CHANCE

ISSN: 0933-2480 (Print) 1867-2280 (Online) Journal homepage: https://www.tandfonline.com/loi/ucha20

Formal Privacy for Modern Nonparametric Statistics

Jordan Awan, Matthew Reimherr & Aleksandra (Seša) Slavković

To cite this article: Jordan Awan, Matthew Reimherr & Aleksandra (Seša) Slavković (2020) Formal Privacy for Modern Nonparametric Statistics, CHANCE, 33:4, 43-49, DOI: 10.1080/09332480.2020.1847959

To link to this article: https://doi.org/10.1080/09332480.2020.1847959

Formal Privacy for Modern Nonparametric Statistics

Jordan Awan, Matthew Reimherr, and Aleksandra (Seša) Slavković

Modern nonparametric (NP) statistics is an increasingly important and expanding tool set in data analytics as more large, complex data are gathered and analyzed. However, corresponding privacy concerns that arise require novel methods to balance privacy guarantees with statistical utility. NP methods present a unique challenge for privacy because the resulting summaries can contain significant amounts of individual-level information.

Modern NP statistics consists of tools to analyze data without assuming the data are distributed according to some pre-specified parametric family, such as assuming the data is distributed normally. Often, the goal of NP statistics is to estimate a function (e.g., probability density or regression function) with only limited assumptions, such as the number of derivatives.

Unlike parametric models, where only a fixed number of parameters are estimated, the number of "parameters" to estimate in NP statistics can be viewed as infinite, because an arbitrary real-valued function requires an infinite amount of data points to specify fully. (For a general introduction to NP and related methods, see *All of Nonparametric Statistics* by Larry Wasserman.)

While NP tools are flexible and powerful, they also have increased privacy risks when applied to sensitive data. Due to the infinite-dimensional nature of the quantities estimated in NP statistics, estimators often capture large amounts of individual-level data, and the value of an outlier can drastically change the shape of an estimated density or regression function (see Figures 1 and 4a).

From another perspective, while parametric methods generally estimate *global* properties, such as means and variances, NP methods often work with *local* information, giving higher priority to fewer data points to determine the shape of the function in a region.

The leading framework for constructing formal privacy methods is differential privacy (DP), proposed in *Calibrating Noise to Sensitivity in Private Data Analysis* (Dwork, McSherry, Nissim, and Smith. 2006), which can be interpreted as offering plausible deniability to data contributors. Many versions of differential privacy differ in various aspects now, but generally fit the intuition that a method that satisfies DP inserts additional randomness into the computations, so the probability of any output being publicly released is similar when an individual's data are changed in the input database. (Dwork and Roth. 2014.)

For certain quantities, such as sample means and medians, the dependence on a single individual is small and only a negligible amount of noise is required to privatize the estimate. However, the magnitude of the noise required for privacy grows significantly with the dimension of the release. This is less of a problem with parametric models, since the number of parameters to estimate is fixed. With NP methods, the goal is to estimate an infinite-dimensional object, and the estimators require estimating a number of parameters that can be viewed either as growing with the sample size n or as infinite.

Altogether, NP methods present a challenging setting for formal privacy methods, because they are much more sensitive to changes in an individual's data, and require more noise to privatize. The goal is to optimize the accuracy of the privatized estimator while offering formal privacy guarantees, which sometimes requires using a different approach from common nonprivate methods.

The issue of satisfying differential privacy while maintaining statistical utility can be highlighted and illustrated by exploring the problems of density estimation and nonparametric regression—two classical NP problems.

There are also some exciting developing areas of NP statistics that have unique privacy challenges.

