Binomial Models of Reference Read Counts

Attila Gulyás-Kovács

March 16, 2016

1 Introduction

1.1 Goals

estimation of π , where π is the expected fraction of monoallelically expressed genes per genome

within-gene variation to what extent and how does exclusion state vary across individuals for any given gene?

regression if there is such variation, how does it depend on age and other measured explanatory 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 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 the relative contribution of within-gene and of across-gene variance to total variance (across all genes and individuals) remains unknown. 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 the lack of clarifying those properties, which has motivated this article to explicitly describe previously intended and/or novel modeling approaches and the way their utility in achieving 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 given individuals and genes, all individuals (population) within genes, and across all genes (genome). This is achieved via local (transcript level) and global (individual and higher levels) 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
frac. π of monoall.	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 (not giving null/alternative distributions) that prevents likelihood calculations.

Even if it were possible to calculate the likelihoods based on S_{ig}, it could only be done with loosing information on exclusion state 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 mean that S_{ig} is not a sufficient statistic[1] for exclusion state. Although they were recognized previously, only partial, post-hoc, corrections could be applied. In contrast, proposed local models operate with bona fide counts and relax the simplifying assumption by considering all possible *allele configurations* thus containing all information on exclusion state and its likelihood[2].

The lack of S_{ig}-based likelihood for exclusion state prevents the estimation of the error rates of classification and that of the expected fraction π because the two are inherently coupled as explained in a previous article[3]. All proposed global models contain π parameter, which can be estimated by maximum likelihood based on the complete dataset. That estimate then can be combined with likelihood ratios representing the odds that the read count data support mono vs. biallelic expression. 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 those based on likelihood ratios.

Previous regression analysis used the vector LOI_R_g as response variable derived from S_{ig} with a data transformation step; some limitations of the former obviously follow from

[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

those of latter (discussed above). More limitations have been found[5] to arise from incorrect use of regression weights. Moreover, the data transformation may only partially remove the observed strong nonlinearity and heteroscedasticity of read count/S_{ig}-based regression thus leading to bias. Finally, the interpretation of LOI_R$_g$-based regression results in terms of exclusion state is 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 awkward 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 permitting selection of the best fitting model based on some objective criterion like AIC or BIC.

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.

TODO: 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.

2.2 Local models of allelic exclusion

The probability model presented here is *local* in the sense that the global models in Section 3 will be based on this local model or a very similar one. However, even though we call this model local, it describes allelic exclusion and read counts at the biologically relevant level of (i, g) pairs in contrast with the previously considered binomial model restricted to the lower, and irrelevant, level of sites v.

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

2.2.1 Binary (Bernoulli) exclusion state

We introduce (allelic) *exclusion state* $\theta_{ig}$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 binary or 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 generally unknown and must be inferred from the data, which results in uncertainty about not only its exact value 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 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 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. However, read counts have been 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 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.

We may call (ϕ_{ig}, ψ_v) 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

	$\phi_{ig} \neq \psi_v$	$\phi_{ig} = \psi_v$
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 2: Definition of exclusion state θ_{ig} of (i, g) based on p_v and the allele configuration (ϕ_{ig}, ψ_v) for site $v \in (i, g)$

[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.

	$\phi_{ig} \neq \psi_v$	$\phi_{ig} = \psi_v$
biallelic, $\theta_{ig} = 0$	$1 - p_v \in \mathcal{P}_0$	$p_v \in \mathcal{P}_0$
weakly monoallelic, $\theta_{ig} = 1$	$1 - p_v \in \mathcal{P}_1$	$p_v \in \mathcal{P}_1$
strongly monoallelic, $\theta_{ig} = 2$	$1 - p_v \in \mathcal{P}_2$	$p_v \in \mathcal{P}_2$

Table 3: Definition of exclusion state θ_{ig} of (i, g) under the multinomial local model with $K = 3$ states.

