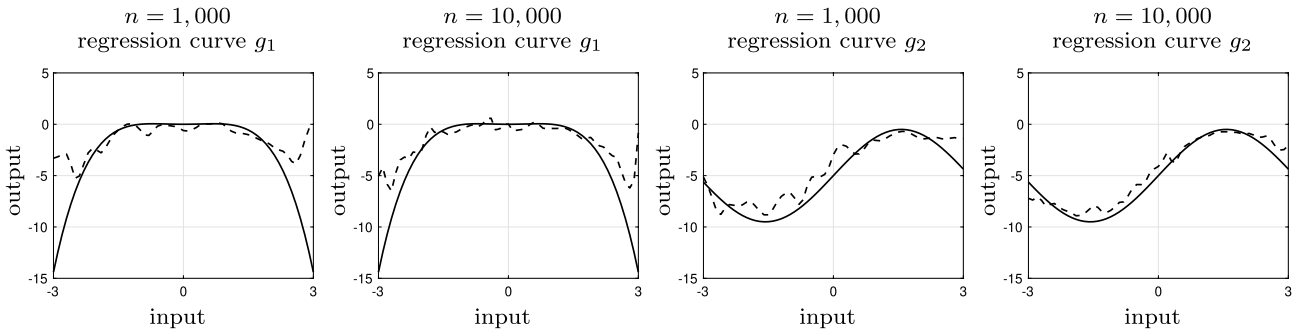

Figure 1. The kernel regression model (dashed black) and true regression curve (solid black) for mixture Gaussian data made differentially private with $\epsilon = 10$.

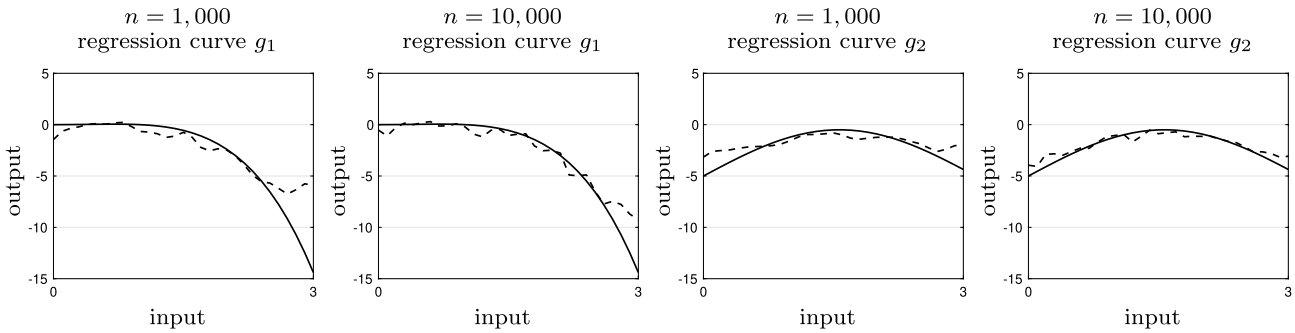

Figure 2. The kernel regression model (dashed black) and true regression curve (solid black) for chi-squared data made differentially private with $\epsilon = 10$.

This approach has been widely used for setting the bandwidth in non-parametric regression[38].

Results

In this section, we demonstrate the performance of the developed methods on multiple datasets. We first use a synthetic dataset for illustration purposes and then utilize real financial and demographic datasets. Throughout this section, we use the following original kernel:

$$K(x) = \frac{1}{\pi}\frac{1}{1+x^2}.$$

Note that $\mathbf{x} = x$ is a scalar as we are only considering a single input. This is the Cauchy distribution. We get the adjusted kernel in

$$\begin{aligned}\widehat{K}_h(x) &= \left(1 - \frac{b^2}{h^2}\frac{\mathrm{d}^2}{\mathrm{d}x}\right)K(x)\\ &= \frac{1}{\pi}\left[\frac{1}{1+x^2} - \frac{b^2}{h^2}\frac{8x^2}{(x^2+1)^3} + \frac{b^2}{h^2}\frac{2}{(x^2+1)^2}\right].\end{aligned}$$

We use the cross-validation procedure in (8) to find the bandwidth in the following simulation and experiments.

