

DATA FUSION, MODEL MIS-SPECIFICATION AND CAUSAL INFERENCE

Ioanna Manolopoulou
University College of London, UK

In (Bayesian) statistics, in the absence of model and prior distribution mis-specification, more data is always better. However, in the presence of model misfit, this is not always the case: adding a data source which suffers from model misfit or unobserved errors may harm your inference. This problem is especially pronounced if there are pieces of the model which are unidentifiable, in which case prior specification also plays a crucial role. A particular example arises in causal inference heterogeneous treatment effect estimation, when combining data with and without unobserved confounding. When the unconfounded data are not representative of the whole population, the effect of de-confounding will be poor for subsets of the population that fall outside the range of these data. Depending on the structure of the model and the nature of the prior distributions used within a Bayesian model, the model will try to address this by borrowing information from other parts of the space. In this work, we highlight the importance of building models that can account for regions of poor identification and corresponding uncertainty, using examples from causal inference as well as Gaussian processes.

Joint work with Charif el Gataa, Ilina Yozova, Nathan McJames, Brieuc Lehmann, Gianluca Baio