Binomial Models of Reference Read Counts

Attila Gulyás-Kovács

March 17, 2016

1 Introduction

1.1 Goals

estimation of $1 - \pi_0$, where π_0 is the genome-wide probability (i.e. the expected fraction) of biallelically expressed genes

within-gene variation how does exclusion state vary population-wide within any given gene?

regression if there is such variation, how is it explained by age and other measured variables?

classification predict exclusion state for each individual–gene pair to learn about species and tissue specificity

To what extent has the previous work achieved these goals? Estimation of π_0 has not yet been achieved. Within-gene variation has been characterized using the conditional distribution of the S_{ig} statistic for any given gene g but it remains unknown what is the relative contribution of within-gene and of across-gene (genome-wide) variance to the total variance (both across individuals and genes). Regression on explanatory variables has been performed but left the generality and statistical significance of the results an open question. Classification has been performed using S_{ig} but without estimated error rates and—inconsistently with the results of regression analysis—also without taking explanatory variables into account.

1.2 Improvement relative to previous approach

As explained in this section, answering the remaining questions is limited by the properties of the previous models and—to even greater extent perhaps—by their incomplete or implicit description. This has motivated the present article, which explicitly describes novel modeling approaches—contrasting them to the previously used ones—, as well as their inferential utility towards the remaining goals (Table 1).

Both the previous and present approach starts out from modeling read counts at heterozygous sites as binomial random variables. However, only the present approach considers their joint distribution at the level of entities that are directly relevant to biology: transcripts for a given individual and gene, population-wide within a given gene, and both population- and genome-wide. This is achieved via local (transcript level, Section 2.2) and global (individual and higher levels, Section 3) joint models of the complete data.

		previous	proposed
local model(s)	read counts at heteroz. sites	binomial	binomial
	sites jointly modeled	no	yes
	direct biol. relevance	no	yes
global models	well-defined	no	yes
	objective selection possible	no	yes
	selection done	inconsistently	not yet
π_0	estimation possible	no	yes
regression	nonlinearity	no	yes
	heteroscedasticity	no	yes
classification	test statistic	S_{ig}	posterior pr.
	likelihood (distrib.) known	no	yes
	sufficiency (given likelihood)	no	yes
	error control	no	yes

Table 1: Salient properties of previous model(s) and the ones proposed in this article, and properties of inferences based on those models.

These new models draw direct, and explicit, link between read counts and allelic exclusion state θ_{ig} by enabling likelihood calculations. The previous approach was both indirect and implicit because it used the S_{ig} statistic derived from read counts to describe exclusion state in a non-probabilistic way since the sampling distribution of S_{ig} for a given exclusion state was not specified. That in turn prevented likelihood calculations.

Were the likelihoods based on S_{ig} expressed, they would lack some information on exclusion state. This is because S_{ig} only considers the proportion of read counts (for one allele) discarding the counts themselves, which enhance confidence. Further information is lost by the simplifying assumption on haplotype phase that all "higher read counts" originate from the same chromosome. These shortcomings imply that S_{ig} is not a sufficient statistic[1] for exclusion state. Although the shortcomings had been recognized earlier, only partial and post-hoc corrections were employed. In contrast, proposed local models operate with counts *per se* and also relax the simplifying assumption on haplotype by considering all possible *allele configurations*. Thus they contain all information on exclusion state and its likelihood[2].

The lack of S_{ig}-based likelihood for exclusion state prevented the estimation of the error rates of classification and that of genome-wide probabilities π for those states because the two are inherently coupled, as explained in an earlier article[3]. On the other hand, all proposed global models contain a π parameter vector, which can be estimated by maximum likelihood based on the complete dataset (Section 4.3.1). That estimate then can be combined with likelihood ratios representing the odds that the read count data support mono vs. biallelic expression (Section 4.4). This yields the posterior probability of monoallelic expression, which naturally incorporates error. Alternatively, likelihood ratios can be used on their own as Bayes factors. Note that the Neyman-Pearson lemma[4] guarantees that there do not exist more powerful tests than that based on likelihood ratios.

