Bachelor Thesis

Inferring the Population Quantity of Multilocus Genotype Data

Supervisors: Manfred Opper,
Olivier Françios,
Michael Blum

Fabian Bergmann, 372918

Pages: 3

Submission Date: January 21, 2019

1 Generating Data

1.1 Biological Background

1.1.1 Key words

- **Chromosome:** A DNA molecule that encodes genetic information.
- **Gene:** A DNA (or RNA) sequence that specifies the structure of a particular functional molecule.
- **Locus:** A particular position on the chromosome, like the position of a specific gene.
- **Allele:** A variant form of a given gene. Different alleles can lead to distinct phenotypic traits.

1.2 Admixture

The subsequent admixture model, follows a model proposed by Pritchard, Stephens, and Donnelly 2000.

2 Theory

2.1 PCA

2.2 RMT

Let $\mathbf{A} = \frac{1}{n}\mathbf{X}\mathbf{X}^T$ with $\mathbf{X}$ being an $m \times n$ matrix and $\lambda_1, \lambda_2, \ldots, \lambda_m$ the corresponding eigenvalues of $\mathbf{A}$, then the empirical spectral distribution (**ESD**) for $\mathbf{M}$ is given by:

$$F^M(x) = \frac{1}{m} \mid \{\lambda_i \leq x \mid i \leq m\} \mid$$

Whereby $| \cdot |$ denotes the size of a set.

By assuming a theoretical setting in which $m, n \to \infty$ while $y = \frac{m}{n} \to (0, \infty)$ the Marchenko-Pastur Law extends the ESD to the continuous case.
Under the assumption that the entries of $\mathbf{X}$ are random variables iid distributed it states that the probability density of the eigenvalues is given by:

$$p^M(x) = \frac{1}{2\pi x y \sigma^2}\sqrt{(\rho_+ - x)(x - \rho_-)}$$

where $\rho_\pm = \sigma^2(1 \pm \sqrt{y})^2$ and σ^2 is the variance of the random variables.
Insert Plot of distribution

3 Gradient Boosting

References

[1] Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. “Inference of population structure using multilocus genotype data”. In: *Genetics* 155.2 (2000), pp. 945–959.