Estimation of static community memberships from multiplex and temporal network data

K. Avrachenkov[1], M. Dreveton[1] and L. Leskelä[2]

[1] *Inria Sophia-Antipolis, France, {k.avrachenkov,maximilien.dreveton}@inria.fr*
[2] *Aalto University, Finland, lasse.leskela@aalto.fi*

Data sets in many application domains consist of pairwise interactions observed over time. Pair interactions are often characterized by the type of interacting objects, and a set of objects with a common type is called a community. Community recovery or clustering is the unsupervised task of inferring the community memberships from the observed pair interactions.

As an information-theoretic benchmark, we study data sets generated by a homogeneous block model where the pairwise interactions take values in a measurable space $\mathcal{S}$. The interactions within a block are distributed according to f_{in}, and interactions between blocks according to f_{out}. We derive a lower bound on the expected number of misclassified nodes made by *any* clustering algorithm. This naturally extends the recent results of [3] to a non-asymptotic setting which makes no regularity assumptions on f_{in}, f_{out} nor on the underlying space $\mathcal{S}$ of interaction types. In particular, we can consider a multiplex (dynamic) network where the number of layers (snapshots) grows with the number of nodes N. Then, we show that this bound is achieved by an *ad-hoc* algorithm.

If we denote by D the Rényi-divergence between and f_{in} and f_{out}, then for same-size clusters almost exact recovery (the expected proportion of misclassified nodes going to zero) is possible if $ND \ll 1$, and is impossible otherwise. This provides a natural extension to known results in block models [4]. We later apply those results to dynamic networks where the interaction kernel has a Markov structure, generalizing the results of [2] on a multiplex SBM with independent layers.

In the second part, we propose several clustering algorithms, both offline and online, which fully utilize the temporal nature of the observed data.

For further details and proofs, we refer to our preprint [1].

References

[1] Avrachenkov, K., Dreveton, M., and Leskelä, L. (2020) Estimation of Static Community Memberships from Temporal Network Data. *arXiv:2008.04790.*

[2] Paul, S., and Chen, Y. (2016). Consistent community detection in multi-relational data through restricted multi-layer stochastic blockmodel. *Electronic Journal of Statistics*, **10(2)**, 3807-3870.

[3] Xu, M., Jog, V., and Loh, P. L. (2020). Optimal rates for community estimation in the weighted stochastic block model. *The Annals of Statistics*, **48(1)**, 183-204.

[4] Zhang, A. Y., and Zhou, H. H. (2016). Minimax rates of community detection in stochastic block models. *The Annals of Statistics*, **44(5)**, 2252-2280.

Networked Federated Multi-Task Learning

Y SarcheshmehPour, Y. Tian, L. Zhang and A. Jung[1]

[1] *Aalto University, Finland, alex.jung@aalto.fi*

Many important application domains generate distributed collections of heterogeneous local datasets. These local datasets are often related via an intrinsic network structure that arises from domain-specific notions of similarity between local datasets. Different notions of similarity are induced by spatio-temporal proximity, statistical dependencies or functional relations. We use this network structure to adaptively pool similar local datasets into nearly homogenous training sets for learning tailored models. Our main conceptual contribution is to formulate networked federated learning using the concept of generalized total variation (GTV) minimization as a regularizer. This formulation is highly flexible and can be combined with almost any parametric model including Lasso or deep neural networks. We unify and considerably extend some well-known approaches to federated multi-task learning. Our main algorithmic contribution is a novel federated learning algorithm which is well suited for distributed computing environments such as edge computing over wireless networks. This algorithm is robust against model misspecification and numerical errors arising from limited computational resources including processing time or wireless channel bandwidth. As our main technical contribution, we offer precise conditions on the local models as well on their network structure such that our algorithm learns nearly optimal local models. Our analysis reveals an interesting interplay between the (information-) geometry of local models and the (cluster-) geometry of their network. For further details, we refer to our preprint [1].

References

[1] SarcheshmehPour, Y., Tian, Y., Zhang, L., and Jung, A., "Networked Federated Multi-Task Learning",*arXiv e-prints*, 2021.

Multiple testing of paired null hypotheses using a latent graph model

Tabea Rebafka[1] and Etienne Roquain[1] and Fanny Villers[1]

[1] *Sorbonne Université, Paris, France, fanny.villers@upmc.fr*

We consider pairwise measures between n entities of interest and we want to decide which pairs of entities are significantly related. This multiple testing problem can also be seen as follows : interactions between entities are described via noisy information, resulting in a dense graph with valued edges. Remove noise to detect significant relationships corresponds to estimate an unobserved underlying binary network.

To this end, we develop a multiple testing procedure that incorporates the graph topology : the interaction structure of the underlying graph is learned with a Noisy version of the Stochastic Block Model. Parameter estimates and a node clustering is then provided via a variational expectation-maximization algorithm. It can be shown that our procedure asymptotically mimics the oracle procedure that controls the false discovery rate (FDR, average proportion of errors among the detected edges) while maximizing the true discovery rate (TDR, average proportion of recovered true edges). Numerical experiments illustrate the performance of our test procedure and show that it outperforms classical methods.

References

[1] Rebafka T., Roquain E., Villers F. Graph inference with clustering and false discovery rate control. *arXiv e-prints*, 2021.