Gene Set Enrichment Analysis

Palle Duun Rohde, Stefan McKinnon Høj-Edwards, Izel Fourie Sørensen, & Peter Sørensen

2022-03-14

Contents

1 Gene Set Enrichment Analysis

After setting up a linear mixed model (chapter **??**) and fitting the model with data (chapter **??**), it is now time to evaluate whether the parameters of the model and the estimated effects are significant. For this chapter, we assume the variance components are estimated, and the relevant algorithm for fitting a model has successfully converged.

The chapter starts with a quick introduction of the different aspects that need to be considered while evaluating an outcome,followed by a quick brush up of the G-BLUP model and, more importantly here, the likelihood. We then finally immerse ourselves in the test statistics used in gene set enrichment analysis (GSEA).

2 Statistical modelling approaches

The test statistics are categorised as belonging to either the *Single-step* or *Two-step* approaches. In the single-step approaches, a genomic feature is modelled by a single model. The estimated effects or variance components are then evaluated, either by the properties of this model (e.g. score based statistics) or by comparing to null hypothesis models.

In the two-step approaches, a single model is used to calculate test statistics on all the markers' effects. These statistics are then aggregated by various means for each genomic feature to test. Goeman and Bühlmann [2007] refer to these as *post hoc* methods, and are similar to popular Gene-set enrichment tests.

In our setting, a genomic feature is basically a set of markers, defined by any external information, and can refer to genes, QTL regions, sequence ontologies (SNPs within genes, in upstream regions, known regulators, etc.), or even chromosomes. The information can be layered, such as when using biological pathways that map to multiple genes that each map to multiple markers.

The two approaches differ in how a genomic feature is modelled. For the single-step approaches, the set of markers are modelled as a *joint* contribution to a phenotypic trait, by including them as an extra random effect (e.g. eq. 3). In the two-step approaches, the markers are modelled independently, and test statistics are calculated for each marker. The test statistics for the genomic features are then different aggregations of the marker-based test statistics.

The test statistics described here all attempt to determine whether a given set of genetic variants contributes to the observed phenotypic trait with anything than mere noise.

2.1 Null hypotheses

We distinguish between two types of null hypotheses, the *competitive* and the *self-contained* [Goeman and Bühlmann, 2007, Maciejewski, 2013]. The self-contained is the easiest and corresponds to determining whether a genomic feature, by it self, does not display any association to the phenotypic trait. This is usually done by defining that the variance component or predicted effect equals zero.

The competitive corresponds to determining whether the degree of association within a genomic feature is the same as outside the genomic feature.

Naturally, the choice of null hypothesis affects the choice of test statistic, but also the biological interpretation of the significance of a finding. The self-contained may be preferable over a competitive, as it has more power [Goeman and Bühlmann, 2007], and the biological interpretation is simpler, as it determines whether there is or there is no association.

2.2 Evaluating the test statistics

Once a test statistic has been calculated, it needs to be evaluated to determine whether the genomic feature of interest is significant. This is done by finding the test statistic's position within a distribution, allowing us to evaluate the probability of finding a test statistic of the given magnitude by chance

We distinguish between three types of distributions; the exact, the approximate, and the empirical found distribution.

The *exact distributions* (e.g. hypergeometric test) are derived from the test statistic itself. They might seem to be the preferred, but only if the test statistic actually does describe the desired property being tested.

The *approximate distributions* (e.g. χ^2) relies on that some distributions approximate each other under certain conditions. We can then replace an intangible expression with a simpler, but when being applied to actual data, the conditions are 'bent' into place.

The *empirical distributions* are the brute-force 'when-all-else-fails' solutions we attend to, when the other distributions are too computational demanding, or the conditions for approximating seem to strongly bent. Usual methods for obtaining these are bootstrapping or permutation routines, but caution should be taken under which conditions the routines are performed.

3 Linear Mixed Models, revisited

We will now refresh our memory on G-BLUP. The ordinary G-BLUP as displayed in (**??**) will be referred to as the null model (M_0), and all assumptions and properties as detailed in section **??** and **??** still hold.

We also introduce the expanded models (see section **??**, page **??**), where the second random factor model the contribution from a subset of markers.

Recall that $\mathbf{g} = \mathbf{Wb}$, and that G-BLUP and M-BLUP are equivalent (see section **??**, page **??**).

$$M_0 : \quad \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zg} + \mathbf{e} \tag{1}$$
$$M_{GWAS} : \quad \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zg} + \mathbf{Zw}_i\mathbf{b}_i + \mathbf{e} \tag{2}$$
$$M_{Set} : \quad \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zg} + \mathbf{Zg}_i + \mathbf{e} \tag{3}$$

where $\mathbf{W}_i$ is the column vector of $\mathbf{W}$ corresponding to the $\{\text{i}^{\text{th}}\}$ marker, $\mathbf{b}_i$ the $\{\text{i}^{\text{th}}\}$ marker effect, and $\mathbf{g}_s = \mathbf{W}_s\mathbf{b}_s$ are the genetic effects modelled exclusively on a subset of markers.

We assume that $\text{Var}(\mathbf{g}) = \mathbf{g}\sigma_g^2$, $\text{Var}(\mathbf{b}_i) = \sigma_{b_i}^2$, $\text{Var}(\mathbf{g}_i) = \mathbf{g}_i\sigma_i^2$, and $\text{Var}(\mathbf{e}) = \mathbf{R} = \mathbf{I}\sigma_e^2$ (note that this differs slightly from previous assumptions).

