PennState
Eberly College of Science

Formal Privacy for Functional Data with Gaussian Perturbations

Matthew Reimherr
Department of Statistics
Pennsylvania State University

Contributors

Current work:

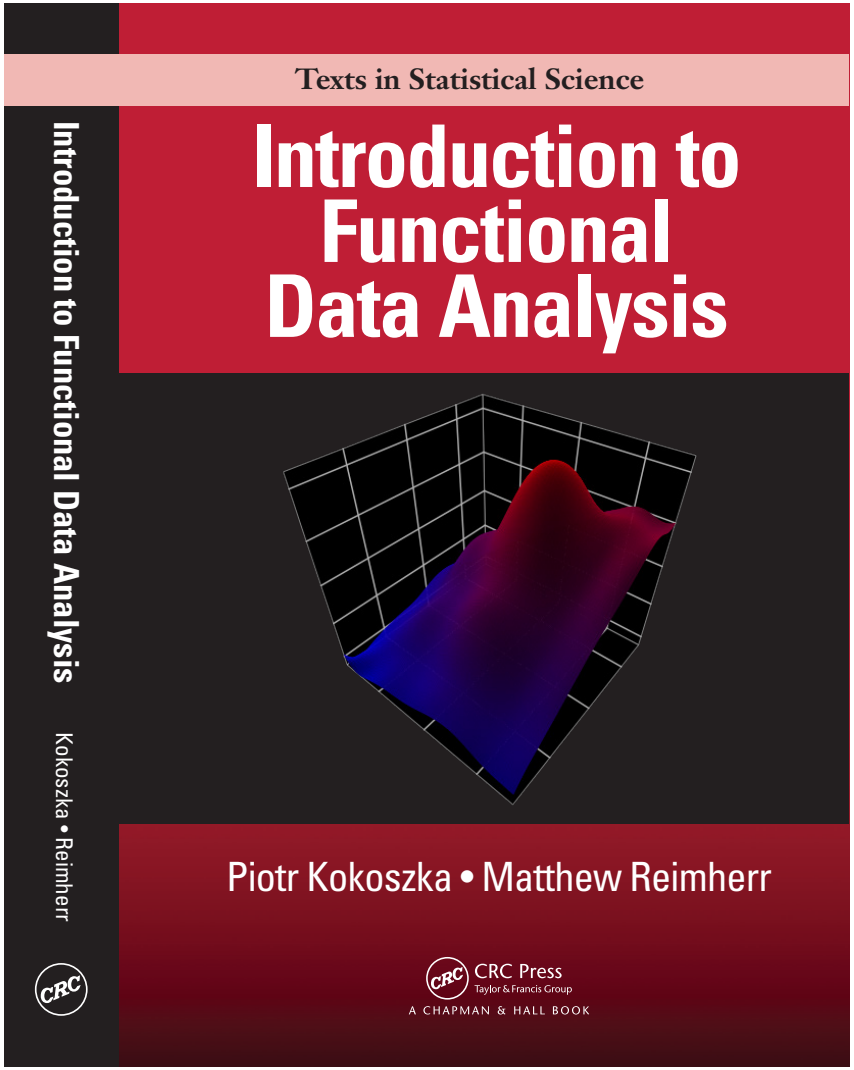
- Aleksandra Slavkovic, Professor of Statistics and Associate Dean for Graduate Education, Penn State
- Ardalan Mirshani, PhD Candidate in Statistics, Penn State

Additional work:

- Ana Kenney, PhD Candidate in Statistics, Penn State
- Jordan Awan, PhD Candidate in Statistics, Penn State

