Binomial Models of Reference Read Counts

Attila Gulyás-Kovács

March 10, 2016

1 Introduction

1.1 Goals

1.2 The modeled data

We have $i = 1, ..., I$ individuals, $g = 1, ..., G$ genes and $v = 1, ..., V$ polymorphic (SNP) sites. With the notation $v \in (i, g)$ we will express that site v is in gene g and it is heterozygous in individual i, and we distinguish v from w if $w \in (j, g)$ and if $i \neq j$ even if both v and w map to the same site in a reference genome (meaning they are homologous).

We assume only one alternative allele at each site v, and write Y_v to denote the read count of the alternative allele at site v. We also define

$$
\begin{aligned}
Y_{ig} &= \{Y_v\}_{v \in (i,g)}, \qquad n_{ig} = \{n_v\}_{v \in (i,g)} &(1)\\
Y &= [Y_{ig}], \qquad n = [n_{ig}], &(2)
\end{aligned}
$$

where $[Y_{ig}]$ denotes a matrix whose rows are indexed by $i = 1, ..., I$ and columns by $g = 1, ..., G$. Moreover, we have a design matrix $X = [X_{ir}],\ r = 1, ..., R$ whose columns x_r are explanatory variables a.k.a. regressors. All proposed inferences in this article will be based on Y and X.

Much of the previous inferences of the MAE project were based on the statistic $S = [S_{ig}]$. The connection between S and Y can be drawn by introducing the "higher read count" $H_v = \max(Y_v, n_v - Y_v)$ and writing $S_{ig} = \left(\sum_{v \in (i,g)} H_v\right) \times \left(\sum_{v \in (i,g)} n_v\right)^{-1}$. The scalar S_{ig} aggregates the vectors Y_{ig} and n_{ig} and, as we will see, the information lost in that aggregation has an impact on all statistical analysis based on the models below.

1.3 Definition of bi and monoallelic expression

A prerequisite of the following definition is the assumption that Y_v is binomially distributed with parameters n_v (the total read counts) and q_{ig}. Thus, for all sites $v \in (i, g)$ the expected fraction $\mathrm{E}[Y_v]/n_v = q_{ig}$. We regard the expected proportion q_{ig} the single direct determinant of allelic exclusion based on which we can define bi and monoallelic expression as follows.

Informally speaking, we define the biallelic case such that the two alleles are expressed equally, so $q_{ig} = 1/2$, and the monoallelic case with q_{ig} close to either 1 or 0 depending on whether the reference or the alternative allele is excluded, respectively. To express our indifference about that last point we introduce $p_{ig} = \max(q_{ig}, 1 - q_{ig})$, which implies that $1/2 \leq p_{ig} \leq 1$.

For the formal definition we introduce variable θ_{ig} indicating the biallelic and monoallelic case ($\theta_{ig} = 0$ and $\theta_{ig} = 1$, respectively). We also fix parameters p_0, p_1 by setting $p_0 = 1/2$ and $p_1 = 0.9$, say. We then define allelic exclusion with the general equation $p_{ig} = p_{\theta_{ig}}$ or, equivalently, with

$$\begin{array}{rcccc} \text{allelic exclusion} & & \text{indicator} & & \text{expected proportion} \\ \hline \text{biallelic exp. of } (i,g) & \Leftrightarrow & \theta_{ig} = 0 & \Leftrightarrow & p_{ig} = p_0 \\ \text{monoallelic exp. of } (i,g) & \Leftrightarrow & \theta_{ig} = 1 & \Leftrightarrow & p_{ig} = p_1 \end{array} \tag{3}$$

A few things deserve mentioning in the context of Eq. 3.

1. By indexing θ and p using both i and g we allow variation in allelic exclusion not only across genes but also across individuals,

2. we define monoallelic expression by a theoretical expectation based on a simple parametric model rather than referring to some previous gold standard data set of (i, g) pairs that have been classified as either bi or monoallelically expressing,

3. the choice of $p_0 = 1/2$ leaves little room for debate but that of p_1 is quite arbitrary, and p_1 will in general influence all outcomes of statistical inference; so the results must be interpreted in light of the definition,

4. using only two classes (bi and monoallelic expression) means only two possible values of p_{ig} so we cannot account for relatively subtle differences among individuals and/or genes by fine-tuning p_{ig}; this constraints the way we can model dependence on age across all individuals for a given gene, or dependence on distance from previously identified imprinted genes across all genes for a given individual.

1.4 Latent and observable variables

Our preference of p_{ig} to q_{ig} motivates the introduction of

$$Z_v = \begin{cases} Y_v & \text{if } p_{ig} \geq 1/2 \\ n_v - Y_v & \text{otherwise.} \end{cases} \tag{4}$$

where $v \in (i, g)$. Then $Z_v \sim \mathrm{Binom}(p_{ig}, n_v)$ if and only if $Y_v \sim \mathrm{Binom}(q_{ig}, n_v)$. Using Z_v facilitates expressing models in the most direct manner (Section 2). However, Z_v is a latent (unobserved) variable because we are uncertain about p_{ig}. For this reason, statistical inference will require using likelihood functions based on Y_v (Section 3 and 4).

