1. FUNCTION TO EXTRACT P-VALUES FROM DGLM

Here we give an R function `dglm.Pvalues(.)` that extracts the P-values for a fitted model `dglm.fit`.

```r
# function to extract p-values from dglm

dglm.Pvalues <- function(dglm.fit){
  P.mean = summary(dglm.fit)$coef[2,4]
  list(P.mean=P.mean, P.disp=P.disp)
}
```

Below we illustrate the use of the function with a simple example where the simulated genetic effects are set to 0.

```r
#An example using dglm.Pvalues(.) with 200 simulated observations
set.seed(123)
require(dglm)

n = 200
a <- rbinom(n, 1, 0.5)  ##Covariate for additive genetic effects
sex <- rbinom(n, 1, 0.5)  ##Covariate for a non-genetic effect
add.effect = 0
sex.effect = 1

res.var = exp( a*add.effect )  ##Residual variance
y <- 10 + a*add.effect + sex*sex.effect + rnorm(n ,0 , sqrt(res.var))

##The additive genetic effect must be given first in the following formula

d.fit <- dglm( formula = y ~ a + sex, dformula = ~ a )
P.values <- dglm.Pvalues( d.fit )
print( P.values )
```

This code gives the following output in R:

```r
$P.mean
```
2. DOUBLE GLM FOR NON-NORMAL TRAITS

For non-normal distributed traits a GLM can be used to model the mean-controlling QTL and the deviance components from this GLM are subsequently used in eq 3 (in the paper) instead of the squared raw residuals $\varepsilon_i^2$ (Smyth 2002; Lee et al. 2006). In the more general GLM setting, the dispersion term $\phi$ is modeled in the second submodel of the double GLM. For the exponential family of distributions the variance of $y$ can be written as (McCullagh and Nelder 1989):

$$Var(y) = \phi V(\mu)$$

where $V(\mu)$ is the “variance function”, which gives the relationship between the variance of $y$ and the mean $\mu$. For the normal distribution $V(\mu) = 1$ (i.e. the variance does not vary with the mean) and the dispersion term is the residual variance, $\phi \equiv \sigma^2$. For the gamma distribution $V(\mu) = \mu^2$ and since $Var(y)/V(\mu) = \phi$ we have that $\phi \equiv CV^2$. Consequently, we can model the coefficient of variation using a gamma GLM (see chapter 8 in (McCullagh and Nelder 1989)). This gives a method to fit CV-controlling QTL. Ansel et al. (2008) referred to such QTL in a gene expression study as “noise QTL”. For these calculations the dglm package fits a digamma distribution (Smyth 1989).

3. THE EFFECT OF UNCERTAINTY IN GENOTYPE STATE ON THE INference OF VQTLs

In a constant variance model, the sampling distribution of phenotype $y_i$ of animal $i$ given the genotype $x_i = g_j$ at the QTL locus is $p(y_i|x_i = g_j) = N(\alpha, \sigma^2)$ with

$$E(y_i|x_i = g_j) = \alpha_j \quad \text{and} \quad Var(y_i|x_i = g_j) = \sigma^2$$
or with $x_i$ representing an indicator vector for the genotype and $\alpha$ the genotype effects,

$$E(y_i|x_i) = x_i^T \alpha \quad \text{and} \quad Var(y_i|x_i) = \sigma^2$$

**The constant variance model with uncertain genotype** When genotype is uncertain, such that $p(x_i = g_j|\text{Data}) = p_{ij}$ and $p_i = [p_{i1}, p_{i2}, ...]$, then the sampling distribution conditional on $p_i$ is the mixture

$$p(y_i|p_i) = \sum_j p_{ij} p(y_i|x_i = g_j)$$

with $p(y_i|x_i = g_j)$ defined as above. Conditional on $p_i$ the expectation is $E(y_i|p_i) = p_i^T \alpha$, much as before, but the variance is now

$$Var(y_i|p_i) = \text{Var}_g(E(y_i|g_j)|p_i) + \text{Var}_g(\text{Var}(y_i|g_j)|p_i)$$

$$= \left[ E_g(E(y_i^2|g_j)) - E_g^2(E(y_i|g_j)) \right] + E_g(\text{Var}(y_i|g_j))$$

