Differentially Private Gaussian Processes

Michael Thomas Smith, Max Zwiessele & Neil D. Lawrence
Department of Computer Science
University of Sheffield
(m.t.smith|m.zwiessele|neil)@sheffield.ac.uk

Abstract

Differential privacy allows algorithms to have provable privacy guarantees. Gaussian processes are a widely used approach for dealing with uncertainty in functions. This paper explores differentially private mechanisms for Gaussian processes. We compare adding noise both pre- and post-regression. For the former we develop a new kernel for use with binned data. For the latter we show that using inducing inputs allows us to reduce the noise scale. For the datasets used, the two strategies have comparable accuracy. Together these methods provide a starter toolkit for combining differential privacy and Gaussian processes.

1 Introduction

Interest is increasing in mechanisms that allow individuals to retain their privacy while inferences can still be drawn through aggregated data. A differentially private (DP) algorithm (Dwork & Roth, 2014) allows queries to be performed while minimising the release of information about individual records, by perturbing the result of a query. This perturbation can be added (Berlioz et al., 2015) to the (i) input data, prior to its use in the algorithm, (ii) to components of the calculation or (iii) to the output of the algorithm.

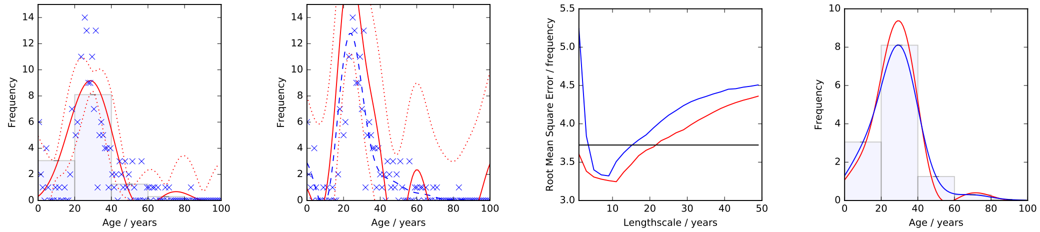
We exploit the ability of the Gaussian process (GP) to assimilate summary measures in Section 2 by deriving the covariance function for binned data. Section 3 finds efficient bounds on the sensitivity of the mean function by reformulating the model with *inducing inputs*, allowing DP to be more generally applicable to a wider array of problems, amenable to GP regression.

2 Privacy for Model Inputs: Data Binning

In this section we develop a kernel for binned data. To proceed via GP regression (see e.g. Lawrence et al., 2006) we assume that there is some *latent function*, $f(t)$. The summary measures (the binned outputs) can then be derived through integrating across the latent function to give us the necessary average. Importantly, if the latent function is drawn from a GP then its integral is also *jointly* drawn from the same GP. This allows us to analytically map between the aggregated measure and the observation of interest. Assume that a second function, $F(s,t)$, describes the integral between the ages s and t of $f(\cdot)$. We are given observations, $y(s,t)$, which are noisy samples from $F(s,t)$.

For a GP, any *linear operator* (such as integration) applied to the original function will lead to a joint GP over the result of that linear operator and the original function. In other words there will be a joint GP between the two functions $f(t')$ and $F(s,t)$. To construct the joint GP posterior we need expressions for the covariance between values of $f(t)$ and $f(t')$, values of $F(s,t)$ and $F(s',t')$, and the 'cross covariance' between the the latent function $f(t')$ and the output of the integral $F(s,t)$. For the underlying latent function we assume that the covariance between the values of the latent function $f(\cdot)$ are described by the exponentiated quadratic (EQ) form, $k_{ff}(u,u') = \alpha\ e^{-\frac{(u-u')^2}{\ell^2}}$,

Private Multi-Party Machine Learning Workshop (NIPS 2016), Barcelona, Spain.