The random effects are assumed uncorrelated.

3.1 The likelihood, and the first and second derivatives.

As stated in the previous chapter, the log-transformed *ordinary* likelihood can be expressed as

$$l(\boldsymbol{\beta}, \mathbf{V}|\mathbf{X}, \mathbf{y}) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{V}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\beta) \tag{??}$$

where $\mathbf{V} = \sum_{i=1}^{r-1}\mathbf{Zg}_i\mathbf{Z}'\sigma_i^2 + \mathbf{I}\sigma_e^2$ and $\mathbf{g}_i = \frac{\mathbf{W}_i\mathbf{W}_i'}{m_i}$.

$$M_0: \quad V_0 = ZGZ'\sigma_g^2 + I\sigma_e^2 \tag{4}$$
$$M_{GWAS} \quad V = ZGZ'\sigma_g^2 + ZZ'w_i w_i'\sigma_{b_i}^2 + I\sigma_e^2 \tag{5}$$
$$M_{Set} \quad V = ZGZ'\sigma_g^2 + ZG_sZ'\sigma_s^2 + I\sigma_e^2. \tag{6}$$

We note that $\frac{W_s W_s'}{m_s} = G_s$ (**??**).

The *first* derivative of the full model with respect to the variance component of interest is written as a function of the $\{\mathrm{i}^{\mathrm{th}}\}$ element of θ, a vector listing the variance components:

$$\begin{aligned} l'(\theta_i) = \frac{\partial l(\boldsymbol{\beta}, \mathbf{V}|\mathbf{X}, \mathbf{y})}{\partial \theta_i} &= -\frac{1}{2}\operatorname{Tr}(\mathbf{PZG}_i\mathbf{Z}') + \frac{1}{2}\mathbf{y}'\mathbf{ZG}_i\mathbf{Z}'\mathbf{Py} \\ &= -\frac{1}{2}\operatorname{Tr}(\mathbf{V}^{-1}\mathbf{Zg}_i\mathbf{Z}') \\ &\quad + \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}\mathbf{Zg}_i\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \end{aligned} \tag{7}$$

where the projection matrix $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}$.

For the *second* derivatives, we have the *observed* and *expected* information matrix, respectively, as

$$\mathbf{I}_O = \frac{\partial^2 l(\theta|\mathbf{y})}{\partial\theta_i \partial\theta_j} \tag{8}$$
$$\mathbf{I}_E = -\operatorname{E}[\mathbf{I}_O] \tag{9}$$

The inverse of the expected information matrix details the uncertainty of the variance components in θ, and we can write $\hat{\theta} \sim N\left(\theta_0, [\mathbf{I}_E(\theta_0)]^{-1}\right)$ [Sorensen and Gianola, 2002, p.179], where $\hat{\theta}$ is the vector of estimated variance components and θ_0 is the vector of true (but unknown) variance components.

The *average information matrix*, as derived from the second derivations, can be calculated as the average of the observed and expected observation matrix:

$$\mathbf{I}_A(\theta_i, \theta_j) = \frac{1}{2}\mathbf{y}'\mathbf{V}^{-1}\mathbf{Zg}_i\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Zg}_j\mathbf{Z}'\mathbf{V}^{-1}\mathbf{y} \tag{10}$$
$$\mathbf{I}_A(\theta_i, \theta_e) = \frac{1}{2}\mathbf{y}'\mathbf{V}^{-1}\mathbf{ZGZ}'\mathbf{V}^{-1}\mathbf{V}^{-1}\mathbf{y} \tag{11}$$
$$\mathbf{I}_A(\theta_e, \theta_e) = \frac{1}{2}\mathbf{y}'\mathbf{V}^{-1}\mathbf{V}^{-1}\mathbf{V}^{-1}\mathbf{y}. \tag{12}$$

Although these definitions differ from (**??**), the similarity to (**??**) and (**??**) should be evident, if you keep (**??**) in mind. As the second derivative is central to the following scores, an interpretation is described in stat box **??**.

In the case of the restricted likelihood, then $\mathbf{V}$ and $\mathbf{V}^{-1}$ are replaced with $\mathbf{P}$.

4 Single-step approaches

In the single-step approaches, we fit the data to the model with AI-REML and get estimates of the variance components, as well as the likelihood. In other words, we only need to fit the data once per model. For likelihood ratio (LR) testing, two models are required; the full model and a reduced model.

