A recursion relaion for online coviarnce of haplotype counts

Robert Vogel

2025-01-15

In this short write-up I document a recursion relation for single-pass, or online, estimation of the unbiased covariance of haplotype counts. This is necessary because our genetic data sets are just too large to load into memory and for inefficient algoritms. Indeed, such a relationship is well established for the covariance, [1], however I was unsure how this procedure would extend to my definition of haplotype count covariance. In what follows I define the haplotype count covariance, hcov for short, and then show the recursion relation that I'll implement in C++.

I begin by introducing the quantities of interest. Denote the expected number of haplotype p copies, at locus m, and for sample i as $x_{mi}^{(p)} \in [0, 2]$. Importantly, the $x_{mi}^{(p)}$ is a real number as it represents an *expected* copy number, and not the actual copy number. The number of distinct haplotypes are the number of K founder strains in our heterogenous stock rat colony making $p \in \{1, 2, \ldots, K\}$. A locus $m \in \{1, 2, \ldots, M\}$ is any integer from 1 to the M total number of loci. Lastly, we assume that the population consists of N animals, i.e. samples.

As we are interested in the single-pass computation of the covariance, we must first define the haplotype mean and covariance statistics.

Definition 1. *The mean of expected haplotype counts for haplotype p is*

$$\bar{x}_{Mi}^{(p)} = \frac{1}{M} \sum_{m=1}^{M} x_{mi}^{(p)}.$$

Next the unbiased covariance over markers and haplotypes.

Definition 2. *The unbiased covariance of expected haplotype counts over p distinct haplotypes and M markers is defined as*

$$cov^{(M)}(x_i, x_j) = \frac{1}{M-1} \sum_{p=1}^{K} \sum_{m=1}^{M} \left(x_{mi}^{(p)} - \bar{x}_{Mi}^{(p)} \right) \left(x_{mj}^{(p)} - \bar{x}_{Mj}^{(p)} \right)$$

for which we will use $C_{ij}^{(M)}$ as shorthand.

The aim for this analysis is to write the mean Def 1 and covariance Def 2 as a recursion relation. That is, suppose there are a total of M' loci in which we want an estimate of the mean and covariance. The statistics will be computed by a single `for` loop over the total number of markers M'. Meaning that for any iteration $1 \leq M \leq M'$ we must figure out how to update the mean and covariance estimates from the previous $M-1$ observations. That is we recursively update our estimates of the mean and covariance. When we have updated our mean and covariance estimates for all M' markers, the `for` loop will terminate and we'll have our desired estimates all without loading the entire data set at one time.

In presenting the recursion, let us define the quantity $\delta_{Mi}^{(p)}$

Definition 3. *The value $\delta_{Mi}^{(p)}$ is a quantity for updating the mean and covariances from iteration $M-1$ to iteration M of sample i and is defined as follows*

$$\delta_{Mi}^{(p)} = x_{Mi}^{(p)} - \bar{x}_{(M-1)i}^{(p)}.$$

Result 1.

$$\bar{x}_{Mi}^{(p)} = \bar{x}_{(M-1)i}^{(p)} + \frac{\delta_{Mi}^{(p)}}{M}$$

Result 2.

$$C_{ij}^{(M)} = \frac{M-2}{M-1} C_{ij}^{(M-1)} + \frac{1}{M} \sum_{p=1}^{K} \delta_{Mi}^{(p)} \delta_{Mj}^{(p)}$$

References

[1] Wikipedia contributors, 2025. Last access 2025-01-15.