(a) GP fit to a perturbed histogram ($\varepsilon = 1$)

(b) perturbed GP fitted to raw data ($\varepsilon = 1$)

(c) Accuracy of different kernel functions

(d) Demonstration of integral kernel vs EQ kernel

Figure 1: (a) Solid line is the mean GP fit to DP histogram, using the integral-kernel. Annual counts blue crosses. (b) GP fit to the original data, then perturbed. Original mean indicated by dashed blue line. DP mean in red. 95% CI are dotted lines and $\varepsilon = 1, \delta = 0.01$ in both. (c) RMSE for individual years, for the EQ kernel (blue) and integral kernel (red). Using histogram frequency as estimate is in black. (d) Comparison of the two kernels, EQ (blue) and the integral kernel (red).

where α is the scale of the output and ℓ is the (currently) one-dimensional length-scale[1] To compute the covariance for the integrated function, $F(\cdot)$ we integrate the original EQ form over both its time variables,

$$k_{FF}((s,t),(s',t')) = \alpha \int_s^t \int_{s'}^{t'} k_{ff}(u,u')\, \mathrm{d}u' \mathrm{d}u$$

substituting in our EQ kernel, and integrating;

$$k_{FF}((s,t),(s',t')) = \alpha \frac{\ell^2}{2} \left[g\left(\frac{t-s'}{l}\right) + g\left(\frac{t'-s}{l}\right) - g\left(\frac{t-t'}{l}\right) - g\left(\frac{s-s'}{l}\right) \right] \tag{1}$$

where we defined $g(z) = z\sqrt{\pi}\mathrm{erf}(z) + e^{-z^2}$ and $\mathrm{erf}(\cdot)$ is the Gauss error function. Similarly we can calculate the cross-covariance between F and f and extend the inputs to multiple dimensions.

We applied the approach to the age distribution of 255 people from the 2011 UK census, grouped into a histogram (10 years per bin). Figure 1c illustrates the improvement over the EQ kernel in which we fix the lengthscale and fit the noise variance and kernel scale. Figure 1a shows a GP fit to the histogram data. Figure 1d illustrates how the integral kernel will ensure the area under its curve is equal to the area of the bin.

Figure 1c shows that for all values of the lengthscale parameter, the integral kernel produces a more accurate estimate of the original data. We believe that, in general, this kernel will be superior for regressing binned or histogram datasets. Both GP regression results were, with the right lengthscale, superior to the use of the binned means.

3 Privacy for Model Outputs

So far we have considered DP on the inputs to the GP: in particular when the data is first aggregated into a histogram of binned means or counts. We now turn our attention to DP guarantees on the output of a GP, which has been trained on the original, non-private, data. For this analysis we assume that the kernel has a maximum value of one, which restricts us to using stationary kernels with normalised outputs. Hall et al. (2013) showed that one can ensure that a version of function f, denoted $\widetilde{f}$, is (ε, δ)-DP by adding a scaled sample from the Gaussian distribution G (which uses the same kernel as f). We scale the sample by $\frac{\Delta c(\delta)}{\varepsilon,}$ where $c(\delta) \geq \sqrt{2 \log \frac{1.25}{\delta}}$. Δ is the sensitivity of the function, which we will need to calculate.

[1] Note that there is a $\sqrt{2}$ difference between our length-scale and that normally defined, this is for convenience in later integrals.

Consider f and its neighbour f' (which has had one data point altered). The distance, $||f - f'||$, between these points is bounded by the sensitivity, $\Delta \geq \sup_{D \sim D'} ||f_D - f_{D'}||_H$. The norm here is defined to be $||g||= \sqrt{\langle g, g \rangle_H}$.

Assume that the inputs are non-private columns, while the outputs are private. The covariance function does not need perturbation as it does not contain direct reference to the output values (ignoring any parameter selection steps). It is just the mean function we need to perturb. We need to know the sensitivity of the output $\boldsymbol{y}$. For the age histogram example, the $\boldsymbol{y}$ values are the result of a histogram query, and thus have a sensitivity, Δ_y, of 1 (Dwork & Roth, 2014).

