Extensions of graphical models with applications in genetics and genomics

Behrouzi, Pariya

Chapter 6

Conclusions

6.1 General overview of thesis

The objectives of this work have been to extend graphical models for different data structures and to enlarge the applicability of graphical models in various fields, particularly in systems genetics. In this thesis, we have developed a method based on undirected graphical models to infer relationships between many components of a system. In addition, we have extended graphical models to high-dimensional time-series data with non-Gaussian structure, where we combined directed and undirected graphical models to explore dynamic and contemporaneous interactions for high-dimensional time series data. We have implemented the proposed methods as R packages on CRAN, which are freely accessible for users.

Genetic systems naturally have a network structure that describes genetic "flow" from DNA to the phenotype through intermediate molecular traits. It is our aim to understand the flow of biological information that underlies complex traits. Powerful methods are required to understand this system. Graph theory has the potential to address long-standing questions in system genetics about complex processes such as epistasis (several loci collectively affecting a single trait) or pleiotropy (a single locus affecting multiple traits). The main challenges are that most models and inference methods are suited for low-dimensional Gaussian data, whereas the genetic system has discrete components, i.e., alleles, and is very high-dimensional. Although non-Gaussian graphical models do exist, they constitute a class that is too large for sensible inference. By making use of a convenient copula approach, we have been able to contain the computational complexity on the one hand and make it suitable for high-dimensional discrete inference on the other.

6.2 Highlight of the results

Chapter 2 introduced a method for reconstructing a conditional independence network from non-Gaussian data, in particular for ordinal and for mixed ordinal-and-continuous data. Such data are common in disciplines like genetics and genomics. In particular, epistasis is the joint effect of various genotype loci on some phenotypic trait. In this chapter, we focused on the trait "survival": we aim to find loci that do not segregate independently conditional on other loci in observed plant samples. These loci show potentially epistatic interactions. Our inference method was completely novel: instead of focusing on recombination fractions, we proposed a penalized Gaussian copula graphical model. This accounts for a large number of markers p and a small number of samples n, which is common in this setting. The method captures the conditionally short- and long-range linkage disequilibrium (LD) dependency structure, which allows us to focus on the unusual marker-marker associations of non-neighboring markers that might be due to epistatic selection rather than gametic linkage. A multi-core implementation of our method has made it feasible to estimate the graph in high-dimensions even when significant portions of data are missing. We demonstrated the efficiency of the proposed method on simulated datasets as well as on genotyping data in *A.thaliana* and maize.

Chapter 3 extended the sparse copula graphical model, as proposed in Chapter 2, for constructing high-quality linkage maps for any biparental diploid and polyploid species. A linkage map contains important genetic information such as the number of chromosomes of a species, the number of markers inside each chromosome, and the order of markers within each chromosome. Until now, recombination fractions are used to estimate linkage maps. This is perhaps straightforward for a haploid population, but for polyploids, this approach starts to suffer more and more from multiple testing problems. Instead, by using the same sparse Gaussian copula graphical models, or via the related nonparanormal skeptic approach, we propose a completely new approach. We discover linkage groups (LGs), typically chromosomes, and the order of markers in each LG by inferring the conditional independence relationships among large numbers of markers in the genome in genotyping studies, such as GWAS. We illustrated the efficiency of the inference method even in the presence of significant amount of missingness of genotype data and genotyping errors. We showed that our method outperforms other available methods in determining the correct number of linkage groups and ordering markers. In addition, we implement the method on real genotype data of barley and potato from diploid and tetraploid populations, respectively. Given that linkage map constructions for most polyploid species like tetraploid

potato have been generated either from diploid populations or from a subset of marker types (e.g. both parents were heterozygous), developing a map construction method based on discrete graphical models makes it possible to construct high-quality linkage maps for any biparental diploid and polyploid species containing all different marker types.

Chapter 4 detailed the R package `netgwas`, which efficiently applies the methods proposed in Chapters 2 and 3. This package contains a set of tools based on undirected graphical models to accomplish three important and interrelated goals in genetics and genomics: linkage map construction, intra– and inter–chromosomal interactions, and high-dimensional genotype-phenotype (and genotype-phenotype-environment) interactions network. More precisely, `netgwas` is able to deal with species with any ploidy level, namely diploid (2 sets), triploid (3 sets), tetraploid (4 sets) and so on. Using the sparse matrix output and the multicore implementation of the `netgwas` package maximizes computational speed and minimizes memory requirements.

Chapter 5 introduced a sparse dynamic chain graph model for network inference in high dimensional non-Gaussian time series data. The proposed method is able to estimate both the (slower) auto-regressive dynamics as well as the (faster) dynamics that happens on a time-scale that is almost instantaneous with respect to the sampling times. The estimation of the parameters in the proposed method relies again on the Gaussian copula graphical models, this time extended in the direction of Bayesian networks under the penalized expectation-maximization (EM) algorithm inference framework. In this chapter we use an efficient coordinate descent algorithm to optimize the penalized log-likelihood with the smoothly clipped absolute deviation penalty. The method is implemented in the R package `tsnetwork`.