The LR test, Wald's test, and Rao's Score test, can be referred to as 'The Holy Trinity' [Rao, 2009], and are all related to the likelihood and the first and second derivation. For a graphical comparison, see figure

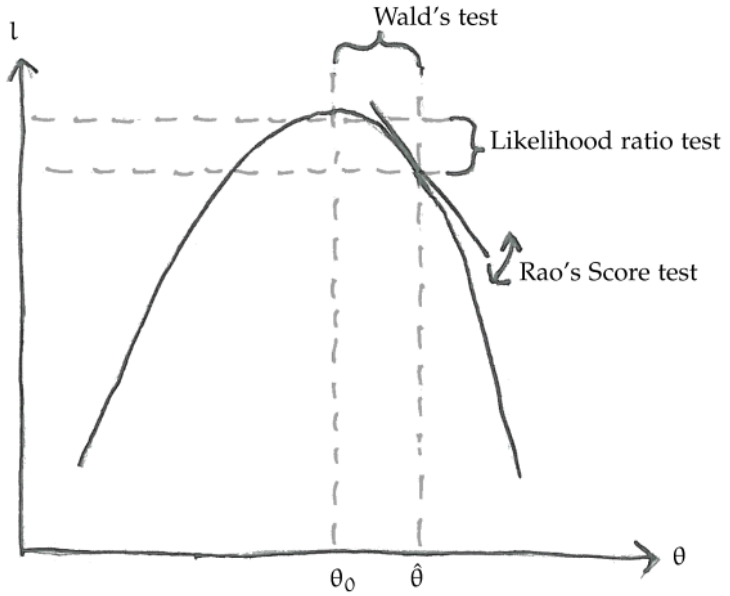

Figure 1: The Holy Trinity of Likelihood Ratio, Wald's and Rao's Score test. The graph displays likelihood as a function of the variance components, maximised at the true value, $\hat{\theta}$.

1. When reviewing these tests, keep in mind that the first derivative gives the slope of the function, and the second derivative is related to the uncertainty of the estimated variance component.

Shortly, the LR test compares the model fit between the full model and the reduced model. In Wald's test, the model parameters are fitted using the full model, and we ask if an estimated variance component is significantly different from a particular value (usually zero). Rao's Score test uses the reduced model (i.e. null model), and estimates the size of improvement in model fit, if we were to plug an additional variance component into the model. Both the Wald and the Rao's score tests are asymptotically equivalent to the LR test, that is, as the sample size becomes infinitely large, the values of the Wald and Rao's score test statistics will become increasingly close to the test statistic from the LR test.

The first derivative of the likelihood is also referred to as the 'score', which is the basis of the Score based statistic. It differs from the previous tests by relying on a simplified expression of the first derivative. The advantage of the score test is that it can be used to search for omitted variables when the number of candidate variables is large.

We conclude this section with test statistics based on estimated effects. The subscript i corresponds to each marker or set of markers.

4.1 Likelihood Ratio Test

The concept of likelihood also provides a framework for testing hypotheses regarding, for example, goodness of fit of models. In particular, the so-called likelihood ratio tests are used to assess whether a reduced model fits the data better than a full model by comparing the likelihoods of the two models. A high likelihood ratio shows that the full model with two (different) variance components is better at explaining the observed genomic variance than the reduced model with only one variance component.

It is fundamental for the reduced model to be nested in the full model, otherwise this approach does not make any sense. When the REML procedure is used, it is also important for the two models being compared to have the same fixed effects, otherwise the two likelihoods are not comparable, as can be easily understood

by looking at the concept of restricted likelihood (see section **??**). The LR test statistic can be derived by using the following formula:

$$T_{\text{LRT}} = 2 \ln \left[\frac{L(\hat{\theta}|\mathbf{y})}{L(\hat{\theta}_r|\mathbf{y})} \right] = -2 \left[l(\hat{\theta}_r|\mathbf{y}) - l(\hat{\theta}|\mathbf{y}) \right] \tag{13}$$

where $l(\hat{\theta}|y)$ is the log-likelihood for the full model, and $l(\hat{\theta}_r|y)$ is the log-likelihood for the reduced model.

When the sample size is sufficiently large, the LR statistic is χ^2 distributed with κ degrees of freedom, where κ parameters that were free in the full model, have been assigned fixed values in the reduced.

At least in theory.We must bear in mind that the above description applies to the reduced model having a parameter *fixed*, i.e. the reduced model has a variance component set to zero. Self and Liang [1987] has shown that when the estimated variance components are on the border of the parameter space (i.e. close to zero), the LR statistic distribution is no longer χ^2 distributed, but instead a *mixture* of χ^2 distributions. **?** comments on a alternative method of estimating the significance, that includes permuting the linkage between the observations and the random variables. But as this might be computational demanding in numerous repeated REML analysis, Visscher suggest permuting the linkage, setting the variance component to a very small value, and then simply calculate the likelihood, and compare this to the model with the variance component set to zero.

4.2 Wald's test

The Wald's test is a parametric test that compares an estimated variance component to some particular value, θ_0, based on some null hypothesis:

$$\frac{(\hat{\theta} - \theta_0)^2}{\text{Var}(\hat{\theta})} \tag{14}$$

The test statistic is assumed χ^2-distributed with one degree of freedom. In our current framework, if our null hypothesis was that the $\{\text{i}^{\text{th}}\}$ variance component was equal to zero, the above can be expressed as a quadratic form by

$$T_{\text{Wald}} = (\hat{\theta}_i - 0)' \left[\mathbf{I}_E(\hat{\theta})^{-1} \right]^{ii} (\hat{\theta}_i - 0) \tag{15}$$

where $\left[\mathbf{I}_E(\hat{\theta})^{-1} \right]^{ii}$ is the $\{\text{i}^{\text{th}}\}$ diagonal element of the inverse expected information matrix. Wald's test has the advantage, that it only requires fitting and estimating the parameters under the full model. If the test fails to reject the null hypothesis, this suggests that removing the corresponding variance component from the model will not substantially harm the fit of that model.