From Williams & Rasmussen (2006), the conditional distribution from a GP at test point $\boldsymbol{x}_*$ has mean, $\bar{f}_* = \boldsymbol{k}_*^\top \left(K' + \sigma_n^2 I\right)^{-1} \boldsymbol{y}$, and covariance, $V[f_*] = k(\boldsymbol{x}_*, \boldsymbol{x}_*) - \boldsymbol{k}_*^\top \left(K' + \sigma_n^2 I\right)^{-1} \boldsymbol{k}_*$, where $\bar{f}_*$ is the mean of the posterior, $k(\boldsymbol{x}_*, \boldsymbol{x}_*)$ is the test point's prior variance. $\boldsymbol{k}_*$ is the covariance between the test and training points, K' is the covariance between (the latent function that describes) the training points, σ_n^2 is the variance of the iid noise added to each observation and $\boldsymbol{y}$ are the output observed values of the training data.

Note that, using the representer theorem, we can rewrite the above expression as the weighted sum of n kernel functions, $\bar{f}(\boldsymbol{x}_*) = \sum_{i=1}^n \alpha_i k(\boldsymbol{x}_i, \boldsymbol{x}_*)$, where $\boldsymbol{\alpha} = \left(K' + \sigma_n^2 I\right)^{-1} \boldsymbol{y}$. For simplicity in the following we replace $K' + \sigma_n^2 I$ with K. We are interested in finding,

$$||f_D(\boldsymbol{x_*}) - f_{D'(\boldsymbol{x_*})}||_H^2 = \left\langle f_D(\boldsymbol{x_*}) - f_{D'}(\boldsymbol{x_*}), f_D(\boldsymbol{x_*}) - f_{D'}(\boldsymbol{x_*}) \right\rangle \tag{2}$$

In our case the inputs are identical. It is the values of y (and hence $\boldsymbol{\alpha}$) that we need to offer privacy.

$$f_D(\boldsymbol{x_*}) - f'_D(\boldsymbol{x_*}) = \sum\nolimits_{i=1}^n \alpha_i k(\boldsymbol{x_*}, \boldsymbol{x}_i) - \sum\nolimits_{i=1}^n \alpha'_i k(\boldsymbol{x_*}, \boldsymbol{x}_i) = \sum\nolimits_{i=1}^n k(\boldsymbol{x_*}, \boldsymbol{x}_i) \left(\alpha_i - \alpha'_i\right) \tag{3}$$

We need to provide a bound on difference between the values of $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}'$.To reiterate, $\boldsymbol{\alpha} = K^{-1}\boldsymbol{y}$. So the difference between the two vectors is $\boldsymbol{\alpha} - \boldsymbol{\alpha}' = K^{-1}\boldsymbol{y} - K^{-1}\boldsymbol{y}' = K^{-1}\left(\boldsymbol{y} - \boldsymbol{y}'\right)$.

DP assumes that all the values of $\boldsymbol{y}$ and $\boldsymbol{y}'$ are equal except for the last element which differs by at most Δ_y. I.e. $\boldsymbol{\alpha} - \boldsymbol{\alpha}'$ is bounded by Δ_y times the sum of the absolute values of the last column in K^{-1}. The use of the last column is arbitrary, and so we should consider the maximum possible sum of any column, the infinity-norm, $||K^{-1}||_\infty$. As K doesn't contain private information itself (it is dependent purely on the input and the features of the kernel) we can find the exact value of $||K^{-1}||_\infty$. We shall call this value $b(K)$. Returning to the calculation of the sensitivity, we can expand equation 3 substituted into 2:

$$||f_D(\boldsymbol{x}_*) - f'_D(\boldsymbol{x}_*)||^2 = \left\langle \sum\nolimits_{i=1}^n (\alpha_i - \alpha'_i) k(\boldsymbol{x}_*, \boldsymbol{x}_i),\ \sum\nolimits_{i=1}^n (\alpha_i - \alpha'_i) k(\boldsymbol{x}_*, \boldsymbol{x}_i) \right\rangle \tag{4}$$

We now use our constraint that the chosen kernel has a maximum value of one, so the *weighted* sum of $\alpha_i - \alpha'_i$ will be less than or equal to the sum of $\alpha_i - \alpha'_i$, which we already know have an upper bound of $\Delta_y b(K)$. This means that an upper bound on the sensitivity is $||f_D(\boldsymbol{x}_*) - f'_D(\boldsymbol{x}_*)||^2 \leq \Delta_y^2 \, b(K)^2$.

