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

$$Y_{ig} = \{Y_v\}_{v\in(i,g)}, \qquad n_{ig} = \{n_v\}_{v\in(i,g)} \tag{1}$$

$$Y = [Y_{ig}], \qquad n = [n_{ig}], \tag{2}$$

where $[Y_{ig}]$ denotes a matrix whose rows are indexed by $i = 1, ..., I$ and columns by $g = 1, ..., G$. Moreover, we have an $I \times R$ design matrix $X = [x_{ir}]$, $r = 0, ..., R-1$ whose columns are explanatory variables a.k.a. regressors except for the 0th column, whose entries $x_{i0} = 1$ for all i. 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

We will refer to bi and monoallelic expression collectively as (allelic) *exclusion state*. Here we define exclusion state for any given (i, g) pair. Suppose q_{ig} is the expected fraction of transcripts[1] from the maternal chromosome and $1 - q_{ig}$ for the paternal chromosome, and let $p_{ig} = \max(q_{ig}, 1 - q_{ig})$ implying that $p_{ig} \geq 1/2$.

We regard the expected proportion p_{ig} as the single direct determinant of allelic exclusion: if p_{ig} is near $1/2$ we call (i, g) biallelically expressed, whereas if p_{ig} is near 1 we classify

[1]The word "expected" implies a probability distribution for maternal transcripts. This can be either binomial if the total number of transcripts is fixed, or else Poisson. In the latter case q_{ig} is to be interpreted as the relative transcription rate on the maternal chromosome.

(i, g) monoallelic. Formally, let $\mathcal{P}_0 = [1/2, p')$ and $\mathcal{P}_1 = [p'', p''']$ disjoint subintervals of $[1/2, 1]$ so that $1/2 \leq p' \leq p'' \leq p''' \leq 1$. Then we define exclusion state of (i, j) as

$$p_{ig} \in \begin{cases} \mathcal{P}_0 & \Leftrightarrow \text{biallelic} \\ \mathcal{P}_1 & \Leftrightarrow \text{monoallelic.} \end{cases} \tag{3}$$

There are some complications with this definition. First, our data consists of reads instead of full-length transcripts. We assume that the read count Y_v for the alternative allele at polymorphic site v is binomially distributed with parameters n_v (the total read counts) and p_v Figure TODO. Second, read counts have confounded by various measurement errors but we assume that they are proportional to allele specific transcription rates. This allows us to write $p_v = p_{ig}$ given the event that the alternative allele is on the chromosome that has the higher expected fraction of transcripts; we denote that event with $\phi_v = 1$. Otherwise $\pi_v = 0$, which implies that $1 - p_v = p_{ig}$. We may call ϕ_v *allele configuration* at site v. Third, some reads may map to multiple polymorphic sites $v_1, v_2, ...$ coupling $\phi_{v_1}, \phi_{v_2}, ...$ We suppose this happens rarely enough to be completely ignored so that all allele configurations ϕ_v for any given (i, g) can be assumed independent.

With the preceding considerations the definition of exclusion state can be based on p_v and ϕ_v as

	biallelic	monoallelic
$\phi_v = 0$	$1 - p_v \in \mathcal{P}_0$	$1 - p_v \in \mathcal{P}_1$
$\phi_v = 1$	$p_v \in \mathcal{P}_0$	$p_v \in \mathcal{P}_1$

Table 1: Definition of exclusion state of (i, g) based on p_v and ϕ_v for site $v \in (i, g)$

2 Models

M1 No influence of explanatory variables

Table 2

	$\theta_{ig} = 0$	$\theta_{ig} = 1$
$\phi_v = 0$	$1 - p_0$	$1 - p_1$
$\phi_v = 1$	p_0	p_1

Table 2: Value of p_v at all four combinations of allelic configuration and exclusion state in model M1.

M1.1 Special cases

As $\nu \to 0$, the beta distribution becomes Bernoully and μ_g will be 1 with probability π and 0 with probability $1 - \pi$. For any gene g this couples the exclusion state for all individuals so that $\theta_{1g} = ... = \theta_{Ig}$. So we can replace the general structure of model M1 with a

probabilistically equivalent but simpler structure by introducing $\theta_g \equiv \theta_{1g}$ and removing μ_g (Figure TODO).

In the limit $\nu \to \infty$ we have $\mu_1 = ... = \mu_G = \pi$. Therefore we can once again simplify the model structure by removing μ_g. But the effect on θ_{ig} is the opposite in that $\theta_{1g}, ..., \theta_{Ig}$ become completely uncoupled in the sense that $\{\theta_{ig}\}_{ig}$ becomes independent and identically distributed (Figure TODO).

The interpretation of the first limiting case is that individuals show no variation in exclusion status for any gene g. Thus it makes sense to speak about bi or monoallelically expressing genes population-wide without the need of looking at individuals. The second limiting case, on the other hand, means that all genes have the same population-wide tendency for bi or monoallelic expression.

M2 Regression of Y_v on explanatory variables

This model is an extension of model M1 when $\nu \to 0$. For a given gene g each individual has the same exclusion state θ_g but the across individual variation in explanatory variables x_i induces variation in the expected fraction in the sense that $p_v \neq p_w$ if $v \in (i, g)$ and $w \in (j, g),\ (i \neq j)$.

