Network modeling and integrative analysis of high-dimensional genomic data

Jonatan Kallus

Akademisk avhandling som för avläggande av filosofie doktorsexamen i matematisk statistik försvaras vid offentlig disputation **onsdagen den 10 juni 2020 klockan 13.15 i sal Pascal**, Matematiska vetenskaper, Chalmers tvärgata 3, Göteborg.

Avhandlingen försvaras på engelska.
Fakultetsopponent är professor Magne Thoresen, institutionen för biostatistik, Universitetet i Oslo.

GÖTEBORGS UNIVERSITET

Institutionen för matematiska vetenskaper
Göteborgs universitet och Chalmers tekniska högskola
SE-412 96 Göteborg
Telefon: 031-772 1000

ISBN 978-91-7833-888-7 (tryckt), ISBN 978-91-7833-889-4 (elektronisk)
Tillgänglig via http://hdl.handle.net/2077/63747

Network modeling and integrative analysis of high-dimensional genomic data

Jonatan Kallus

Division of Applied Mathematics and Statistics
Department of Mathematical Sciences
University of Gothenburg and Chalmers University of Technology

Abstract

Genomic data describe biological systems on the molecular level and are, due to the immense diversity of life, high-dimensional. Network modeling and integrative analysis are powerful methods to interpret genomic data. However, network modeling is limited by the requirement to select model complexity and due to a bias towards biologically unrealistic network structures. Furthermore, there is a need to be able to integratively analyze data sets describing a wider range of different biological aspects, studies and groups of subjects. This thesis aims to address these challenges by using resampling to control the false discovery rate (FDR) of edges, by combining resampling-based network modeling with a biologically realistic assumption on the structure and by increasing the richness of data sets that can be accommodated in integrative analysis, while facilitating the interpretation of results. In paper I, a statistical model for the number of times each edge is included in network estimates across resamples is proposed, to allow for estimation of how the FDR is affected by sparsity. Accuracy is improved compared to state-of-the-art methods, and in a network estimated for cancer data all hub genes have documented cancer-related functions. In paper II, a new method for integrative analysis is proposed. The method, based on matrix factorization, introduces a versatile objective function that allows for the study of more complex data sets and easier interpretation of results. The power of the method as an explorative tool is demonstrated on a set of genomic data. In paper III, network estimation across resamples is combined with repeated community detection to compensate for the structural bias inherent in common network estimation methods. For estimation of the regulatory network in human cancer, this compensation leads to an increased overlap with a database of gene interactions. Software implementations of the presented methods have been published. The contributed methods further the understanding that can be gained from high-dimensional genomic data, and may thus help to devise new treatments and diagnostics for cancer and other diseases.

Keywords: graphical modeling, biomolecular interactions, sparsity, model selection, resampling, stability selection, community detection, matrix factorization, Euler parametrization, bi-clustering

ISBN 978-91-7833-888-7 (print)
ISBN 978-91-7833-889-4 (electronic)
Available at http://hdl.handle.net/2077/63747