[1] https://en.wikipedia.org/wiki/Sufficient_statistic

[2] Note that, trivially, the likelihood is always a sufficient statistic

[3] Feb 10, 2016: Project on Monoallelic Expression: a Statistical View

[4] http://mathworld.wolfram.com/Neyman-PearsonLemma.html

Previous regression analysis used the vector LOI_R$_g$ as a response variable, which was derived from S_{ig} with a data transformation step. Some limitations of LOI_R$_g$ obviously follow from those of S_{ig} (discussed above). More limitations have been found[5] to have arisen from the previous incorrect use of regression weights. Moreover, the data transformation may introduce biased estimation of regression parameters due to insufficient removal of the observed strong nonlinearity and heteroscedasticity of read count/S_{ig}-based regression. Finally, the interpretation of LOI_R$_g$-based regression results in terms of exclusion state remained unclear. All these shortcomings are now removed by the proposed logistic regression approach using directly read counts or, alternatively, exclusion state as response variables.

The above complications might have contributed to the inconsistency in previous analysis that conflicting models were used in different inferences: LOI_R$_g$-based regression model finding dependence on some explanatory variables (like age) and a S_{ig}-based non-regression model for classification that ignores any such dependence. The proposed approach is consistent because the observed variable is read counts in all alternative models. Moreover, the likelihood under all proposed global models can be calculated, which permits the objective selection of the best fitting model using e.g. the Akaike or the Bayesian information criterion (AIC, BIC, see Section 4.3).

2 Data and local models

2.1 The modeled data: read counts

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 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 $S_{ig} = \left(\sum_{v\in(i,g)} H_v\right) \times \left(\sum_{v\in(i,g)} n_v\right)^{-1}$. As the vectors Y_{ig} and n_{ig} are aggregated into the scalar S_{ig} some information is inevitably lost, which in turn leads to the insufficiency of S_{ig} mentioned in Section 1.2.

2.2 Local models of allelic exclusion

The probability models presented here is *local* in the sense that they describe only hierarchically lower levels of parameters, i.e. those on which the observed read count data directly

[5] lab-notebook post from Mar 2, 2016: Repeating Ifat's Regression Analysis with 5 More Genes

symbol	name/description	type	specific to
Y_v	read count at site v, altern. variant	observed	
n_v	read count at site v, total	fixed	
P	multinomial proportions, Eq. 6	deterministic	M1, M3
B	regression parameters, Eq. 9	deterministic	M2
ψ_v	indicates altern. var. on maternal allele	unobserved	
ϕ_{ig}	indicates paternal allele exclusion (if any)	unobserved	
κ	probability of paternal allele exclusion	fixed	
θ_{ig}	exclusion state indicator	unobserved	
θ_g	exclusion state i. (zero var. within genes)	unobserved	M1.1, M2
μ_g	exclusion state probabilities for gene g	unobserved	
π	exclusion state probabilities genome-wide		
ν	"pseudocount"	unobserved	M1, M3
x_i	explanatory variables	fixed	M2, M3

Table 2: Parameters and other components of the proposed models

or relatively directly depends. Despite the qualifier "local", these models are sufficiently high level to describe for a given (i, g) pair the biologically relevant allelic exclusion state introduced below. The global models in Section 3 will be based on these local models or very similar ones. Parameters are summarized by Table 2.

2.2.1 Binary exclusion state

Suppose there are only two (allelic) *exclusion states*: biallelic and monoallelic expression. We introduce the exclusion state indicator θ_{ig} for any given (i, g) pair such that biallelic expression of gene g in individual i is indicated by $\theta_{ig} = 0$ and monoallelic by $\theta_{ig} = 1$. Thus θ_{ig} is a Bernoulli random variable.

Suppose p_{ig} is the expected fraction of transcripts[6] from the maternal chromosome and $1-p_{ig}$ for the paternal chromosome, and let $q_{ig} = \max(p_{ig}, 1-p_{ig})$ implying that $q_{ig} \geq 1/2$.

