Fusion

Methodology / Theory / Privacy / Genetics / Criminality

Supervisors: Murray Pollock[*], Gareth Roberts (Warwick)
Project Co-Supervisors: Paul Jenkins (Warwick), Jim Q. Smith (Warwick)
Collaborators: Hongsheng Dai (Essex), Louis Aslett (Durham), Cyril Chimisov (Google)

Interested students should schedule a meeting to discuss this project prior to selection
Find updated project listing / availability at https://warwick.ac.uk/mpollock/projects

Overview

A common problem arising in statistical inference is the need to unify distributed analyses and inferences on shared parameters from multiple sources, into a single coherent inference. This unification (which we term 'fusion') problem can arise either explicitly due to the nature of a particular application, or artificially as a consequence of the approach a practitioner takes to tackling an application.Typically there will exist no closed form analytical approach to unifying distributed inferences, and so we focus on a Monte Carlo approach. Stated generally, we are interested in sampling (without error) the following d-dimensional (fusion) target density,

$$\pi(x) \propto f_1(x) \cdots f_C(x), \tag{1}$$

where each $f_c(x)$ $(c \in \{1, \ldots, C\})$ is a density (up to a multiplicative constant) representing one of the C distributed inferences we wish to unify. Each $f_c(x)$ (which we term a sub-posterior) may in practice itself be represented by a Monte Carlo sample $(\tilde{f}_1(x))$, and we may suppose in the simplest setting we are able to sample (directly and exactly) from each $f_c(x)$. The 'Monte Carlo Fusion' approach introduced by the PI and co-authors in their recent work [Dai et al., 2018], allow for the direct (and perfect) sampling exactly from (1) using a rejection sampling scheme on an extended space.

A number of interesting theoretical and methodological projects within Fusion exist, some of which are listed below. Interested student(s) should speak to one or more of {Murray Pollock, Gareth Roberts} (and if applicable the other listed person) for more details on the proposed topic(s) and whether the topic(s) align with their interests.

Project 1: Methodology

Within [Dai et al., 2018] a simple Monte Carlo approach was outlined in order to sample from the extended target density discussed above. However, the methodology to do this is cumbersome and inefficient, and could be considerably optimised. In particular, for tractability proposals based on diffusions were made, but this results in considerable computational complications which in full generality requires methodology known as 'path-space rejection sampling' (PSRS).

We would like to develop more practical approaches for simulating from (1), which do not use the full machinery of PSRS. It is anticipated that this will involve the development of alternative (diffusion) proposal mechanisms for increased efficiency. Considerable practical (and philosophical) issues will also require

[*] Email: m.pollock@warwick.ac.uk

addressing. For instance, dimensionally mis-matched sub-posteriors, and how to split the prior.

Truly efficient sampling mechanisms for (1) may require the development of approximations to the diffusion and jump diffusion proposals, and consequently an understanding of the error in resulting approximations. Alternatively, other Monte Carlo approaches as an alternative to the pure simulation of diffusions (for instance, SMC and MCMC) to underpin Fusion methodology could be explored.

Project 2: Theory

Beyond [Dai et al., 2018] very little theory is established. As such, preliminary work will be in understanding the scaling of core Fusion methodology. This will include the computational complexity when considering an increasing number of parties to be unified, the effect of increasing dimensionality in the sub-posteriors, and the effect on efficiency of poorly matched sub-posteriors.

Concurrently with the development of further Fusion methodology, theory will need to be developed for adaptions to the current Fusion approach. This includes developing theory to understand the effect of any simplifying approximations made.

Project 3: Privacy Application (MP /+ GOR with Louis Aslett)

A direction of particular interest for broad societal impact is in Statistical Cryptography. In the simplest application setting we will have a number of trusted parties who wish to securely share their distributional information on a common parameter space and model (by means of (1)), but would prefer not to reveal their individual level distributions. To make traction on this challenging problem we will exploit a technique in the cryptography literature known as 'Homomorphic Secret Sharing'.

