Formal Privacy for Functional Data with Gaussian Perturbations

Ardalan Mirshani [1] **Matthew Reimherr** [1] **Aleksandra Slavkovic** [1]

Abstract

Motivated by the rapid rise in statistical tools in *Functional Data Analysis*, we consider the Gaussian mechanism for achieving differential privacy (DP) with parameter estimates taking values in a, potentially infinite-dimensional, separable Banach space. Using classic results from probability theory, we show how densities over function spaces can be utilized to achieve the desired DP bounds. This extends prior results of Hall et al. (2013) to a much broader class of statistical estimates and summaries, including "path level" summaries, nonlinear functionals, and full function releases. By focusing on Banach spaces, we provide a deeper picture of the challenges for privacy with complex data, especially the role regularization plays in balancing utility and privacy. Using an application to penalized smoothing, we highlight this balance in the context of mean function estimation. Simulations and an application to diffusion tensor imaging are briefly presented, with extensive additions included in a supplement.

1. Introduction

New studies, surveys, and technologies are resulting in ever richer and more informative data sets. Data being collected as part of the "big data revolution" have dramatically expanded the pace of scientific progress over the last several decades, but often contain a significant amount of personal or subject level information. These data and their corresponding analyses present substantial challenges for preserving privacy as researchers attempt to understand what information can be publicly released without impeding scientific advancement and policy making (Lane et al., 2014).

One type of big data that has been heavily researched in the statistics community over the last two decades is *functional data*, with the corresponding branch of statistics called *functional data analysis*, FDA. FDA is concerned with conducting statistical inference on samples of functions, trajectories, surfaces, and other similar objects. Such tools have become increasingly necessary as our data gathering technologies become more sophisticated. FDA methods have proven very useful in a wide variety of fields including economics, finance, genetics, geoscience, anthropology, and kinesiology, to name only a few (Ramsay & Silverman, 2002; 2005; Ferraty & Romain, 2011; Horváth & Kokoszka, 2012; Kokoszka & Reimherr, 2017). Indeed, nearly any data rich area of science will eventually come across applications that are amenable to FDA techniques. However, functional and other high dimensional data are also a rich source of potentially personally identifiable information (Kulynych, 2002; Erlich & Narayanan, 2014; Schadt et al., 2012).

Related Work: To date, there has been very little work concerning FDA and statistical data privacy, in either *Statistical Disclosure Limitation*, SDL or *Differential Privacy*, DP. SDL is the branch of statistics concerned with limiting identifying information in released data and summaries while maintaining their utility for valid statistical inference, and has a rich history for both methodological developments and applications for "safe" release of altered (or masked) microdata and tabular data (Dalenius, 1977; Rubin, 1993; Willenborg & De Waal, 1996; Fienberg & Slavković, 2010; Hundepool et al., 2012). DP has emerged from theoretical computer science with a goal of designing privacy mechanisms with mathematically provable disclosure risk (Dwork, 2006; Dwork et al., 2006b). Hall et al. (2013) provide the most substantial contribution to statistical privacy with FDA to date, working within the DP framework and the Gaussian mechanism for releasing a finite number of point-wise evaluations, with applications to kernel density estimation and support vector machines. They provide a limiting argument that establishes DP for certain sets of functions. One of the major findings of Hall et al. (2013) is the connection between DP and Reproducing Kernel Hilbert Spaces, which we extend more broadly to *Cameron-Martin Spaces*. Recently, Aldà & Rubinstein (2017) extended the work of Hall et al. (2013) by considering a Laplace (instead of a Gaussian) mechanism and focused on releasing an approximation based on Bernstein polynomials, exploiting their close connection to point-wise evaluation on a grid or mesh.

*Equal contribution [1]Department of Statistics, Pennsylvania State University, State College, PA, USA. Correspondence to: Matthew Reimherr <mreimherr@psu.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

Other related contributions include Alvim et al. (2018) who consider privacy over abstract metric spaces assuming one has a sanitized dataset, and Smith et al. (2018) who examine how to best tailor the mechanism from Hall et al. (2013).

Our Contribution: In this work, we move beyond Hall et al. (2013), Aldà & Rubinstein (2017) and Smith et al. (2018) by developing a DP mechanism for functional data much more broadly. We first show that the Gaussian mechanism achieves DP for a large class of linear functionals and then show that this mechanism offers seemingly complete protection against any summary imaginable, covering any "path level" summaries (such as integrals and derivatives), nonlinear transformations, or even a full function release, though the later is usually not computationally feasible without some additional structure (e.g., continuous time Markov chains). Such extensions are critical for working with transformations that are not simple point wise evaluations, such as basis expansions, norms, and derivatives or when the objects exhibit complex nonlinear dynamics. We also provide an interesting negative result, that shows that not all Gaussian noises are capable of achieving DP for a particular summary, regardless of how the noise is scaled. In particular, we introduce a concept called *compatibility*, and show that if a particular summary is not compatible with a Gaussian noise, then it is impossible to achieve DP with that particular process. To establish the necessary probabilistic bounds for DP we utilize functional densities via the *Cameron-Martin Theorem*. This is also of independent interest in FDA as densities for functional data are rarely utilized due to the lack of a natural base measure (Berrendero et al., 2018). Most attempts at utilizing or defining densities for functional data involve some work-around to avoid working in infinite dimensions (Delaigle & Hall, 2010; Dai et al., 2017). Lastly, we demonstrate these tools by considering mean function estimation via penalized smoothing, where we also provide guarantees on the utility of the sanitized estimate.

One of the major findings of this work is the interesting connection between *regularization* and privacy. We show that by slightly over smoothing, one can achieve DP with substantially less noise, thus better preserving the utility of the release. This is driven by the fact that a great deal of personal information can reside in the "higher frequencies" of a functional parameter estimate, while the "lower frequencies" are typically shared across subjects. To more fully illustrate this point, we demonstrate how a cross-validation for choosing smoothing parameters can be dramatically improved when the cross-validation incorporates the function to be released. Previous works concerning DP and regularization have primarily focused on performing shrinkage regression in a DP manner (e.g. Kifer et al., 2012; Chaudhuri et al., 2011) and model selection with linear regression (e.g., Lei et al. (2018)), not exploiting the regularization to recover some utility as we propose here.

