Learning causal graphs from continuous or mixed datasets of biological or clinical interest

Mots clés :
- Directeur de thèse : Hervé Isambert
- Co-encadrant(s) :
- Unité de recherche : Physico-Chimie Curie
- École doctorale : École Doctorale Informatique, Télécommunications, Électronique de Paris
- Domaine scientifique principal: Divers

Résumé du projet de recherche (Langue 1)

Information-theoretic methods have become ubiquitous to quantitatively analyze information-rich data of biological or clinical interest. However, most information theoretic analyses can only consider categorical datasets, such as binary data or alphabet-encoded sequences. Yet, many large scale data of practical interest are not readily categorized but rather continuous or continuous-like in nature, such as gene expression levels in single cells, quantitative image analysis of cellular tissues during development, or consist of mixed datasets combining continuous and categorical variables, such as clinical records of hospitalized patients. The objective of this interdisciplinary PhD project is to extend and implement novel information-theoretic methods to learn causal graphs from continuous and mixed datasets from our biologist and clinician collaborators at Institut Curie. We will analyse, in particular, morphogenetic networks shaping the early embryonic development and clinical records of breast cancer patients from Institut Curie. The Isambert team has recently developed a novel inference method to reconstruct causal networks from large scale, yet categorical datasets. This method can learn a broad class of ancestral graph models that include undirected (\(\sim\)), directed (\(\rightarrow\)) and possibly bidirected (\(\sim\rightarrow\)) edges originating from latent common causes, \(\mathcal{L}\), unobserved in the available data (\(\mathcal{L}\)). Affeldt et al. 2016 & 2015, Verny et al. 2016. The statistical and computational approach is based on the analysis of\(^{2}\) and (unifies causal and non-causal network learning frameworks) while including the effects of unobserved \(\mathcal{L}\) which can be seen as hidden confounding factors. In brief, this information-theoretic method starts from a complete graph and iteratively removes dispensable edges, by uncovering significant information contributions from indirect paths, and assesses edge specific confidences from randomization of available data. The remaining edges are then oriented based on the signature of causality in observational data. This information-theoretic approach outperforms existing methods on a broad range of benchmark networks, achieving significantly better results with only ten to hundred times fewer samples and running ten to hundred times faster than the state-of-the-art methods. The method has been applied at different biological scales, from gene regulation in single cells to whole genome duplication in tumor development as well as long term evolution of vertebrates. In all these applications, we provided new insights and testable predictions, Affeldt et al. 2016, Verny et al. 2016. (A limitation of this information theoretic approach remains the treatment of continuous data), for which the estimate of mutual information is notoriously difficult, as mutual information is usually defined on discrete rather than continuous data. Traditional approaches to analyze continuous datasets are typically limited to the specific case of gaussian distributed data, for which a correspondence exists between mutual information and correlation coefficient. However, this approximation is not satisfactory for mixed datasets or datasets of continuous variables with multimodal distributions. (In this PhD project, we will develop and implement a more robust analysis of continuous data based on information theory results). In principle, multivariate information can be defined on continuous as well as discrete data, where continuous information is simply obtained as the limit of infinite dataset discretization (Cover and Thomas 2009). Yet, as actual datasets are always finite, the computation of continuous information need regularization, in practice. This can be done through\(^{3}\) in a polynomial (quadratic) time with respect to the size of the dataset (Kontkanen 2007) by mapping the discretization problem onto an iterative optimization problem of bin number and size distribution. The proposed approach will actually enable the reconstruction of causal networks from datasets consisting either of only continuous variables (e.g. morphogenetic networks in early embryonic development) or datasets with a combination of continuous and discrete variables (e.g. breast cancer clinical data from Institut Curie). The implementation of this unique feature will require however some advanced computation strategies to perform the necessary optimization and we plan to adapt the current code to enable\(^{4}\) using multithreading techniques and/or graphic cards (GPU). GPU computing has in fact attracted a renewed interest in recent years with the implementation of powerful algorithms to perform\(^{5}\) from very large datasets (such as with "deep learning" methods applied to millions of images from the internet). The PhD candidate will also apply the\(^{6}\) (Pearl 2009), recently adapted to genomics data by the team (Singh et al. Cell Rep 2012, Singh et al. PloS Comp Biol 2015, PloS Comp Biol 2014), to analyze the\(^{7}\) in the reconstructed networks from continuous or mixed datasets.