6.3 Discussion

Fields such as systems genetics, systems biology, epidemiology, and bioinformatics often involve large-scale models in which thousands of components are linked in complex ways. What is perhaps most distinctive about the graphical model approach is its suitability in formulating probabilistic models of complex phenomena in applied fields, while maintaining control over the computational cost associated with these models. Graphical models use a graph to encode the conditional independence structure between components of a system. It provides a general framework for representing high-dimensional data. Here, we summarize a number of discussion points.

6.3.1 Gaussian copula

Construction of linkage maps is a fundamental step required in a detailed genetic study of diseases and traits. In particular, constructing linkage maps in polyploids, with outcrossing behavior, is a challenging task. So far, based on our experience, no method has been developed to construct polyploid linkage maps for a large number of different marker types without any manual adjustment and visual inspection. We used a penalized Gaussian copula graphical model for the construction of high-quality and high-density linkage maps in diploid species (like humans, containing two copies of each chromosome) and polyploid species (like potato including many other plants, containing more than two copies of each chromosome). The linkage map is inferred through conditional independence relationships among genetic markers in the genome, modeled via a Gaussian copula.
The Gaussian copula implies that we employ merely one parameter to model the interaction between a pair of markers and no parameters are available for higher order interactions between three or more markers. This is computationally convenient, clearly, but this may not correspond to reality. In fact, it is always sensible to do post-hoc checks to see if the fitted model provides a satisfactory fit to the data.

6.3.2 Ordering markers

In our mapping algorithm, we introduce a two-step approach. After fitting the penalized copula graphical model, we use ordering algorithms to project the original high-dimensional space in a one-dimensional map. This two-step approach is computationally efficient, but there are obvious drawbacks. If our objective is to find a one-dimensional map, we could limit our inference to finding the best permuted band-diagonal precision matrix. Although this is in principle possible, it is not a submodel of a penalized copula graphical model and therefore the proposed inference algorithm would not be directly useful. Moreover, the number of permutations scales quadratically in the number of markers and for large genome-wide association studies this method would be practically impossible to implement.

6.3.3 Interpretation of multi-trait networks

To date, most genetic studies designed to map trait loci have focused on single traits. However, both small-scale studies of experimental crosses of model organisms and large-scale studies in humans, animal, and plants often include data collection for multiple traits. For example, studies of human obesity might include multiple measures of obesity, such as the

body mass index, percent fat mass and waist circumference, that are moderately to strongly correlated. Studies might also have measures for related traits, such as obesity, diabetes, and kidney disease. In this thesis, we have developed a method based on graphical models to construct genotype-phenotype networks in GWAS in presence of multiple traits. The proposed genotype-phenotype conditional dependencies network uncovers networks of interactions between loci and traits as well as networks of interactions among traits, and among loci. Biological advantages of performing the proposed graphical-based joint analysis of correlated/related multiple traits include the ability to address the issue of pleiotropy (one locus influencing multiple correlated traits) vs. tight linkage (linked loci each influencing one of the traits) as well as the ability to understand epistasis and how biochemical pathways relate to complex traits.

In genome-wide association study era the greatest challenge lies in combining GWAS findings with multi-trait and multi-environment data to functionally characterize the associations. Multi-trait and multi-environment analyses help geneticists and epidemiologists to fully understand the behavior of complex traits. For example, data in plant breeding science often have a multi-trait multi-environment structure, but until now limited statistical methodology was available to infer genetic and phenotypic interactions in these data. The common approach to analyze multi-trait and multi-environment data is to perform a series of single-trait and single-environment analyses and then combine the results. Methods of analysis for single traits in single environments often have the form of regression models with single error terms combined with least squares procedures for parameter estimation. But, the advances in graph-based models have made it possible to jointly investigate the effect of environmental factors on phenotypes and loci on a genome. In this thesis, we developed a method that involves inferring a web of interactions among the effects of environmental factors on multi-loci and multi-phenotypes data. The proposed network-based method allows a more realistic analysis of data, as accounts for simultaneous associations and interactions among them.

Nevertheless, the interpretation of genotype-phenotype networks is not straightforward. If one trait mediates the interaction of between another trait and a number of markers, then what does this mean? Does this really mean that these markers do not have a direct effect on the second trait or does it merely suggest that the first trait is a more “primitive” trait than the second? And in what sense should this be understood? For the moment, we introduced the genotype-phenotype networks as convenient exploratory tools, but it is clear that the underlying biomedical scientific understanding should provide a possible explanation for such structures that go beyond mere genotyping.