Synthetic dataset. We use a simulation study is to illustrate the performance of the Nadaraya–Watson kernel regression in (7) for privacy-preserving data. We consider multiple scenarios. We use two distributions for $\{x[i]\}_{i=1}^n$. The first one is a Gaussian mixture $(1/3)\mathcal{N}(-1,1) + (2/3)\mathcal{N}(3/2,1/2)$ truncated over $[-3,3]$. The second distribution is a chi-squared distribution with three degrees of freedom distribution $\chi^2(3)$ truncated over $[0,3]$. The truncation in both cases is for satisfaction of Assumption 1. We also consider two regression curves: $g_1 : x \mapsto x^2(1-x^2)/5$ and $g_2 : x \mapsto 4.5\sin(x) - 5$. Finally, we assume a Gaussian measurement noise $\mathcal{N}(0,1)$, i.e., $y[i] = g_j(x[i]) + v[i]$ for $j = 1,2$, where $v[i]$ is a zero mean Gaussian random variable with unit variance.

Figure 1 shows the kernel regression model (dashed black) and true regression curve (solid black) for mixture Gaussian data made differentially private with $\epsilon = 10$. Here, we consider two dataset size of $n = 1000$ and $n =$ 10,000 and two regression curves of g_1 and g_2, introduced earlier. The Nadaraya–Watson kernel regression using differentially-private provides fairly accurate predictions. The accuracy of the prediction improves as the dataset gets larger. Figure 2 illustrates the kernel regression model (dashed black) and true regression curve (solid black)

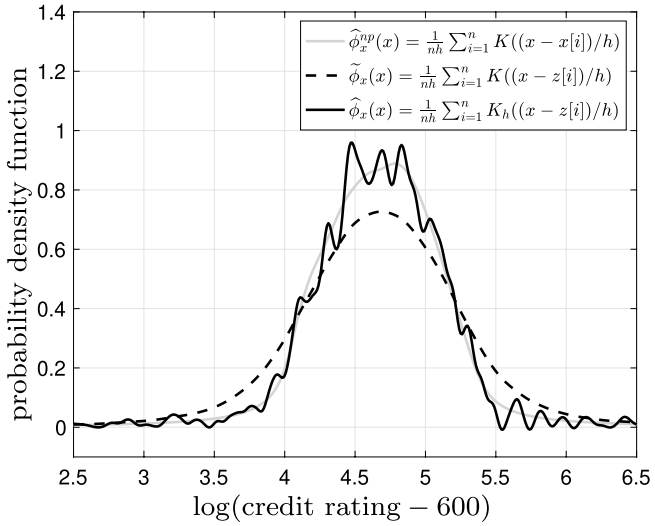

Figure 3. Estimates of probability density function of the credit score using original noiseless data with original kernel $\widehat{\phi}_x^{np}(x) = \frac{1}{nh}\sum_{i=1}^{n} K((x - x[i])/h)$ (solid gray), ϵ-locally differential private data with original kernel $\widetilde{\phi}_x(x) = \frac{1}{nh}\sum_{i=1}^{n} K((x - z[i])/h)$ (dashed black), and ϵ-locally differential private data with adjusted kernel $\widehat{\phi}_x(x) = \frac{1}{nh}\sum_{i=1}^{n} K_h((x - z[i])/h)$ (solid black) for $\epsilon = 5.0$ and bandwidth $h = 0.1$.

for chi-squared data made differentially private with $\epsilon = 10$. This shows that the fitness of the Nadaraya-Watson kernel regression is somewhat independent of the underlying distribution of the data.

