

1 Situation

For the past years, data science has been increasingly used by management companies, a lot of industry sectors are benefiting from the expansion of computer performance, we are able to build more complex models. The volume of available data is increasing, hence a need for better inference.

Often, when trying to find a model for data, we have the *small n, large p* situation. This is the most common type of data. A valuable tool to identify the dependencies across many variables may be much larger than the number of observations. This situation is called *small n, large p*. Traditional techniques are limited by computational constraints.

In this thesis, we will focus on the *small n, large p* situation. We will study high-dimensional regression in the Bayesian framework. This is a difficult problem, which often dissuades users from adopting

2 Motivation

Current technology allows us to numerically represent genomic data. We can study the influence of the genome on diseases or phenotypes. *Genetic variants*, changes at specific locations on the genome, are of interest. We will focus on the most common category of genetic variants (SNPs), i.e., variations in the nucleotides that are found at specific positions. Combinations of SNPs are inherited together, which is called linkage disequilibrium. This is a complex problem, which often dissuades users from adopting