$$= \left[ \sum_j p_{ij} \alpha_{ij}^2 - \left( \sum_j p_{ij} \alpha_{ij} \right)^2 \right] + \sigma^2$$

$$= (p_i^T \alpha^2) - (p_i^T \alpha)^2 + \sigma^2$$

$$v_i = \alpha_{\text{extra}}^2 + \sigma^2.$$

A quick alternative to fitting the mixture by maximum likelihood is the regression model used for Haley-Knott or HAPPY where it is assumed

$$p_{HK}(y_i|p_i) = N(p_i^T \alpha^2, \sigma^2)$$

or, as described by BROMAN and SEN (2009), an “extended Haley-Knott” regression fitted by ML to provide a closer approximation to the mixture model

$$p_{EHK}(y_i|p_i) = N(p_i^T \alpha^2, v_i)$$

which is slower than the HK model but faster than a full ML fit.
The vQTL model with known genotype. In the double GLM model, the sampling distribution of phenotype $y_i$ of animal $i$ given the genotype $x_i = g_j$ at the QTL locus is

$$p(y_i|g_j) = N(\alpha_j, \exp(\theta_j)),$$

where $\alpha_j$ and $\theta_j$ are specific to genotype $g_j$. The expectation and variance of $y_i$ are

$$E(y_i|x_i = g_j) = \alpha_j \quad \text{and} \quad Var(y_i|x_i = g_j) = \exp(\theta_j)$$

The vQTL model with uncertain genotype. With uncertain genotype, the mean is the same as the constant variance model, $E(y_i|\mathbf{p}_i) = \mathbf{p}_i^T \mathbf{\alpha}$, but the variance is

$$Var(y_i|\mathbf{p}_i) = Var_y(E(y_i|g_j)|\mathbf{p}_i) + E_y(Var(y_i|g_j)|\mathbf{p}_i)$$

$$= [\mathbf{p}_i^T(\mathbf{\alpha}^2) - (\mathbf{p}_i^T \mathbf{\alpha})^2] + \mathbf{p}_i^T \exp(\mathbf{\theta})$$

$$v_i = v_i^{\text{extra}} + v_i^{\text{vQTL}}$$

Since we are interested in how $x_i$ affects the variance, our only measure of $x_i$ is $\mathbf{p}_i$ and since $v_i^{\text{extra}} = f(\mathbf{p}_i, \mathbf{\alpha})$ such that $v_i$ can potentially change with $x_i$ in the absence of a vQTL effect, there is the potential for genotype uncertainty to mimic the effect of a vQTL. For low informative markers the dglm approach captures only part of the vQTL effect, as $Var(y_i|\mathbf{p}_i) = \exp(\mathbf{p}_i^T \mathbf{\theta})$, so that

$$v_i = v_i^{\text{vQTL}} + v_i^{\text{extra}} + (\mathbf{p}_i^T \exp(\mathbf{\theta}) - \exp(\mathbf{p}_i^T \mathbf{\theta}))$$

A Taylor expansion shows that the last term is small for small to moderate vQTL effects ($\theta << 1$). This term does not depend on the effects in the mean model (i.e. $\mathbf{\alpha}$). Its contribution to $v_i$ should only reduce the power of the model, whereas the term $v_i^{\text{extra}}$ may increase the Type I error.
4. TRANSFORMATION AND ASSESSMENT OF MEAN-VARIANCE RELATIONSHIP

Maximum likelihood can be used to estimate the parameter $\lambda$ in a Box-Cox transformation (Box and Cox 1964) for the linear model $z = X\beta + \varepsilon$ where the response is

$$z = \begin{cases} \frac{y^{\lambda} - 1}{\lambda} & \lambda \neq 0 \\ \log(y) & \lambda = 0 \end{cases}$$

and the residuals $\varepsilon$ are assumed normal i.i.d. The function `boxcox(.)` in the MASS library in R can be used for this estimation. We can thereby obtain a response $z$ that is approximately normal. See Pawitan (2001, page 178) for a description of the method. For the case of a pleiotropic gene controlling both the mean and the variance of a trait, the Box-Cox transformation of the raw data might be over-conservative by removing large parts of the vQTL effects. A potential remedy to this problem could be to include the Box-Cox parameter in the likelihood and to estimate it simultaneously with all other parameters in our model. A similar solution for mixed effect models has previously been developed using Bayesian techniques (Yang 2010) but we do not investigate this possibility in our current paper.