Lending club dataset. The dataset contains information of 2,260,701 accepted and 27,648,741 rejected loans application on Lending Club, a peer-to-peer lending platform, over 2007 to 2018. The dataset is available for download on Kaggle[50]. For the accepted loans, dataset contains interest rates of the loans per annum and loan attributes, such as total loan size, and borrower information, such as number of credit lines, credit rating, state of residence, and age. Here, we only focus on data from 2010 (to avoid possible yearly fluctuations of the interest rate), which contains 12,537 accepted loans. We also focus on the relationship between the FICO (https://www.fico.com/en/products/fico-score.) credit score (low range) and the interest rates of the loan. This is an interesting relationship pointing to the value of credit rating reports[51]. The FICO credit score is very sensitive (as it relates to the financial health of an individual) and possesses a significant commercial value (as it is sold by a for-profit corporation). Thus, we assume that is is made available publicly in a privacy-preserving manner using (3). Note that the original data in[50] provides this data in an anonymized manner without privacy-preserving noise.

Figure 3 illustrates estimates of probability density function of the credit score $\phi_x(x)$ using original noiseless data with original kernel $\widehat{\phi}_x^{np}(x)$ in (2) (solid gray), ϵ-locally differential private data with original kernel $\widetilde{\phi}_x(x) = \frac{1}{nh}\sum_{i=1}^{n} K((x - z[i])/h)$ (dashed black), and ϵ-locally differential private data with adjusted kernel in (5) (solid black) for $\epsilon = 5.0$ and bandwidth $h = 0.1$. Note that $\widetilde{\phi}_x(x) = \frac{1}{nh}\sum_{i=1}^{n} K((x - z[i])/h)$ is a naive density estimate as it does not try to cancel the effect of the privacy-preserving noise. Clearly, using the original kernel for the noisy privacy-preserving data flattens the density estimate $\widetilde{\phi}_x(x)$. This is because we are in fact observing a convolution of the original probability density with the probability density of the Laplace noise. Upon using the adjusted kernel $\widehat{K}_h(x)$ the estimate of the probability density using the noisy privacy-preserving data matches the estimate of the probability density with the original data (with additional fluctuations due to the presence of noise). This provides a numerical validation of (6).

Now, let us focus on the regression analysis. Figure 4 shows the kernel regression model (solid black) and the linear regression model (dashed black) based on the original data with bandwidth $h = 0.02$ superimposed on the original noiseless data (gray dots). The mean squared error for the kernel regression model is 4.42 and the mean squared error for the linear regression model is 4.61. The kernel regression model is thus slightly superior (roughly 4%) to the linear regression model; however, the gap is narrow. Figure 5 illustrates the kernel regression model (solid black) and the linear regression model (dashed black) based on the ϵ-locally differential private data with $\epsilon = 5$ and bandwidth $h = 0.20$ superimposed on the original noiseless data (gray dots). The mean squared error for the kernel regression model is 5.70 and the mean squared error for the linear regression model is 7.11. In this case, the kernel regression model is considerably (roughly 20%) better. In Fig. 6, we observe the mean squared error for the kernel regression model and the linear regression model based on the ϵ-locally differential private data versus privacy budget ϵ. Clearly, the kernel regression model is consistently superior to the linear regression model. As ϵ grows larger, the performance of the kernel regression model and the linear regression model based on the ϵ-locally differential private data converge to the performance of the kernel regression model and the linear regression model based on original noiseless data. This intuitively makes sense as, by increasing the privacy budget, the magnitude of the privacy-preserving noise becomes smaller.