Density Estimation

Density estimation is a classical area of nonparametric statistics, and has been tackled using tools as simple as histogram estimators, as well as more-sophisticated tools such as

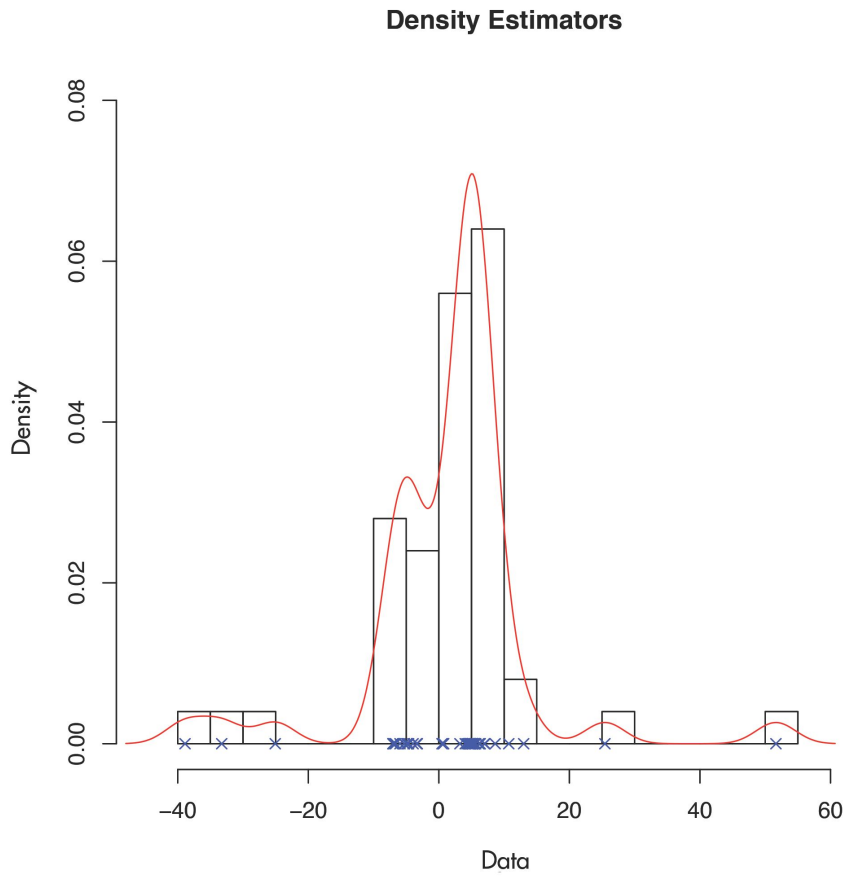


Figure 1. Sample of size 50 drawn from a mixture of Cauchy(−5, 1) and Cauchy(5, 1) with probabilities 1/3 and 2/3, respectively. Kernel density estimator using normal kernel and bandwidth ≈ 3, and histogram estimator with bin width 5.

kernel density estimators. Figure 1 provides an example of a histogram and kernel density estimator.

The typical setup for density estimation is: Let $X_1, \ldots, X_n$ be i.i.d. real valued random variables drawn from a density $f(x)$; f is known to be a nonnegative valued function that integrates to 1. Often, additional structure for f may be assumed, such as continuity or differentiability.

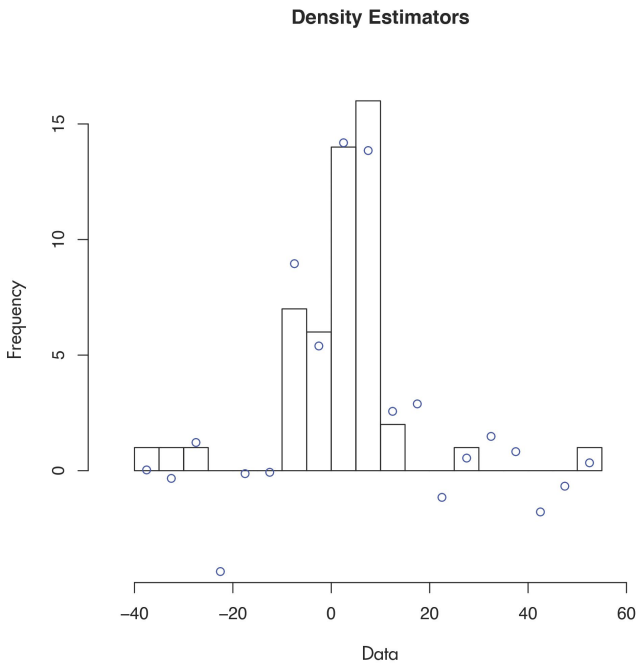
Histogram estimates are among the simplest and most-intuitive estimators for densities, especially when there are very limited assumptions. For histogram estimators, the number of bins typically increases in n, or conversely, the bin width decreases with n. Histogram estimators can be privatized by adding noise to each bin, appropriately scaled to obscure the contribution of one individual. Figure 2 illustrates a histogram estimator and an example of a privatized estimator by adding Laplace noise to each bin count.