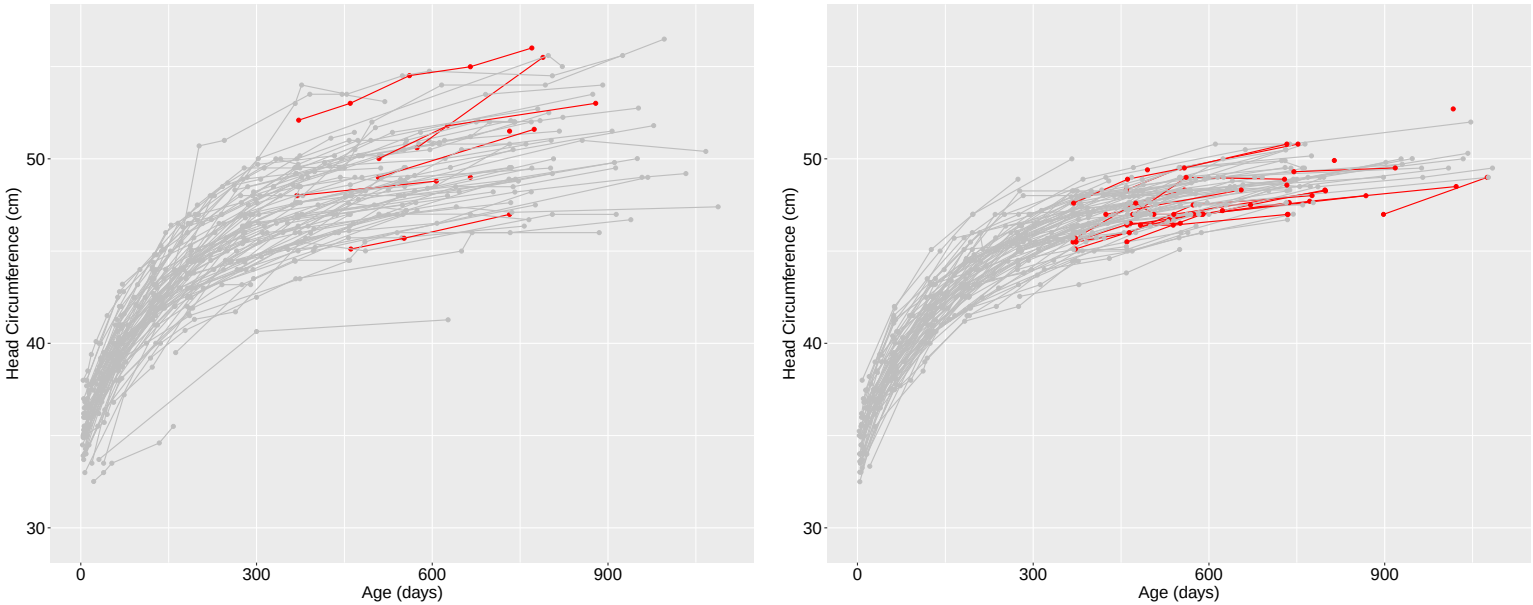
I am also grateful to NSF (DMS 1712826) for their support.

Textbook on FDA



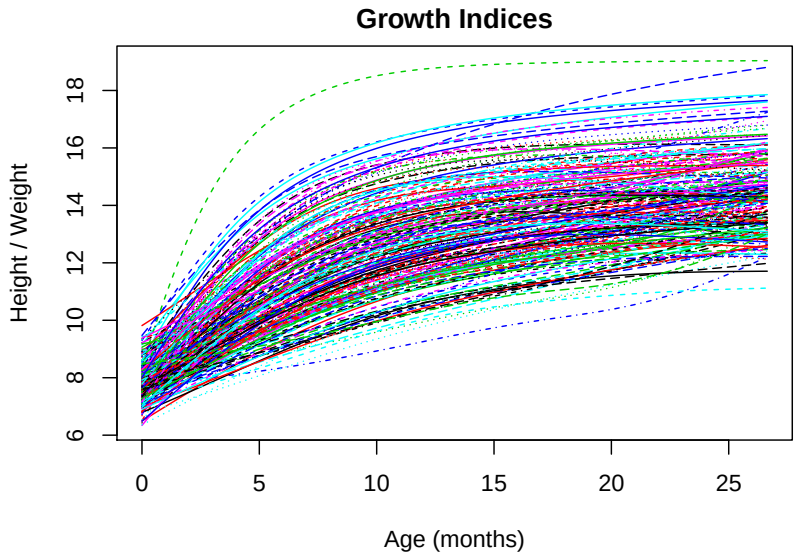
- For advanced undergraduate and MS courses
- Advanced chapters for PhD courses
- Introduces fundamental tools and perspectives in FDA, alongside applications and computational tools in R.
- Subjects covered include functional regression, sparse FDA, functional GLMs, spatial functional data, and functional time series.
- Each chapter includes problem sets which can be used for homework or practice.