2 Models

M1

3 Likelihood functions

M1 Sampling distribution for read counts y_{ig}

$$f_v(y_v | n_v, p_a) = \binom{n_v}{y_v} p_a^{y_v} (1 - p_a)^{n_v - y_v} \tag{5}$$

strategy	conditional (sequential)		joint
inference task(s)	model selection, parameter estimation	classification	all
required prior info	training set	known model	basic assumptions

Table 1: Two basic strategies for carrying out inference tasks relevant to the project.

The p.m.f. for y_{ig}

$$f_{ig}(y_{ig}|n_{ig}, p_a, \kappa) = \prod_{v \in (i,g)} [\kappa f_v(y_v|n_v, p_a) + (1 - \kappa) f_v(y_v|n_v, 1 - p_a)] \tag{6}$$

M1 Marginal likelihood for π

model M.I.1; the marginal likelihood $L(\pi; y, n, p, \kappa) \equiv f(y|n, p, \kappa, \pi)$ for π equals

$$L(\pi) = \prod_{i,g} [(1 - \pi) f_{ig}(y_{ig}|n_{ig}, p_0, \kappa) + \pi f_{ig}(y_{ig}|n_{ig}, p_1, \kappa)] \tag{7}$$

model M.I.2; the marginal likelihood $L(\pi; y, n, p, \kappa, \nu) \equiv f(y|n, p, \kappa, \nu, \pi)$ for π is given by

$$L(\pi) = B^{-1} \prod_g \int_0^1 \mu^{\pi\nu} (1 - \mu)^{(1-\pi)\nu} \prod_i u_{ig}(\mu) \, \mathrm{d}\mu \tag{8}$$

$$u_{ig}(\mu) = (1 - \mu) f_{ig}(y_{ig}|n_{ig}, p_0, \kappa) + \mu f_{ig}(y_{ig}|n_{ig}, p_1, \kappa) \tag{9}$$

where B is the beta function evaluated at $(\pi\nu, (1 - \pi)\nu)$.

4 Inference

Given the models in Section 2 and their parameters, the goals of the study can be framed in the following statistical inference tasks:

1. assess dependence on explanatory variables via two tightly linked tasks:
   - *select the model*[1] that best fits both the data and some prior information such as definitions or theoretical considerations
   - *estimate* regression parameters β_h (Eq.???)
2. assess the fraction of monoallelically expressed genes by finding an *estimate* $\hat{\pi}_1$ for π_1
3. call novel monoallelically expressed genes: depending on the selected model *classify* each (i, g) or g by hypothesis testing (Eq. 3)

[1] When several models are nearly equally good, it is preferred to avoid selecting only one of them and discard the rest. In that case Bayesian model averaging provides a normative solution.

Depending on what prior information we wish to take advantage of, we may choose between two major strategies, summarized by Table 1. The conditional strategy requires prior information beyond the basic assumptions, where the latter correspond to the constraints of the most general model we consider (??? in Section 2).

One such piece of prior information is a *training set* of (i, g) pairs (or of genes g) that are labeled either as mono or biallelically expressing. Given the training set the best model can be selected and most parameters (like β) can be estimated. Parameter π_1, however, is special in the sense that it can only be estimated from the genome-wide test data (or its addressable subset).

The conditional strategy is also sequential in that in the first step model selection and the estimation of β must be achieved, then based on that the estimation of π_1 together with classification.

In principle it is possible to evade the discomforting uncertainty that may surround prior information by ignoring those completely. This, however, requires a joint inference strategy that is both challenging to implement and validate and may lead to high errors in all three tasks depending on how valuable the discarded prior information are.

M1 Classification

5 Appendix

If we want to base inference on the scalar S_{ig} instead of the vector Y_{ig}, we need to derive likelihood functions for S_{ig} using Eq.???. Let $\mathcal{S} = \{(i, g) : n_{ig} s_{ig} = y_{ig}\}$, that is the set of all (i, g) pairs leading to the observed s_{ig}. Then the likelihood functions h_{ig} and h'_{ig} for S_{ig} can be expressed in terms of $\{f_{ig}\}_{(i,g)\in\mathcal{S}}$:

$$\begin{aligned} h_{ig}(s_{ig}|n_{ig}, p_h) &= \sum_{(i,g)\in\mathcal{S}} f_{ig}(y_{ig}|n_{ig}, p_h) && (10) \\ h'_{ig}(s_{ig}|p_h) &= \sum_{(i,g)\in\mathcal{S}} f_{ig}(y_{ig}|n_{ig}, p_h)\, q_{ig}(n_{ig}|p_h). && (11) \end{aligned}$$

The difference between h_{ig} and h'_{ig} is whether or not we condition the distribution of S_{ig} on the observed n_{ig}. If we don't take advantage of the observations on n_{ig} (Eq. 11), we must then treat it as a random variable and specify a distribution for it, say q_{ig}. In either case we need *some* kind of information or assumption on n_{ig}. This holds regardless we want to use h_{ig} (or h'_{ig}) in simulations, in parameter estimation or in classification with error control.