A problem with histogram estimators is that they always result in discontinuous estimators. When the density is assumed to be smooth, kernel density estimators are often used instead. Let $K: \mathbb{R} \to \mathbb{R}$ be a function, called a kernel, that integrates to 1. A kernel density estimator is of the form

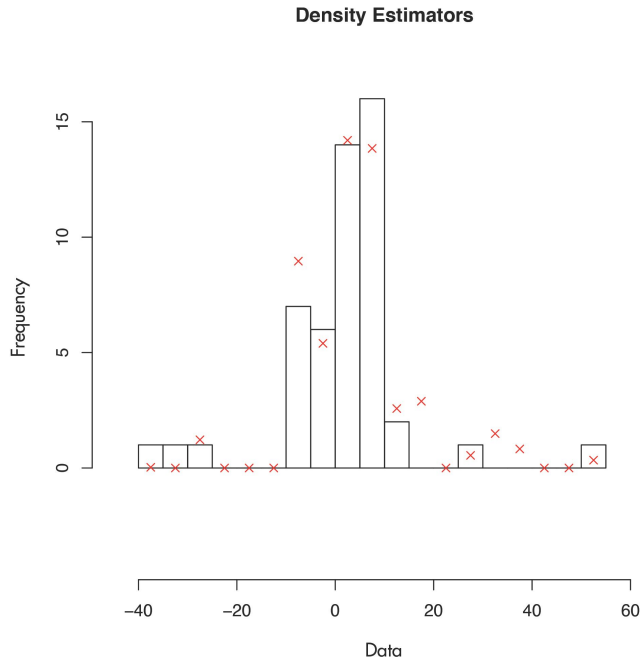
$$\hat{f}(t) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{t - X_i}{h}\right),$$

where $h > 0$ is the *bandwidth*, which decreases with n. Intuitively, a kernel density estimator places mass around each data point X_i, with decreasing influence the farther t is from X_i; Figure 1 provides an example using a Gaussian kernel.

Figure 1 illustrates one of the problems for privacy when using NP methods: They contain high amounts of individual-level information. Note that the outlying data points in Figure 1 can be



(a) Histogram estimator privatized by adding independent Laplace noise to each bin. Largest features are preserved, but noise results in some negative counts. Blue circles represent privatized counts.



(b) Post-processing of the privatized histogram in subfigure (a), where all negative counts are set to zero. Red crosses are privatized counts. Histogram estimator is privatized by adding independent Laplace noise to each bin.

Figure 2. Privatized histogram without post-processing. Noise results in some negative counts. Blue circles represent privatized counts.

identified by inspecting the "bumps" in the tails of the density estimator. Because one individual can have a large impact on the shape of the density estimator, much more noise is required to protect the privacy of the sample than is required for parametric approaches. It is a challenge to protect privacy while maintaining statistical utility.

The simplest method of preserving privacy while estimating a density is to add noise to each data point. However, this approach results in excessive noise, since each data point has high sensitivity. Rather than adding noise to the data, it is preferable to add noise to the resulting estimator.