Motivation - Macrocephaly



- Head circumference (cm) as a function of age (days)
- Cases (left) vs subset of controls (right)
- Electronic medical records, 74,000 children (only 85 cases)
- Up to 3 years of age, varying observations per child (1-20+)
- Binary response indicating presence of pathologies related to Macrocephaly (large heads)

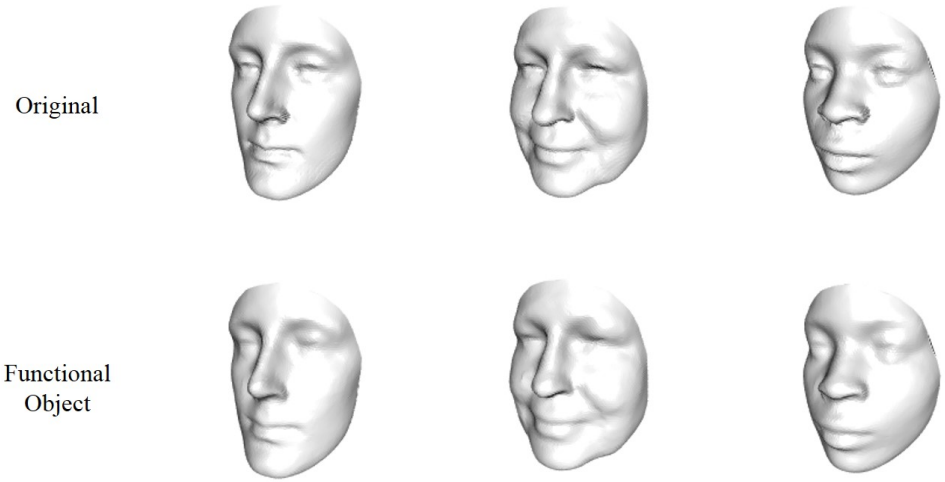
Motivation - INSIGHT



INSIGHT is an ongoing study led by Dr. Ian Paul's Lab at Penn State Hershey, with microbiome and genetic data processed by Dr. Kateryna Makova's lab at Penn State. Features of the data include

- over 200 children and growing (family/siblings)
- 7 clinical visits over two years
- bacterial abundances in the gut and mouth
- recently genotyped families.

Motivation - ADAPT



ADAPT[a] is an ongoing study let by Dr. Mark Shriver's lab at Penn State. Features of the data include

- over 6000 faces and growing
- 7150 points per face, as x,y,z coordinates
- demographic variables measured for each subject
- nearly 3000 subjects genotyped and growing (most with 1M+ snps).

[a] Anthropology, DNA, and the Appearance and Perception of Traits

Functional Data Models

Models are usually defined off of the complete trajectories, though they might be estimated from incomplete (called sparsely sampled) data.

- Mean + Covariance only

$$\mathrm{E}[X_i(t)] = \mu(t) \qquad \mathrm{Cov}(X_i(t), X_i(s)) = C(t,s).$$

- Function-on-scalar regression

$$\mathrm{E}[X_i(t)|Z_i] = \alpha(t) + \beta(t)Z_i.$$

- Scalar-on-function regression

$$\mathrm{E}[Z_i|\{X_i(t) : t \in \mathcal{T}\}] = \alpha + \int_{\mathcal{T}} \beta(t)X_i(t)\ dt.$$

And there are many many more. The major idea is that certain variables and parameters can be viewed as functions. Our goal is to develop tools for constructing differentially private output for such objects; today we focus on the mean function.

DP Setup (Dwork, McSherry, Nissim, and Smith, 2006)

Here we follow a classic setup for (ϵ,δ) differential privacy.

- **D**: population of units.
- $\mathcal{D}_n$: collection of all subsets of **D** of size n.
- $f : \mathcal{D}_n \to \mathbb{H}$ is a summary and $\mathbb{H}$ is a real separable Hilbert space.
- $\tilde{f} : \mathcal{D}_n \to \mathbb{H}$ is a privatized version of f (and random).

