Bayesian Fusion

Unifying Distributed Analyses

Supervisors: Murray Pollock*, Gareth Roberts, Hongsheng Dai (Essex)

Overview

A (surprisingly difficult) challenge in many application settings is to conduct inference on a parameter set of interest by combining information or inferences available on those parameters from multiple sources. This can arise both naturally as a consequence of the idiosyncrasies of a given application, or, through the design and manner in which the inference is conducted. Examples of this challenge naturally occurring include the elicitation of multiple expert opinions and the (product) pooling of those opinions into a single consensus view, and multi-view learning, in which one attempts to coherently combine disparate data sets providing partial insight into a problem (for instance, in genetic applications one way may have access to separate socio-economic, lifestyle and genome data sets and analyses). A prominent, and highly topical, example of this challenge arising through design appears in the context of 'big data' in the recent work of [Scott et al., 2016]. In this setting, for reasons of computational feasibility and data storage, MCMC can not be conducted on a single machine. Instead, the data is artificially split between a large cluster of machines, inference is conducted on each in isolation, and attempts are made at recombining the resulting analyses. Unfortunately, in all but a small number of settings, it is unclear how to appropriately recombine.

Recent work on 'Bayesian Fusion' [Dai et al., 2017, Dai et al., 2018] (which is ongoing joint work by the project supervisors), provides an exciting exact and parallelisable approach to unifying such distributed analyses. This approach is currently underpinned by the theoretical properties of classes of measure preserving diffusions, and the ability to use the extensive recent methodology on path-space rejection sampling to simulate from such diffusions without error [Pollock et al., 2016b].

A number of possible future research directions in Bayesian Fusion exist. This includes extending and strengthening the theoretical foundations of the work (incorporating related ideas, such as those found in [Pollock et al., 2016a]), extending the methodology to consider related problems (such as addressing the approximation in Approximate Bayesian Computation (ABC)), and considering the design of problems in other application areas (such as those found in the calculation of likelihoods in graphical models). There is also considerable scope in focussing on the computational aspects of the challenge and the current big data application. Interested students should contact one of the supervisors above to discuss in more detail the project, and how it aligns (or could align) to their research interests.

Selected References

[Dai et al., 2017] Dai, H., Pollock, M., and Roberts, G. (2017). Distributed Monte Carlo. *Preprint available.*

[Dai et al., 2018] Dai, H., Pollock, M., and Roberts, G. (2018). Bayesian Fusion: An exact and parallelisable approach to unifying distributed analyses. *In preparation.*

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

[Pollock et al., 2016a] Pollock, M., Fearnhead, P., Johansen, A., and Roberts, G. (2016a). The Scalable Langevin Exact Algorithm: Bayesian Inference for Big Data. *arXiv:1609.03436*.

[Pollock et al., 2016b] Pollock, M., Johansen, A., and Roberts, G. (2016b). On the exact and ε-strong simulation of (jump) diffusions. *Bernoulli*, 22(2):794–856.

[Scott et al., 2016] Scott, S., Blocker, A., and Bonassi, F. (2016). Bayes and Big Data: The Consensus Monte Carlo Algorithm. *International Journal of Management Science and Engineering Management*.