Exact Estimation of Multiple Directed Acyclic Graphs via Integer Linear Programming

C. J. Oates, J. Q. Smith
Department of Statistics
University of Warwick

S. Mukherjee
MRC Biostatistics Unit and
CRUK Cambridge Institute
University of Cambridge

J. Cussens
Department of Computer Science and
York Centre for Complex Systems Analysis
University of York

Directed acyclic graphical models (DAGs) are widely used to make causal inferences and predictions concerning interplay in multivariate systems. In many applications, data are collected from multiple related but non-identical units whose causal mechanisms may differ but are likely to share many features. For example, in biology, units may correspond to different patients or cell lines and the DAGs themselves to gene regulation or protein signalling pathways. Interplay in cellular systems can depend on the genetic and epigenetic state of the units, such that even for a well-defined system, such as signalling in response to ligand stimulation, or regulation of a gene by transcription factors, details may differ between even closely related units. Furthermore, the availability of increasingly high-throughput biochemical assays has led to an increase in experimental designs that include panels of potentially heterogeneous units. In such settings there is scientific interest in unit-specific DAGs as well as their similarities and differences. We note that much recent research has focussed on joint estimation for multiple (undirected) Gaussian graphical models (GGMs); however, unlike DAG models, GGMs do not admit an associated theory of inferred causation and are not suitable for the explicitly causal scientific purposes that motivate our research.

Joint statistical estimation for multiple DAGs is complicated by the requirement that all DAGs are simultaneously acyclic. We propose a novel Bayesian joint model for multiple DAGs and prove that the maximum a posteriori (MAP) estimate is characterised as the solution to an integer linear program (ILP) [1]. Consequently the MAP estimate may be computed exactly using advanced computational techniques such as constraint propagation and cutting planes. In addition to a general framework for joint learning, we allow for a complex dependency structure on the collection of units, including group and subgroup structure. This dependency structure can itself be efficiently learned from data ("bi-clustering for structure learning") and we prove that the corresponding joint estimator is the solution to an augmented ILP. Special cases of our methodology facilitate (i) efficient inference for mixture

models where the number of mixture components is itself determined from data, and (ii) an analogue of k-means clustering for DAGs that permits within-cluster variability.

The methodology is motivated by, and demonstrated on, the analysis of a multi-subject functional magnetic resonance imaging (fMRI) experiment [2]. Here it is desired to share information between subjects in order to improve the estimation of subject-specific DAGs $G^{(j)}$ that capture neural connectivity. Our methodology is able to exactly estimate the joint MAP DAGs $G^{(j)}$ under a varying amount of information sharing within a hierarchical model, as displayed in Figure 1. Here a prior distribution over all DAGs is given by $p(G^{(1)}, \ldots, G^{(6)}) \propto \exp(-\lambda \sum_{i,j} \|G^{(i)} - G^{(j)}\|)$, where $\|\cdot\|$ is structural Hamming (or "edit") distance, so that higher values of the hyperparameter λ force the DAGs $G^{(j)}$ to be more similar.

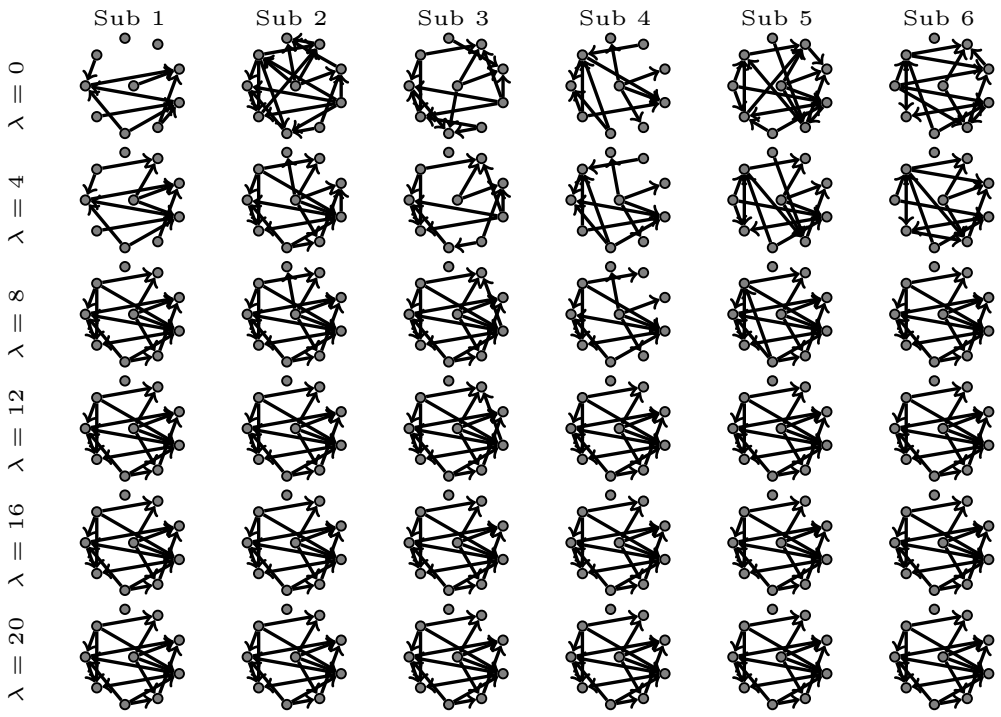

Figure 1: Analysis of fMRI data on six subjects. Here we jointly learn subject-specific DAGs by sharing information within a hierarchical statistical model. Exact estimation is used to identify the joint MAP DAGs for various values of a hyperparameter λ that controls the similarity of the DAGs. [Here the nodes represent 10 activity regions of the motor cortex identified by independent component analysis, as described in [2].]

References

[1] Oates, C.J., Smith, J.Q., Mukherjee, S., Cussens, J. (2014) Exact Estimation of Multiple Directed Acyclic Graphs. *arXiv*:1404.1238.

[2] Oates, C.J., Carneiro da Costa, L., Nichols, T. (2014) Towards a Multi-Subject Analysis of Neural Connectivity. *arXiv*:1404.1239.