We say that $\tilde{f}$ achieves (ϵ-δ) differential privacy if for any two samples $D_n, D_n' \in \mathcal{D}_n$ that differ in only one record/unit, $D_n \sim D_n'$, we have

$$P(\tilde{f}(\mathcal{D}_n) \in A) \leq e^{\epsilon} P(\tilde{f}(\mathcal{D}_n') \in A) + \delta \qquad \text{for all measurable } A.$$

The primary challenge here is that $\mathbb{H}$ can be (in our case usually is) infinite dimensional, and thus the A are collections of functions.

Previous Works

- Hall, Rinaldo, and Wasserman (2013): release finite number of function evaluations; based on RKHS and Gaussian noise.
- Holohan, Leith, and Mason (2015): DP in abstract metric spaces assuming a sanitized database.
- Aldá and Rubinstein (2017): Function release based on Bernstein polynomials and Laplace noise.
- Smith, Alvarez, Zwiessele, Lawrence (2018): Examined how to tailor the noise in the Hall et al (2013) framework.

Gaussian Noise

Definition

We say that $Z \in \mathbb{H}$ is a Gaussian noise if, for any $h \in \mathbb{H}$ we have that $\langle h, Z \rangle$ is normal in $\mathbb{R}$.

A classic result from probability theory states that for any Z there exists $\mu \in \mathbb{H}$ and a positive definite nuclear operator C such that

$$\langle h, Z \rangle \sim \mathcal{N}(\langle \mu, h \rangle, \langle Ch, h \rangle).$$

We thus denote $Z \sim \mathcal{N}(\mu, C)$. Using the same strategy as Hall et al (2013) we use the following mechanism

$$\tilde{f}(D_n) = f(D_n) + \sigma Z,$$

where δ is some constant. However, there are some issues that arise when $\mathbb{H}$ is infinite dimensional.

Compatibility

However, not just any noise can be chosen, even if the covariance operator, C, has full rank. Let $\mathbb{K} \subset \mathbb{H}$ consist of all elements h such that $\langle C^{-1}h, h\rangle < \infty$. Equipped with the inner product $\langle \cdot, \cdot \rangle_{\mathbb{K}} = \langle C^{-1}\cdot, \cdot \rangle$, $\mathbb{K}$ is also a Hilbert space.

Definition (Mirshani, R, and Slavkovic, 2019)

We say a summary f is *compatible* with a noise $\mathcal{N}(0, C)$ if $f(D_n)$ is in the Hilbert Space, $\mathbb{K}$, generated by C for all D_n.

The major problem is that without compatibility, it is possible to "project" the noise out of the summary.

Theorem (Mirshani, R, and Slavkovic, 2019)

If f is not compatible with $\mathcal{N}(0, C)$, then there exists no $\sigma \in \mathbb{R}$ such that $\tilde{f}(D_n) := f(D_n) + \sigma Z$ achieves DP in $\mathbb{H}$.

Global Sensitivity

In the infinite dimensional setting, it does not appear possible to define a notion of global sensitvity that doesn't explicitly take into acount the noise.

Definition (Hall, Rinaldo, and Wasserman (2013))

The global sensitivity of a summary f with respect to noise $\mathcal{N}(0, C)$ is given by

$$\Delta^2 := \sup_{D_n \sim D_n'} \|f(D_n) - f(D_n')\|_{\mathbb{K}}^2,$$

where $\|\cdot\|_{\mathbb{K}}$ is the inner product norm over $\mathbb{K}$.

Theorem (Hall, Rinaldo, and Wasserman, 2013)

If $f(D_n)$ is a real valued function over an interval $\mathcal{T}$, denoted as $f_{D_n}(t)$, then

$$(f_{D_n}(t_1) + \sigma Z(t_1), \ldots, f_{D_n}(t_m) + \sigma Z(t_m))$$

achieves (ϵ, δ) DP in $\mathbb{R}^m$ for any choice of $\{t_i\}$ and any finite m if

$$\sigma^2 = \frac{2\log(2/\delta)}{\epsilon^2}\Delta^2.$$

More General Formulation

This result can be extended much more generally.