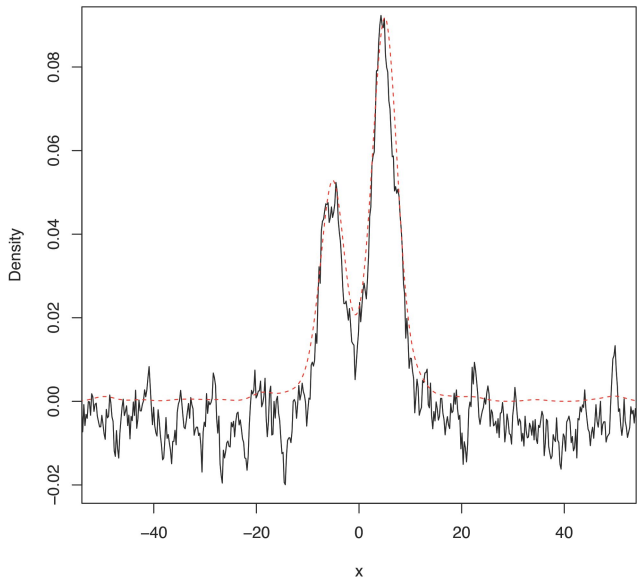
In multivariate settings, a Laplace or Gaussian random vector could be added; in this setting, a stochastic process, such as a Gaussian process, can be added to the resulting estimator.

Techniques to apply Gaussian processes to privatize function-valued parameters were developed in *Differential Privacy for Functions and Functional Data* (Hall, Rinaldo, and Wasserman. 2013) and *Formal Privacy for Functional Data with Gaussian Perturbations* (Mirshani, Reimherr, and Slavkovic. 2019). They discovered a deep connection between privacy and a space defined by the covariance function of the Gaussian process, classically known as the Cameron-Martin Space, which can be viewed as a reproducing kernel hilbert space (RKHS), whose kernel is given by the corresponding covariance function.

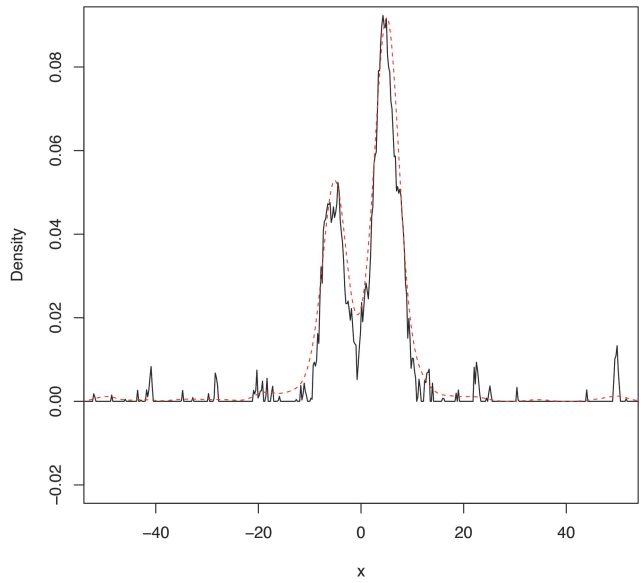
It turned out that it was exactly in that space that the sensitivity

Function-Valued Parameters: A parameter is a population quantity that can be used to specify the distribution. In nonparametric statistics, the parameters of interest are whole functions (such as the density function, or regression function), instead of finite-dimensional quantities. Such function-valued parameters are inherently infinite-dimensional.

Gaussian Process: A collection of random variables, indexed by time or space, is a Gaussian process if every finite subset of the variables has a multivariate Gaussian distribution. As in finite dimensions, a Gaussian process is characterized by its mean and covariance.



(a) Privatized density without post-processing.



(b) Privatized density, setting negative values to zero.

Figure 3. Data generated as in Figure 1, but with sample size of $n = 500$. Kernel density estimator using Gaussian kernel is fit to data with bandwidth of ≈ 2.01, plotted in red. Gaussian process is added to density estimator using exponential kernel so (1, .01)-DP is satisfied, plotted in black. Bandwidth assumed to be public.