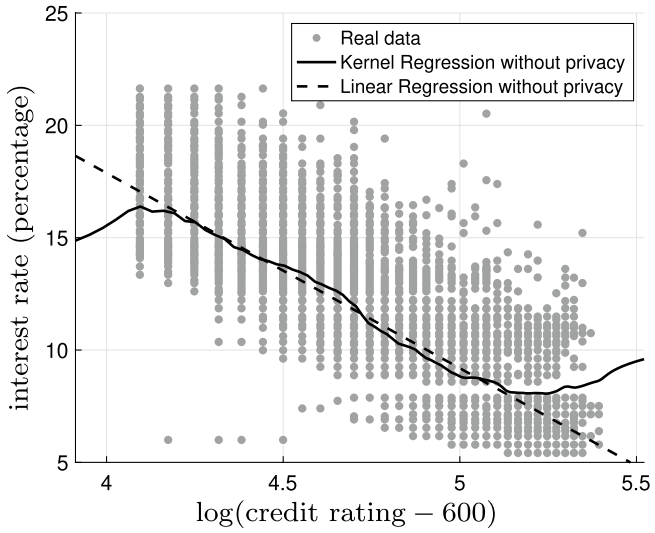

Figure 4. The kernel regression model (solid black) and the linear regression model (dashed black) based on the original data with bandwidth $h = 0.02$ superimposed on the original noiseless data (gray dots). The mean squared error for the kernel regression model is 4.42 and the mean squared error for the linear regression model is 4.61.

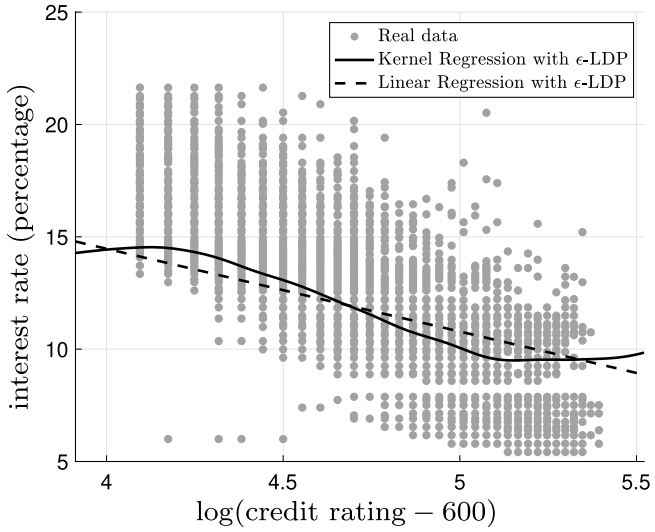

Figure 5. The kernel regression model (solid black) and the linear regression model (dashed black) based on the ϵ-locally differential private data with $\epsilon = 5$ and bandwidth $h = 0.20$ superimposed on the original noiseless data (gray dots). The mean squared error for the kernel regression model is 5.70 and the mean squared error for the linear regression model is 7.11.

Adult dataset. The dataset contains information of 32,561 individuals from the 1994 Census database. The dataset is available for download on UCI[52]. The dataset contains attributes, such as education, age, work type, gender, race, and a binary report whether the individual earns more than 50,000$ per year. We also focus on the relationship between the education (in years) and the individual ability to earn more than 50,000$ per year. The education is assumed to be made public in a privacy-preserving form following (3). This information can be considered private as it can be used in conjunction with other information to de-anonymize the dataset.

Figure 7 The kernel regression model (solid black) and the logistic regression model (dashed black) based on the original data with bandwidth $h = 0.17$. The logarithm of the likelihood for the kernel regression model is -0.49 and the logarithm of the likelihood for the logistic regression model is -0.50. The kernel regression model is thus slightly superior (roughly 2%) to the logistic regression model; however, the gap is almost negligible. Figure 8 illustrates the kernel regression model (solid black) and the logistic regression model (dashed black) based on the ϵ-locally differential private data with $\epsilon = 5.0$ bandwidth $h = 2.98$. The logarithm of the likelihood for the kernel regression model is -0.51 and the logarithm of the likelihood for the logistic regression model is -0.53.