A considerable literature exists in Homomorphic Secret Sharing (HSS), which considers a variety of (non-statistical) settings [Shamir, 1979, Benaloh, 1986, Feldman, 1987, Franklin and Yung, 1992]. These range from the secret sharing of single integer-valued secrets, to the sharing of multiple real-valued secrets, and sharing under variations in the primary assumptions (for instance, the trust-worthiness of the parties). Considering the simplest setting of a single secret then a (k, n) threshold secret sharing scheme describes a method in which a secret S of interest which is divided into n pieces $(S_1, \dots, S_n)$ in such a way that knowledge of any k pieces makes S computable, but knowledge of fewer than k pieces leaves S uncomputable (and furthermore knowledge of $k-1$ pieces provides no more information than knowledge of 0 pieces).

It transpires that the Fusion methodology that has been developed only requires point-wise evaluation of individual sub-posteriors at a finite (deterministic) number of locations. As such, one could view the (full) posterior at each point-wise location to be the secret which has been shared among the multiple parties, and the secret shares are each individual party's sub-posterior evaluated at the same point-wise location. In principle, this application is direct in that now a gold-standard information theoretic security of the raw data can be achieved (another party with unbounded compute power could not determine secret information of any other party) – so called 'CONfidential Fusion (Confusion)'.

Project 4: Genetic Application (MP with Paul Jenkins)

In the context of population genetics, DNA sequence datasets might be analysed separately and their inferences combined only later. Massively increasing volumes of DNA sequence data are providing unprecedented opportunities for inference about parameters of biological and evolutionary importance. For instance, we can treat the demographic history of a population – that is, its size $N(t)$ as a function of time t into the past – as a parameter to be inferred. We might have separate data for different chromosomes say, but each dataset provides an independent insight into the same history of the population. It is possible in principle to infer the timings of past population bottlenecks, expansions, migration events, and so on [Schraiber and Akey, 2015], as well as other parameters such as rates of mutation and recombination. Existing methods typically struggle with whole genome data, but splitting such data into independent genetic loci is a commonly-used and expedient solution. The problem is confounded by the fact that the likelihood for each genetic locus depends on an unobserved, latent genealogical history relating a sample, and this must itself be integrated over.

Another setting in which genetic analyses require unifying from the perspective of other communities of researchers and practitioners is to use the inference obtained on DNA sequences and supplement this with further inferences obtained from other works or data sets, involving other covariates (for instance, lifestyle). An example of such a 'multi-view' learning (in which the raw data is unavailable for practical reasons but the inference is) could be a researcher determining the risk of a particular individual developing cancer.

Project 5: Criminality Application (MP with Jim Q. Smith)

This project would study public health and criminal networks from the perspective of national security, using data available to the supervisor from collaborative work at the ATI. For instance, for the application to criminal networks one may look at combing inferences which could include network analysis of criminal organisations, large data obtained for social media, criminal studies in published scientific studies.

Selected References

[Benaloh, 1986] Benaloh, J. (1986). Secret sharing homomorphisms: Keeping shares of a secret secret. In *Conference on the Theory and Application of Cryptographic Techniques*, pages 251–260. Springer.

[Dai et al., 2018] Dai, H., Pollock, M., and Roberts, G. (2018). Monte Carlo Fusion. *Under invited revision to the Journal of Applied Probability.*

[Feldman, 1987] Feldman, P. (1987). A practical scheme for non-interactive verifiable secret sharing. In *Foundations of Computer Science, 1987., 28th Annual Symposium on*, pages 427–438. IEEE.

[Franklin and Yung, 1992] Franklin, M. and Yung, M. (1992). Communication complexity of secure computation (extended abstract). In *Proc. of the 24th Annual ACM STOC*, pages 699–710.

[Schraiber and Akey, 2015] Schraiber, J. G. and Akey, J. M. (2015). Methods and models for unravelling human evolutionary history. *Nature Reviews Genetics*, 16(12):727–740.

[Shamir, 1979] Shamir, A. (1979). How to share a secret. *Communications of the ACM*, 22(11):612–613.