Reproducing Kernel Hilbert Space (RKHS): Some infinite-dimensional vector spaces have properties very different from finite-dimensional Euclidean spaces. A subset of infinite- dimensional spaces that are much better behaved are called reproducing kernel Hilbert spaces (RKHSs). A space is an RKHS if every evaluation functional is continuous—the functions in the space are smooth. An RKHS is in one-to-one correspondence with a positive-definite kernel, which encodes how smooth the functions are.

of the estimator had to be computed to ensure differential privacy was satisfied. At a high level, their results showed that the Gaussian process noise must be "rougher" than the nonprivate estimator to obscure any personal information that may be captured in the high frequencies. Figure 3 illustrates the use of Gaussian process noise with the exponential kernel to privatize the kernel density estimator, resulting in a very rough density estimator.

Nonparametric Regression

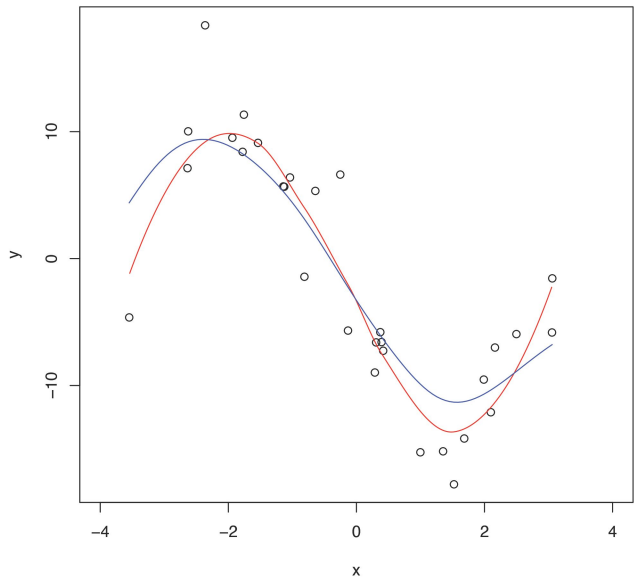
One of the most-common and natural questions that scientists encounter is understanding how one independent variable or more affects a dependent variable. This question is often addressed by the use of regression techniques. The classical approach of linear regression assumes that a linear predictor function captures the relationship between the independent/predictor/covariate and dependent/outcome variables, and typically imposes distributional assumptions for the errors.

These assumptions are often either are unreasonable or the sample size is large enough to explore deeper relationships between the variables and, thus, more-flexible techniques are desirable. Nonparametric regression techniques offer this increased flexibility.

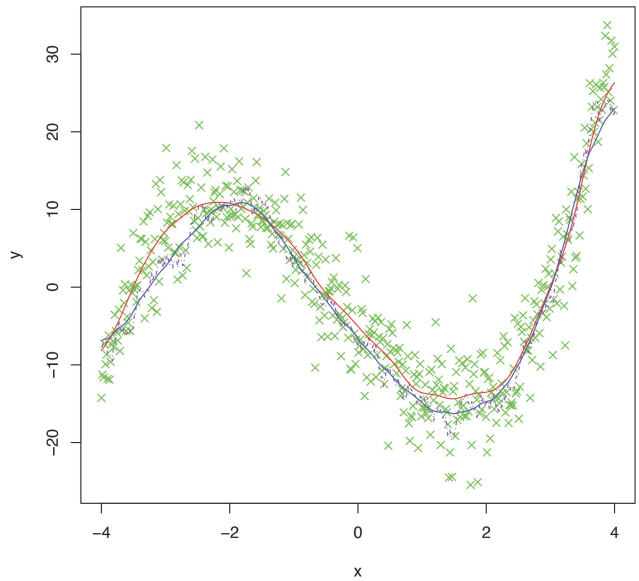
Suppose that $(X_1, Y_1),\ldots, (X_n, Y_n)$ are i.i.d. bivariate random vectors so that $Y_i = f(X_i)+e_i$ for some function f and mean-zero errors e_i. The goal is to estimate the function f. Depending on the setting, the X_i could be assumed to be deterministic in a controlled experiment, or random such as in observational studies.