Résumé du projet de recherche (Langue 2)

Information-theoretic methods have become ubiquitous to quantitatively analyze information-rich data of biological or clinical interest. However, most information theoretic analyses can only consider categorical datasets, such as binary data or alphabet-encoded sequences. Yet, many large scale data of practical interest are not readily categorized but rather continuous or continuous-like in nature, such as gene expression levels in single cells, quantitative image analysis of cellular tissues during development, or consist of mixed datasets combining continuous and categorical variables, such as clinical records of hospitalized patients. The objective of this interdisciplinary PhD project is to extend and implement novel information-theoretic methods to learn causal graphs from continuous and mixed datasets from our biologist and clinician collaborators at Institut Curie. We will analyse, in particular, morphogenetic networks shaping the early embryonic development and clinical records of breast cancer patients from Institut Curie. The Isambert team has recently developed a novel inference method to reconstruct causal networks from large scale, yet categorical datasets. This method can learn a broad class of ancestral graph models that include undirected (\(\sim\)), directed (\(\rightarrow\)) and possibly bidirected (\(\sim\rightarrow\)) edges originating from latent common causes, \(\mathcal{L}\), unobserved in the available data (\(\mathcal{L}\)). Affeldt et al. 2016 & 2015, Verny et al. 2016. The statistical and computational approach is based on the analysis of\(^{2}\) and (unifies causal and non-causal network learning frameworks) while including the effects of unobserved \(\mathcal{L}\) which can be seen as hidden confounding factors. In brief, this information-theoretic method starts from a complete graph and iteratively removes dispensable edges, by uncovering significant information contributions from indirect paths, and assesses edge specific confidences from randomization of available data. The remaining edges are then oriented based on the signature of causality in observational data. This information-theoretic approach outperforms existing methods on a broad range of benchmark networks, achieving significantly better results with only ten to hundred times fewer samples and running ten to hundred times faster than the state-of-the-art methods. The method has been applied at different biological scales, from gene regulation in single cells to whole genome duplication in tumor development as well as long term evolution of vertebrates. In all these applications, we provided new insights and testable predictions, Affeldt et al. 2016, Verny et al. 2016. (A limitation of this information theoretic approach remains the treatment of continuous data), for which the estimate of mutual information is notoriously difficult, as mutual information is usually defined on discrete rather than continuous data. Traditional approaches to analyze continuous datasets are typically limited to the specific case of gaussian distributed data, for which a correspondence exists between mutual information and correlation coefficient. However, this approximation is not satisfactory for mixed datasets or datasets of continuous variables with multimodal distributions. (In this PhD project, we will develop and implement a more robust analysis of continuous data based on information theory results). In principle, multivariate information can be defined on continuous as well as discrete data, where continuous information is simply obtained as the limit of infinite dataset discretization (Cover and Thomas 2009). Yet, as actual datasets are always finite, the computation of continuous information need regularization, in practice. This can be done through\(^{3}\) in a polynomial (quadratic) time with respect to the size of the dataset (Kontkanen 2007) by mapping the discretization problem onto an iterative optimization problem of bin number and size distribution. The proposed approach will actually enable the reconstruction of causal networks from datasets consisting either of only continuous variables (e.g. morphogenetic networks in early embryonic development) or datasets with a combination of continuous and discrete variables (e.g. breast cancer clinical data from Institut Curie). The implementation of this unique feature will require however some advanced computation strategies to perform the necessary optimization and we plan to adapt the current code to enable\(^{4}\) using multithreading techniques and/or graphic cards (GPU). GPU computing has in fact attracted a renewed interest in recent years with the implementation of powerful algorithms to perform\(^{5}\) from very large datasets (such as with "deep learning" methods applied to millions of images from the internet). The PhD candidate will also apply the\(^{6}\) (Pearl 2009), recently adapted to genomics data by the team (Singh et al. Cell Rep 2012, Singh et al. PloS Comp Biol 2015, PloS Comp Biol 2014), to analyze the\(^{7}\) in the reconstructed networks from continuous or mixed datasets.
Informations complémentaires (Langue 1)

The results of this PhD project will be presented at international conferences and submitted for publication in international journals.

Informations complémentaires (Langue 2)