For GLM distributions in general, the concept of extended quasi-likelihood (Nelder and Pregibon 1987) can be used to assess the mean-variance relationship. Using e.g. the `eql(.)` function in the EQL package in R, the parameter $\psi$ for the relationship $\text{Var}(y) = \phi \mu^\psi$ with $\phi$ being a constant dispersion parameter. So that for a normal distribution $\psi = 0$, for a Poisson distribution $\psi = 1$, and for a gamma distribution $\psi = 2$. See Pawitan 2001 for a recent and clear description of the method.

Below we investigate the use of Box-Cox transformation to remove false vQTL from non-normal data and thereafter specify guidelines to avoid detecting false vQTL due to scale effects.
Using Box-Cox transformation to remove false vQTL from non-normal data. We simulated data with a homoscedastic (ie, constant) variance as

\[ y \sim N(\mu + \alpha q, 1) \]

setting \( \mu = 1 \), \( \alpha = 10 \) and fixing \( q \) as a vector of 50 zeros and 50 ones to give 100 observations in total. For each condition described below we simulated 20 replicates. For each simulated \( y \) we obtained four versions of an “observed phenotype” \( w \). We obtained these versions by applying four alternative transformations that set \( w \) to \( y \) (ie, the identity transformation), \( y^2 \), \( \sqrt{y} \), or \( \frac{1}{y} \) (as listed in Table A1 below).

The Box-Cox parameter \( \lambda \) was estimated by means of MLE (see, eg, page 178 in Pawitan 2001) for the model

\[
\frac{w_i^\lambda - 1}{\lambda} = \mu + \alpha q_i + \varepsilon_i
\]

where residuals are normally distributed and \( \sigma^2 \) is constant (i.e. for a model with homoscedastic residual variance). Two models were fitted using \textit{dglm} (Dunn and Smyth 2009):

\[
w_i = \mu + \alpha q_i + \varepsilon_i
\]

\[
\log(\sigma^2_i) = m + \theta q_i
\]

and

\[
\frac{w_i^\lambda - 1}{\lambda} = \mu + \alpha q_i + \varepsilon_i
\]

\[
\log(\sigma^2_i) = m + \theta q_i
\]

where \( \mu \) and \( \alpha \) is an intercept term and regression coefficient for the mean model, and \( m \) and \( \theta \) is an intercept term and regression coefficient for the variance model.
Table A1: Estimated Box-Cox parameter $\lambda$ and estimated regression parameter $\theta$ in the variance part of the model (standard errors within brackets)

<table>
<thead>
<tr>
<th>$w$</th>
<th>$\hat{\lambda}$</th>
<th>$\hat{\theta}$ with response $w$</th>
<th>$\hat{\theta}$ with response $\frac{w^\lambda-1}{\lambda}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>0.96 (0.015)</td>
<td>0.113 (0.047)</td>
<td>-0.006 (0.018)</td>
</tr>
<tr>
<td>$y^2$</td>
<td>0.48 (0.007)</td>
<td>2.92 (0.074)</td>
<td>-0.006 (0.018)</td>
</tr>
<tr>
<td>$\sqrt{y}$</td>
<td>1.92 (0.031)</td>
<td>-1.39 (0.039)</td>
<td>-0.006 (0.018)</td>
</tr>
<tr>
<td>$\frac{1}{y}$</td>
<td>-0.96 (0.015)</td>
<td>-6.47 (0.045)</td>
<td>-0.006 (0.018)</td>
</tr>
</tbody>
</table>

For $y$, generated from normal random numbers with a homogeneous variance, the average estimate of $\theta$ over the 20 replicates was close to 0 as expected ($\hat{\theta} = 0.113$). By chance the samples of $y$ were slightly skewed and the average estimate of $\lambda$ was 0.96 with $\hat{\theta} = -0.006$ after Box-Cox transformation. Note that the average estimate of $\theta$ was -0.006 for all four cases and that the risk of detecting false vQTL due to scale effects therefore can be avoided using Box-Cox transformation. Note also that all three transformations gave estimates of $\theta$ significantly different from 0, implying that we can expect non-normal data to give false vQTL.