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

$$q_{ig} \equiv \max(p_{ig}, 1 - p_{ig}) \in \begin{cases} \mathcal{P}_0 & \Leftrightarrow \theta_{ig} = 0, \text{ biallelic} \\ \mathcal{P}_1 & \Leftrightarrow \theta_{ig} = 1, \text{ monoallelic.} \end{cases} \tag{3}$$

There are some complications with this definition. First, p_{ig} is unobserved and so must be inferred from the observed read counts. This raises uncertainty about not only exact value of p_{ig} but also whether $p_{ig} \geq 1/2$ and therefore $q_{ig} = p_{ig}$, or else $< 1/2$ and therefore $q_{ig} = 1 - p_{ig}$. Let $\phi_{ig} = 1$ indicate the former event and $\phi_{ig} = 0$ the latter with prior probability κ and $1 - \kappa$, respectively. Thus κ quantifies the tendency of the paternal allele to be excluded. In the present models κ is not specific to individuals and genes but it is

[6] 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 p_{ig} is to be interpreted as the relative transcription rate on the maternal chromosome.

straight-forward to extend the models in that direction at the expense of introducing many more parameters. It may be reasonable to set $\kappa = 1/2$.

Several further complications arise because our data consists of read counts instead of the count 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. Read counts are known to be confounded by various measurement errors but we assume here that they are proportional to allele specific transcription rates. This allows us to write $p_v = p_{ig}$ given the random event that the alternative allele is on the maternal chromosome; we denote that event with $\psi_v = 1$. Otherwise $\psi_v = 0$, which implies that $1 - p_v = p_{ig}$. We will assume $1/2$ prior probability for $\psi_v = 1$ for all v. Moreover, some reads may map to multiple polymorphic sites $v_1, v_2, ...$ coupling $\psi_{v_1}, \phi_{v_2}, ...$. We suppose this happens rarely enough to be completely ignored so that all allele configurations ψ_v for any given (i, g) can be assumed independent.

The pair (ϕ_{ig}, ψ_v) will denote an *allele configuration* at site v. With the preceding considerations the definition of exclusion state θ_{ig} can be based on p_v and the allele configuration as follows:

		$\phi_{ig} \neq \psi_v$	$\phi_{ig} = \psi_v$
biallelic	$\theta_{ig} = 0$	$1 - p_v \in \mathcal{P}_0$	$p_v \in \mathcal{P}_0$
monoallelic	$\theta_{ig} = 1$	$1 - p_v \in \mathcal{P}_1$	$p_v \in \mathcal{P}_1$

Table 3: Definition of the binary exclusion state and its indicator θ_{ig} based on p_v and the allele configuration (ϕ_{ig}, ψ_v)

2.2.2 Multiple exclusion states

The binary local model may be generalized to a K-ary one, in which there are $K \geq 2$ exclusion states. Then $\{\mathcal{P}_k : k = 0, ..., K-1\}$ is a sequence of disjoint subintervals of $[1/2, 1]$ and the definition of exclusion states are as follows:

		$\phi_{ig} \neq \psi_v$	$\phi_{ig} = \psi_v$
biallelic	$\theta_{ig} = 0$	$1 - p_v \in \mathcal{P}_0$	$p_v \in \mathcal{P}_0$
mildly monoallelic	$\theta_{ig} = 1$	$1 - p_v \in \mathcal{P}_1$	$p_v \in \mathcal{P}_1$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
strongly monoallelic	$\theta_{ig} = K - 1$	$1 - p_v \in \mathcal{P}_K - 1$	$p_v \in \mathcal{P}_K - 1$

Table 4: Definition of the general K-ary exclusion state and its indicator θ_{ig}

We will symbolically represent Table 4 by writing

$$\begin{aligned} p_v &= P[\theta_{ig}, \delta_{\phi_{ig}\psi_v}] && (4) \\ P &= \begin{pmatrix} 1 - \mathcal{P}_0 & \mathcal{P}_0 \\ \vdots & \vdots \\ 1 - \mathcal{P}_{K-1} & \mathcal{P}_{K-1} \end{pmatrix}, && (5) \end{aligned}$$

where δ_{ab} is the Kronecker delta function, which is 1 if $\phi_{ig} = \psi_v$ and 0 otherwise.

