Statistical methods for the analysis of complex omics data

Tissier, R.

Author: Tissier, R.
Title: Statistical methods for the analysis of complex omics data
Issue Date: 2018-12-04

Summary

This dissertation focuses on the development of new statistical methods designed to take into account existing structures inside omic datasets. The major challenge in analysing omic datasets is the strong dependencies which are present. Taking into account and modelling the different dependency structures can lead to further improvements of our knowledge of the biological mechanisms. Therefore, improving our ability to predict diseases.

Chapter 1 provides a general introduction to the existing dependency structures possibly faced when studying omic datasets. First, the most common measure of dependence is described, i.e. the Pearson correlation coefficient. Next, the different dependency structures are described. Namely, dependencies between individuals, between outcome measures and between omic features. For each of these dependency levels the challenges faced and the commonly used methods are described.

Chapters 2 and 3 present methods for the analysis of secondary phenotypes in ascertained family studies. Chapter 2 presents a new approach to analyse secondary phenotype for the multiple case family design. Where families are selected when they have at least a specific number of cases. The proposed method is illustrated by a data example obtained from the Leiden Longevity Study, which is a multiple-cases family study that investigates human longevity (primary phenotype). Here the association between, triglyceride levels and glucose (secondary phenotypes), and genetic markers was estimated. Chapter 3 presents methods used in the literature for secondary phenotype analysis for the proband family design. This design comprises family members of specific probands (often cases with the primary outcome). These methods are then compared with the method previously developed in Chapter 2. The real data analysis presented in this chapter is part of the Social Anxiety Disorder (SAD) family study, and aims to identify possible endophenotypes of SAD.

Chapter 2 develops an approach to obtain unbiased association estimates between secondary phenotypes and biomarkers as well as unbiased heritability estimates of the available secondary phenotypes. This method accommodates the ascertainment process while explicitly modelling the familial relationships. To do so, Our approach uses the retrospective likelihood in order to correct for the ascertainment process with existing methods for mixed-effects models. The retrospective likelihood approach automatically corrects for the ascertainment. A multivariate probit model is used to capture the association between the mixed type primary (binary variable) and secondary phenotypes (continuous variable).

Estimates are then obtained by maximizing the log-likelihood.

An important empirical finding is that the heritability estimates for the secondary traits can be severely underestimated unless the sampling mechanism is taken into account. Extensive simulations show that the presented method preserves the type I error at nominal level and provides accurate estimates irrespective of the disease prevalence, the strength of the association between the genetic variants and the primary phenotype, and the ascertainment mechanism. Currently, a key limitation of this approach is the computational time of multivariate integrals, especially in case of large pedigrees.

Chapter 3 investigates the performances of the previous method, from Chapter 2, for the analysis of proband family study design. Theses performances are compared with methods currently used in the litterature. Namely, ignoring the ascertainment process and modelling the conditional distribution of the secondary phenotype values of the families given the secondary phenotypes of the probands. Furthermore, we propose an extension of the latter approach, by modelling the joint conditional distribution of the primary and secondary phenotype values of the families given the joint distribution of the primary and secondary phenotypes of the probands.

Extensive simulations show that only the retrospective likelihood approach developed in Chapter 2 is able to obtain unbiased heritability estimates of the secondary phenotype as well as association estimates of the secondary phenotypes with genetic markers. Furthermore, conditioning on the secondary phenotype values of the proband can severely underestimate heritability estimates and therefore limiting the identification of candidate endophenotypes of primary phenotypes. Only the retrospective likelihood approach could identify a candidate endophenotypes of SAD in the real data analysis . Another important key point of this chapter is that current methods provide biased estimates when the proband information is missing. Therefore, the use of such study design should not, at this time, be considered.

Chapter 4 considers the problem of conducting gene co-expression network analysis for family studies. A large between-family variation in expression levels could severely bias the network structure obtained if the pedigree structure is not taken into account. To overcome this issue, we propose a meta-analytic approach. We first build the omic network for each pedigree to identify clusters of correlated microarray probes. The eigengene (first principal component) of each cluster of each pedigree are then tested for association with a phenotype of interest. After identification of the most strongly associated cluster, clusters presenting the largest overlap with this cluster in each family are then combined with this one. Finally, the eigengene of the combined cluster is then tested for association with the phenotype. This method was used for analysis of the simulated dataset provided for the Genetic Analysis Workshop 18. This method was compared with methods such as: single probe analysis, ignoring the pedigree structure, and build the network on "decorrelated" omic variables.

Chapter 5 and Chapter 6 presents new methods to incorporate grouping information in prediction models in order to obtain more stable and possibly interpretable models. All the analyses shown in these chapters are using data from the DIetary, Lifestyle, and Genetic determinants of Obesity and Metabolic syndrome study (DILGOM) and the publicly available breast cancer cell lines pharmacogenomics dataset for illustration.

In Chapter 5, a new strategy for model selection based on three steps is presented : Network construction of omic features, empirical derivation of modules of related feature via clustering, and construction of prediction model incorporating the grouping information. This approach aims to overcome issues caused by the presence of strong correlations. Several methods are considered to performs steps 1 and 3 of the developed approach. We compare the performance of this strategy with standard regularized regression such as lasso, ridge regression, and elastic net via simulations.

Simulation and data application results show that this strategy provide more stable prediction models and can perform, in terms of prediction accuracy, as well as standard regularized regression. Indeed methods such as lasso or elastic net tend to select randomly one variable from group of strongly correlated variable leading to unstable models and, therefore, the results are hard to reproduce. Comparisons in prediction performance of the various combinations of network approaches and prediction models allows us to provide guidelines in which combination of methods to use. The combination of graphical lasso and group lasso is overall the best performing approach. However, in large datasets the use of WGCNA instead of graphical lasso is preferred due to the intensive computations needed for graphical lasso.

Chapter 6 studies how to use different omics datasets simultaneously in prediction models. Combining several omic sources in one prediction model is challenging due the presence of strong heterogeneity between omic sources. Heterogeneity in terms of dimensionality, normalization procedures, and error structures. In this chapter we propose several strategies to integrate two omic sources in one prediction model. Specifically, we propose three strategies: 1) stacking both omic sources together and applying the approach proposed in Chapter 5, 2) performing network construction and clustering on each omic source separately and build the prediction model, 3) performing network construction and clustering on each omic source separately, identifying correlation between clusters and between omic sources, and incorporation of this information in the prediction model. The data examples in this chapters comprise metabolomics and transcriptomics datasets from Dilgom and, and Copy number variants and gene expression from the breast cancer cell lines pharmacogenomics dataset.

The key components of our proposed approach are to capture groups of correlated features within and between omic datasets and to include this information by a group penalization model. Simulations results showed that naively stacking datasets is usually not a good strategy as it often perform worse than a model based on a single omic datasets. Including information about the correlation between the omic datasets might improve the prediction accuracy. When the noise structures from both omic sources are different, performing the network analysis and clustering on each omic sources separately proved to be more robust in terms of predictive accuracy than stacking the datasets together.