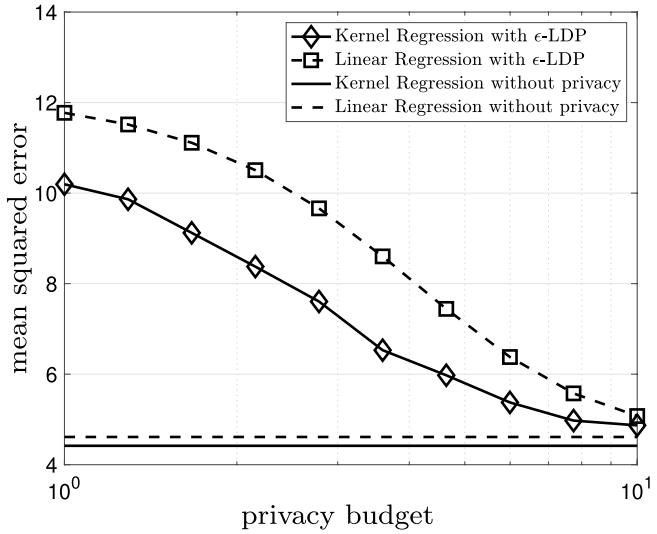

Figure 6. The mean squared error for the kernel regression model and the linear regression model based on the ϵ-locally differential private (ϵ-LDP in the legend) data versus privacy budget ϵ. The horizontal lines show the mean squared error for the kernel regression model and the linear regression model based on original noiseless data.

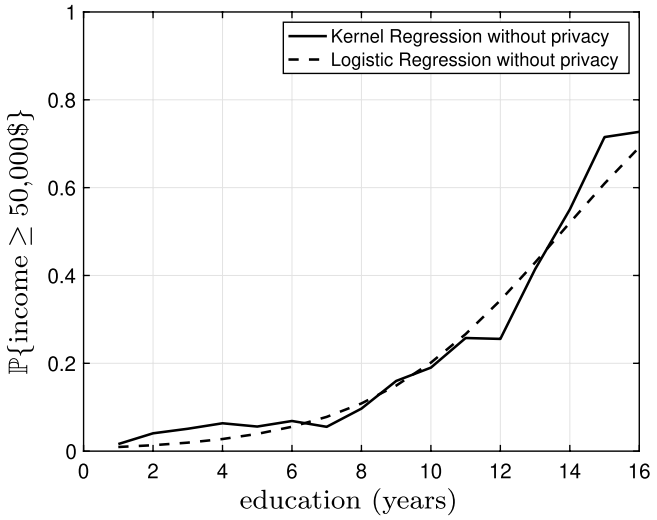

Figure 7. The kernel regression model (solid black) and the logistic regression model (dashed black) based on the original data with bandwidth $h = 0.17$. The logarithm of the likelihood for the kernel regression model is -0.49 and the logarithm of the likelihood for the logistic regression model is -0.50.

In this case, the kernel regression model is slightly (roughly 4%) better. In Fig. 9, we observe the logarithm of the likelihood for the kernel regression model and the logistic regression model based on the ϵ-locally differential private data versus privacy budget ϵ. The horizontal lines show the logarithm of the likelihood for the kernel regression model and the logistic regression model based on original noiseless data. Again, the kernel regression model is consistently superior to the logistic regression model. However, the effect is not as pronounced as the linear regression in the previous subsection. Finally, again, as ϵ grows larger, the performance of the kernel regression model and the logistic regression model based on the ϵ-locally differential private data converge to the performance of the kernel regression model and the linear regression model based on original noiseless data.

Discussion

The density of privacy-preserving data is always flatter in comparison with the density function of the original data points due to convolution with privacy-preserving noise density function. This is certainly a cause for concern due to addition of differential-privacy noise in 2020 US Census. This unfortunate effect is always present irrespective of how many samples we gather because we observe the convolution of the original probability density with the probability density of the privacy-preserving noise. This can result in miss-estimation of the