We will symbolically represent Table 2 by writing

$$p_v = P[\theta_{ig}, \delta_{\phi_{ig}\psi_v}] \tag{4}$$

$$P = \begin{pmatrix} 1 - \mathcal{P}_0 & \mathcal{P}_0 \\ 1 - \mathcal{P}_1 & \mathcal{P}_1 \end{pmatrix}, \tag{5}$$

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. Based on the data we have some uncertain knowledge on p_v, which we want to use to infer θ_{ig}. Suppose we know 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 general we are uncertain about the allele configuration and we need to take expectation (i.e. average) over all four configurations using the prior probabilities κ and $1/2$. Moreover, if the number s_{ig} of polymorphic sites is > 1 then we will base the inference of θ_{ig} on all $p_v : v \in (i, g)$ jointly, taking expectation over all $4^{s_{ig}}$ configurations.

2.2.2 Multinomial exclusion state

A generalization of the binary local model is the multinomial with K exclusion states including biallelic expression (for the binary model $K = 2$).

Then $\{\mathcal{P}_k : k = 0, ..., K-1\}$ is a sequence of disjoint subintervals of $[1/2, 1]$. For instance, if $K = 3$ then Table 2 changes to Table 3 and the P matrix of the binary model (Eq. 5) needs to be extended accordingly.

3 Global models

Several global models are formulated in this article, which can be classified by two aspects (Table 4):

1. the variance of exclusion state θ_{ig} within each gene g and

2. the response variable to the explanatory variables in X

M1 No influence of explanatory variables

TODO: plate diagram

In this model, denoted as M1, both subintervals in Eq. 3-5 consist of a single point such that $\mathcal{P}_0 = \{1/2\}$ and $\mathcal{P}_1 = \{p_1\}$, where p_1 is some fixed number, say 0.9. Then Eq. 4 remains the same but Eq. 5 changes to

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

For example, 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.

A key aspect of M1

Now we will consider two special cases of M1.

TODO: DAG with theta leaves

M1.1 Zero variance within each gene

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 $\theta_{1g} = ... = \theta_{Ig}$. This means that 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).

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).

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.

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 across individual variation in explanatory

		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 4: Overview of the global models in this article. For each global model the exclusion state θ_{ig} may be binary or multinomial, in which case an “m” is appended to the name, (e.g. M1m) to distinguish from the binary cases (e.g. M1) indicated in the table.

variables x_i induces variation in p_v. For this the local models introduced in Section 2.2 must be extended with the regression of Y_v on X.

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 Eq. 4 modifies to

$$p_v = \max\left(\operatorname{logit}^{-1}(x_i\, b_v), \frac{1}{2}\right) \tag{7}$$

$$b_v = B[\theta_{ig}, \delta_{\phi_{ig}\psi_v}], \tag{8}$$

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:

$$B = \begin{pmatrix} (0, ..., 0)^\top & (0, ..., 0)^\top \\ -\beta & \beta \end{pmatrix} \tag{9}$$

$$\beta = (\beta_0, \beta_1, ..., \beta_{R-1})^\top \tag{10}$$

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 $\operatorname{logit}^{-1}(0) = 1/2$ 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 $\beta_0 = \operatorname{logit}(p_1)$ also holds, then M2 is probabilistically equivalent to M1.1. So, for consistency between models, we should set $\beta_0 = \operatorname{logit}(p_1)$, which has the additional advantage of having one less unknown parameters.

TODO: logit function

M3 Regression of θ_{ig} on explanatory variables

TODO: plate diagram

TODO: link function

4 Inference

4.1 Local models and classification

4.1.1 Likelihood

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

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

$$\begin{aligned} L_{ig}^a &\equiv f_{ig}(y_{ig} | n_{ig}, P, \kappa, \theta_{ig} = a) && (12) \\ &= \frac{1}{2} \sum_{\phi_{ig}=0}^{1} \kappa^{\phi_v} (1-\kappa)^{1-\phi_v} \prod_{v \in (i,g)} \sum_{\psi_v = 0}^{1} f_v(y_v | n_v, P, \phi_{ig}, \psi_v, \theta_{ig} = a), && (13) \end{aligned}$$

where a is 0 or 1, and L_{ig}^a 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}^a shall be used model M2 as well; its specific semantics shall be clear from the context.