Organization: The remainder of the paper is organized as follows. In Section 2 we give the necessary background on DP and FDA. In Section 3 we present our chief results concerning releasing a finite number of linear functionals followed by full function and nonlinear releases. Section 4 has an application on penalized smoothing for mean estimation, which is especially amenable to our privacy mechanism. In Section 5 simulations highlight the role of different parameters, while Section 6 contains an application of Diffusion Tensor Imaging of Multiple Sclerosis patients. In Section 7 we discuss our results and present concluding remarks.

2. Background

2.1. Differential Privacy

Differential Privacy, DP, was introduced in Dwork et al. (2006b). Let $\mathbb{D}$ be a (potentially infinite) population of records, and denote by $\mathcal{D}$ the collection of all n-dimensional subsets of $\mathbb{D}$. Throughout we let D and D' denote elements of $\mathcal{D}$. Notationally, we omit the dependence on n for ease of exposition. We work with (ϵ, δ)-DP, where $\epsilon \in \mathbb{R}^+$ and $\delta \in \mathbb{R}^+$ are parameters representing the *privacy budget* with smaller values indicating stronger privacy; when $\delta = 0$ one has pure or ϵ-DP. DP is a property of the privacy mechanism applied to the data summary, in this case $\theta_D := \theta(D)$, prior to release. For simplicity, we will denote the application of this mechanism using a tilde; so $\tilde{\theta}_D := \tilde{\theta}(D)$ is the *sanitized* version of θ_D. Probabilistically, we view $\tilde{\theta}_D$ as a random variable indexed by D (which is not treated as random). This criteria can be defined for any probability space.

Definition 2.1 (Dwork et al. (2006b); Wasserman & Zhou (2010))**.** *Let $\theta : \mathcal{D} \to \Omega$, where $(\Omega, \mathcal{F})$ is some measurable space. Let $\tilde{\theta}_D$ be random variables, indexed by D, taking values in Ω and representing the privacy mechanism. The privacy mechanism is said to achieve $(\epsilon, \delta)-$DP if for any two datasets, D and D', which differ in only one record, we have*

$$P(\tilde{\theta}_D \in A) \leq P(\tilde{\theta}_{D'} \in A)e^{\epsilon} + \delta,$$

for any measurable set $A \in \mathcal{F}$.

In Section 3.1 we take $\Omega = \mathbb{R}^N$, corresponding to releasing N linear functionals of a functional object, while in 3.2 we consider a real separable Banach space when $\Omega = \mathbb{B}$. In Hall et al. (2013), they consider the space of real valued functions over $\mathbb{R}^d$, i.e., the product space $\Omega = \mathbb{R}^T$ with $T = \mathbb{R}^d$ (or some compact subset), by clever limiting arguments of cylindrical sets; they thus considered DP over $\mathbb{R}^T$ equipped with the cylindrical σ-algebra (i.e. the smallest σ-algebra that makes point-wise evaluations measurable). However, in most cases we are actually interested in a subspace of $\mathbb{R}^T$, such as the space of continuous functions, square integrable functions, differentiable functions, etc. It turns out that the resulting σ-algebras (and thus the protection offered by

DP) are in general quite different, and that working directly with $\mathbb{R}^T$ can result is some fairly glaring holes. Chapter 7 of Billingsley (2008) or Section 3.1 of Bogachev (1998) discuss these issues, but it is interesting to note that the cylindrical σ-algebra on $\mathbb{R}^T$ is missing the sets of linear functions, polynomials, constants, nondecreasing functions, functions of bounded variation, differentiable functions, analytic functions, continuous functions, functions continuous at a given point, and Borel measurable functions. To avoid this issue, we work directly with the Borel σ-algebra on the function space of interest, which in our case is always a Banach space, though in principle this approach can be extended to handle any locally convex vector space.

At a high level, achieving $(\epsilon, \delta)-$DP means that the object to be released changes relatively little if the sample on which it is based is perturbed slightly. This change is related to what Dwork (2006) called *sensitivity*. Another nice feature is that if $\tilde{\theta}_D$ achieves DP, then so does any measurable transformation of it; see Dwork et al. (2006a;b) for the original results, Wasserman & Zhou (2010) for its statistical framework, and Dwork & Roth (2014) for a more recent detailed review of relevant DP results.

2.2. Functional Data Analysis

Much of FDA is built upon the *Hilbert space* approach to modeling, viewing data and/or parameters as elements of a complete inner product space (most commonly $L^2[0,1]$ after possibly rescaling). However, we take a more general approach by allowing for arbitrary separable Banach spaces, i.e., a complete normed vector space, which will dramatically increase the application of our results, while requiring only a small amount of more technical work. All of the concepts/tools from this section are classic probability theory results that might be of interest in the FDA and privacy communities. We refer the interested reader to Bogachev (1998) for a nearly definitive treatment of Gaussian measures. Throughout we let $\mathbb{B}$ denote a real separable Banach space; we always implicitly assume that $\mathbb{B}$ is equipped with its Borel σ-algebra, which is the smallest σ-algebra containing the open sets.

Let $\theta : \mathcal{D} \to \mathbb{B}$ denote the particular summary of interest and for notational ease, we define $\theta_D := \theta(D)$. In Section 3.1 we consider the case where the aim is to release a finite number of linear functionals of θ_D, whereas in Section 3.2 we consider releasing sanitized versions of the entire function or some nonlinear transformation of it (such as a norm or basis expansion).

The backbone of our privacy mechanism is the same as in Hall et al. (2013), and is used extensively across the DP literature. In particular, we add Gaussian noise to the summary and show how the noise can be calibrated to achieve DP. A random process $X \in \mathbb{B}$ is called *Gaussian* if $f(X)$ is Gaussian in $\mathbb{R}$, for any continuous linear functional $f \in \mathbb{B}^*$ (Bogachev, 1998, Def. 2.2.1). Throughout we use $*$ to denote the corresponding topological dual space. Equipped with the norm $\|f\|_{\mathbb{B}^*} = \sup_{\|h\|_{\mathbb{B}} \leq 1} f(h)$, the dual space is also a separable Banach space. The pair $(\mathbb{B}, \nu)$ is often called an *abstract Weiner space* (Bogachev, 1998, Sec. 3.9), where ν is the probability measure over $\mathbb{B}$ induced by X. Every Gaussian process is uniquely parametrized by a mean, $\mu \in \mathbb{B}$, and a covariance operator $C : \mathbb{B}^* \to \mathbb{B}$, which for every $f \in B^*$ satisfies

$$\mathrm{E}[f(X)] = f(\mu), \qquad C(f) = \mathrm{E}[f(X-\mu)(X-\mu)]$$

(Laha & Rohatgi, 1979). One can equivalently identify C as a bilinear form $C(f,g) = \mathrm{Cov}(f(X), g(X))$, and we will use both notations whenever convenient. It follows that

$$f(X) \sim \mathcal{N}(f(\mu), C(f,f)),$$

for any $f \in \mathbb{B}^*$. We use the short hand notation $\mathcal{N}$ to denote the Gaussian distribution over $\mathbb{R}$, but include subscripts for any other space, e.g., $\mathcal{N}_{\mathbb{B}}$ for $\mathbb{B}$.