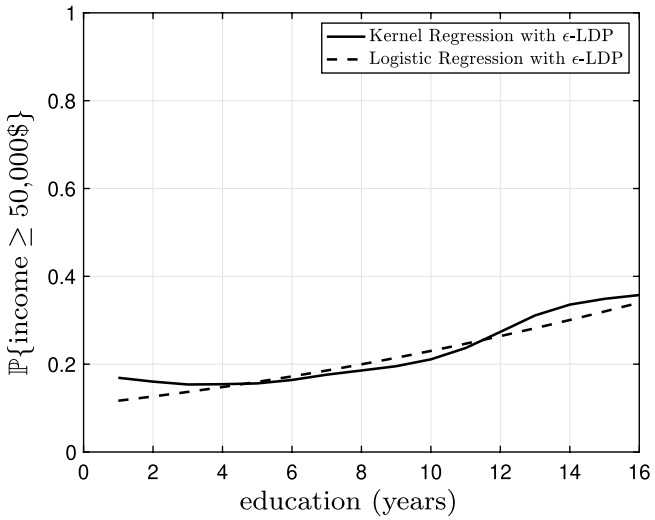

Figure 8. The kernel regression model (solid black) and the logistic regression model (dashed black) based on the ϵ-locally differential private data with $\epsilon = 5.0$ bandwidth $h = 2.98$. The logarithm of the likelihood for the kernel regression model is -0.51 and the logarithm of the likelihood for the logistic regression model is -0.53.

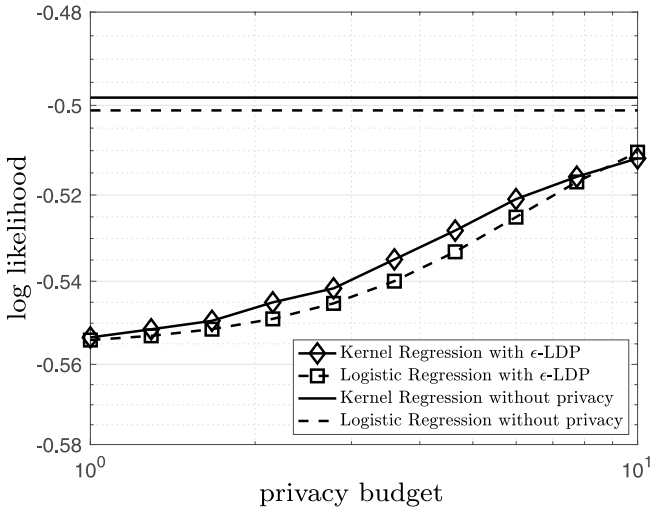

Figure 9. The logarithm of the likelihood for the kernel regression model and the logistic regression model based on the ϵ-locally differential private data versus privacy budget ϵ. The horizontal lines show the logarithm of the likelihood for the kernel regression model and the logistic regression model based on original noiseless data.

heavy-hitters that often play an important role in social sciences due to their ties to minority groups. We developed density estimation methods using smoothing kernels and used the framework of deconvoluting kernel density estimators to remove the effect of privacy-preserving noise. This can result in a superior performance both for estimating probability density functions and for kernel regression in comparison to popular regression techniques, such as linear and logistic regression models. In the case of estimating the probability density function, we could entirely remove the flatting effect of the privacy-preserving noise at the cost of additional fluctuations. The fluctuations however could be reduced by gathering more data.

Received: 11 September 2020; Accepted: 17 November 2020
Published online: 07 December 2020

References

1. Bennett, C. J. & Raab, C. D. Revisiting the governance of privacy: contemporary policy instruments in global perspective. *Regul. Govern.* **14**(3), 447–464 (2020).
2. Dwork, C. Differential privacy: a survey of results. In *Theory and Applications of Models of Computation, Vol. 4978 of Lecture Notes in Computer Science* (eds Agrawal, M. *et al.*) 1–19 (Springer, Berlin, 2008).

