Inference for stochastic processes under privacy constraints

Cristina Butucea (ENSAE) et Jean-François Delmas (ENPC)

One of the many new challenges for statistical inference in the information age is the increasing concern of data privacy protection. Therefore, data are often available not in their original version but after the employment of privacy preserving release mechanisms.

Differential privacy as introduced by Dwork *et al.* (2006) provides a simple way to quantify the privacy induced by a privacy mechanism through a parameter $\alpha > 0$ that is required to be small for higher privacy. Consider n individuals who possess data $X_1, \ldots, X_n$. The statistician does not get to see the original data, but only a *privatized* version of observations Z. The conditional distribution of Z given $X = (X_1, \ldots, X_n)$ is denoted by Q and referred to as a channel distribution or a privatization scheme, i.e. $Pr(Z \in A|X = x) = Q(A|x)$. For $\alpha \in (0, \infty)$, the channel Q is said to provide α-differential privacy if

$$\sup_A \sup_{x,x':d_0(x,x')=1} \frac{Pr(Z \in A|X = x)}{Pr(Z \in A|X = x')} \quad \leq \quad e^{\alpha}.$$

Privacy mechanisms may act on a whole sample (globally) or on each individual (locally). A rich literature establishes that we may expect to recover information on the underlying population using globally privatized data at similar rates as if we had used the original sample, see e.g. Wasserman and Zhou (2010). In nonparametric and high-dimensional inference it is usually the case that a loss of rate occurs when using locally privatized data (Duchi *et al.* 2013a, Butucea *et al.*, 2019, etc.). It is therefore necessary to show optimality of the proposed method over all local differentially private mechanisms and over all estimation procedures using the published data. Information theoretic inequalities by Duchi *et al.* (2013b) are usually used to prove such optimality results, but they are not always optimal as seen in Butucea *et al.* (2021).

Previous results have been proposed for i.i.d. sensitive data. However, financial or biological data present often dependence structures. The current project aims at addressing high-dimensional and non-parametric inference problems for stochastic processes under differential privacy constraints. Also, problems specific to the setup of stochastic processes will be raised as e.g. estimating periods of stationary regimes or detecting change-points.

References

DWORK, C., MCSHERRY, F., NISSIM, K. and SMITH, ADAM (2006). Calibrating noise to sensitivity in private data analysis. *Theory of Cryptography*, 265–284.

BUTUCEA C., DUBOIS A., KROLL M. and SAUMARD, A. (2019) Local differential privacy: elbow effect in optimal density estimation and adaptation over Besov ellipsoids. *Bernoulli*, **26**, 3, 1727-1764;

BUTUCEA C., ROHDE A. and STEINBERGER, L. (2021) Interactive versus non-interactive locally differentially private estimation: Two elbows for the quadratic functional;

DUCHI, J. C., JORDAN, M. I. and WAINWRIGHT, M. J. (2013a) Local privacy and statistical minimax rates. *IEEE 54th Annual Symposium on Foundations of Computer Science*, 429-438

DUCHI, J. C., JORDAN, M. I. and WAINWRIGHT, M. J. (2013b) Local privacy and minimax bounds: Sharp rates for probability estimation. *Adv. Neural Inf. Process Systems*, 1529–1537

WASSERMAN, L. and ZHOU, S. (2010) A statistical framework for differential privacy. JASA, **105**, 489, 375-389