A key object concerning privacy will be the *Cameron-Martin space* (Bogachev, 1998, Sec. 2.4) of X (or equivalently of $(\mathbb{B}, \nu)$). Using C one can equip $\mathbb{B}^*$ with an inner product

$$\begin{aligned}\langle f, g\rangle_{\mathcal{K}} &:= \mathrm{Cov}(f(X), g(X)) \\ &= \int f(x-\mu) g(x-\mu)\, d\nu(x).\end{aligned}$$

However, $\mathbb{B}^*$ is no longer complete under this inner product; denote the completed space as $\mathcal{K}$. Finally, consider the set of all $h \in \mathcal{H} \subset \mathbb{B}$ such that the mapping, $f \to f(h)$, is continuous in the $\mathcal{K}$ topology. Intuitively, these functions are ones that are "nicer" than arbitrary elements of $\mathbb{B}$. In particular, they must be regular enough to ensure that $f(h)$ is finite for any $f \in \mathcal{K}$, which are much "uglier" functionals than those in $\mathbb{B}^*$. By the Riesz representation theorem, we can associate each element $h \in \mathcal{H}$ with a $T_h \in \mathcal{K}$ such that $\langle T_h, f\rangle_{\mathcal{K}} = f(h)$. The set $\mathcal{H}$ equipped with the inner product

$$\langle x, y\rangle_{\mathcal{H}} = \langle T_x, T_y\rangle_{\mathcal{K}},$$

is called the *Cameron-Martin Space*, and is itself a separable Hilbert space. Note that, slightly less abstractly, we have $C(T_h) = h$ (Bogachev, 1998, Lemma 2.4.1). One can also view $\mathcal{K}$ as being a type of *Reproducing Kernel Hilbert Space* (Bogachev, 1998, pg. 44) in a very broad sense since we have $\langle T_h, f\rangle_{\mathcal{K}} = f(h)$, for any $f \in \mathcal{K}$. In infinite dimensions the Cameron-Martin space does not contain the sample paths of X, but they can be thought of as "living at the boundary" of $\mathcal{H}$. While the Cameron-Martin space is introduced via Gaussian processes, it is determined entirely by the covariance operator C.

2.3. Hilbert Space Example

While working with a general Banach space allows for a broader impact, it is also conceptually much more challenging. We can gain additional insight by considering what happens when $\mathbb{B} = \mathbb{H}$ is a Hilbert space. By the Riesz Representation Theorem, which characterizes continuous linear functionals, $\mathbb{H}$ is isomorphic to $\mathbb{H}^*$ so we can always identify $\mathbb{H}^*$ with $\mathbb{H}$ and de-emphasize the linear functionals.

We can obtain very convenient expressions if we take a basis $\{v_i : i = 1, 2, \dots\}$ of $\mathbb{H}$ consisting of the eigenfunctions of C (recall in Hilbert spaces C must be nonnegative definite and trace class). In this case we have that

$$C(v_i) = \lambda_i v_i \quad \text{where} \quad \lambda_i > 0.$$

Assuming that that there are no zero eigenvalues ($\lambda_i \neq 0$), define $e_i = \lambda_i^{-1/2} v_i$, then these form an orthonormal basis of $\mathcal{K}$ as

$$\langle e_i, e_j \rangle_{\mathcal{K}} = \lambda_i^{-1/2} \lambda_j^{-1/2} \operatorname{Cov}(\langle v_i, X \rangle_{\mathbb{H}}, \langle v_j, X \rangle_{\mathbb{H}}) = \delta_{ij},$$

where δ_{ij} is 1 if $i = j$ and zero otherwise. The space $\mathcal{K}$ consists of all linear combinations of the e_i whose coefficients are square summable. The inner product on Cameron-Martin space, $\mathcal{H}$, is given by

$$\langle x, y \rangle_{\mathcal{H}} = \sum \frac{\langle x, v_i \rangle_{\mathbb{H}} \langle y, v_i \rangle_{\mathbb{H}}}{\lambda_i},$$

so that

$$\mathcal{H} := \left\{ x \in \mathbb{H} : \sum_{i=1}^{\infty} \frac{\langle x, v_i \rangle_{\mathbb{H}}^2}{\lambda_i} < \infty \right\}. \tag{1}$$

In other words, those elements of $\mathcal{H}$ are the functions whose coefficients in the v_i basis decrease sufficiently quickly. Note that the case where some λ_i are actually zero (meaning C has a nontrivial null space) can be easily handled by restricting $\mathcal{H}$ to the range of C.[1]

The space $\mathcal{H}$ is a Hilbert space when equipped with the inner product $\langle x, y \rangle_{\mathcal{H}} = \sum \lambda_i^{-1} \langle x, v_i \rangle \langle y, v_i \rangle$. When $\mathbb{H} = L^2[0, 1]$ and C is an integral operator with continuous kernel $c(t, s)$, then $\mathcal{H}$ is isomorphic to a Reproducing Kernel Hilbert Space, RKHS (Berlinet & Thomas-Agnan, 2011) (one has to be slightly careful as L^2 consists of equivalence classes), meaning $c_t(s) \in \mathcal{H}$ for all t when viewed as a function of s and $\langle c_t, f \rangle_{\mathcal{H}} = f(t)$ for all $f \in \mathcal{H}$.

3. Privacy Enhanced Functional Data

In this section we present our main results. The mechanism we use for guaranteeing DP is to add a Gaussian noise before releasing θ_D; our release is based on a private version $\tilde{\theta}_D = \theta_D + \sigma Z$, where Z is a Gaussian process and σ is a constant determined by the sensitivity and privacy budget. However, it turns out that not just any Gaussian noise, Z, can be used. In particular, the options for choosing Z depend heavily on the summary θ. This is made explicit in Definition 3.1.

