Bachelor Thesis

Inferring the Population Quantity of Multilocus Genotype Data

Supervisors: Manfred Opper,
Olivier Françios,
Michael Blum

Fabian Bergmann, 372918

Pages: 5

Submission Date: January 31, 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 **pritchard2000inference**.

2 Theory

2.1 ABC

To infer the number of populations K expressed in a given dataset X the conditional probability $P(K|X)$ with respect to K is maximized. Since the large dimensionalities of the used datasets pose substantial computational difficulties, the datasets are summarized in an effective manner, such that the approximation $P(K|X) \approx P(K|sum(X))$ is sufficient for the intended inference. Bayes' theorem then yields

$$P(K|sum(X)) = \frac{P(sum(X)|K)P(K)}{P(sum(X))}$$

The calculation of the likelihood $P(sum(X)|K)$ however is intractable because **???** To circumvent this problem the likelihood is implicitly calculated by employing a supervised learning method to estimate the posterior, such as a neural network or boosting decision trees. These methods are trained by trying to link summary statistics of datasets to the corresponding values for the number of populations. The prior $P(K)$ can be implicitly adjusted by changing the proportions of K in the training data. Gradient boosting with decision trees is chosen in this case, for it has demonstrated good results for various classification problems citation needed. **??? Furthermore, some inuitive reasoning exists, as explained later on, for the use of decision trees in this particular case.**

2.2 Choosing the Summary Statistics

The choice of an adequate summary statistics is essentiell to obtain significant results. On the one hand a summary of the data is necessary to handle its large dimensionality, on the other hand the manner of summarisation has to be chosen carefully to sustain the vital information necessary for the inference, because each summarisation consequently forfeits some of the principal information.

$$\begin{aligned}
h(x) &= -\int_{-\infty}^{\infty} N(x|\mu,\Sigma) ln(N(x|\mu,\Sigma))dx \\
&= E[ln(N(x|\mu,\Sigma)] \\
&= E[ln(det(2\pi\Sigma)^{-\frac{1}{2}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)})] \\
&= \frac{1}{2} ln(det(2\pi\Sigma)) + \frac{1}{2} E[(x-\mu)^T\Sigma^{-1}(x-\mu)] \\
&= \frac{1}{2} ln(det(2\pi\Sigma)) + \frac{1}{2} E[trace(\Sigma^{-1}(x-\mu)^T(x-\mu))] \\
&= \frac{1}{2} ln(det(2\pi\Sigma)) + \frac{1}{2} E[trace(I)] \\
&= \frac{1}{2} ln(det(2\pi e\Sigma))
\end{aligned}$$

The only non-constant factor influencing the entropy of a multivariate gaussian is the determinant of the respective correlation matrix. Since any real symmetric matrix is diagonalisable, $det(\Sigma)$ breaks down to $det(\Sigma) = det(\mathbf{Q}^{-1}) \cdot det(\mathbf{\Lambda}) \cdot det(\mathbf{Q}) = \prod_{i=1} \lambda_i$, thus revealing that actually the eigenvalues of the correlation matrix are responsible for the magnitude of the entropy. Furthermore, the multivariate Gaussian for a given mean and variance is the maximum entropy distribution, for a proof the reader is referred to **cover2012elements**.

Assume Data is linearly correlated??? connection of normal distribution to PCA

2.2.1 PCA

introduction Principle component analysis is a statistical method that performs a basis transformation on a given data set, such that no linear correlations are anymore expressed by the data. Since the direction of a linear correlation corresponds to the direction of the highest variance in a concerning subspace, a new axis musst be aligned according that particular direction. This construction of the axes is done by requiering in an iterative manner each new axis to align with the direction that captures the most variance in the data, which however has not been captured by already established axes.
Variance can be utilised to **measure (a mathematical term, is variance a measure in mathematical sense???)**

$$\begin{aligned} u^T S u &= \frac{1}{N} u^T X X^T u - u^T \overline{XX}^T u \\ &= \frac{1}{N} u^T X (u^T X)^T - \overline{u^T X} (\overline{u^T X})^T \\ &= \frac{1}{N} (u^T X)^2 - (\overline{u^T X})^2 \end{aligned}$$

$\max_u u^T \Sigma u - \lambda(u^T u - 1)$

$$\begin{aligned} \frac{d}{du}(u^T \Sigma u - \lambda(u^T u - 1)) &= 0 \\ \Sigma u - \lambda u &= 0 \\ \Sigma u &= \lambda u \end{aligned}$$

2.3 RMT

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

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

Whereby $\mid \cdot \mid$ 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 with mean 0, 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