In either case, nonparametric regression estimators are often *linear smoothers* that can be expressed as:

$$\hat{f}(t) = \sum_{i=1}^{n} \ell_i(t) Y_i, \qquad (1)$$



(a) n = 30. Red curve is a local quadratic regression and blue curve is a kernel-smoothing regression. Left-most x value in both has great influence on fitted curves.



(b) n = 500. Kernel regression with Gaussian kernel, bandwidth = .5. Privatized regression estimator using Gaussian process with exponential kernel, which achieves (1, .01)-DP in dashed purple. Range of Yi is assumed to be [20, 35], and sensitivity was numerically estimated. Post-processed curve is in solid blue, achieved by applying 50 pt moving average to privatized curve. Nonprivate kernel regression estimator is in red.

Figure 4. Example of nonparametric regression methods; y values are produced by equation $y_i = (x_i - 3)(x_i + 3.5)(x_i + 1/2) + e_i$, where $e_i \overset{\text{iid}}{\sim} N(0, 4)$. On left, $X_i \overset{\text{iid}}{\sim} U[-4, 4]$. On right, X_i is equally spaced in [−4, 4].

where $\ell_1, \ldots, \ell_n$ are real-valued functions, depending on the data

$$(X_i, Y_i)_{i=1}^n.$$

In most cases, the functions ℓ_i are normalized so that

$$\sum_{j=1}^n \ell_j(t) = 1 \text{ for all } t.$$

Some examples of linear smoothers include Nadarya-Watson kernel regression, more-general local polynomial regression, Reproducing Kernel Hilbert Space regression, and basis function regression. Kernel regression has the form

$$\ell_i(t) = K\left(\tfrac{t-X_i}{h}\right) / \left(\sum_{j=1}^n K\left(\tfrac{t-X_j}{h}\right)\right)$$

for a kernel $K(\cdot)$ and bandwidth h.

Based on Equation (1), there are a few challenges when it comes to privatizing these estimators. First, releasing the nonprivate $\hat{f}(x)$ often allows aspects of the original sample to be reconstructed. As noted earlier, adding noise to the X_i and Y_i before regressing would satisfy privacy, but introduces an excessive amount of noise, destroying statistical utility.

While the problem of producing private NP regression estimators may seem similar to density estimation, it has an additional challenge compared to density estimation. With either the histogram estimator or the kernel density estimator, it was easy to measure the *sensitivity* of the estimator; that is, how much the estimator changed when one person's data changed. However, with regression estimators of the form (1), calculating the sensitivity is much more, since changing one pair (X_i, Y_i) and affects both the contribution to the sum and the normalization due to the constraint

$$\sum_{j=1}^n \ell_j(x) = 1.$$

This aspect makes it much more difficult to measure the sensitivity of the function accurately. An overestimate of the sensitivity will result in an excessive amount of noise.

The literature currently presents limited methods that have successfully produced accurate DP estimates for NP regression problems. One solution is to simplify the problem by treating the independent variables as public, and only protecting the dependent variable. Figure 4b is an example of this approach.

