Of Zeros and Ones

Rauschenberger, A.

2021

document version
Publisher's PDF, also known as Version of record

"And the earth was full of zeros and ones."
(cf. Genesis 1:2)

Summary

This thesis is about association and prediction in genomics. The first part is about association (Chapters 2, 3 and 4) and the second part is about prediction (Chapters 5 and 6). The methods we developed have very practical applications.

- **Negative binomial global test** (Chapter 2)

  The aim is to test for association between an overdispersed count outcome and high-dimensional features, e.g. between RNA-Seq gene expression and SNPs (eQTLs). Instead of testing each feature separately, we test all features simultaneously.

  Here comes a little story to explain this idea: Bicycle safety is a top priority in the Netherlands. After eating breakfast and putting on a helmet, the typical Dutch first checks whether his or her bike is safe to ride. Does the bell ring loud enough to chase away tourists? Are the brakes working correctly? (This is crucial for steep descents.) Are the tyres properly inflated to safely carry a shopping bag on the right arm and two children on the left arm? Then everything is safe. Instead of testing these things one by one, we propose to test them all at once. If everything is fine, nothing stops our brave cyclist. And if there is a problem somewhere, we can still check later where the problem lies.

- **Semi-supervised mixture test** (Chapter 3)

  The aim is to test for main and interactive effects of high-dimensional binary features on a continuous or discrete outcome, e.g. of SNPs on a quantitative trait (GWAS). Instead of testing all pairwise interactions between features, we test for each feature whether it has a main or interactive effect.

Let's explain this in other words: Imagine the average statistician supervising an exam. As students should work on their own, the statistician doesn't simply sit around and read some stats news, but he actively searches for cheaters. His numerous years of studies allow him to adopt a systematic approach. He therefore checks whether the first and the second student cooperate, whether the first and the third student cooperate, and so on. Then he checks whether the second and the third student cooperate, whether the second and the fourth student cooperate, and so on. But once the exam is finished, he is only halfway through. Here we propose a more efficient approach. Instead of verifying for each pair of students whether they cooperate, we verify for each student whether he or she cooperates with another student.

- **Multinomial global test** (Chapter 4)

  The aim is to test for association between a multivariate multinomial outcome and high-dimensional features, e.g. between RNA-Seq exon expressions of a gene and SNPs (sQTLs). We are interested in effects on the relative (not absolute) frequencies of the categories.

  PhD students shouldn't make fun of their supervisor's research. This is why I won't write about absolute and relative amounts of ice cream but merely refer the reader to the scientific summary.

- **Paired lasso** (Chapter 5)

  The aim is to predict an outcome from two sets of features, where each feature in one set forms a pair with a feature in the other set, e.g. two transformations of one molecular profile. We combine information from both sets and account for their paired structure.

  Now a simple example: Belgium is a neatly organised country with five (known) layers of government: one nation, three regions, three communities, ten provinces, and many municipalities. An ambitious team of German researchers is currently trying to understand the details. But they always get lost on field trips.

The problem is that most Belgian towns have (among others) one Dutch and one French name, such as Luik/Liège, Bergen/Mons or Mechelen/Malines. If our researchers used a Dutch map, they would get lost in the South. If they used a French map, they would get lost in the North. And if they used both maps side by side, they would be lost anyway. To solve this problem, we propose using both maps on top of each other.

- **Stacked elastic net** (Chapter 6)

  The aim is to predict an outcome from high-dimensional features, where it is unclear whether the effects are dense or sparse. For elastic net regression, we do not search for the best compromise between ridge and lasso by tuning, but we combine all compromises by stacking.

  In plain language: What is the best compromise between catching some big fish and many little fish? Of course, catch them all! This is the spirit of the stacked elastic net. For those who do not fish: there are different kinds of nets. Dense nets are useful for catching many little fish (#RidgeRocks). And sparse nets are useful for catching some big fish (#LassoLosesLast). There is also something called the elastic net. We can stretch it from dense to sparse, so we can adjust it to the fish. Here we introduce the stacked elastic net. It combines different nets. Then we do not need to decide whether we want to catch little fish or big fish, but simply catch all fish.