To assess the risk of Box-Cox transformation removing true vQTL, we simulated data with heteroscedastic variances:

$$y \sim N(\mu + \alpha q, \log(m + \theta q_d))$$

where $\mu = 1$, $\alpha = 10$, $m = 0$, and $\theta$ was set to either 1 or -1. As above, $q$ was a vector of 50 zeros and 50 ones. We studied two cases for each simulated value of $\theta$:

- Case 1: $q_d = q$
- Case 2: $q_d$ and $q$ are uncorrelated

The Box-Cox parameter $\lambda$ was estimated by means of MLE as described above, and two
models were fitted using \textit{dglm}

\[ y_i = \mu + \alpha q_i + \varepsilon_i \]
\[ \frac{y_i^\lambda - 1}{\lambda} = \mu + \alpha q_i + \varepsilon \]
\[ \log(\sigma_i^2) = m + \theta q_{d,i} \]

From the results for Case 1 ($q_d = q$) in Table A2, we can see that Box-Cox transformation removes the vQTL effects if the covariates in the mean and variance parts of the model are highly correlated, i.e. if the vQTL is located at, or very close to, an ordinary QTL. From Case 2, we can see that Box-Cox transformation does not remove the vQTL effect if the vQTL and the ordinary QTL are not linked.

Table A2: Estimated Box-Cox parameter $\lambda$ and estimated regression parameter $\theta$ in the variance part of the model (standard errors within brackets)

<table>
<thead>
<tr>
<th>Simulated $\theta$</th>
<th>Simulated case</th>
<th>$\hat{\lambda}$</th>
<th>$\hat{\theta}$ for response $y$</th>
<th>$\hat{\theta}$ for response $\frac{y^\lambda - 1}{\lambda}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$q_d = q$</td>
<td>0.65 (0.013)</td>
<td>1.11 (0.047)</td>
<td>0.070 (0.018)</td>
</tr>
<tr>
<td>-1</td>
<td>$q_d = q$</td>
<td>1.28 (0.018)</td>
<td>-0.89 (0.047)</td>
<td>-0.073 (0.020)</td>
</tr>
<tr>
<td>1</td>
<td>$q_d \neq q$</td>
<td>0.98 (0.023)</td>
<td>1.03 (0.060)</td>
<td>1.00 (0.065)</td>
</tr>
<tr>
<td>-1</td>
<td>$q_d \neq q$</td>
<td>0.95 (0.022)</td>
<td>-1.00 (0.061)</td>
<td>-0.95 (0.059)</td>
</tr>
</tbody>
</table>

Guidelines for fitting vQTL when the data generation process is not known or expected to be approximately normal.

1. Fit ordinary QTL with a homoscedastic linear model and include ordinary QTL with large effects in this preliminary linear model.

2. Do a QQplot of the residuals and estimate the Box-Cox transformation parameter $\lambda$ for the preliminary linear model.
3. If the QQplot is satisfactory and $\lambda$ is close to 1 then perform dglm to detect both ordinary QTL and vQTL.

4. If the QQplot is not satisfactory or $\lambda$ is not close to 1. Use the Box-Cox transformed response in a dglm to detect both QTL and vQTL.

Guidelines for fitting vQTL when the data is expected to have been generated by a non-normal GLM, e.g. binomial, Poisson or gamma.

1. Fit ordinary QTL with a GLM and include QTL with large effects in a preliminary linear model.

2. Perform GLM diagnostics plots, including QQplot of for the standardized deviance residuals, and estimate the EQL parameter $\psi$.

3. If the diagnostics plots are satisfactory and the estimate of $\psi$ is close to the assumed variance function for $\mu^\psi$ then use dglm to detect both ordinary and vQTL.

4. If the diagnostics plots are not satisfactory, or the EQL estimate of $\psi$ is not close to the expected value, then compare different distributions and link functions prior to the DGLM scan.

The interpretation of vQTL varies with different GLM distributions. Generally, vQTL are effects explaining differences in the dispersion parameter of the GLM, such that for a gamma distribution a vQTL explains differences in CV (coefficient of variation), whereas for a Poisson distribution a vQTL explains differences in the ratio of variance/mean.

LITERATURE CITED