3. Dwork, C., McSherry, F., Nissim, K. & Smith, A. Calibrating noise to sensitivity in private data analysis. in *Theory of Cryptography Conference*, 265–284 (2006).
4. Abowd, J. M. The US Census Bureau adopts differential privacy. in *Proc. 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2867 (2018).
5. Mervis, J. Researchers object to census privacy measure. *Science* **363**(6423), 114 (2019).
6. Dewri, R. Local differential perturbations: location privacy under approximate knowledge attackers. *IEEE Trans. Mobile Comput.* **12**(12), 2360–2372 (2013).
7. Duchi, J. C., Jordan, M. I. & Wainwright, M. J. Local privacy and statistical minimax rates. in *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, 429–438 (2013).
8. Kairouz, P., Oh, S. & Viswanath, P. Extremal mechanisms for local differential privacy. in *Advances in Neural Information Processing Systems*, 2879–2887 (2014).
9. Liu, P. *et al.* Local differential privacy for social network publishing. *Neurocomputing* **391**, 273–279 (2020).
10. Warner, S. L. Randomized response: a survey technique for eliminating evasive answer bias. *J. Am. Stat. Assoc.* **60**(309), 63–69 (1965).
11. Kuk, A. Y. Asking sensitive questions indirectly. *Biometrika* **77**(2), 436–438 (1990).
12. Blair, G., Imai, K. & Zhou, Y.-Y. Design and analysis of the randomized response technique. *J. Am. Stat. Assoc.* **110**(511), 1304–1319 (2015).
13. Boruch, R. F. Assuring confidentiality of responses in social research: a note on strategies. *Am. Sociol.* **6**, 308–311 (1971).
14. Fox, J. A. & Tracy, P. E. Measuring associations with randomized response. *Soc. Sci. Res.* **13**(2), 188–197 (1984).
15. Kearns, M. & Roth, A. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design* (Oxford University Press, Oxford, 2019).
16. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K. & Zhang, L. Deep learning with differential privacy.' in *Proc. 2016 ACM SIGSAC Conference on Computer and Communications Security*, 308–318 (2016).
17. Friedman, A. & Schuster, A. Data mining with differential privacy. in *Proc. 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 493–502 (2010).
18. Wezerek, G. & Riper, D. V. Changes to the census could make small towns disappear. *The New York Times* (accessed 27 Aug 2020). https://www.nytimes.com/interactive/2020/02/06/opinion/census-algorithm-privacy.html.
19. Acharya, J., Sun, Z. & Zhang, H. Hadamard response: estimating distributions privately, efficiently, and with little communication. in *The 22nd International Conference on Artificial Intelligence and Statistics*, 1120–1129 (2019).
20. Bassily, R. & Smith, A. Local, private, efficient protocols for succinct histograms. in *Proc. Forty-Seventh Annual ACM Symposium on Theory of Computing*, 127–135 (2015).
21. Erlingsson, Ú., Pihur, V. & Korolova, A. Rappor: randomized aggregatable privacy-preserving ordinal response. in *Proc. 2014 ACM SIGSAC Conference on Computer and Communications Security*, 1054–1067 (2014).
22. Wang, T., Blocki, J., Li, N. & Jha, S. Locally differentially private protocols for frequency estimation. in *26th USENIX Security Symposium (USENIX Security 17)*, 729–745 (2017).
23. Ye, M. & Barg, A. Optimal schemes for discrete distribution estimation under locally differential privacy. *IEEE Trans. Inf. Theory* **64**(8), 5662–5676 (2018).
24. McSherry, F. & Talwar, K. Mechanism design via differential privacy. in *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, 94–103, IEEE (2007).
25. Li, Z., Wang, T., Lopuhaa-Zwakenberg, M., Li, N. & Skoric, B. Estimating numerical distributions under local differential privacy. in *Proc. 2020 ACM SIGMOD International Conference on Management of Data*, 621–635 (2020).
26. Parzen, E. On estimation of a probability density function and mode. *Ann. Math. Stat.* **33**(3), 1065–1076 (1962).
27. Rosenblatt, M. Remarks on some nonparametric estimates of a density function. *Ann. Math. Stat.* **27**(3), 832–837 (1956).
