Carine Legrand
Dr. sc. hum.

Exploring and controlling for underlying structure in genome and microbiome case-control association studies

Fach/Einrichtung: DKFZ
Doktorvater: Prof. Dr. rer. nat. Frank Lyko

Case-control association studies in human genetics and microbiome pave the way to personalized medicine by enabling a personalized risk assessment, improved prognosis, or allowing an early diagnosis. However, confounding due to population structure, or other unobserved factors, can produce spurious findings or mask true associations, if not detected and corrected for. As a consequence, underlying structure improperly accounted for could explain lack of power or some unsuccessful replications observed in case-control association studies. Besides, points considered as outliers are commonly removed in such studies although they do not always correspond to technical errors. A wealth of methods exist to determine structure in genetic and microbiome association studies. However, there are few systematic comparisons between these methods in the frame of genetic or microbiome association studies, and even less attempts to apply robust methods, which produce stable estimates of confounding underlying structure, and which are able to incorporate information from outliers without degrading estimates quality.

Consequently, the aim of this thesis was to detect and control robustly for underlying confounding structure in genetic and microbiome data, by comparing systematically the most relevant standard and robust forms of principal components analysis (PCA) or multidimensional scaling (MDS) based methods, and by contributing new robust methods. Own contributions include robustification of existing methods, adaption to the genetic or to the microbiome framework, and a dimensionality exploration and reduction method, nSimplices. Analysed datasets include a first synthetic example with a low-variance 2-groups confounding structure, a second synthetic example with a simple linear underlying structure, genome-wide single nucleotide polymorphism (SNP) from 860 case and control individuals enrolled in the European Prospective Investigation into Cancer and nutrition (EPIC prostate), and finally, 2 255 microbiome samples from the human microbiome project (HMP). Synthetic or real outliers were added in the second example and in EPIC and HMP datasets. All meaningful existing and contributed methods were applied to the EPIC and HMP datasets, while a restricted set was applied to the synthetic, illustrative examples. The 10 principal components or top axes resulting from each method were kept for further analysis. Quality of a method was assessed by how well these axes summarized the underlying structure (using Akaike's information criterion -AIC- from the regression of the 10 axes on known underlying structure in the data), and by how robust the estimates stayed in the presence of outliers (adjusted R^2 from the regression of each outlier-disturbed axis on the original axis).

In synthetic example 1, only ICA was able to uncover the low-variance confounding structure, whereas PCA or MDS failed to do so, in agreement with the fact that these methods detect large rather than small variance or distance components. In synthetic example 2, non-metric MDS remained the most representative and robust method when distance outliers are included, while nSimplices combined with classical MDS was the only method to stay

representative and robust if contextual outliers are present. In the EPIC dataset, Eigenstrat was the most representative method (AIC of 782.8) whereas sample ancestry was best captured by new method gMCD (unbiased genetic relatedness estimates used in a Minimum Covariance Determinant procedure). Methods gMCD, spherical PCA, IBS (MDS on Identity-by-State estimates) and nSimplices were more robust than Eigenstrat, with a small to moderate loss in terms of representativity (AIC between 789.6 and 864.9). Association testing yielded p-values comparable with published values on candidate SNPs. Further SNPs rs8071475, rs3799631, rs2589118 with lowest p-value were identified, whose known role in other disorders could point to an indirect link with prostate cancer. In the HMP dataset, the new method nSimplices combined to data-driven normalization method qMDS mirrored best the underlying structure. The most robust method was qMDS (with nSimplices or alone), followed by CSS and MDS. Lastly, the original method nSimplices performed in all settings at least comparably (except ancestry in EPIC), and in some cases considerably better than other methods, while remaining tractable and fast in high-dimensional datasets.

The improved performance of gMCD and qMDS agrees with the fact that these methods use adapted measures (genetic relatedness, selected model distribution, respectively) and recognized robust approaches (minimum covariance determinant and quantiles). Conversely, wMDS is likely to have failed because variance is not an adequate parameter for microbiome data. More generally, different methods report the underlying structure differently and are advantageous in different settings, for example PCA or non-metric MDS were best in some settings but failed in other. Finally, the original method nSimplices proved useful or markedly better in a variety of settings, with the exception of highly noisy datasets, and provided that distance outliers are corrected.

Current genetic case-control association studies tend to integrate several types of data, for example clinical and SNP data, or several omics datasets. These approaches are promising but could be subject to increased inaccuracies or replication issues, by the mere combination of several sources of data. This motivates a reinforced use of robust methods, which are able to mirror accurately and steadily genetic information, such as gMCD, nSimplices or spherical PCA. Nevertheless, results on Eigenstrat show this stays a reasonable method. Results in microbiome confirmed that MDS based on proportions is a suboptimal method, and suggested the exponential distribution should be considered instead of multinomial-based distributions, certainly because the exponential better represents the inherent competitiveness between phylogenies in the microbiome. Moreover, illustrative and real world examples showed that methods could capture relevant, but different information, encouraging to apply several complementary methods when starting to explore a dataset. In particular, a low-variance confounder could stay undetected in some methods. Additionally, methods based on least absolute residuals revealed several shortcomings in spite of their utility in a univariate frame, but their expected benefit in a multivariate setting should motivate the development of more tractable implementations.

Finally, SPH, IBS, gMCD are recommended methods in a genetic SNP dataset, while Eigenstrat should perform best if no more than 2% outliers are present. To mirror structure in a microbiome dataset, nSimplices (combined with qMDS, or with CSS) can be expected to perform best, whereas MDS on proportions is likely to underperform. Method nSimplices proved beneficial or largely better in various situations and should therefore be considered to analyse datasets including, but not limited to, genetic SNP and microbiome abundances.