To see the utility of P, consider the following example with binary exclusion state ($K = 2$). Based on the data we have some uncertain knowledge on p_v, which we want to use to infer θ_{ig}. Suppose we know that the allele configuration $(\phi_{ig}, \psi_v) = (0, 1)$. Then $\delta_{\phi_{ig}\psi_v} = 0$ and so we need to consider only the first column of P. If the data supports $p_v = P[0, 0] = 1 - \mathcal{P}_0$ better than $p_v = P[1, 0] = 1 - \mathcal{P}_1$, we can conclude that $\theta_{ig} = 0$ (biallelic expression) is more likely than $\theta_{ig} = 1$ (monoallelic expression).

In practice the allele configuration is unobserved so we are uncertain about it. However, using our probability model we can take the expectation (i.e. average) over all four configurations. If the number s_{ig} of polymorphic sites is > 1 then we can base the inference of θ_{ig} on all $p_v : v \in (i, g)$ jointly, taking expectation over all $4^{s_{ig}}$ configurations.

3 Global models

Several global models are formulated in this article, which can be classified by two aspects (Table 5):

1. the population-wide variation of exclusion state θ_{ig} within each gene g and

2. how that variation is explained by the measured variables in X

M1 No influence of explanatory variables

A key aspect of M1 is that it accounts for population-wide variation within a given gene g through a hierarchy of parameters. The indicator θ_{ig} of exclusion state is a K-ary multinomial random variable (or Bernoulli variable when $K = 2$) with a K-length parameter vector $\mu_g = (\mu_{g0}, ..., \mu_{gK-1}$ containing the probabilities for each exclusion state. This setup permits population-wide variation within any given gene g. μ_g is itself a random variable with Dirichlet distribution with parameters π, ν, which models the genom-wide variation of allelic exclusion. $\pi = (\pi_0, ..., \pi_{K-1}$ are the prior probabilities for the K exclusion states and the scalar ν controls the density of μ_{gk} at π_k, and so $\pi_k \nu$ may be considered a pseudocount with the interpretation as the number of genes that were found in state k in prior studies.

TODO: plate diagram

Turning to the local properties of M1, each of the K subinterval (Section 2.2.2) consists of a single point such that $\mathcal{P}_0 = \{1/2\}$ and $\mathcal{P}_k = \{p_k\}$ $(k > 0)$, where p_k is some fixed number. Taking binary state ($K = 2$) for instance, p_1 may be fixed at 0.9. Then Eq. 4

		variance of exclusion state θ_{ig} within each gene g		
		any	zero	maximum
	none	M1	M1.1	M1.2
response var.:	read counts Y_{ig}		M2	
	exclusion state θ_{ig}	M3		

Table 5: Overview of the global models in this article.

remains the same but Eq. 5 changes to

$$P = \begin{pmatrix} 1/2 & 1/2 \\ 1 - p_1 & p_1 \\ \vdots & \vdots \\ 1 - p_{K-1} & p_{K-1} \end{pmatrix}. \tag{6}$$

Let's take for example the binary local model represented by P in Eq. 6 with $K = 2$. If the allele configuration at site v is $(0, 0)$ and the data supports $p_v = p_1$ stronger than $p_v = 1/2$ then we conclude, based only on v, that the exclusion state $\theta_{ig} = 1$ (i.e. monoallelic expression) is more likely.

Now we will consider two special cases of M1 named M1.1 and M1.2. We give their interpretation before their mathematical definition. The interpretation of M1.1 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. Model M1.2, on the other hand, means that all genes have the same population-wide tendency for bi or monoallelic expression and the only source of variation is the one within the population.

TODO: DAG with theta leaves

M1.1 Zero variance within each gene

As $\nu \to 0$, the Dirichlet distribution becomes multinomial and μ_{gk} will be 1 with probability π_k. For any gene g this couples the exclusion state for all individuals so that $\theta_{1g} = ... = \theta_{Ig}$. This means that we can replace the general structure of model M1 with a probabilistically equivalent[7] but simpler structure by introducing $\theta_g \equiv \theta_{1g}$ and removing μ_g (Figure TODO).

M1.2 Maximum variance within each gene

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).

M2 Regression of Y_v on explanatory variables

TODO: plate diagram

The global structure of this model is the same as M1.1. So, for a given gene g all individuals have the same exclusion state θ_g but the population-wide variation in explanatory variables x_i induces variation in p_v. To this end the local models introduced in Section 2.2 must be extended with the regression of Y_v on X. For simplicity we describe this model assuming binary exclusion state and briefly sketch the general K-ary case at the end of this section.