Theorem (Mirshani, R, and Slavkovic, 2019)

If $f(D_n) \in \mathbb{H}$ *then*

$$(\langle f_{D_n} + \delta Z, h_1 \rangle, \ldots, \langle f_{D_n} + \delta Z, h_m \rangle)$$

achieves ϵ-δ *DP in* $\mathbb{R}^m$ *for any choice of* $\{h_i \in \mathbb{H}\}$ *and any finite* m *if*

$$\sigma^2 = \frac{2\log(2/\delta)}{\epsilon^2}\Delta^2.$$

And we can even obtain it for the entire process.

Theorem (Mirshani, R, and Slavkovic, 2019)

If $f(D_n) \in \mathbb{H}$ *then*

$$\tilde{f}(D_n) = f(D_n) + \sigma Z$$

achieves ϵ-δ *DP in* $\mathbb{H}$ *if*

$$\sigma^2 = \frac{2\log(2/\delta)}{\epsilon^2}\Delta^2.$$

Theoretical Tools

Our work is buil upon classic results from the 1960s and 1970s concerning the equivalence/orthogonality of Gaussian measures.

Theorem (Bogachev, 1998)

Let P^1 and P^2 be two Gaussian measures over a separable Hilbert space, $\mathbb{H}$, with means/covariances (μ_1, C) and (μ_2, C) respectively, and assume C has full rank. Then the two meausres are equivalent (in a probablistic sense) if

$$\|C^{-1/2}(\mu_1 - \mu_2)\|_{\mathbb{H}}^2 < \infty,$$

and orthogonal otherwise.

With equivalent measures, you can also define densities for the Gaussian processes. Proving our results then depends on using these densities to extend arguments/bounds from the multivariate setting.

Example - Mean Function Estimation

Let $X_i(t)$ be an iid sample from $\mathbb{H} = L^2[0,1]$ with $\|X_i\|_{\mathbb{H}} \leq \tau$. Define $\mu(t) = \mathrm{E}[X_i(t)]$ and let $C(t,s)$ be a continuous positive definite function. Then the $\mathbb{K}$ is a reproducing kernel Hilbert space with kernel C. We can estimate μ using

$$\hat{\mu} := \mathrm{argmin}_{\mu \in \mathbb{H}} \frac{1}{n} \sum_{i=1}^{n} \|X_i - \mu\|_{\mathbb{H}}^2 + \phi \|C^{-\eta/2}\mu\|_{\mathbb{H}}^2.$$

Theorem (Mirshani, R, and Slavkovic (2017))