Definition 3.1. *We say that the summary θ is* **compatible** *with a Gaussian noise, $Z \sim \mathcal{N}_{\mathbb{B}}(0, C)$, if $\theta_D := \theta(D)$ resides in the Cameron-Martin space of Z for every $D \in \mathcal{D}$.*

Intuitively, this means that the noise must be "rougher" than the summaries. Our next definition is a generalization of one from Hall et al. (2013), which focused on functions in RKHS only.

Definition 3.2. *The global sensitivity of a summary θ, with respect to a Gaussian noise $Z \sim \mathcal{N}_{\mathbb{B}}(0, C)$ is given by*

$$\Delta^2 = \sup_{D' \sim D} \|\theta_D - \theta_{D'}\|_{\mathcal{H}}^2,$$

where $D' \sim D$ means the two sets differ at one record only, and $\|\cdot\|_{\mathcal{H}}$ is the norm on the Cameron-Martin space of Z.

The *global sensitivity (GS)* is a central quantity in the theory of DP; the amount of noise, σZ, depends directly on the global sensitivity. Here we focus on the global sensitivity that typically leads to the *worst case* definition of risk under DP; for a detailed review of DP theory and concepts, including other notions of "sensitivity", such as local sensitivity, see Dwork & Roth (2014). If a summary is not compatible with a noise, then it is possible to make the global sensitivity infinite, in which case no finite amount of noise would be able to preserve privacy in the sense of satisfying DP. Interestingly, sensitivity is computed with the Cameron-Martin norm, which can be convenient as it is a Hilbert space norm and avoids the original Banach space norm.

Theorem 3.1. *If a summary θ is not compatible with a noise $Z \sim N_{\mathbb{B}}(0, C)$ then for any $\sigma > 0$, $\tilde{\theta}_D := \theta_D + \sigma Z$* **will not** *satisfy DP.*

Proof. This is a direct consequence of the Cameron-Martin Theorem, which characterizes the equivalence/orthogonality of Gaussian measures. Two measures are said to be equivalent if they agree on sets of measure zero and orthogonal if they concentrate on disjoint sets. If the summary is not compatible with the noise, then there exists a $D \sim D'$ such that $\|\theta_D - \theta_{D'}\|_{\mathcal{H}} = \infty$, which implies that the distributions $\tilde{\theta}(D)$ and $\tilde{\theta}(D')$ are orthogonal. Since the measures are orthogonal, it means that there exists a set A such that $P(\tilde{\theta}(D) \in A) = 1$ and $P(\tilde{\theta}(D') \in A) = 0$, which means that $\tilde{\theta}_D$ is not differentially private for $\delta < 1$. □

Intuitively, if the summary is not compatible with the noise, then one can pool even small amounts of information from

[1] In fact, such a game can be played quite broadly as any Radon measure over a *Fréchet space* will concentrate on a reflexive separable Banach space (Bogachev, 1998, Thm 3.6.5).

across an infinite number of dimensions to produce a disclosure. An example where one would have $\|\theta_D - \theta_{D'}\|_{\mathcal{H}} = \infty$ would be if $\mathbb{B} = L^2[0,1]$, θ_D only possessed one derivative, but the paths of Z possessed two derivatives. However, we stress that this is very specific to Gaussian processes; other privacy mechanisms may have other forms of compatibility and sensitivity that become critical in infinite dimensions.

3.1. Releasing Finite Projections

We begin with the comparatively simpler task of releasing a finite vector of linear functionals of θ_D. In particular, we aim to release $f(\theta_D) = \{f_1(\theta), \ldots, f_N(\theta_D)\}$, for $f_i \in \mathcal{K} \supset \mathbb{B}^*$ and some fixed N. While placed in a more general context, the core concepts involved are the same as in Hall et al. (2013) (they focused on point-wise evaluations, which are continuous linear functionals over an appropriate space). Since we are using the Cameron-Martin space, we can actually release more than just continuous linear functionals; we can release any functional from $\mathcal{K}$, which is, in general, much larger than $\mathbb{B}^*$.

Theorem 3.2. *Assume θ is compatible with $Z \sim \mathcal{N}(0, C)$, $\epsilon \leq 1$, and define*

$$\tilde{\theta}_D = \theta_D + \sigma Z \qquad \textit{with} \qquad \sigma^2 \geq \frac{2\log(2/\delta)}{\epsilon^2}\Delta^2.$$

Now define $f(\theta_D) = \{f_1(\theta), \ldots, f_N(\theta_D)\}$ and $\tilde{f}(\theta_D) = \{f_1(\tilde{\theta}_D), \ldots, f_N(\tilde{\theta}_D)\}$, for $\{f_i \in \mathcal{K}\}_{i=1}^N$. Then $\tilde{f}_D$ achieves (ϵ, δ)-DP in $\mathbb{R}^N$.

Theorem 3.2 can be viewed as an extension of Hall et al. (2013) who focus on point-wise releases. If $\mathbb{B}$ is taken to be the space of continuous functions with an appropriate topology, then Theorem 3.2 implies point-wise releases are protected as well. However, this theorem allows the release of any functional in $\mathcal{K}$. This dramatically increases the release options and applications as compared to Hall et al. (2013) or Aldà & Rubinstein (2017).

3.2. Full Function and Nonlinear Releases

While Section 3.1 covers a number of important cases, it does not cover all releases of potential interest. In particular, full function releases are not protected and neither are nonlinear releases, such as norms or derivatives. A full function release is not often practically possible. However in some situations, such as continuous time Markov chains, full paths can be completely summarized using a finite number of values, but these values are not simple point-wise evaluations or linear projections and thus not covered under Hall et al. (2006); Aldà & Rubinstein (2017) or our results from Section 3.1. Still, there is a certain comfort in knowing that one has a complete protection that holds regardless of whatever special structures one might be able to exploit or new computational tools that might become available. In addition, one can obtain a great deal of insight by considering the infinite dimensional problem, as it highlights the fundamental role smoothing plays when trying to maintain utility while achieving DP.