Asymptotically under the null hypothesis ($\hat{\theta} = 0$) the large-sample distribution of the maximum likelihood estimates of the parameters is multivariate normal with mean vector $\mathbf{0}$ and variance-covariance matrix $\mathbf{V} = \mathbf{I}_E(\hat{\theta} = 0)^{-1}$ [Sorensen and Gianola, 2002, p. 179]. Consequently, the large-sample distribution of the quadratic form of T_{Wald} is χ^2 with number of parameters in $\hat{\theta}_i$ as degrees of freedom. This is evident if we consider $(\hat{\theta}_i - 0)$ as a random variable and apply the theorem in section **??**, page **??**, and assuming that $[\mathbf{I}_E(\hat{\theta})^{-1}]^{ii}$ is full rank. It is important to note that Wald's test can be applied to both variance components, as above, and predicted marker effects.

The Wald's test is computed as the parameter estimate divided by its asymptotic standard error. The asymptotic standard errors are computed from the inverse of the second derivative matrix of the likelihood with respect to each of the covariance parameters. The Wald's test is valid for large samples, but it can be unreliable for small data sets. When used on correlated variance components, I_E might not be full rank and

therefore not invertible. If one of the variance components is close to zero, we may also experience issues with the information matrix. Which are known to have a skewed or bounded sampling distribution (which be the case for REML).

4.3 Rao's Score test

The Rao's score test requires estimating only a single model that does not include the parameter(s) of interest. This means we can test whether adding the variance component to the model will result in a significant improvement in model fit, without fitting additional models. The test statistic is based on the slope (or score) of the likelihood function, using model parameters estimated under the null model. If the null model is true, then we would expect that the slope of the likelihood function is close to zero. If the null model is not true, fixing a variance component to a value will penalise the likelihood.

Instead of calculating likelihoods for both the null and the full model, we can utilise the first and second derivatives in the same way as in the Newton-Raphson method (**??**), page **??**, to get an indication of the produced change. The Rao's Score test statistic can therefore be formulated as

$$T_{\text{Rao}} = \left(l'(\theta_i = 0, \hat{\theta}_{-1})\right)' \left[\mathbf{I}_E(\theta_i = 0, \hat{\theta}_{-1})^{-1}\right]^{ii} \left(l'(\theta_i = 0, \hat{\theta}_{-1})\right) \tag{16}$$

where $\left(l'(\theta_i = 0, \hat{\theta}_{-i})\right)$ is the first derivative of the *full model's* likelihood function, calculated using the parameters estimated with the *null model* and the parameter of interest (θ_i) fixed cf. null model. $\left[\mathbf{I}_E(\theta_i = 0, \hat{\theta}_{-1})^{-1}\right]^{ii}$ is the $\{\text{i}^{\text{th}}\}$ diagonal element of the inverse expected information matrix, under same conditions as the first derivative. It is possible to use the average between the expected and observed information matrix, i.e. the average information matrix, as it may be easier to compute, cf. section **??** [Freedman, 2007, Johnson and Thompson, 1995, Madsen et al., 1994, Jensen et al., 1997].

The Rao's Score test has an asymptotic distribution of χ^2 with number of parameters in $\hat{\theta}_i$ as degrees of freedom when the null hypothesis is true. Some issues related to the test statistic may occur if the information matrix is not positive definite which can happen if the null hypothesis is true [Freedman, 2007].

4.4 Score based statistics

There are several alternate score based statistics that are also derived from the first derivative of the likelihood (7). The last term form the basis of a number of score based test statistics [Goeman et al., 2004, Wu et al., 2011, Wang et al., 2013] from an argument that this is the only part that involves the data [Huang and Lin, 2013]. The score statistic can therefore be written as

$$T_{\text{Score}} = \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}\mathbf{Z}\mathbf{g}_i\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{17}$$

which under the null hypothesis $H_0 : \sigma_i^2 = 0$ should be close to zero. If the parameters are estimated under the null model, we have the score statistic for a group of markers i as

$$T_{\text{Score}} = \frac{1}{2}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\hat{\mathbf{V}}_0^{-1}\mathbf{Z}\mathbf{g}_i\mathbf{Z}'\hat{\mathbf{V}}_0^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \tag{18}$$

and by utilizing $\hat{\mathbf{P}}\mathbf{y} = \hat{\mathbf{V}}_0^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \hat{\mathbf{e}}$, T_i can be computed as

$$T_{Score} = \frac{1}{2}\hat{\mathbf{e}}'\mathbf{Z}\mathbf{G}_i\mathbf{Z}'\hat{\mathbf{e}} = \frac{1}{2}\hat{\mathbf{e}}'\mathbf{Z}\frac{\mathbf{W}_i\mathbf{W}_i'}{m_i}\mathbf{Z}'\hat{\mathbf{e}} \tag{19}$$

where the latter expansion is done cf. (**??**), page **??**, for the subset of markers. This is computational simple, and also easy to derive an empirical distribution of the score statistic under both the competitive and self-contained null hypothesis. For a subject-randomisation, where the link between observations and genotyped animals are permuted, this can be obtained by shuffling $\hat{\mathbf{e}}$. For gene-randomisation that approximates the competitive null hypothesis, only $\mathbf{g}_i$ needs to be recalculated. In both cases, the 'randomised' test statistic is re-calculated by a series of sums without the inconveniences of matrix inversions.