The estimator $\hat{\mu}$ has a global sensitivity bounded by

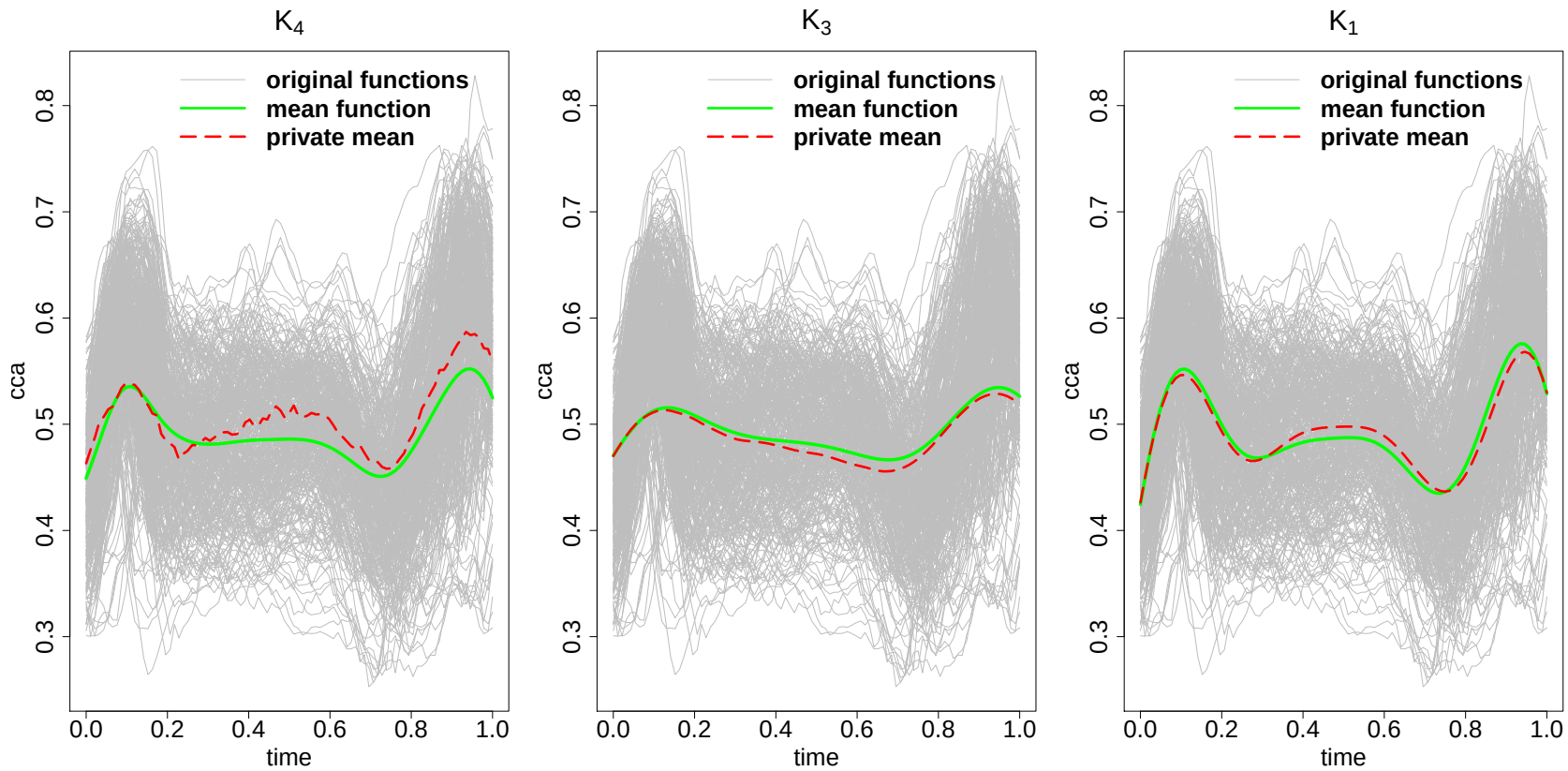
$$\Delta^2 \leq \frac{\tau^2}{n^2 \phi^{1/\eta}}.$$

If the tuning parameter is taken such that $\phi \gg n^{-\eta}$ then the privacy noise is asymptotally negligible

$$\|\tilde{\mu} - \hat{\mu}\|_{\mathbb{H}}^2 = o_P(n^{-1}).$$

Numerical Illustration - Corpus Callusum

One nice aspect is that one can choose the noise $\mathcal{N}(0, C)$ and by forcing the f to lie in $\mathbb{K}$ one can oversmooth a bit to dramatically reduce the noise one has to add to achieve DP. Below are 3 different kernels: exponential, Gaussian, Matern (3/2) respectively.



Overall Conclusions

- Lot's of interesting challenges for privacy with highly complex objects.
- In high/infinite dimensional settings, careful regularization appears key to maintaining utility.
- Main privacy challenge seems to reside in the "higher frequencies" of the processes.
- Lack of densities in infinite dimensional settings produces a substantial hurdle.
- Paper considers more general Banach spaces as well.

Overall Conclusions

- Lot's of interesting challenges for privacy with highly complex objects.
- In high/infinite dimensional settings, careful regularization appears key to maintaining utility.
- Main privacy challenge seems to reside in the "higher frequencies" of the processes.
- Lack of densities in infinite dimensional settings produces a substantial hurdle.
- Paper considers more general Banach spaces as well.

Thank you for your time!

`www.personal.psu.edu/~mlr36`
`mreimherr@psu.edu`