To guarantee privacy for these types of releases, we need to establish (ϵ, δ)-DP for the entire function, not just finite projections. This means that in Definition 2.1, the space is taken to be $\mathbb{B}$, which is infinite dimensional. Previous works, e.g., Dwork et al. (2014); Hall et al. (2013), establish the probability inequalities as in Definition 2.1, using bounds based on multivariate normal densities. This presents a serious problem for FDA and infinite dimensional spaces as it becomes difficult to work with such objects (there is very little FDA literature that does so). For example, Delaigle & Hall (2010) define densities only for finite "directions" of functional objects, and Bongiorno & Goia (2015) define psuedo-densities by carefully controlling "small ball" probabilities. Both works claim that for a functional object the density "generally does not exist." However, this turns out to be a *technically* incorrect claim, while still often being true in spirit. The correct statement is that, in general, it is difficult to define a *useful* density for functional data. In particular, to work with likelihood methods, a family of probability measures should all have a density with respect to the same base measure, which, at present, does not appear to be possible in general for functional data.

The difficulty in defining densities in infinite-dimensional spaces comes from the fact there is no common *base* or *reference* measure (Cuevas, 2014), such as Lebesgue measure, however our goal in using densities is more straightforward. We require densities (with respect to the same base measure) for the family of probability measures induced by $\{\theta_D + \sigma Z : D \in \mathcal{D}\}$, where Z is a mean zero Gaussian process in $\mathbb{B}$ with covariance operator C. It turns out that this is in fact possible because we are adding the exact same type of noise to each element. We give the following lemma, which is a rephrasing of the classic Cameron-Martin formula (Bogachev, 1998, Corollary 2.4.3).

Lemma 3.1. *Assume that the summary θ is compatible with a noise Z. Denote by Q the probability measure over $\mathbb{B}$ induced by σZ, and by $\{P_D : D \in \mathcal{D}\}$ the family of probability measures over $\mathbb{B}$ induced by $\theta_D + \sigma Z$. Then every measure P_D has a density over $\mathbb{B}$ with respect to Q, which is given by*

$$\frac{dP_D}{dQ}(x) = \exp\left\{-\frac{1}{2\sigma^2}\left(\|\theta_D\|_{\mathcal{H}}^2 - 2T_{\theta_D}(x)\right)\right\},$$

Q almost everywhere. Recall that $\theta_D = C(T_{\theta_D})$ and that the density is unique up to a set of Q measure zero.

At this point we stress that the noise is chosen by the user; it is not a property of the data. The primary hurdle for the user is ensuring that the summary is compatible with the selected

noise. As we will see in Section 4, one can accomplish this by using specific estimators. Lemma 3.1 implies that, for any Borel measurable set $A \subset \mathbb{B}$ we have

$$P_D(A) = \int_A \frac{dP_D}{dQ}(x)\, dQ(x),$$

which we exploit in our proofs later on.

Now that we have a well defined density we can establish differential privacy for entire functions.

Theorem 3.3. *Assume θ is compatible with a noise Z and that $\epsilon \leq 1$, then $\tilde{\theta}_D := \theta_D + \sigma Z$ achieves (ϵ, δ)-DP over $\mathbb{B}$ (with the Borel σ-algebra), with σ defined in Theorem 3.2.*

We also have the following simple corollary, which is a consequence of the post-processing inequality (Dwork & Roth, 2014).

Corollary 3.1. *Let θ be compatible with a noise Z, and let f be any measurable transformation from $\mathbb{B} \to \mathcal{S}$, where $\mathcal{S}$ is a measurable space. Then $f(\theta_D + \sigma Z)$ achieves (ϵ, δ)-DP over $\mathcal{S}$, where σ is defined in Theorem 3.2.*

Together, Theorem 3.3 and Corollary 3.1 imply that the Gaussian mechanism gives very broad privacy protection for functional data and other infinite dimensional objects, as nearly any transformation or manipulation of the privacy enhanced release is guaranteed to maintain DP; this is known as a *post-processing* property (e.g., see Dwork & Roth (2014)).

4. Privacy for Mean Function Estimation

In this section we consider the problem of estimating a mean function μ from a sample $X_1, \ldots, X_n$ that are iid elements of $\mathbb{H}$ with $\mathrm{E}\, X_i = \mu \in \mathbb{H}$ and $\|X_i\|_{\mathbb{H}} \leq \tau < \infty$ for all i. We derive a bound on the global sensitivity as well as utility guarantees. In Section 5 and in the Supplemental we will illustrate how to produce private releases of mean function estimates based on RKHS smoothing in more specific settings. In Hall et al. (2013) one can also find examples for kernel density estimation and support vector machines.

As is usual in the DP literature, we assume that the data is standardized so that it is bounded, usually with $\tau = 1$. In this case, the sample mean $\hat{\mu} = n^{-1}\sum_{i=1}^n X_i$ is root-n consistent and asymptotically normal (Kokoszka & Reimherr, 2017). There are a multitude of methods for estimating smooth functions, however, a penalized approach is especially amenable to our privacy mechanism. In this case we define a penalty using the covariance of the noise, C. However, the penalty and noise kernels need not be exactly the same, and in particular, we assume that penalty uses C^η for some $\eta \geq 1$. Here C^η has the same eigenfunctions as C, but the eigenvalues have been raised the power η. This allows for greater flexibility in terms of smoothing and it is helpful for deriving utility guarantees. We define the penalized estimate of the mean μ

$$\hat{\mu} = \operatorname*{argmin}_{m \in \mathcal{H}} \frac{1}{N}\sum_{i=1}^N \|X_i - m\|_{\mathbb{H}}^2 + \phi \|m\|_\eta^2,$$

where ϕ is the *penalty parameter*. The norm $\|\cdot\|_\eta$ is defined as the Cameron-Martin norm of C^η. While the most natural candidate is $\eta = 1$, taking something slightly larger can actually help with statistical inference as we will see later on. Here, we can see the advantage of a penalized approach as it forces the estimate to lie in the space $\mathcal{H}$ which means that the compatibility condition, as discussed in theorems 3.1 and 3.2, is satisfied. A kernel different from the noise could be used, but one must be careful to make sure that the compatibility condition is met. If (λ_j, v_j) are the eigenvalue/function pairs of the C and $\{X_i = \sum_{j=1}^\infty x_{ij} v_j : i = 1, \ldots, N\}$, with $x_{ij} = \langle X_i, v_j \rangle_{\mathbb{H}}$, then the estimate can be expressed as

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^N \sum_{j=1}^\infty \frac{\lambda_j^\eta}{\lambda_j^\eta + \phi} x_{ij} v_j, \tag{2}$$

We then have the following result.