4.5 Approximate distribution for T_{Score}

As an alternative to an empirical distribution of the score based test statistics, it is possible to derive an approximate distribution of T_i as follows [Wang et al., 2013]. We can rewrite (19) as

$$T_{\text{Score}} = \frac{1}{2}\tilde{\mathbf{y}}'\mathbf{V}_0^{\frac{1}{2}}\mathbf{M}\mathbf{V}_0^{\frac{1}{2}}\tilde{\mathbf{y}} \tag{20}$$

where $\mathbf{M} = \mathbf{V}_0^{-1}\mathbf{Z}\mathbf{g}_i\mathbf{Z}'\mathbf{V}_0^{-1}$, and $\tilde{\mathbf{y}} = \mathbf{V}_0^{-\frac{1}{2}}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ which entails $\tilde{\mathbf{y}} \sim N(\mathbf{0}, \mathbf{I})$.

We consider Satterthwaite's procedure of moment matching to approximate the null distribution of T_{Score} by a Gamma distribution $\text{GAMMA}(a, b)$. The two parameters in the approximate distribution are calculated by matching the first and second moments (mean and variance) with those of the score statistic. Taking a Gamma distribution as an example,
we attempt to obtain the mean, $\mu_T = ab$, and the variance $\nu_T = ab^2$. $\Rightarrow a = \mu_T^2/\nu_T$ and $b = \nu_T/\mu_T$. Due to its quadratic form, it is easy to obtain the mean and variance of T_{Score}:

$$\begin{aligned}\mu_T &= \frac{1}{2}\,\text{Tr}(\mathbf{V}_0^{-1}\mathbf{Z}\mathbf{g}_i\mathbf{Z}') \\ \nu_T &= \frac{1}{2}\,\text{Tr}\left((\mathbf{V}_0^{-1}\mathbf{Z}\mathbf{g}_i\mathbf{Z}')^2\right)\end{aligned}$$

The derivation of this is shown in appendix section **??**.

4.6 Decomposing the score based statistic

Goeman et al. [2004, 2006] showed that the score based statistics have a number of interpretations. This can be illustrated by rewriting the test statistic. First, the influence of the markers can be seen by rewriting (19) as

$$\begin{aligned}T_{\text{Score}} &= \frac{1}{2}\frac{1}{m_i}\sum_{j=1}^{m_i}\left(\mathbf{W}_i'\mathbf{Z}'\hat{\mathbf{V}}_0^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})\right)^2 \\ &= \frac{1}{2}\frac{1}{m_i}\sum_{j=1}^{m_i}(\mathbf{W}_i'\mathbf{Z}'\hat{\mathbf{e}})^2\end{aligned} \tag{21}$$

where the expression $T_{ij} = (\mathbf{W}_{ij}Z'\hat{\mathbf{V}}_0^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}))^2$ is the contribution of the $\{\text{j}^{\text{th}}\}$ marker to the test statistic. Therefore the test statistic T_i for a set of m_i markers is just the average of the statistics $T_1, ..., T_{m_i}$ calculated for the m_i single markers that the genomic feature consists of. From this expression it can also be seen that each T_i can again be written as (a multiple of) the squared covariance between the genetic markers and the adjusted phenotype. Because the averaging is done at this squared covariance level, markers with a large variance have much more influence on the outcome of the test statistic T_i than genetic markers with a small variance.

Second, the influence from each of the q subjects can be seen by rewriting T_i as

$$T_i = \frac{1}{2}\frac{1}{m_i}\sum_{j=1}^{q}\sum_{k=1}^{q}\left((\mathbf{g}_i\mathbf{Z}')_{jk}\hat{\mathbf{e}}_j\hat{\mathbf{e}}_k\right)^2 \tag{22}$$

The statistic T_i therefore has a high value whenever the terms of these two matrices are correlated, that is when the covariance structure of the genetic markers between subject resembles the covariance structure between their phenotypes. The score test can therefore be seen as a test to see whether subjects with similar genetic profiles also have similar phenotypes.

4.7 Deriving a set test statistic for predicted marker effects

As we can evaluate the significance of a model fit by using properties of the likelihood ratio, it is also possible to evaluate whether a predicted marker effect is significant. By using G-BLUP, it is possible to 'backsolve' the predicted marker effects from predicted genetic values, by using (**??**), page **??**:

$$\hat{\mathbf{b}} = \mathbf{W}'(\mathbf{W}\mathbf{W}')^{-1}\hat{\mathbf{g}} \tag{??}$$

and we can derive the variance in a similar fashion as

$$\mathrm{Var}(\hat{\mathbf{b}}) = \mathbf{W}'(\mathbf{W}\mathbf{W}')^{-1}\,\mathrm{Var}(\hat{\mathbf{g}})(\mathbf{W}\mathbf{W}')^{-1}\mathbf{W}' \tag{23}$$

where $\mathrm{Var}(\hat{\mathbf{g}}) = \mathbf{g} - \mathbf{C}^{gg}$ (**??**). By solving G-BLUP (instead of M-BLUP) it is now possible to predict the marker effects, as well as estimates of the variance of the predicted effects. Assuming a null hypothesis $H_0 : \hat{\mathbf{b}}_j = 0$, i.e. that the marker effects are equal to zero, we can simply apply Wald's test (14), this time as

