# Hierarchical inference for genome-wide association studies

Claude Renaux[1], Laura Buzdugan[1], Markus Kalisch[1], Peter Bühlmann[1]
[1]Seminar for Statistics, D-MATH, ETH Zürich

## Seminar for Statistics

## 1 Introduction

The goal is to perform high-dimensional statistical inference for genome-wide association studies (GWAS). We develop meta analysis for multiple studies and novel software in terms of an R-package hierinf. Inference and assessment of significance is based on very high-dimensional multivariate (generalized) linear models: in contrast to often used marginal approaches, this provides a step towards more causal-oriented inference.

## 2 Method overview

Hierarchical inference is a key technique for computationally and statistically efficient hypothesis testing and multiple testing adjustment. It address the problems of high pairwise absolute empirical correlation between covariates, or near linear dependence among a small set of covariates.

To summarize the method, one starts by clustering the data hierarchically. This means that the clusters can be represented by a tree. The main idea is to pursue testing top-down and successively moving downwards until the null-hypotheses cannot be rejected. The p-value of a given cluster is calculated based on the multiple sample splitting approach and aggregation of those p-values

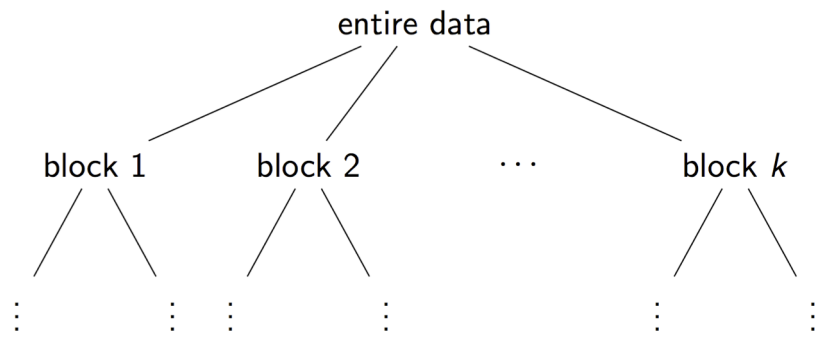


**Fig. 1.** From a computational perspective, it makes sense to pre-define dis-joint sets of SNPs, say, chromosomes which can be clustered separately.

## 3 R package hierinf

The R package hierinf will be available on bioconductor. There are several advantages:

- only two function calls
- works for multiple studies / data sets (aggregate p-values)
- package can be applied generally (not specific for GWAS)
- easy to run testing in parallel (set seed for reproducibility, multicore or snow)

```
# Cluster the SNPs
dendr <- cluster_var(x = sim.geno, block = block)

# Test the hierarchy using multi sample split
result <- test_hierarchy(x = sim.geno, y = sim.pheno,
                         covar = sim.covar, dendr = dendr,
                         block = block, family = "binomial",
                         parallel = "multicore", ncpus = 2)
```

## 4 Aggregating p-values compared to pooling

The results in the figure show that there are situations where aggregating the p-values is clearly better than pooling (while pooling is never substantially better than aggregating p-values). This is to be expected: pooling is conceptually wrong and aggregating p-values should not be much worse even in the homogeneous case where the multiple datasets have the same distribution (the same parameters).

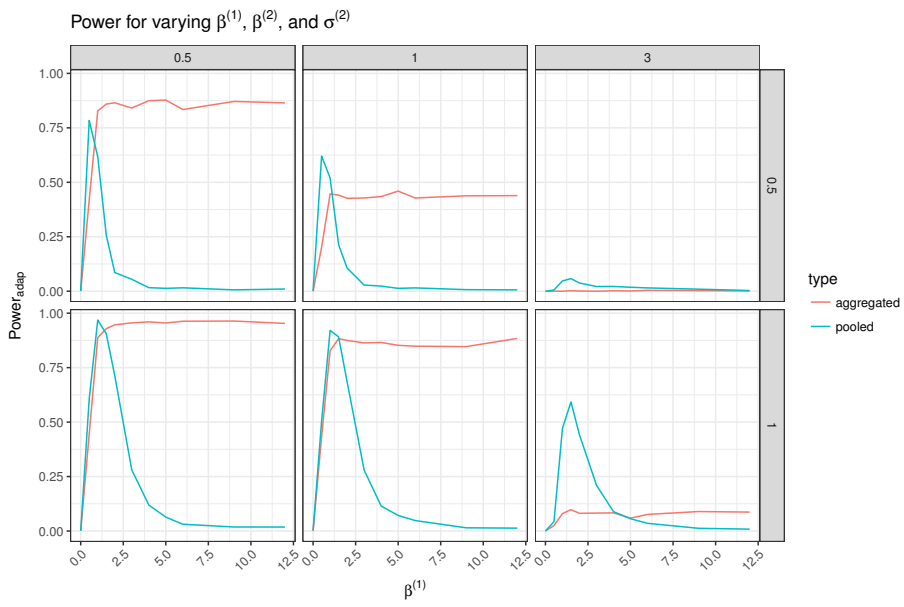


**Fig. 2.** We simulate two studies using a linear model and semi-synthetic data based on raw genotyping data from openSNP.org (under CC0 license). Each data set has 300 observations and 1000 covariates. Both data sets share the same 10 active variables. The adaptive power is calculated to compare aggregating p-values and simply pooling the data sets.

## 5 Conclusion

Inferring statistical significance in such high-dimensional settings of genome-wide association studies is very challenging.

- Hierarchical inference is a very natural and powerful approach towards better and more reliable inference.
- We promote the use of multivariate models.
- We advocate the use of meta-analysis within a single hierarchical structure which is simple and coherent.
- New implementation in the R-package hierinf provides many possibilities: two options for constructing hierarchical structures, fitting linear and logistic linear response models with possible additional adjustment for external control variables, and efficient parallel computation.

## 6 References

1. Claude Renaux, Laura Buzdugan, Markus Kalisch, Peter (2018). Hierarchical inference for genome-wide association studies: a view on methodology with software. Preprint arXiv:1805.02988
2. Renaux C, Buzdugan L, Kalisch M, Bühlmann P (2018). hierinf: Hierarchical Inference. R package version 0.99.3. https://bioconductor.org/packages/devel/bioc/html/hierinf.html