Theorem 4.1. *If the $\mathbb{H}$ norm of any element of the population is bounded by a constant $0 < \tau < \infty$ then the GS of $\hat{\mu}$ for $\eta \geq 1$ is bounded by*

$$\Delta_n^2 \leq \frac{4\tau^2}{N^2} \sup_j \frac{\lambda_j^{2\eta-1}}{(\lambda_j^\eta + \phi)^2}$$

or more simply

$$\Delta_n^2 \leq \frac{\tau^2}{N^2 \phi^{1/\eta}} \left[\frac{(2\eta - 1)^{2-1/\eta}}{\eta^2}\right] \leq \frac{4\tau^2}{N^2 \phi^{1/\eta}}.$$

The resulting bound is practically very useful. Data can be rescaled so that their $\mathbb{H}$ bound is, for example, 1, and then the remaining quantities are all tied to the used noise/RKHS. Thus, the bound can be practically computed and the corresponding releases are guaranteed to achieve DP.

We conclude with a final theorem that provides a guarantee on the utility of $\hat{\mu} + \sigma Z$. One interesting note is that in finite dimensional problems, the magnitude of the noise added for privacy is often of a lower order than the statistical error of the estimate. However, in infinite dimensions, this is no longer true unless $\eta > 1$. This is driven by the fact that the squared bias is of the order ϕ, and thus ϕ must go to zero like N^{-1} if it is to balance the variance of $\hat{\mu}$. However, in that case the magnitude of the noise added for privacy is of the order $\sigma^2 \asymp N^{-2+1/\eta}$. If $\eta = 1$, then σ^2 is also of the order N^{-1}, while if $\eta > 1$, then it is of a lower order and thus asymptotically negligible. We remind the reader that the noise and thus C is arbitrary, so η can be chosen in a way that is appropriate for μ by using a rougher noise.

Theorem 4.2. *Assume the X_i are iid elements of $\mathbb{H}$ with norm bounded by $\tau < \infty$. Define*

$$\tilde{\mu} := \hat{\mu} + \sigma Z,$$

where

$$\sigma^2 = \left[\frac{2\log(2/\delta)}{\epsilon^2}\right] \times \left[\frac{\tau^2(2\eta-1)^{2-1/\eta}}{N^2\phi^{1/\eta}\eta^2}\right].$$

If the tuning parameter, ϕ, satisfies $\phi \propto N^{-1}$ and if $\|\mu\|_\eta < \infty$ then we have

$$\mathrm{E}\,\|\tilde{\mu}-\hat{\mu}\|_{\mathbb{H}}^2 = o(N^{-1}) \qquad \textit{and} \qquad \mathrm{E}\,\|\tilde{\mu}-\mu\|_{\mathbb{H}}^2 = O\left(N^{-1}\right),$$

while $\tilde{\mu}$ achieves (ϵ-δ) DP in $\mathbb{H}$.

5. Empirical Study

Here we briefly present simulations with $\mathbb{B} = L^2[0,1]$ to explore the impact of parameters on the utility of sanitized releases. We consider the problem of estimating the mean function from a random sample of functional observations using RKHS smoothing, as discussed in Section 4.

For the RKHS, $\mathcal{H}$, we consider the Gaussian (squared exponential) kernel :

$$C_1(t,s) = \exp\left\{\frac{-|t-s|^2}{\rho}\right\} \tag{3}$$

We simulate data using the Karhunen-Loeve expansion, a common approach in FDA simulation studies. In particular we take

$$X_i(t) = \mu(t) + \sum_{j=1}^{m} j^{-p/2} U_{ij} v_j(t) \qquad t \in [0,1], \tag{4}$$

where the scores, U_{ij}, are drawn iid uniformly between $(-0.4, 0.4)$. The functions, $v_j(t)$, are taken as the eigenfunctions of C_1 and m was taken as the largest value such that λ_m was numerically different than zero in `R` (usually about $m = 50$). All of the curves are generated on an equally spaced grid between 0 and 1, with 100 points and the RKHS kernel and the noise kernel will be taken to be the same (i.e. $\eta = 1$). The range parameter for the kernel used to define $\mathcal{H}$ is taken $\rho = 0.001$ and the smoothness parameter of the $X_i(t)$ is set to $p = 4$. The mean function, sample size and DP parameters will also be set as $\mu(t) = 0.1\sin(\pi t)$, $N = 25$, $(\epsilon = 1, \delta = 0.1)$, respectively. We vary the penalty, ϕ, from 10^{-6} to 1 to consider its effect.

Note that we take $\tau = \sup \|X_i\|_{\mathbb{H}}$ for any $i \in 1, \ldots, N$ and thus all qualities needed for Theorem 4.1 are known. The risk is fixed by choosing the ϵ and δ in the definition of DP. We thus focus on the utility of the privacy enhanced curves by comparing them graphically to the original estimates. Ideally, the original estimate will be close to the truth and the privacy enhanced version will be close to the original estimate. What we will see is that by compromising slightly on the former, one can makes substantial gains in the latter.

In Figure 1 we plot all of the generated curves in gray, the RKHS smoothed mean in green, and the sanitized estimate in red. We can see that as the penalty increases, both estimates shrink towards each other and to zero. There is a clear "sweet spot" in terms of utility, where the smoothing has helped reduce the amount of noise one has to add to the estimate while not over smoothing. Further simulations that explore the impact of different parameters can be found in the supplemental B.

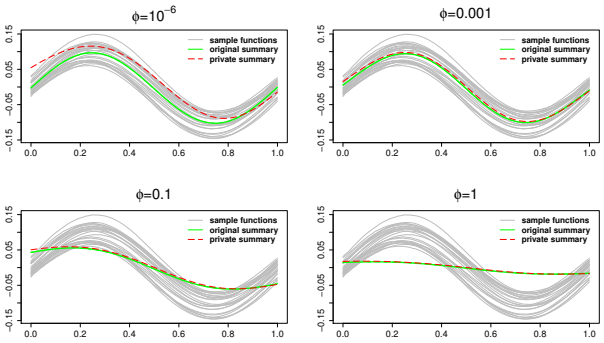

Figure 1. Original and private RKHS smoothing mean with Gaussian Kernel (C_1) for different values of penalty parameter ϕ

6. Applications

In this section we illustrate our method on an application involving brain scans (*diffusion tensor imaging*, DTI) that give fractional anisotropy (FA) tract profiles for the corpus callosum (CCA) and the right corticospinal tract (RCST) for patients with multiple sclerosis as well as controls; data are part of the `refund` (Huang et al., 2016) `R` package. Each profile/function consists of thickness measurements taken along the tract of the corresponding tissue. This type of imaging data is becoming more common and the privacy concerns can be substantial. Images of the brain or other major organs might be quite sensitive source of information, especially if the study is related to some complex disease such as cancer, HIV, etc. Thus it is useful to illustrate how to produce privacy enhanced versions of function valued statistics such as mean functions. We focus on the *CCC* data, which includes 382 patients measured at 93 equally spaced locations along the CCA.