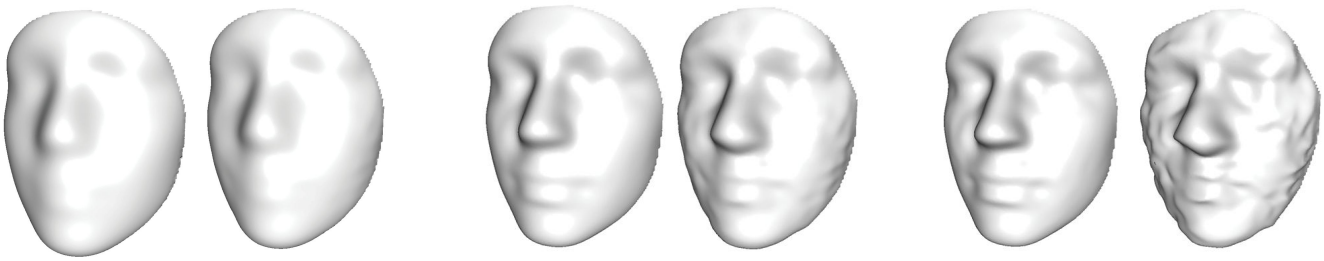


Figure 5. Pairs of RKHS estimates (first, third, fifth) and their sanitized versions (second, fourth, sixth) for three different levels of smoothing. Images produced by Ardalan Mirshani.

This approach was employed in *Differentially Private Regression with Gaussian Processes* (Álvarez, Zwiessele, and Lawrence. 2018) and enables a tight sensitivity calculation, more similar to that of a kernel density estimator. They use a Gaussian process regression, which treats all independent variables as public, and the tools of Hall, Rinaldo, and Wasserman to privatize the regression function.

This approach applies when the independent variables are publicly known (such as spatial locations), and only the dependent variables are sensitive. However, more-sophisticated techniques are required when all variables are at risk.

Frontiers of Nonparametric Statistics

While the problems of density estimation and nonparametric regression tackle a wide variety of settings, these problems can still be viewed as working in (infinite-dimensional) linear spaces. A significant increase in complexity arises when working in *nonlinear* spaces. Some tools for working on such problems come from shape analysis, manifold learning, and topological data analysis.

To understand these problems, consider two examples: covariance matrix estimation and a data set of 3-D images of human faces.

Covariance Estimation: Recall that for an m-dimensional data set, the covariance matrix is a positive-semidefinite symmetric $m \times m$ matrix. Many differential privacy methods designed for the release of a privatized covariance matrix include the addition of noise to the empirical covariance estimator. However, a significant limitation of this approach, especially in smaller sample sizes, is that this noise can result in a matrix that is no longer positive-semidefinite.

An approach that has yet to be pursued is to produce a DP family of distributions on the *manifold* of positive-semidefinite matrices. A manifold is a space that behaves locally like Euclidean space, but may not be generally closed under addition or scalar multiplication. The space of covariance matrices is closed under addition, but not under subtraction or multiplication by negative scalars.

Another difference between working in the manifold of covariance matrices and standard linear spaces is the choice of metric. Typically, standard matrix norms are used to evaluate the distance between covariance matrices. While this metric is of some use, it may not be the best choice to truly capture the similarity of two covariance matrices. When working with manifolds, there is a natural distance called a *Riemannian metric*, which is better at capturing the geometry of the space.

Incorporating tools specifically designed for manifolds could result in better-performing DP covariance estimation methods.

Shape Data: An exciting data set highlights many of the possibilities of privacy tools for complex data structures. Mark Shriver at the Pennsylvania State University collected a database of 3-D scans of human faces, along with extensive demographic and genomic data.

This data set is inherently sensitive, with stringent data management regulations. While Gaussian processes can be used to privatize the faces, such noise can distort the faces so significantly that they no longer look like real people (see Figure 5). Here it is particularly important to be able to develop privacy tools that remain in the *manifold of faces*.

This data set may require the incorporation of tools from shape analysis, manifold learning, and differential geometry, pushing privacy tools to the frontier of NP statistics.

Discussion

Nonparametric methods provide flexible and powerful tools for scientists and statisticians in working with complex data structures. However, these methods have unique privacy challenges.

One aspect of nonparametric methods is their dependence on tuning parameters. Histograms depend on the bin width, and both kernel density estimators and linear smoothers depend on a bandwidth. In practice, these parameters are often chosen based on cross-validation, which enhances the choice of bandwidth to optimize the fit. However, for differential privacy, even these tuning parameters must be chosen in a formally private way.

To simplify the examples, it has been assumed that the tuning parameters were publicly known, but in practice, these parameters are usually data-dependent. While there have been a few DP mechanisms to estimate these parameters, this is an area of research that requires more attention.