28. Murthy, V. K. Nonparametric estimation of multivariate densities with applications. In *Multivariate Analysis* (ed. Krishnaiah, P. R.) 43–58 (Academic, New York, 1966).
29. Loftsgaarden, D. . O. & Quesenberry, C. . P. . A nonparametric estimate of a multivariate density function. *Ann. Math. Stat.* **36**, 1049–1051 (1965).
30. Mokkadem, A., Pelletier, M. & Slaoui, Y. The stochastic approximation method for the estimation of a multivariate probability density. *J. Stat. Plann. Inference* **139**(7), 2459–2478 (2009).
31. Slaoui, Y. Bias reduction in kernel density estimation. *J. Nonparametr. Stat.* **30**(2), 505–522 (2018).
32. Slaoui, Y. Data-driven deconvolution recursive kernel density estimators defined by stochastic approximation method. *Sankhya A*. https://doi.org/10.1007/s13171-019-00182-3 (2019).
33. Carroll, R. J. & Hall, P. Optimal rates of convergence for deconvolving a density. *J. Am. Stat. Assoc.* **83**(404), 1184–1186 (1988).
34. Stefanski, L. A. & Carroll, R. J. Deconvolving kernel density estimators. *Statistics* **21**(2), 169–184 (1990).
35. Neumann, M. H. & Hössjer, O. On the effect of estimating the error density in nonparametric deconvolution. *J. Nonparametr. Stat.* **7**(4), 307–330 (1997).
36. Delaigle, A. *et al.* On deconvolution with repeated measurements. *Ann. Stat.* **36**(2), 665–685 (2008).
37. Dwork, C. & Roth, A. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* **9**(3–4), 211–407 (2014).
38. Delaigle, A. & Meister, A. Nonparametric regression estimation in the heteroscedastic errors-in-variables problem. *J. Am. Stat. Assoc.* **102**(480), 1416–1426 (2007).
39. Ioannides, D. & Alevizos, P. Nonparametric regression with errors in variables and applications. *Stat. Probab. Lett.* **32**(1), 35–43 (1997).
40. Fan, J. & Truong, Y. K. Nonparametric regression with errors in variables. *Ann. Stat.* **21**, 1900–1925 (1993).
41. Wand, M. P. & Jones, M. C. *Kernel Smoothing* (Chapman & Hall/CRC, Cambridge, 1994).
42. Rudemo, M. Empirical choice of histograms and kernel density estimators. *Scand. J. Stat.* **9**, 65–78 (1982).
43. Robert, P. W. On the choice of smoothing parameters for parzen estimators of probability density functions. *IEEE Trans. Comput.* **25**(11), 1175–1179 (1976).
44. Koontz, W. L. & Fukunaga, K. Asymptotic analysis of a nonparametric clustering technique. *IEEE Trans. Comput.* **100**(9), 967–974 (1972).
45. Delaigle, A. & Gijbels, I. Estimation of integrated squared density derivatives from a contaminated sample. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **64**(4), 869–886 (2002).
46. Delaigle, A. & Gijbels, I. Practical bandwidth selection in deconvolution kernel density estimation. *Comput. Stat. Data Anal.* **45**(2), 249–267 (2004).
47. Wang, C.-C. & Lee, W.-C. A simple method to estimate prediction intervals and predictive distributions: summarizing meta-analyses beyond means and confidence intervals. *Res. Synth. Methods* **10**(2), 255–266 (2019).
48. Es, B. V. & Uh, H.-W. Asymptotic normality of kernel-type deconvolution estimators. *Scand. J. Stat.* **32**(3), 467–483 (2005).
49. Härdle, W. *Appl. Nonparametr. Regress.* (Cambridge University Press, Cambridge, 1990).
50. George, N. All Lending Club loan data: 2007 through current Lending Club accepted and rejected loan data (accessed 20 Aug 2020). https://www.kaggle.com/wordsforthewise/lending-club.

51. Czarnitzki, D. & Kraft, K. Are credit ratings valuable information?. *Appl. Finan. Econom.* **17**(13), 1061–1070 (2007).
52. Dua, D. & Graff, C. University of California (UCI) machine learning repository (accessed 20 Aug 2020). http://archive.ics.uci.edu/ml.

Acknowledgements
The work of F.F. is in part supported by a startup grant from Melbourne School of Engineering at the University of Melbourne.

Author contributions
F.F. is the sole author of the paper.

Competing interests
The author declares no competing interests.

Additional information
Correspondence and requests for materials should be addressed to F.F.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020