Given that Y_v is binomial, logistic regression appears as a natural framework, where the logit function links the expected fraction p_v of Y_v to the ith row of design matrix X unless

[7] https://en.wikipedia.org/wiki/Convergence_of_random_variables

the resulting $p_v < 1/2$. Therefore Eq. 4 modifies to

$$\begin{aligned} p_v &= \max\left(\text{logit}^{-1}(x_i\, b_v),\, \frac{1}{2}\right) && (7) \\ b_v &= B[\theta_{ig}, \delta_{\phi_{ig}\psi_v}], && (8) \end{aligned}$$

where b_v is the R-length vector $(b_{v0}, ..., b_{vR-1})^\top$ and plays the role of regression coefficient in Eq. 7. As Eq. 8 says, b_v is an entry of matrix B of regression parameters, which is indexed by the exclusion state θ_{ig} and the allelic configuration (ϕ_{ig}, ψ_v).

Analogously to P under M1 (Eq.6), B under the present model M2 facilitates the inference of θ_{ig} based on y_v and (ϕ_{ig}, ψ_v). But because b_v is a vector, B has a more complex structure than P, consisting of four R-length vectors:

$$\begin{aligned} B &= \begin{pmatrix} (0, ..., 0)^\top & (0, ..., 0)^\top \\ -\beta & \beta \end{pmatrix} && (9) \\ \beta &= (\beta_0, \beta_1, ..., \beta_{R-1})^\top && (10) \end{aligned}$$

β is a vector of regression parameters consisting of the intercept β_0 and a "slope" parameter β_r for each explanatory variable x_r, $0 < r < R$. The bottom left entry represents a reflection of the regression curve defined by the bottom right entry accross the horizontal straight line defined by $p_v = 1/2$, which is analogous to the "reflection" in P of the point p_1 across the same horizontal line resulting in $1 - p_1$. That the $1, ..., R - 1$ elements of top right entry are 0 expresses the assumption that when $\theta_{ig} = 0$ (biallelic expression) then the explanatory variables have no impact on p_v (Eq. 7); that the 0th element is also 0 follows from the equality $\text{logit}^{-1}(0) = 1/2$ showing that exclusion state $\theta_{ig} = 0$ under both M1 and the present M2 is defined by $p_v = 1/2$.

The connection between M1 and M2 can be made even more explicit by considering the special case of M2 that $\beta_1, ..., \beta_{R-1} = 0$ so that explanatory variables have no impact on p_v also when $\theta_{ig} = 1$ (monoallelic expression). Furthermore, if $\beta_0 = \text{logit}(p_1)$ also holds, then M2 is probabilistically equivalent to M1.1. So, for consistency between models, we should fix $\beta_0 = \text{logit}(p_1)$, which has the additional advantage of having one less unknown parameters.

It is conceptually straight-forward to extend above model from binary to general K-ary exclusion state. In that case the B matrix (Eq. 9) has K rows; the 0th row is identical to the binary case, whereas rows $k = 1, ..., K - 1$ have distinct β_k vectors of the form of $(\beta_{k0}, \beta_{k1}, ..., \beta_{kR-1})^\top$.

TODO: logit function

M3 Regression of θ_{ig} on explanatory variables

Model M3 is very similar to M1. The key difference is the replacement of μ_g by $x_i \beta_g$ so that the allelic state θ_{ig} is a response to the explanatory variables in x_i with regression coefficient β_g. β_g in turn is a random variable parametrized by π and some other parameters.

With this model structure the variation of exclusion state has three components: the genome-wide variation of their effect mediated by β_g, a systematic within-gene variation due to the measured explanatory variables x_i, and the remaining within-gene variation unexplained by x_i.

TODO: plate diagram

4 Inference

Likelihood functions play a central in both frequentist (classical) and Bayesian inference. In this section we present various likelihood functions for the local and global models introduced in Section 3.

