P-value modeling

Renée X. de Menezes
Mark van de Wiel

January 8, 2018

Background

Suppose we have data from $i = 1, \ldots, p$ genes for $j = 1, \ldots, n$ individuals. For the individuals several covariates (e.g. age, case/control label, etc) are known and we are interested in which genes are associated to those covariates. For that the expression of a gene (a relative quantity that is proportional of the the number of copies of a gene) is measured and denoted by Y_{ij}. A simple gene-wise (i) linear regression model then is:

$$Y_{ij} = \beta_{0i} + \sum_{k=1}^{K} \beta_{ki} X_{jk}, \tag{1}$$

where X_{jk} is the value of the kth covariate for individual j.

For simplicity assume the case $K = 1$, so with one covariate. When fitting many (independent) linear regressions, the effect of the covariate can be summarised by the t-test statistic corresponding to the covariate's regression coefficient. A graph of the sorted p-values yielded by this t-test is then a useful way of starting to look at regression results. Indeed, if none of the 2×10^4 genes has expression associated with the covariate, each of the $\times 10^4$ t-test statistics will follow a Student-t distribution with $n - 2$ degrees of freedom, where n is the number of gene expressions observed. In such cases, the null hypothesis of the regression coefficient being zero is true in all cases, and each of the p-values follows a uniform distribution in $[0, 1]$. So, the sorted p-values should approximately follow a straight line with intercept 0 and slope 1.

If a proportion π_1 of the 2×10^4 genes does display an association with the covariate, then the sorted p-values will involve a mixture of those $(1 - \pi_1)2 \times 10^4$ that are generated under the null hypothesis, and those $\pi_1 2 \times 10^4$ that are

not. If the model represents the data well in all cases, the sorted p-values will follow a curve that is dominated by the straight line with intercept 0 and slope 1, since they will involve more small p-values than would have been observed under the null hypothesis only. So, results of the 2×10^4 regression fits are reflected in the empirical p-values distribution.

It turns out that model misspecifications are also reflected in the empirical p-values distribution, although the precise way in which they do so, and the impact this has on multiple testing correction, has not yet been fully worked out. Some examples of typical p-value plots we see in practice are given in Figure 1.

nder H_0, a p-value p follows a uniform distribution on $[0, 1]$, which is a special case of a Beta distribution with parameters $\alpha = 1$ and $\beta = 1$. If a gene's expression is known to affect the response, then H_0 does not hold and the distribution of p is modeled by a Beta distribution, for example with parameters $\alpha = 1$ and $\beta = 4$. In practice, it is not known if the p-value is generated under H_0 or not: if model (1) is misspecified, it may lead to a non-uniform distribution. If no model misspecification or assumptions violation take place, each p-value has its distribution represented by a mixture such as

$$p \sim (1 - \pi_1)\mathcal{U}[0, 1] + \pi_1\mathcal{B}(\alpha, \beta), \tag{2}$$

where $\mathcal{U}[0, 1]$ represents the uniform distribution and $\mathcal{B}(\alpha, \beta)$ represents a Beta distribution parameterized by $\alpha, \beta > 0$. Typical values for the parameters in this case are $\alpha = 1$ and $\beta > 1$.

Model misspecifications and assumption violations will affect these distributions. Considering first that H_0 is true, it may be that the distribution of p no longer is $\mathcal{U}[0, 1]$, but rather larger p-values are more likely than smaller ones, for example when a covariate is missing from the model. Such cases may lead to a p-value distribution similar to a Beta with parameters $\alpha = 4$ and $\beta = 1$, say. Also, over-fitting due to outliers can also affect this null distribution.

Subsequently, the mixture needs to be fitted to the data (a list of p-values in this case), and we should be able to decide on the basis of this result whether or not the mixture distribution model with departure represents the data better than the simpler mixture distribution as given above.

Specific objectives

- Put together an algorithm that can be used to model and fit the distribution of a list of p-values. This will involve expressing the p-value