In Figure 1b we apply this method to the disaggregated census dataset. It shows that the private mean function appears to trace a path that describes the data set reasonably, suggesting that for such data sets with small sensitivities, this method may be useful. Note that negative counts appear in the private function, but that this is expected under a DP mechanism.

3.1 Inducing Variables and Sensitivity

In this section we show how the function sensitivity can be reduced by the introduction of *inducing variables* (IVs). To get some intuition as to why the inducing-inputs reduce the sensitivity, it is useful to consider what the inverse covariance (precision) matrix represents. Between training inputs the diagonal elements, (i, i), of the precision matrix are the inverse variance (precision) of each variable, controlling for the remaining variables. If another input k explains most of the variance of i then the remaining variance will be very small, causing this inverse-variance to be very large. The off-diagonal elements (i, j) are the negative partial correlations scaled by the root product of the two corresponding diagonal elements, (i, i) and (j, j), which we've seen will be large if k explains much

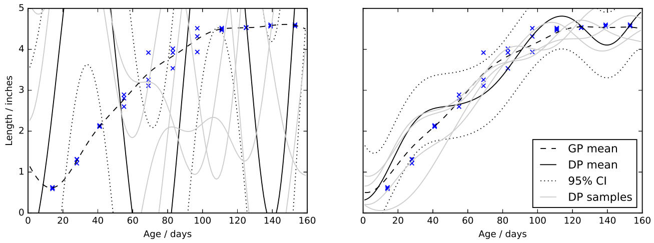


Figure 2: Fish lengths and ages. Left: standard GP. Right: IV GP. Posterior means indicated by dashed black lines. A DP sample is indicated with the solid black line and is surrounded by dotted black lines, indicating the 95% confidence interval, when both the DP noise and GP variance is combined. Four other DP mean samples are included, in light grey. ($\varepsilon = 20$)

	Input Noise		Output Noise	
ε	Integral	EQ	GP	IV GP
0.1	36.3 ± 0.3	$\mathbf{35.8} \pm 0.6$	726.5 ± 0.7	59.73 ± 0.06
0.2	$\mathbf{34.1} \pm 0.4$	37.4 ± 0.7	363.6 ± 0.3	37.4 ± 0.03
0.5	31.6 ± 0.6	36.5 ± 1.0	146.1 ± 0.13	$\mathbf{28.1} \pm 0.014$
2.0	26.22 ± 0.6	32.6 ± 1.6	39.0 ± 0.04	$\mathbf{26.15} \pm 0.004$
∞	25.2	31.8	**14.3**	26.0

Table 1: RMSE for the Citibike data set, using the four methods. The CIs are one SE.

of i. In the inducing-input case, we are interested in how much inducing-inputs explain each other. Placed far apart, there is relatively little correlation, so the remaining unexplained variance for a given input is large, causing the diagonals of the precision matrix to be small. In this paper we show that using a small number of m inducing-inputs can reduce the bound on the sensitivity of the function, potentially substantially, allowing DP to be used for a larger range of data sets. In section 3, we rewrote the expression for the posterior mean, $\mu_* = \boldsymbol{k}_*^\top K^{-1}\boldsymbol{y}$, as $\sum_i \alpha_i \; k(x_i, x_*)$ in which $\boldsymbol{\alpha} = K^{-1}\boldsymbol{y}$. To find the sensitivity we considered the infinity-norm of K^{-1}. We can rewrite the expression for the mean of an IV GP (see Snelson & Ghahramani (2005); Quiñonero Candela & Rasmussen (2005); Titsias (2009); Damianou et al. (2016)) the same way, and find the infinity-norm again. In summary we want the infinity norm of $Q_{uu}^{-1} K_{uf} \left(\Lambda + \sigma^2 I\right)^{-1}$, where $Q_{uu} = K_{uu} + K_{uf} \left(\Lambda + \sigma^2 I\right)^{-1} K_{fu}$, the matrix $\Lambda = \text{diag}(\boldsymbol{\lambda})$ in which the diagonal elements are $\lambda_i = K_{ii} - \boldsymbol{k}_i^\top K_{uu}^{-1} \boldsymbol{k}_i$. We define K_{fu} to be the covariance matrix between training and inducing variables, K_{uu} is the covariance between IVs and $\boldsymbol{k}_i$ the covariance between the ith training input and the inducing variables. Figure 2 demonstrates this improvement with fish length and age data from Freund & Minton (1979), illustrating the improved output using the IVs.