Instead of using a Guassian process to privatize an NP estimator directly, an alternative approach is to use \emph{objective perturbation}. Chaudhuri and Monteleoni introduced objective perturbation in *Privacy Preserving Logistic Regression*, and it has been adapted and modified in several other works. Instead of adding noise to the resulting estimator directly, objective perturbation mechanisms add noise to the objective function before optimizing. This approach has seen much success for regression problems, where the estimators have complex formulae but are the solutions to a well-behaved objective function.

Indeed, in the case of local polynomial regression, the objective function is of a form that allows for a simple sensitivity calculation. Formalizing this approach to allow the release of the full NP regression function is a matter for future work, but may be a promising direction.

Another direction for future work is investigating different noise-adding distributions. The examples in this article use Gaussian process noise with the exponential kernel to privatize the density and regression estimators. This approach can be modified slightly by choosing a different kernel to preserve smoothness properties of the original estimator. Future research may quantify the benefits and drawbacks of different noise-generating distributions in multivariate settings, as well as function spaces.

Further Reading

Chaudhuri, K., and Monteleoni, C. 2009. Privacy-preserving logistic regression. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems* 21, 289–296. Red Hook, NY: Curran Associates, Inc.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. 2006. *Calibrating Noise to Sensitivity in Private Data Analysis*, 265–284. Berlin and Heidelberg, Germany: Springer Berlin Heidelberg.

Dwork, C., Roth, A., et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science* 9(3–4):211–407.

Hall, R., Rinaldo, A., and Wasserman, L. 2013. Differential privacy for functions and functional data. *Journal of Machine Learning Research* 14(1): 703–727.

Mirshani, A., Reimherr, M., and Slavkovic, A. 2019. Formal privacy for functional data with gaussian perturbations. In *International Conference on Machine Learning*, 4,595–4,604.

Smith, M., Álvarez, M., Zwiessele, M., and Lawrence, N.D. 2018. Differentially private regression with gaussian processes. In *International Conference on Artificial Intelligence and Statistics* 1,195–1,203.

Wasserman, L. 2006. *All of nonparametric statistics*. Berlin, Heidelberg, and Dordrecht, Germany; and New York: Springer Science & Business Media.

About the Authors

Jordan Awan is an assistant professor in the Department of Statistics at Purdue University. He studied at Clarion University from 2011–14, earning a BS in mathematics, and completed an MA in mathematics at Brandeis University in 2016 under the advisement of Olivier Bernardi. He obtained his PhD in statistics at Penn State University, advised by Aleksandra Slavković and Matthew Reimherr. His research focuses on differential privacy, with applications to functional data analysis, classical statistical inference, and empirical risk minimization. He has also collaborated on several projects related to pitch tracking for the analysis of physiological signals.

Matthew Reimherr is an associate professor in the Department of Statistics at Penn State. He obtained a PhD in statistics from the University of Chicago and an MS in Statistics and a BS in mathematics from the University of Utah. He is currently an associate editor of *Statistical Modelling*, the *Journal of Multivariate Analysis*, *Annals of Applied Statistics*, and the *Journal of the Royal Statistical Society: Series B*. Reimherr's work focuses on problems in functional data analysis, longitudinal data analysis, statistical genetics, high dimensional regression and screening, data privacy, and shape analysis.

Aleksandra (Seša) Slavković is a professor of statistics and associate dean for graduate education in the Eberly College of Science at Penn State. She earned her PhD in statistics from Carnegie Mellon University. Her research focuses on methodological developments in data privacy and confidentiality in the context of small and large-scale surveys, health, genomic, and network data, including differential privacy and broad data access. Her other research interests include evaluation methods for human performance in virtual environments, application of statistics to information sciences and social sciences, algebraic statistics, and causal inference. She served as a chair of the ASA Privacy and Confidentiality Committee, and is chair-elect 2021 for the ASA Social Statistics Section.