4.1.2 Classification

4.2 Selection of a global model, estimation of π and β

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 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 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

4.2.1 Marginal likelihood for π

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

$$L_{M1}(\pi, \nu) = b^{-1} \prod_g \int_0^1 \mu^{\pi\nu} (1-\mu)^{(1-\pi)\nu} \prod_i \left[(1-\mu) L_{ig}^0 + \mu L_{ig}^1 \right] \mathrm{d}\mu \tag{14}$$

where b is the beta function evaluated at $(\pi\nu, (1-\pi)\nu)$. L_{M1} is marginal in the sense that expectation was taken over not only ψ_v (as in Eq. 12) but also θ_{ig} and μ_g.

L_{M1} in Eq. 14 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 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. 14 simplifies to

$$L_{M1.1}(\pi) \quad = \quad \prod_{g} \left[(1-\pi) \prod_{i} L_{ig}^{0} + \pi \prod_{i} L_{ig}^{1} \right] \tag{15}$$

$$L_{M1.2}(\pi) \quad = \quad \prod_{i,g} \left[(1-\pi) L_{ig}^{0} + \pi L_{ig}^{1} \right], \tag{16}$$

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

Turning to model M2, we assume that the matrix B of regression parameters is known (preset and/or estimated). 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. 15; of course in this case the semantics of L_{ig}^{a} is connected to M2 (recall remark below Eq. 12).

4.3 Estimating regression parameters from training data

This estimation we need

- to accept model M2, which implies, for each gene, uniformity of exclusion state across all individuals
- a training set of genes known to be expressed monoallelically, collected from I' individuals

$$L_{M2}(B) = \frac{1}{2} \prod_{g} \prod_{i=1}^{I'} \sum_{\phi_{ig}=0}^{1} \kappa^{\phi_v} (1-\kappa)^{1-\phi_v} \prod_{v \in (i,g)} \binom{n_v}{y_v} \sum_{\psi_v=0}^{1} p_v^{y_v} (1-p_v)^{n_v - y_v} \tag{17}$$

where p_v is given by Eq. 7-8. The outer and inner summation together reflect marginalization over all allelic configurations (ϕ_{ig}, ψ_v), whereas the three running products represent the data aggregation over individual polymorphic sites v over individuals i over monoallelically expressed genes g to the level of the complete training data set.

4.4 Classification

We formulate the task of classification as the statistical test of two simple hypotheses: $H_0 : \ \theta_{ig} = 0$ versus $H_1 : \ \theta_{ig} = 1$ (biallelic versus monoallelic expression). According to the Neyman-Pearson lemma the likelihood ratio $\Lambda_{ig} = L_{ig}^{1} / L_{ig}^{0}$ provides the test statistic for the most powerful test at a given significance level, so it is 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. But in the present case 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 $\pi/(1-\pi)$ the corresponding prior odds. Let's write $\pi(\theta_{ig} = 1) \equiv \pi$ to emphasize that π is the prior probability that $\theta_{ig} = 1$; and likewise

$\pi(\theta_{ig} = 0) \equiv 1 - \pi$. Then the posterior probability of $H_1 : \theta_{ig} = 1$ given n_{ig} and after observing that $Y_{ig} = y_{ig}$ is

$$\pi(\theta_{ig} = 1 | n_{ig}, y_{ig}) = \frac{L^1_{ig} \pi(\theta_{ig} = 1)}{L^1_{ig} \pi(\theta_{ig} = 1) + L^0_{ig} \pi(\theta_{ig} = 0)} \quad (18)$$

4.5 Thoughts on simulations

Simulations are helpful in 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 across individuals within any given gene.