$$T_{\hat{\mathbf{b}}_j} = \frac{\hat{\mathbf{b}}_j^2}{\mathrm{Var}(\hat{\mathbf{b}}_j)} \tag{24}$$

where $\mathrm{Var}(\hat{\mathbf{b}}_j)$ is the estimate of variance of the $\{\mathrm{j^{th}}\}$ element of $\hat{\mathbf{b}}$, obtained from the $\{\mathrm{j^{th}}\}$ diagonal of (23). We previously defined this to be χ^2-distributed with one degree of freedom (p. 6). For a set of markers, the predicted effects is a vector, $\hat{\mathbf{b}}_{set}$, and the above equation is rewritten to

$$T_{\hat{\mathbf{b}}_{Set}} = \hat{\mathbf{b}}'_{Set}\left(\mathrm{Var}(\hat{\mathbf{b}}_{Set})\right)^{-1}\hat{\mathbf{b}}_{Set} \tag{25}$$

and for large sets of genetic markers would require inverting a large matrix which may be computationally difficult. An alternative to (24) is

$$T_{\hat{\mathbf{b}}_j} = \frac{\hat{\mathbf{b}}_j}{\sqrt{\mathrm{Var}(\hat{\mathbf{b}}_j)}} \tag{26}$$

which follows a Student t-distribution with $(n-m)$ degrees of freedom [Cule et al., 2011][1]. However, in our scenarios $m \gg n$, so we refer to an alternate calculation to obtain the effective number of degrees of freedom for ordinary linear regressions [Cule et al., 2011]:

$$\kappa = n - \mathrm{Tr}(\mathbf{H}) \tag{27}$$

[1] When the number of degrees of freedom approaches infinity, the Student t-distribution approximates the standard normal distribution.

where **H** is the hat matrix for the conditional prediction as defined in (**??**). As a final comment on the hat matrix, Cule et al. [2011] wrote:

> Hastie and Tibshirani define Tr(**H**) as the degrees of freedom taken up by the penalised model fit. (...) In the case of large sample size, as is typically the case in genetic data, the distribution of the test statistic under the null hypothesis is asymptotically (standard) normal.

5 Two-step approaches

In the first step, a test statistic for the association (e.g. t-statistics) of individual markers with the trait phenotype is obtained from traditional single-marker (M_{GWAS}), or all-marker statistical models (e.g. M_{Set}). In the second step, for each set of markers being tested, a summary statistic is obtained. For each set we construct an appropriate summary statistic that measures the degree of association between the set of markers and the phenotypes.

5.1 Summary statistics

The first summary statistic is based on the idea to identify the association between two types of classification of the markers: 1) being in a predefined set of markers, and 2) being associated to the trait phenotype. The test is then based on counting the number of markers that are in and outside the set as well as the number of markers that are associated or not.

Determination of association of individual markers is based on a single marker test statistic such as the t-statistics and a threshold for this statistic.

Let m denote the total number of markers tested, m_F is the total number of markers belonging to the set of interest, m_A is the number of associated markers, and m_{AF} is the number of associated markers belonging to the feature. Thus m, m_A, and m_{AF} are fixed.

We consider two properties of a marker; 1) to be associated to the phenotypic trait, and 2) belong to the genomic feature of interest. Let H_0 denote the null hypothesis, that the two properties of a marker are independent, or equivalently that the associated markers are picked at random from the total population of tested markers. Rivals et al. [2007] show that this can be formulated and tested in a number of ways.The different tests can be evaluated using an exact (Hypergeometric), approximate (χ^2), or empirical distribution (T_{Sum}) under the null hypothesis.

5.2 Hypergeometric test

The total number of markers that belong to the genomic feature of interest and that are associated to the trait phenotype can be computed as

$$T_{\text{Count}} = m_{AF} = \sum_{i=1}^{m_F} \mathbf{I}(t_i > t_0) \tag{28}$$

where t_i is the $\{\text{i}^{\text{th}}\}$ single marker test statistics, t_0 is an arbitrary chosen threshold for the single marker test statistics, and I is an indicator function that takes the value 1 if the argument $(t_i > t_0)$ is satisfied.

The number of associated markers that belong to a genomic feature, m_{AF}, can be modelled using a Hypergeometric distribution that has a discrete probability distribution that describes the probability of m_{AF} successes in m_F draws without replacement (can only be drawn one time) from a finite population of size m containing exactly m_A successes. Thus if the null hypothesis is true (associated markers are picked at random from the total population of tested markers), then the observed value m_{AF} is a realization of the

random variable M_{AF} having a hypergeometric distribution with parameters m, m_A, and M_F, which we denote by $M_{AF} \sim \text{HYPER}(m, m_A, m_F)$.

However, the hypergeometric test assumes that the markers being sampled are independent, a rather strong assumption in genetic data. Therefore, the hypergeometric test might not correctly identify significant association, but instead associated markers that are strongly correlated [Goeman and Bühlmann, 2007].

5.3 χ^2 test

The second summary statistic is based on a χ^2 test. Let the observed data be presented in a contingency table where each observation is allocated to one cell of a two-dimensional array of cells according to the values of the two outcomes:

%\begin{table}%

Table 1: Contingency table for χ^2 test in two-step approach.