4.1 Likelihood of local models

Since the read count Y_v of the alternative variant at any given heterozygous site v, the lowest-level likelihood functions are

$$\binom{n_v}{y_v} p_v^{y_v} (1-p_v)^{n_v - y_v} = \begin{cases} f_v(y_v | n_v, P, \phi_{ig}, \psi_v, \theta_{ig}), & p_v = \text{Eq. } 4 \quad (M1, M3) \\ f_v(y_v | n_v, x_i, B, \phi_{ig}, \psi_v, \theta_g), & p_v = \text{Eq. } 7, 8 \quad (M2). \end{cases} \tag{11}$$

As mentioned in Section 2.2 the allelic configurations $\{(\phi_{ig}, \psi_v) : v \in (i,g)\}$ are neither observed nor informative and so must be considered nuisance parameters to be removed by marginalization. So we take the expectation over all possible configurations. This yields the following likelihood function under model M1 and M3:

$$L_{ig}^k \equiv f_{ig}(y_{ig} | n_{ig}, P, \kappa, \theta_{ig} = k) \tag{12}$$

$$= \frac{1}{2} \sum_{\phi_{ig}=0}^{1} \kappa^{\phi_v} (1-\kappa)^{1-\phi_v} \prod_{v \in (i,g)} \sum_{\psi_v = 0}^{1} f_v(y_v | n_v, P, \phi_{ig}, \psi_v, \theta_{ig} = k), \tag{13}$$

where $k = 0, ..., K-1$ and L_{ig}^k is a convenient shorthand. Under model M2 f_{ig} has the same form except that P is replaced by x_i, B and θ_{ig} by θ_g as in Eq. 11. The same shorthand L_{ig}^k shall be used model M2 as well; its specific semantics shall be clear from the context.

4.2 Regression parameters under M2 from training data

For this estimation we need a training set of genes known to be expressed monoallelically ($\theta_g = 1$), collected from I' individuals. Unless we have training data distinguishing between different strengths of allelic exclusion, we must use the binary version of model M2. Then the likelihood for the B matrix of regression parameters based on the training data is

$$L_{M2}(B) = \prod_g \prod_{i=1}^{I'} L_{ig}^1 \tag{14}$$

where L_{ig}^1 is given by the modification of Eq. 12-13 as described in Section 4.1. The two running products represent the data aggregation over individuals i and monoallelically expressed genes g describing the complete training data set.

4.3 Model selection and estimation of π

Consider a set $\mathcal{M}$ of alternative models like the proposed ones in this article. Model selection means to find the best model(s) according to a criterion reflecting its fit to the observed data. The general frequentist[8] procedure goes as follows:

[8] The presented frequentist model selection procedure has a Bayesian equivalent with the advantage that similarly well scoring models may be averaged together.

1. express the *marginal likelihood* L_M for π based on y and X under all models $M \in \mathcal{M}$, by taking expectations (over nuisance parameters such as μ_{ig}, ψ_v or over unknown θ_{ig})
2. maximize L_M with respect to π obtaining the ML estimate $\hat{\pi}_M = \arg\max_\pi L_M(\pi)$
3. for each $M \in \mathcal{M}$ evaluate model fit using some objective, likelihood-based, criterion such as AIC, BIC and select the highest scoring model M^* and the corresponding $\hat{\pi}_{M^*}$

In the following section (4.3.1) we will express the marginal likelihoods for π (and possibly other parameters). Computational points on optimization and implementation are not discussed. In Section 4.5 we consider model comparison using simulated data.

4.3.1 Marginal likelihood

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

$$L_{M1}(\pi, \nu) = \frac{\Gamma(\nu)}{\prod_k \Gamma(\pi_k \nu)} \prod_g \int_0^1 \prod_k \mu_k^{\pi_k \nu - 1} \prod_i \sum_k L_{ig}^k \,\mathrm{d}\mu. \tag{15}$$

The summation of L_{ig}^k terms corresponds to taking expectation over all K exclusion states, whereas the integral marginalizes over μ. The running product over i follows from the conditional independence of exclusion states among individuals given gene g, and the one over g from the assumed independence of genes.

