

University of Groningen

Extensions of graphical models with applications in genetics and genomics

Behrouzi, Pariya

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version

Publisher's PDF, also known as Version of record

Publication date:

2018

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Behrouzi, P. (2018). *Extensions of graphical models with applications in genetics and genomics*. [Thesis fully internal (DIV), University of Groningen]. University of Groningen.

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Summary

In this thesis, we address several problems related to modeling complex systems. 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. In real world, not all datasets are continuous. Discrete data or mixed discrete-and-continuous datasets routinely arise in above-mentioned fields.

In Chapter 2 we introduce 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 systems genetics, where the main focus is to understand the flow of biological information that underlies complex traits. In this chapter, we focus on the trait “survival”: we aim to find loci – locations on a genome – that do not segregate independently conditional on other loci. The network estimation relies on penalized Gaussian copula graphical models; this accounts for a large number of markers p and a small number of individuals n .

In Chapter 3 we extend 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. In the proposed map construction method we discover linkage groups, typically chromosomes, and the order of markers in each linkage group by inferring the conditional independence relationships among large numbers of markers in the genome in genotyping studies, such as genome-wide association studies.

In Chapter 4, we introduce an 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.

In Chapter 5 we introduce sparse dynamic chain graph models for network inference in high-dimensional non-Gaussian time series data. The proposed method is parametrized by a precision matrix that encodes intra time-slice conditional independences among variables at a fixed time point, and an autoregressive coefficient that contains dynamic conditional independence interactions among time series components across consecutive time steps. We apply our method to a Netherlands Study of Depression and Anxiety (NESDA) dataset to determine the psychological factors that influence the development and long-term prognosis of anxiety and depression.