Abstract

Combining brain imaging and genetic data using fast and efficient multivariate correlation analysis

Claudia Grellmann

(Dissertation, 2017)

Many human neurological and psychiatric disorders are substantially heritable and there is growing inter-est in searching for genetic variants explaining variability in disease-induced alterations of brain anatomy and function, as measured using neuroimaging techniques. The standard analysis approach in genetic neuroimaging is the mass-univariate linear modeling approach, which is disadvantageous, since it cannot account for dependencies among collinear variables and has to be corrected for multiple testing. In con-trast, multivariate methods include combined information from multiple variants simultaneously into the analysis, and can therefore account for the correlation structure in both the neuroimaging and the genetic data. Partial Least Squares Analysis and Canonical Correlation Analysis are common multivariate ap-proaches and different variants have been established for genetic neuroimaging. However, a compre-hensive comparison with respect to data characteristics and strengths and weaknesses of these methods was missing to date. This thesis elaborately compared three multivariate techniques, Sparse Canonical Correlation Analysis (Sparse CCA), Bayesian Inter-Battery Factor Analysis (Bayesian IBFA) and Partial Least Squares Corre-lation (PLSC) in order to express a clear statement on which method in to choose for analysis in genetic neuroimaging. It was shown that for highly collinear neuroimaging data, Bayesian IBFA could not be recommended, since additional post-processing steps were required to differentiate between causal and non-informative components. In contrast, Sparse CCA and PLSC were suitable for genetic neuroimaging data. Among the two, the use of Sparse CCA was recommended in situations with relatively low-dimensional neuroimaging and genetic data, since its predictive power was higher when data dimension-ality was below 400 times sample size. For higher dimensionalities, the predictive power of PLSC ex-ceeded that of Sparse CCA. Thus, for multivariate modeling of high-dimensional neuroimaging-genetics-associations, a preference for the usage of PLSC was indicated. The remainder of this thesis dealt with the improvement of the computational efficiency of multivariate statistics in genetic neuroimaging, since it can be expected that there will be a growth in cost- and time-efficient DNA sequencing as well as neuroimaging techniques in the coming years, which will result in excessively long computation times due to increasing data dimensionality. To accommodate this large number of variables, a new and computational efficient statistical approach named PLSC-RP was pro-posed, which incorporates a method for dimensionality reduction named Random projection (RP) into traditional PLSC in order to represent the originally high-dimensional data in lower dimensional spaces. Subsequently, PLSC is used for multivariate analysis of compressed data sets. Finally, the results are transformed back to the original spaces to enable the interpretation of original variables. It was demon-strated that the usage of PLSC-RP reduced computation times from hours to seconds compared to its state-of-the-art counterpart PLSC. Nonetheless, the accuracy of the results was not impaired, since the results of PLSC-RP and PLSC were statistically equivalent. Furthermore, PLSC-RP could be used for inte-grative analysis of data sets containing high-dimensional neuroimaging data, high-dimensional genetic data or both, and was therefore shown to be independent of the statistical data type. Thus, PLSC-RP opens up a wide range of possible applications.

Keywords
genetics, neuroimaging, multivariate, correlation, dimensionality