Our aim is to release a sanitized RKHS estimate of the mean function. We consider three kernels C_1, C_3 and C_4 which correspond to the Gaussian kernel, Matérn kernel with $\nu = 3/2$, and the exponential kernel, respectively. Each kernel

is from the Matérn family of covariances (Stein, 2012). The exact forms are given in (3) in the supplement, where a fourth kernel C_2 is also considered that is "in between" C_1 and C_3 (hence the odd numbering). In all settings we take $(\epsilon, \delta) = (1, 0.1)$ and select the penalty, ϕ, and range parameter, ρ, according to two different approaches. The first is regular *Cross Validation*, CV, and the second we call *Private Cross Validation*, PCV. In CV we fix ϕ and then take the ρ that gives the minimum 10-fold cross validation score. We do not select ϕ based on cross validation because, based on our observations, the minimum score is always obtained at the minimum ϕ for this data. In PCV we take nearly the same approach, however, when computing the CV score we take the *expected* difference (via Monte-Carlo) between our privacy enhanced estimate and the left out fold from the original data. In other words, we draw a sample of privacy enhanced estimates, compute a CV score for each one, and then average the CV scores. In our simulations we use 1000 draws from the distribution of the sanitized estimate. We then find both the ϕ and ρ which give the optimal PCV score based on a grid search.

For the CV-based results, for each of the kernels, we fixed a value for $\phi \in \{0.0001, 0.001, 0.01, 0.03\}$ and then vary the ρ between $[0.01, 2]$. We use the optimal parameter values in Table 1 to produce the privacy enhanced estimates for C_1 in Figure 2. We see that the utility of the privacy enhanced versions increases as ϕ increases, however, the largest values of ϕ produce estimates that are over smoothed. There is a good trade-off between privacy and utility with $\phi = 0.01$ for C_1. The results for other kernels are reviewed in supplemental C.

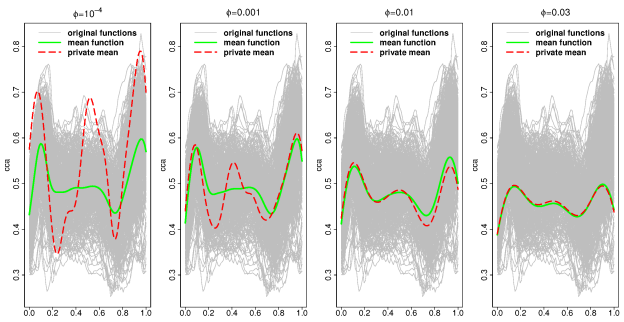

Figure 2. Mean estimate for cca and its private release using Gaussian kernel (C_1) with CV.

Turning to PCV, we varied ϕ in range $[10^{-4}, 0.1]$ for each of the kernels but ρ will be varied in $[0.01, 0.1]$, $[0.05, 0.5]$ and $[0.2, 1]$ for C_1,C_3 and C_4 respectively. Here we use the optimal parameters in Table 2 to generate privacy enhanced estimates, given in Figure 3. Here we see that the utility of the privacy enhanced estimates is excellent for C_1. Using PCV tends to over smooth the original estimates (green lines), however, by slightly over smoothing we make substantial gains in utility as we add less noise.

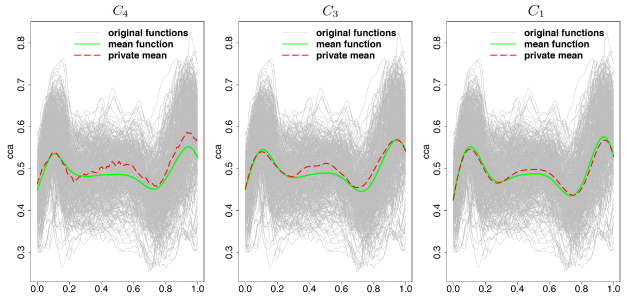

Figure 3. Mean estimate of CCA and its private release for Exponential (C_4), Matérn$(3/2)$ (C_3) and Gaussian kernels (C_1) using PCV.

7. Conclusions

In this work we provide a (ϵ, δ)-DP mechanism for a wide range of summaries related to functional parameter estimates. This work expands upon Hall et al. (2013), (Aldà & Rubinstein, 2017), and (Smith et al., 2018), who explored this topic in the context of point-wise releases of functions. Our work covers theirs as a special case, but also includes path level summaries, full function releases, and nonlinear releases quite broadly.In general, functional data can be highly identifiable compared to scalar data. In biomedical settings, for example, a study may collect and analyze many pieces of information such as genomic sequences, biomarkers, and biomedical images, which either alone or linked with each other and demographic information, lead to greater disclosure risk (Lippert et al., 2017).

The heart of our work utilizes densities for functional data in a way that has not yet been explored in the functional data literature. Previously, usable densities for functional data were thought not to exist (Delaigle & Hall, 2010) and researchers instead relied on various approximations to densities. We showed how useful forms for densities can be constructed and utilized. However, it is still unclear how extensively these densities can be used for other FDA problems.

The literature on privacy for scalar and multivariate data is quite extensive, while there has been very little work done for FDA and related objects. Therefore, there are many opportunities for developing additional theory and methods for such complicated data. One issue that we believe will be especially important is the role of smoothing and regularization in preserving the utility of privacy enhanced releases. As we have seen, a bit of extra smoothing can go a long way in terms of maintaining privacy, however, the type of smoothing may need to be properly tailored to the application for much complicated objects.

Acknowledgments

This research was supported in part by the following grants to the Pennsylvania State University: NSF Grant SES-1534433, NSF Grant DMS-1712826, and NIH UL1 TR002014. Part of this work was done while the second and third authors were visiting the Simons Institute for the Theory of Computing.

References

Aldà, F. and Rubinstein, B. I. The bernstein mechanism: Function release under differential privacy. In *AAAI*, pp. 1705–1711, 2017.

Alvim, M. S., Chatzikokolakis, K., Palamidessi, C., and Pazii, A. Metric-based local differential privacy for statistical applications. *arXiv preprint arXiv:1805.01456*, 2018.