L_{M1} in Eq. 15 depends on the parameter ν, which may be of some interest because it quantifies the population-wide variation of exclusion state within 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 choosing M1.1 or M1.2 a priori, i.e. without evaluating how well they fit the data. To obtain the likelihood for those cases let us denote $L_{M1}(\pi) \equiv f(y|n, p, \kappa, \pi)$ and recall Eq 12. Then Eq. 15 simplifies to

$$L_{M1.1}(\pi) = \prod_g \sum_k \prod_i L_{ig}^k \tag{16}$$

$$L_{M1.2}(\pi) = \prod_{i,g} \sum_k L_{ig}^k, \tag{17}$$

where L_{ig}^k is used in the sense of M1-M3 (Eq. 12).

Turning to model M2, let us assume that the matrix B of regression parameters is known (preset and/or estimated as in Section 4.2). Write $L_{M2}(\pi) \equiv f(y|n, X, B, \kappa, \pi)$. It is easy to see that $L_{M2}(\pi)$ has the same form as Eq. 16; of course in this case the semantics of L_{ig}^k is connected to M2 (recall remark below Eq. 12-13).

4.4 Classification

For binary exclusion state ($K = 2$) we can formulate the task of classification for a given (i, g) as the statistical test of two simple hypotheses. The null hypothesis is that of biallelic expression: $B_{ig} : \theta_{ig} = 0$ and the alternative monoallelic expresssion $M_{ig} : \theta_{ig} = 1$.

As mentioned in Section 1.2 the Neyman-Pearson lemma ensures that the likelihood ratio $\Lambda_{ig} = L_{ig}^1 / L_{ig}^0$ used as test statistic affords the most powerful test at a given significance level, so it is highly preferable to use Λ_{ig}. For nested hypotheses $H_0 \subset H_1$ the asymptotic distribution of twice the log-likelihood ratio is χ^2 with degrees of freedom given by the increase in unknown parameters from H_0 to H_1. However, in the present case $H_0 = B_{ig}$ and $H_1 = M_{ig}$ because we defined exclusion states based on disjoint intervals ($\mathcal{P}_0$ and $\mathcal{P}_1$, Section 2.2.1). For this reason H_0 is not $\subset H_1$ so the asymptotic χ^2 distribution doesn't hold.

Fortunately, however, the present case lends itself to Bayesian hypothesis testing with Λ_{ig} playing the role of Bayes factor and π_1/π_0 the corresponding prior odds. Let's write $\Pr(M_{ig}) \equiv \pi_1$ to emphasize that π_1 is the prior probability of monoallelic expression. Likewise, let $\Pr(B_{ig}) \equiv \pi_0$. Then the posterior probability of monoallelic expression given n_{ig} and after observing that $Y_{ig} = y_{ig}$ is

$$\Pr(M_{ig}|n_{ig}, y_{ig}) = \frac{L^1_{ig}\Pr(M_{ig})}{L^1_{ig}\Pr(M_{ig}) + L^0_{ig}\Pr(B_{ig})}. \tag{18}$$

This Bayesian hypothesis testing easily extends to the general case of K-ary allelic state ($K \geq 2$), where $B_{ig} : \theta_{ig} = 0$ as in the binary case but $M_{ig} : \theta_{ig} > 0$. Then $\Pr(\theta_{ig} = k)$ can be calculated as $L^1_{ig}\Pr(\theta_{ig} = k) / \sum_{k'} L^{k'}_{ig}\Pr(\theta_{ig} = k')$. Therefore $\Pr(M_{ig}|n_{ig}, y_{ig}) = \sum_{k:k>0} \Pr(\theta_{ig} = k)$.

4.5 Thoughts on simulations

Simulations are helpful in many ways such as comparing performance of alternative approaches in some inference task. Two important choices must be made prior to a simulation experiment: the inference task and the model (the sampling distribution). Testing under all relevant tasks (classification or estimation of parameters such as π) is desirable. However, a single model that presumed to be true should be selected based on mechanistic arguments and/or model fit to real data.

In the present case, what should be that presumed true model? As pointed out in Section 1.2, the previous approaches do not allow objective, likelihood-based, evaluation of model fit. Turning to mechanistic arguments, how should allelic exclusion depend on the measured explanatory variables like age or gender? Suppose we have a reason to exclude such dependence. Then it still remains to be specified how allelic exclusion varies population-wide within any given gene.