	In feature	Not in feature	Total
Associated	m_{AF}	m_{AnF}	m_A
Not associated	m_{nAF}	m_{nAnF}	m_{nA}
Total	m_F	m_{nF}	m

Let again H_0 denote the null hypothesis that the property to belong to the genomic feature of interest, and that to be associated, are independent. If the occurrence of these two outcomes are statistically independent, we expect the number in the $\{\text{ij}^{\text{th}}\}$ cell to be $f_{ij} = \frac{m_i m_j}{m^2}$. Based on this expectation we can compute the following summary statistic:

$$T_{\chi^2} = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(m_{ij} - m \cdot f_{ij})^2}{m \cdot f_{ij}} \tag{29}$$

where f_{ij} is the observed frequency in the contingency table. This is called the χ^2 test for independence and it has been shown that the T_{χ^2} variable is asymptotically χ^2 distributed with one degree of freedom [Wackerly et al., 1996, Rivals et al., 2007]. The alternative hypothesis corresponds to the variables having an association or relationship, where the structure of this relationship is not specified.

In summary, under the null hypothesis that the probability of a marker belonging to a genomic feature is independent of being associated to the trait phenotype (i.e. $p_{AF} = p_{nAF}$), the exact distribution of M_{AF} is the hypergeometric distribution $M_{AF} \sim \text{HYPER}(m, m_A, m_F)$. This distribution can, if m is large, be approximated with the bionomial distribution $M_{AF} \sim Bi(m_A, m_F/m)$. If the two samples, are large, it is also possible to exhibit an approximately normal variable Z or its square $D^2 = Z^2$, the latter being hence approximately χ^2 distributed with one degree of freedom.

One of the differences between the hypergeometric and χ^2 test statistic is that the latter implicitly distinguishes between over- or under-representation, i.e. the squared difference between the expected and observed counts for all the 4 cells contribute to the T_{χ^2} test statistic. It is possible to test for both over-representation ($p_{AF} > p_{nAF}$) or under-representation ($p_{AF} < p_{nAF}$).

Both tests are potentially of interest for understanding the genetic basis of complex traits. If the number of associated markers is very small in the genomic feature then it may be interpreted as selection/highly conserved region. If the number of associated markers is large in the genomic feature then this may indicate we have identified an important feature underlying the genomic variance of the trait.

In cases where both over-representation and under-representation of genomic features are of interest then it is generally most appropriate to consider a two-sided test. It is also possible to define more detailed and specific hypothesis such as testing whether the associated markers contribute negatively or positively to the trait of interest.

However, there is the arbitrariness of the threshold for determining 'significantly associated', no matter how it is chosen and markers whose test statistics differ by a tiny amount may be treated completely differently. By design this test will have high power to detect association if the genomic feature harbour markers with large effects, but it will not detect a situation where there are many markers with small to moderate effects [Newton et al., 2007]. In this case, it is more powerful to use a summary statistic such as the mean or sum of the test statistics for all markers belonging to the same genomic feature.

5.4 $T_{\mathbf{Sum}}$

As noted above, if the phenotypic trait of interest is governed by many markers with small to moderate effects, counting 'significantly associated' markers neglects a lot of information. We therefore consider the third summary statistic

$$T_{\text{Sum}} = \sum_{i=1}^{m_F} t_i \tag{30}$$

where t_i is a test statistic for the $\{\text{i}^{\text{th}}\}$ marker. There are number of choices for t_i such as likelihood ratio, the score based statistic, or the predicted marker effects, and they might be transformed by e.g.squaring. The nature of T_{Sum} is therefore difficult to describe in terms of exact or approximate distributions, and is included here as an intuitive example where empirical distributions are useful.

6 Permutation versus exact and asymptotic test

If we can derive an exact distribution of test statistic under the null hypothesis then we can use this to determine the level of statistical significance for the observed test statistic. The advantage of this is that it is computationally fast and that it works better if the sample size (i.e. n number of observation) is small. However, many of the test statistics are derived based on an asymptotic distribution. If the sample size is small the asymptotic formula's used to calculate the p-value may not be correct. In this case a different approach could be to find the p-value using a permutation method.

A drawback of the permutation method is that it is hard to demonstrate very low p-values. Showing that a p-value is lower than 10^{-7} for example, needs at least 10^7 permutations. Often if the sample size is small, the total number of permutations is not large enough to attain very low significance levels.

The manner of which we permute the data is not arbitrary, but depends on the nature of the null hypothesis being tested. Goeman and Bühlmann [2007] classified the null hypotheses as either *self-contained* or *competitive*.

A self-contained null hypothesis assumes that the marker, or set of markers, is not associated to the phenotypic trait, or has an effect without comparison to other markers of sets. I.e. the similarity between observations and genetics is incidental. To obtain an empirical distribution of the test statistic under a self-contained null hypothesis, we can shuffle the observations thus breaking the link between observations and genetics. This can be referred to as a subject-randomisation approach [Goeman and Bühlmann, 2007], but we refer to it as a 'permutation' approach. However, if using models with multiple random effects (such as M_{Set}, eq. 3), where the association between only one of the random effects is in question, shuffling the observations would break the link for all random effects, rendering the permutations useless. In this case, care should be taken to permute the link between the observations and the random effect in question.