6.3.4 Granger causality

While graphical models originally have been developed for variables that are sampled with independent replications, they have been applied more recently also to the analysis of time dependent data. In this thesis, we introduced a graphical time series model for the analysis of dynamic relationships among variables in multivariate time series. The proposed model derived from combining the main features of Gaussian graphical models and dynamic Bayesian networks. More precisely, the model is based on the notion of Granger causality that can be applied to non-Gaussian (and discrete) high-dimensional time series data. Under Granger causality framework, a time series $x_{i,t}$ is Granger causal of another time series $x_{j,t}$ if inclusion of the history of x_i improves prediction of x_j over knowledge of the history of x_j alone. In these graphs each component series is represented by a single vertex and directed edges indicate possible Granger-causal relationships between variables while undirected edges are used to map the contemporaneous dependence structure. Natural applications of the proposed Granger-causality method are, for instance, in time-series survey data, neuroscience, and neuroimaging.

Although it is tempting to interpret Granger causality in a functional, mechanistic way there are a number of factors that should be taken into consideration. In temporal chain graphical models, such as the one defined in Chapter 5, the arrows do not have the same causal interpretation as the arrows in a causal graphical model, as defined e.g. in Pearl (2000). One issue is that the underlying process operates in continuous time, whereas the sampling time of the process is discrete. A number of the undirected arrows in the chain graphical model, therefore, are really directed arrows for which the direction is unknown. Therefore, in the true underlying causal graph, say, gene one and gene two may both affect gene three in a time scale much shorter than the observation times. This v-structure $1 \rightarrow 3 \leftarrow 2$ will now induce a conditional dependence between genes 1 and 3 fixing 2. Therefore, the "instantaneous" conditional independence structure will be a complete graph between genes 1, 2 and 3, but the link between gene one and gene two is not causal at all.

6.4 Future Work

This section mentions a few of the possible directions for extending and building upon this work.

6.4.1 Epistatic interactions network

Another interesting application of the proposed method in chapter 2 is to study a multi-loci genetic incompatibility. The simplest form of epistatic incompatibility involves the interaction between only two loci, say L_1 and L_2; using the proposed method makes it possible to study high-dimensional incompatibility networks which can be viewed as multi-locus extensions of the classical two-locus Dobzhansky–Muller model. Furthermore, we remark that correcting for population structures inside the network inference of chapter 2 would be biologically interesting.

6.4.2 Linkage map

Regarding linkage map construction for diploids and polyploid species, we plan for further improvements in: (i) constructing a one-step linkage map rather than a two-step procedure of reconstructing marker-marker networks and ordering markers, (ii) calculating the physical position of genetic markers on a genome, particularly for polyploid species, (iii) providing a better quality of ordering markers for polyploids, (iv) making our method computationally faster for large datasets $(p \geq 5000)$.

6.4.3 Directed graphs for mixed discrete-continuous data

Undirected graphs are used to model symmetric relationships between variables, whereas directed acyclic graphs (DAG) are used to model asymmetric cause-effect relationships. The proposed Gaussian copula undirected graph model can be extended to a partially directed acyclic graph — a graph which contains both directed and undirected edges, with no directed cycle in its directed subgraph. We can think of the extension as a two-step approach for estimating the causal structure underlying a Gaussian copula model on multivariate mixed data. The essence is to estimate the precision matrix in the latent space, which can then be given to any causal discovery algorithm to search for its underlying structure. We can gain a reliability of structure estimates by generating samples, and running the algorithm several times to gain an insight into the reliability of structure estimates. Similar procedures can be done by bootstrapping the original dataset.

A wide range of applications can benefit from the extension of our method to DAGs including the problem of inferring causal relationships among non-Gaussian phenotypic traits and gene expression networks.

6.4.4 Nonlinear dynamic time-series network

The proposed dynamic chain graph model in Chapter 5 is based on a linear relation between time-series components. We would like to extend the methodology to a higher-oder autoregressive process of order d ($d \geq 2$).

6.4.5 Network inference and modeling networks data

Networks are a powerful way of describing the complex relationships among a large number of variables. Many biological networks (such as gene-gene and protein-protein interactions networks) and neuroscience networks (such as brain-connectivity networks) are complex. Understanding relations encoded in complex networks is a challenging task. Both methodologically and practically, there is great interest in understanding the interactions in complex networks to an extent that enables to summarize and simplify these networks.

Up until now, there have been two classes of network modeling: the first regarded only the nodes as given in order to discover the network, ignoring the underlying structure; the second regarded the network as given in order to model its underlying structure, ignoring the uncertainty of the network. An extension of this work would be to combine the power of these two classes to better understand complex systems such as systems genetics, systems biology, and brain connectivity. This extension can be built on the methodology in Chapter 2, where the proposed sparse Gaussian copula graphical model can be combined with exponential random graph models.

A wide range of applications can benefit from combining these two classes, mainly in genetics and neuroscience. The possible applications would be to infer gene enrichment analysis and enrichment analysis of brain connectivity to treat Alzheimer disease.