Given that Y_v is binomial, logistic regression appears as a natural framework, although some shortcomings will be discussed below TODO. In this framework the logit function links the expected fraction p_v of Y_v to the ith row of design matrix X so that

$$p_v = \text{logit}^{-1}(x_i\, b_v). \tag{4}$$

The regression coefficient $b_v = (b_{v0}, ..., b_{vR-1})^\top$ is an R-length vector and is given by the deterministic "switch" function s such that $d = s(\beta, \theta_g, \phi_v)$, where $\beta = [\beta^0, \beta^1]$ is a $R \times 2$ parameter matrix and $\beta^a = (\beta^a_0, ..., \beta^a_{R-1})^\top$. We give s by specifying the value of b_{vr} in terms of $\theta_g, \phi_v, \beta^0_r$ or β^1_r in Table 3.

	$\theta_g = 0$	$\theta_g = 1$
$\phi_v = 0$	$1 - \beta^0_0$	$1 - \beta^1_0$
$\phi_v = 1$	β^0_0	β^1_0

	$\theta_g = 0$	$\theta_g = 1$
$\phi_v = 0$	β^0_r	β^1_r
$\phi_v = 0$	$-\beta^0_r$	$-\beta^1_r$

Table 3: Value of the intercept $b_{vr} = s(\beta, \theta_g, \phi_v)$ at all four θ_g, ϕ_v combinations. *Left:* $r = 0$; *right:* $r = 1, ..., R-1$

We can set $\beta^0_r = 0,\ r \geq 1$ expressing no dependence of y_{ig} on explanatory variables for all biallelically expressed genes. When the same is true for monoallelically expressed genes, we have $\beta^1_r = 0,\ r \geq 1$, and we end up with the special case of model M1 given by $\nu \to 0$. This result motivates fixing $\beta^a_0 = \text{logit}(p_a)$. When we want to study how, for monoallelically expressed genes, y_{ig} depends on the explanatory variables, we need to infer $\beta^1_r,\ r = 1, ..., R-1$.

M2.1 Shortcomings

M3 Regression of θ_{ig} on explanatory variables

3 Likelihood functions

3.1 Model selection and estimation of π

The general frequentist procedure goes as follows:

1. express the marginal likelihood L_m for π based on y and X under all models $m = 1, ...,$ by taking expectations over nuisance parameters such as $\mu_{ig}, \theta_{ig}, \phi_v$
2. maximize L_m with respect to π obtaining the ML estimate $\hat{\pi}_m = \arg\max_\pi L_m(\pi)$
3. for each m evaluate model fit using a criterion based on the maximized likelihood (such as AIC, BIC) and select the highest scoring model m^* and the corresponding $\hat{\pi}_{m^*}$

TODO: note on Bayesian procedure

3.2 Sampling distribution for read counts y_{ig}

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

Under model M1 and M3 p.m.f. for y_{ig}

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

where a is 0 or 1 reflecting $\theta_{ig} = 0$ or $= 1$. It will be convenient to use the shorthand $L_{ig}^a \equiv f_{ig}(y_{ig}|n_{ig}, p_a, \kappa)$.

3.3 Marginal likelihood for π

Under model M1 the marginal likelihood $L_1(\pi, \nu) \equiv f(y|n, p, \kappa, \nu, \pi)$ for π and ν is given by

$$L_1(\pi, \nu) = B^{-1} \prod_g \int_0^1 \mu^{\pi\nu}(1-\mu)^{(1-\pi)\nu} \prod_i \left[(1-\mu)L_{ig}^0 + \mu L_{ig}^1\right] \mathrm{d}\mu \tag{7}$$

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

L_1 in Eq. 7 depends on the parameter ν, which may be of some interest because it quantifies the tendency of genes to be monoallelically expressed across all individuals (so that individuals tend not to vary for any given gene). We may decide not to care about ν or take it to the limit $\nu \to 0$ or $\nu \to \infty$ by recalling the two special cases of model M1. To obtain the likelihood for those cases let us denote $L_1(\pi) \equiv f(y|n, p, \kappa, \pi)$ and recall Eq 6. Then Eq. 7 simplifies to

$$L_1(\pi) = \begin{cases} \prod_g \left[(1-\pi)\prod_i L_{ig}^0 + \pi \prod_i L_{ig}^1\right] & \text{if } \nu \to 0 \\ \prod_{i,g} \left[(1-\pi)L_{ig}^0 + \pi L_{ig}^1\right] & \text{if } \nu \to \infty. \end{cases} \tag{8}$$

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

Table 4: Two basic strategies for carrying out inference tasks relevant to the project.

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*[2] 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)

Depending on what prior information we wish to take advantage of, we may choose between two major strategies, summarized by Table 4. 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.

4.1 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

[2]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.

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 $\{f_{ig}\}_{(i,g)\in\mathcal{S}}$:

$$h_{ig}(s_{ig}|n_{ig}, p_h) = \sum_{(i,g)\in\mathcal{S}} f_{ig}(y_{ig}|n_{ig}, p_h) \tag{9}$$

$$h'_{ig}(s_{ig}|p_h) = \sum_{(i,g)\in\mathcal{S}} f_{ig}(y_{ig}|n_{ig}, p_h)\, q_{ig}(n_{ig}|p_h). \tag{10}$$

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.