Berlinet, A. and Thomas-Agnan, C. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

Berrendero, J. R., Cuevas, A., and Torrecilla, J. L. On the use of reproducing kernel hilbert spaces in functional classification. *Journal of the American Statistical Association*, 2018.

Billingsley, P. *Probability and measure*. John Wiley & Sons, 2008.

Bogachev, V. I. *Gaussian measures*. Number 62. American Mathematical Soc., 1998.

Bongiorno, E. and Goia, A. Classification methods for hilbert data based on surrogate density. *arXiv preprint arXiv:1506.03571*, 2015.

Chaudhuri, K., Monteleoni, C., and Sarwate, D. Differentially private empirical risk minimization. In *Journal of Machine Learning Research*, volume 12, pp. 1069–1109, 2011.

Cuevas, A. A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference*, 147:1–23, 2014.

Dai, X., Müller, H.-G., and Yao, F. Optimal bayes classifiers for functional data and density ratios. *Biometrika*, 104 (3):545–560, 2017.

Dalenius, T. *Statistik Tidskrift*, 15:429–444, 1977.

Delaigle, A. and Hall, P. Defining probability density for a distribution of random functions. *The Annals of Statistics*, 38:1171–1193, 2010.

Dwork, C. Differential privacy. In *33rd International Colloquium on Automata, Languages and Programming, part II (ICALP 2006)*, volume 4052, pp. 1–12, Venice, Italy, July 2006. Springer Verlag. ISBN 3-540-35907-9. URL https://www.microsoft.com/en-us/research/publication/differential-privacy/.

Dwork, C. and Roth, A. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, August 2014. ISSN 1551-305X. doi: 10.1561/0400000042. URL http://dx.doi.org/10.1561/0400000042.

Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., and Naor, M. Our data, ourselves: Privacy via distributed noise generation. In *Proceedings of the 24th Annual International Conference on The Theory and Applications of Cryptographic Techniques*, EUROCRYPT'06, pp. 486–503, Berlin, Heidelberg, 2006a. Springer-Verlag. ISBN 3-540-34546-9, 978-3-540-34546-6. doi: 10.1007/11761679_29. URL http://dx.doi.org/10.1007/11761679_29.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. *Calibrating Noise to Sensitivity in Private Data Analysis*, pp. 265–284. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006b. ISBN 978-3-540-32732-5. doi: 10.1007/11681878_14. URL https://doi.org/10.1007/11681878_14.

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

Erlich, Y. and Narayanan, A. Routes for breaching and protecting genetic privacy. *Nature Reviews Genetics*, 15 (6):409–421, 2014.

Ferraty, F. and Romain, Y. *The Oxford Handbook of Functional Data Analaysis*. Oxford University Press, 2011.

Fienberg, S. and Slavković, A. *Data Privacy and Confidentiality*, pp. 342–345. International Encyclopedia of Statistical Science. Springer-Verlag, 2010.

Hall, P., Müller, H., and Wang, J. Properties of principal component methods for functional and longitudinal data analysis. *The Annals of Statistics*, 34(3):1493–1517, 2006.

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

Horváth, L. and Kokoszka, P. *Inference for functional data with applications*, volume 200. Springer Science & Business Media, 2012.

Huang, L., Scheipl, F., Goldsmith, J., Gellar, J., Harezlak, J., McLean, M. W., Swihart, B., Xiao, L., Crainiceanu, C., and Reiss, P. *refund: Regression with Functional Data*, 2016. URL https://CRAN.R-project.org/package=refund. R package version 0.1-14.

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

Kifer, D., Smith, A., and Thakurta, A. Private convex empirical risk minimization and high-dimensional regression. In Mannor, S., Srebro, N., and Williamson, R. C. (eds.), *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pp. 25.1–25.40, Edinburgh, Scotland, 25–27 Jun 2012. PMLR. URL http://proceedings.mlr.press/v23/kifer12.html.

Kokoszka, P. and Reimherr, M. *Introduction to functional data analysis*. CRC Press, 2017.

Kulynych, J. Legal and ethical issues in neuroimaging research: human subjects protection, medical privacy, and the public communication of research results. *Brain and cognition*, 50(3):345–357, 2002.

Laha, R. and Rohatgi, V. *Probability Theory*. Wiley, New York, 1979.

Lane, J., Stodden, V., Bender, S., and Nissenbaum, H. *Privacy, big data, and the public good: Frameworks for engagement*. Cambridge University Press, 2014.

Lei, J., Charest, A.-S., Slavkovic, A., Smith, A., and Fienberg, S. Differentially private model selection with penalized and constrained likelihood. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181 (3):609–633, 2018.

Lippert, C., Sabatini, R., Maher, M. C., Kang, E. Y., Lee, S., Arikan, O., Harley, A., Bernal, A., Garst, P., Lavrenko, V., Yocum, K., Wong, T., Zhu, M., Yang, W.-Y., Chang, C., Lu, T., Lee, C. W. H., Hicks, B., Ramakrishnan, S., Tang, H., Xie, C., Piper, J., Brewerton, S., Turpaz, Y., Telenti, A., Roby, R. K., Och, F. J., and Venter, J. C. Identification of individuals by trait prediction using whole-genome sequencing data. *Proceedings of the National Academy of Sciences*, 114(38):10166–10171, 2017. doi: 10.1073/pnas.1711125114. URL http://www.pnas.org/content/114/38/10166.abstract.

Ramsay, J. O. and Silverman, B. W. *Applied functional data analysis: methods and case studies*, volume 77. Springer New York, 2002.

Ramsay, J. O. and Silverman, B. W. *Functional data analysis*. Springer New York, 2005.

Rubin, D. B. Statistical disclosure limitation. *Journal of official Statistics*, 9(2):461–468, 1993.

Schadt, E. E., Woo, S., and Hao, K. Bayesian method to predict individual snp genotypes from gene expression data. *Nature genetics*, 44(5):603–608, 2012.

Smith, M., Álvarez, M., Zwiessele, M., and Lawrence, N. Differentially private regression with gaussian processes. In *International Conference on Artificial Intelligence and Statistics*, pp. 1195–1203, 2018.

Stein, M. L. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 2012.

Wasserman, L. and Zhou, S. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010. doi: 10.1198/jasa.2009.tm08651. URL http://dx.doi.org/10.1198/jasa.2009.tm08651.

Willenborg, L. and De Waal, T. *Statistical disclosure control in practice*. Number 111. Springer Science & Business Media, 1996.