Combining information

7

General Summary

In this thesis, some aspects of statistical methods for combining information are examined. There are two main contexts in general: meta-analysis and combining p-values. A common issue present in most of the dissertation is the combination of non-independent data. Furthermore, the thesis include an examination of model selection techniques in meta-regression.

In **Chapter 2**, we examined model selection techniques in meta-regression. Meta-regression associates the characteristics of the primary studies whose results are synthesized in the meta-analysis to the effect sizes using regression-type methods. Addressing such associations helps to explain the heterogeneity among the effect sizes that arises from varying study characteristics. However, it is crucial to identify the true model that provides the best approximation to the mechanism that underlies the data. Although common model selection techniques based on testing are applicable, their quality is debatable mostly due to their multiple testing nature. On the other hand, information-theoretic techniques are also applicable in the context of meta-regression which may circumvent the issues in the applications of selection via testing techniques. Furthermore, there is a debate on the likelihood estimation techniques. Some argue that restricted maximum likelihood (REML) estimation is not appropriate for comparing models that differ in their fixed effects; thus, maximum likelihood (ML) estimation should be preferred. **Chapter 2** demonstrates the applications of all model selection methods in an illustrative example and provides a simulation study to compare their performance under a variety of conditions. The results imply that selection via testing techniques is largely outperformed by information-theoretic approaches. The latter identify the true model with higher chances than conventional techniques. Moreover, REML estimation has a similar or better probability of identifying the true model than ML estimation. Therefore, the results of this chapter suggest the use of information-theoretic model selection techniques more widely than the conventional techniques in the context of meta-regression.

Next, **Chapter 3** presents a simulation study to investigate meta-analytic models on correlated effect sizes, specifically in ecology. Regular meta-analytic models do not address two issues that commonly arise in ecology. First, ecology studies mostly examine multiple species that share a common evolutionary history known as a phylogeny which violates the independence assumption in regular meta-analytic models. Second, ecology studies usually report multiple effect sizes which introduces within-study dependence. As regular meta-analytic models assume one effect size per study, this phenomenon is also ignored, and summarizing effect sizes within studies is somehow a questionable strategy that may lead to severe information loss. A phylogenetic multilevel meta-analytic

model has been proposed in the literature that potentially addresses these issues by i) decomposing the between-species variance into non-phylogenetic and phylogenetic random-effects components and ii) adopting a multilevel framework to incorporate a study-level random effect. Although both of these issues can be addressed within this framework, whether it can provide an advantage over regular meta-analytic models is unknown. For example, the complexity of such a model may pose a threat to model convergence which is why some meta-analytic studies include only the phylogenetic variance component to reduce the model's complexity. Our simulation results show that the overall mean can be estimated with a little or no bias with both regular and complex models. However, it is essential to employ the most complex model to estimate the uncertainty in the overall mean estimation to derive rejection rates close to the nominal level. In addition, the complex model is also successful in estimating the variance components unbiasedly given that there is at least a moderate level of phylogeny. An important finding regarding this issue is that the simplified model that does not include the non-phylogenetic variance component tends to overestimate the phylogenetic variance component, leading to highly conservative results. Finally, our results show that the complex model can be fitted with little or no convergence issues, making it easily applicable. Therefore, the study in **Chapter 3** shows that the complex multilevel model should be preferred in ecological meta-analyses.

Chapter 4 switches the focus to the combination of dependent p-values and its application in genetics. Genome-wide association studies (GWAS) examine the genetic contribution to a disease by examining the association between single-nucleotide polymorphisms (SNPs) and some phenotype of interest. Nowadays, more than a million SNPs are easily genotyped in GWAS, and testing such a large number of SNPs simultaneously inflates the Type I error rate quickly. Controlling the Type I error rate with conventional methods such as the Bonferroni correction reduces the study power severely due to the large number of simultaneous tests. An alternative analysis to circumvent this issue is gene-based testing which shifts the focus of the study to the gene-level where the number of simultaneous tests can be decreased dramatically. Gene-based testing combines the evidence of individual SNPs that belong to a gene which can statistically be performed by combining their p-values. Although there are well-known methods for combining p-values, they usually assume independence among the p-values which is violated due to the non-random associations among SNPs known as linkage disequilibrium (LD). Several techniques are available to adjust the methods for combining p-values for dependence among the p-values. In **Chapter 4**, we review the most well-

known statistical methods for combining p-values and adjustment techniques for dependency and provide a simulation study to compare their performance under a variety of conditions. Our results show that incorporating the LD information is essential for controlling the Type I error rate. The methods based on multiple testing adjustments (i.e., the Bonferroni and Tippett methods) are overly conservative, while the other techniques under examination are liberal. A generalization of Fisher's method, known as Brown's method, is the most successful technique to control the Type I error rate while retaining sufficient power; however, it is also essential to use its proper generalization for two-sided tests. Brown's method also has a robust performance regarding the number of SNPs and the degree of LD between them in a gene. On the other hand, empirical techniques that do not require a full permutation procedure also perform adequately and, along with Brown's method, provide a large benefit compared to the commonly used Bonferroni method.

Finally, **Chapter 5** introduces the **poolr** package for the open-source statistical analysis software `R` that implements the methods for combining dependent p-values investigated in **Chapter 4**. Furthermore, the **poolr** package implements a method based on the inverse chi-square distribution. The package implements adjustment techniques based on empirical distributions (also included in the simulation in **Chapter 4**) and shows that these adjustments can mimic the results that would be obtained by proper permutation-based methods much quicker. **Chapter 5** presents the use of the **poolr** package with illustrative examples which can be used as a tutorial for researchers.