4 Numerical Comparison, Conclusions and Further Work

We finally use data from 'citibike' (Citibike, since 2013) in New York, considering just hire time and duration. Table 1 shows the integral kernel does more poorly than the EQ at low ε. This suggests this kernel has poorer noise immunity. With increasing ε the integral kernel becomes the best choice of the two kernels; the binning masks some of the peaks in the data, which the integral kernel will represent better. For intermediate ε values the IV GP does best. For larger ε the normal GP has the lowest error as it has access to the real data. The relative accuracy of the methods outlined for introducing DP noise depend on the scale of the noise. As the noise level reduces the loss of fidelity due to binning starts to dominate and the output noise methods become the most optimum.

We have presented novel methods for combining DP and GPs. In the longer term we believe a comprehensive set of methodologies could be developed. This first paper has given a flavour of some of the challenges and their potential solutions. We developed a new kernel for use with any cuboid binned data set (DP or not), applied the DP for functions theory developed by Hall et al. (2013) to GPs, and shown that using IVs can massively reduce the sensitivity of the mean function.

References

Berlioz, A., Friedman, A., Kaafar, M. A., Boreli, R., & Berkovsky, S. (2015). Applying differential privacy to matrix factorization. In *Proceedings of the 9th ACM Conference on Recommender Systems*, (pp. 107–114). ACM.

Citibike (since 2013). Citibike system data. `https://www.citibikenyc.com/system-data`.

Damianou, A., Titsias, M. K., & Lawrence, N. D. (2016). Variational inference for latent variables and uncertain inputs in Gaussian processes. *Journal of Machine Learning Research*, *17*.

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

Freund, R. J., & Minton, P. D. (1979). Regression methods: a tool for data analysis (new york). *Dekker*, (p. 111).

GPy (since 2012). GPy: A gaussian process framework in python. `http://github.com/SheffieldML/GPy`.

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

Kusner, M. J., Gardner, J. R., Garnett, R., & Weinberger, K. Q. (2015). Differentially private bayesian optimization. *arXiv preprint arXiv:1501.04080*.

Lawrence, N. D., Sanguinetti, G., & Rattray, M. (2006). Modelling transcriptional regulation using gaussian processes. In *Advances in Neural Information Processing Systems*, (pp. 785–792).

McSherry, F., & Talwar, K. (2007). Mechanism design via differential privacy. In *Foundations of Computer Science, 2007. FOCS'07. 48th Annual IEEE Symposium on*, (pp. 94–103). IEEE.

Quiñonero Candela, J., & Rasmussen, C. E. (2005). A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, *6*, 1939–1959.

Snelson, E., & Ghahramani, Z. (2005). Sparse gaussian processes using pseudo-inputs. In *Advances in neural information processing systems*, (pp. 1257–1264).

Titsias, M. K. (2009). Variational learning of inducing variables in sparse Gaussian processes. In D. van Dyk, & M. Welling (Eds.) *Proceedings of the Twelfth International Workshop on Artificial Intelligence and Statistics*, vol. 5, (pp. 567–574). Clearwater Beach, FL: JMLR W&CP 5.

Varah, J. M. (1975). A lower bound for the smallest singular value of a matrix. *Linear Algebra and its Applications*, *11*(1), 3–5.

Williams, C. K., & Rasmussen, C. E. (2006). Gaussian processes for machine learning. *the MIT Press*, *2*(3), 4.