A competitive null hypothesis assumes that the marker, or set of markers, is not *more* associated than any other marker or set of markers. An empirical distribution for a competitive null hypothesis is then obtained by sampling random sets of markers. However, all parameters that might influence the test statistic must be the same. I.e. if the number of markers influence the test statistic, the same number of markers must be sampled repetitively to form the random sets. And if there is an inherent structure between the markers in the set, this structure should be present for the random sets.

References

Erika Cule, Paolo Vineis, and Maria De Iorio. Significance testing in ridge regression for genetic data. *BMC Bioinformatics*, 12(1):372, 2011. ISSN 1471-2105. doi: 10.1186/1471-2105-12-372. URL http://www.biomedcentral.com/1471-2105/12/372.

David A Freedman. How Can the Score Test Be Inconsistent? *Am. Stat.*, 61(4):291–295, November 2007. ISSN 0003-1305. doi: 10.1198/000313007X243061. URL http://dx.doi.org/10.1198/000313007X243061.

J. J. Goeman, S. a. van de Geer, F. de Kort, and H. C. van Houwelingen. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93–99, December 2004. ISSN 1367-4803. doi: 10.1093/bioinformatics/btg382. URL http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btg382.

Jelle J. Goeman and Peter Bühlmann. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987, February 2007. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btm051. URL http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btm051.

Jelle J. Goeman, Sara a. van de Geer, and Hans C. van Houwelingen. Testing against a high dimensional alternative. *J. R. Stat. Soc. Ser. B (Statistical Methodol.*, 68(3):477–493, June 2006. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2006.00551.x. URL http://doi.wiley.com/10.1111/j.1467-9868.2006.00551.x.

Yen-Tsung Huang and Xihong Lin. Gene set analysis using variance component tests. *BMC Bioinformatics*, 14(1):210, January 2013. ISSN 1471-2105. doi: 10.1186/1471-2105-14-210. URL http://www.biomedcentral.com/1471-2105/14/210.

Just Jensen, Esa A. Mantysaari, Per Madsen, and Robin Thompson. Residual Maximum likelihood Estimation of (Co) Variance Components in Multivariate Mixed Linear Models Using Average Information. *J. Indian Soc. Agr. Stat.*, 49:215–236, 1997. URL http://isas.org.in/jisas/jsp/abstract.jsp?title=Residual.

D.L. Johnson and Robin Thompson. Restricted Maximum Likelihood Estimation of Variance Components for Univariate Animal Models Using Sparse Matrix Techniques and Average Information. *J. Dairy Sci.*, 78(2):449–456, February 1995. ISSN 00220302. doi: 10.3168/jds.S0022-0302(95)76654-1. URL http://www.journalofdairyscience.org/article/S0022-0302(95)76654-1/abstract.

Henryk Maciejewski. Gene set analysis methods: statistical models and methodological differences. *Brief. Bioinform.*, pages bbt002–, February 2013. ISSN 1477-4054. doi: 10.1093/bib/bbt002. URL http://bib.oxfordjournals.org/content/early/2013/02/09/bib.bbt002.full.

Per Madsen, Just Jensen, and Robin Thompson. Estimation of (co)variance components by REML in multivariate mixed linear models using average of observed and expected information. In *5th WCGALP*, pages 455–462, Guelph, Canada, 1994.

Michael A. Newton, Fernando A. Quintana, Johan A. den Boon, Srikumar Sengupta, and Paul Ahlquist. Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. *Ann. Appl. Stat.*, 1(1):85–106, June 2007. ISSN 1932-6157. doi: 10.1214/07-AOAS104. URL http://projecteuclid.org/euclid.aoas/1183143730.

C. Radhakrishna Rao. Rao score test. *Scholarpedia*, 4(10):8220, 2009. doi: 10.4249/scholarpedia.8220. Revision #121946.

Isabelle Rivals, Léon Personnaz, Lieng Taing, and Marie-Claude Potier. Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics*, 23(4):401–7, February 2007. ISSN 1367-4811. doi: 10.1093/bioinformatics/btl633. URL http://www.ncbi.nlm.nih.gov/pubmed/17182697.

Steven G. Self and Kung-Yee Liang. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Am. Stat. Assoc.*, 82(398):pp. 605–610, 1987. ISSN 01621459. URL http://www.jstor.org/stable/2289471.

Daniel Sorensen and Daniel Gianola. *Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics*. Springer, 2002. ISBN 0-387-95440-6.

Dennis D. Wackerly, William Mendenhall, III, and Richard L. Scheaffer. *Mathematical Statistics with Applications*. Duxbury Press, Belmont, 5 edition, 1996. ISBN 0-534-20916-5.

Xuefeng Wang, Nathan J. Morris, Xiaofeng Zhu, and Robert C. Elston. A variance component based multimarker association test using family and unrelated data. *BMC Genet.*, 14(1):17, January 2013. ISSN 1471-2156. doi: 10.1186/1471-2156-14-17. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3614458&tool=pmcentrez&rendertype=abstract.

Michael C Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics*, 89(1):82–93, July 2011. ISSN 1537-6605. URL http://www.cell.com/AJHG/fulltext/S0002-9297(11)00222-9.