doi 10.18699/bgrs2024-12.3-19

Inferring posterior weight distribution in hierarchical mixture models

Meshcheryakov G.[1]*, Kovanova A.V.[2]
[1] *Institute of Protein Research, RAS, Puschino, Russia*
[2] *MediaLine Group, St. Petersburg, Russia*
** georgy@vega.protres.ru*

Key words: genotype phasing; mixture; special functions; dynamic programming; hierarchy

Motivation and Aim: Mixture models are used to model heterogenous data that is formed by sampling from multiple known distributions. The exact proportion of the data belonging to one each of the specified distributions is usually unknown and is a parameter to be estimated. In Bayesian treatment of the problem, it is assumed that mixture parameters follow a prior distribution, typically the Beta (or Dirichlet in a multivariate case) distribution. Then, posterior weight distribution is trivially inferred for each data point with a maximum-a-posteriori value usually being interpreted as a probability of a data point belonging to a particular mixture component. However, in some applications it is known that subsets of data points jointly originate from the same mixture component. Then, the posterior distribution attains a seemingly intractable form that is tough to compute. For instance, this problem arises when dealing with unphased data in high-throughput sequencing experiments. There, multiple samples/observations of the same SNP have the same unknown phase [1, 2]. The number of SNPs can be in tens of thousands and the average number of observations per SNP can be as high as a hundred, making applying stochastic algorithms such as MCMC infeasible. Here, we infer an analytical form for the posterior weight distribution of a mixture model and provide an algorithm to compute it in $O(n^2)$, making its computation possible in most practical cases.

Methods and Algorithms: Assume that we are dealing with a stratified dataset X of size N and p strata/groups $\{X^{(i)}\}^{n}_{i=1}$. For each strata it it is known that a data sample x from i-th strata originated either from F or G distributions with a probability of the former being w_i. That's it, the PDF of each sample being $h(x|w_i) = w_i f(x) + (1 - w_i)g(x)$. We assume that $w_i \sim Beta(\alpha, \beta)$ and we are interested in the posterior distribution $z(w_i|X^{(i)}) = \frac{L(X^{(i)}|w_i)Beta_{w_i}(\alpha,\beta)}{L(X^{(i)})}$, where $L(X^{(i)}|w_i)$ is the conditional likelihood and $L(X^{(i)})$ is the marginal likelihood. The latter is troublesome to compute when $n > 1$:

$$L(X^{(i)}) = \int_0^1 \prod L(X^{(i)}|w_i)Beta_{w_i}(\alpha,\beta)dw_i \propto \int_0^1 \sum_{0_n<k<1_n} (\frac{1-w}{w})^{|k|}t^k \propto S_t(n,\alpha,\beta),$$

where we used the multi-index notation for the multidimensional sum is a vector of ratios $\frac{f(x_i)}{g(x_i)}$, $S_t(n,\alpha,\beta) = \sum_{0_n<k<1_n} t^k\Gamma(n - |k| + \alpha)\Gamma(|k| + \beta)$. Straightforward

computation of S_t requires computation of 2^n terms. Fortunately, the S_t has properties that alleviate this problem.

Theorem 1. The following equation holds:

$$S_t(n, \alpha, \beta) = S_t(n-1, \alpha+1, \beta) + t_p * S_t(n-1, \alpha, \beta+1).$$

Theorem 2. The following equation holds:

$$S_t(n, \alpha+1, \beta) = (n+\alpha+\beta)S_t(n, \alpha, \beta) - S(n, \alpha, \beta+1).$$

Theorem 3. $\{S_t(n, \alpha, \beta+i)\}^{n+1}{}_{i=0}$ are linearly dependent, i. e. there exist such $\{C_i\}^{n+1}{}_{i=0}, C_i \neq 0$, that

$\sum_{i=0}^{n+1} C_i S(n, \alpha, \beta+i) = 0.$

Theorem 4. $\{C_i\}^{n+1}{}_{i=0}$ attain the form

$$C_i = (-1)^{n-i+1} \beta^{(n+1-i)} \frac{n+1}{i},$$

where $\beta^{(n+1-i)}$ is a raising factorial [3].

Algorithm of complexity $O(n^3)$ follows from the first theorem, whereas the second one simplifies it down to $O(n^2)$. Theorems 3 and 4 are further used to decrease the number of computations twofold, see Figure 1.

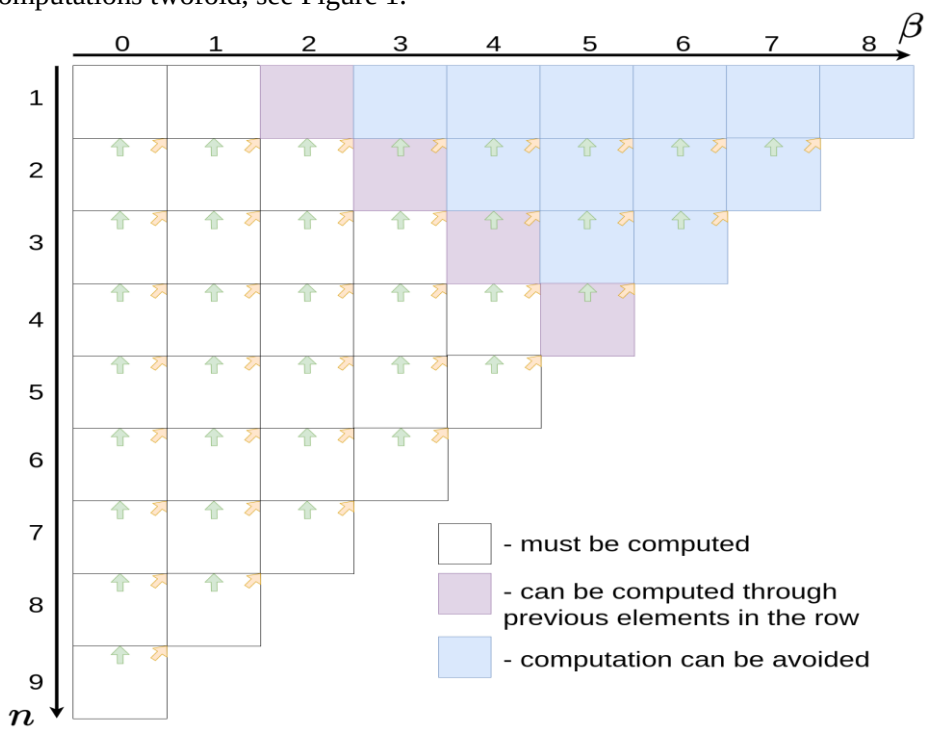

Fig. 1. Dynamic programming matrix of the algorithm that follows from theorems 1–4

References

1. Meshcheryakov G. et al. Beta negative binomial mixture model facilitates identification of allele-specific gene regulation in high-throughput sequencing data. In: Bioinformatics of Genome Regulation and Structure/Systems Biology (BGRS/SB-2022). Novosibirsk, 2022;1107. doi 10.18699/SBB-2022-664
2. Meshcheryakov G. et al. MIXALIME: MIXture models for ALlelic IMbalance Estimation in high-throughput sequencing data. *arXiv*. 2023. doi 10.48550/arXiv.2306.08287
3. Abramowitz M., Stegun I. (Eds.). Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. NBS, 1964