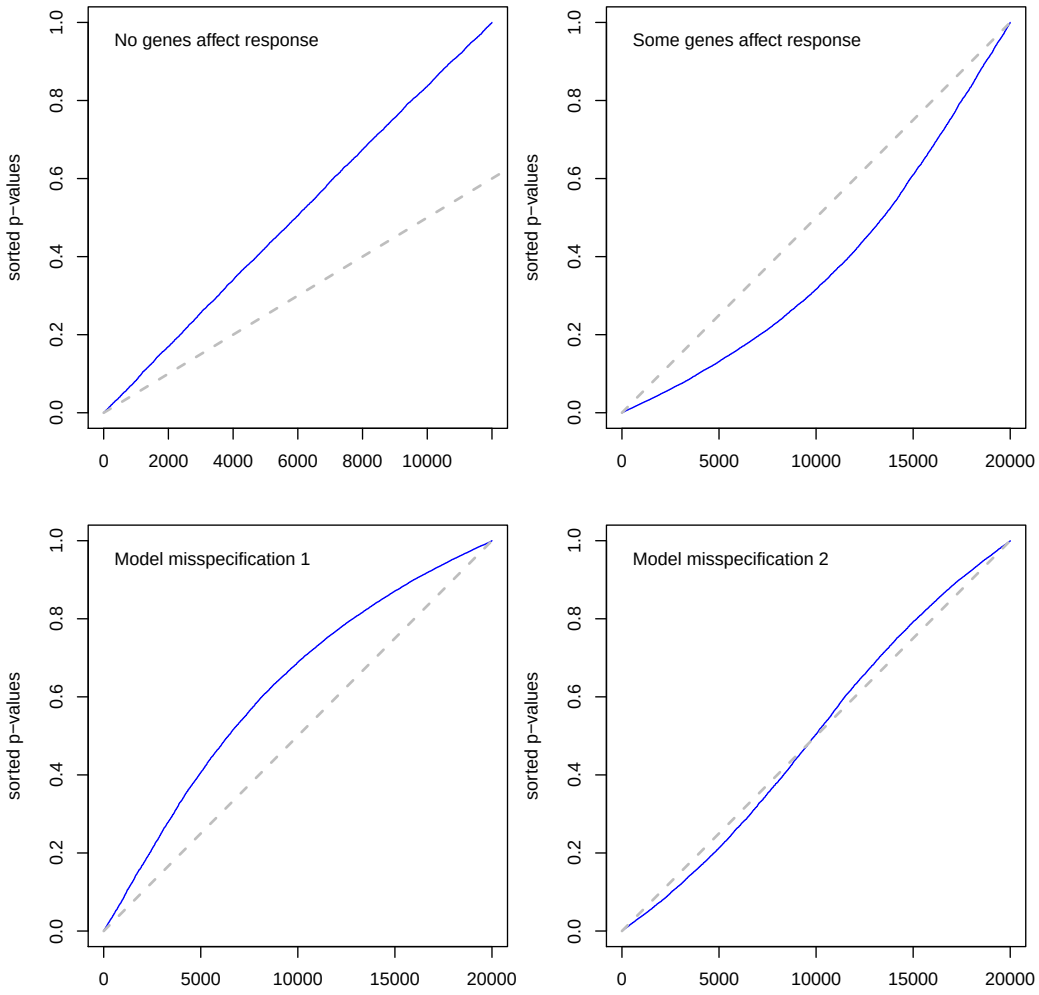

Figure 1: Examples of p-value plots that can be made when examining the covariate effect for many linear regressions.

distribution as a mixture distribution with one component corresponding to the test results under H_0, and one component corresponding to the test results under the alternative. Apply the algorithm for all four lists of p-values.

- Assess whether mixture model (2) suffices to fit the p-value data
- For those p-value lists for which model (2) does not suffice: propose extensions of the model that might fit better
- (Optional) Consider models with more than two mixture components. Can we assess how many mixture components we need to correctly fit the p-value data?

Distribution of p-values under no effect

Let S be a test statistic for a certain null hypothesis H_0, and let $F_S(s) \equiv P\{S \leq s\}$ represent its cumulative distribution function (cdf) under H_0. Assume for simplicity that the alternative hypothesis leads to a one-sided test, so that the rejection region is of the form $S > s_0$. Then after a value s is observed, the p-value for the test can be obtained via $p = P\{S > s\} = 1 - F_S(s)$. Of course, if s is not observed it can be also considered to be random, so that $P = P\{S > s\}$ is also a random variable. We can compute its cdf in the following way:

$$\begin{aligned} F_P(p) &= P\{P \leq p\} \\ &= P\{1 - F_S(s) \leq p\} \\ &= P\{F_S(s) \geq 1 - p\} \\ &= 1 - P\{F_S(s) < 1 - p\} \\ &= 1 - P\{S < F_S^{-1}(1-p)\} \\ &= 1 - F_S\left[F_S^{-1}(1-p)\right] \\ &= 1 - (1-p) \\ &= p. \end{aligned}$$

Thus, if the test statistic S does follow the null hypothesis H_0, the distribution of the test's p-value is uniform in $[0, 1]$.