



## Fitting generalized linear mixed-effects models using lme4

Steven C. Walker  
McMaster University

Rune Haubo Bojesen Christensen  
Technical University of Denmark

Douglas Bates  
University of Wisconsin - Madison

Ben Bolker  
McMaster University

Martin Mächler  
ETH Zurich

---

### Abstract

*abstract goes here*

*Keywords:* sparse matrix methods, linear mixed models, penalized least squares, Cholesky decomposition.

---

## 1. Introduction

*equations are all defined in part I ...*

## 2. Generalized Linear Mixed Models

The generalized linear mixed models (GLMMs) that can be fit by the **lme4** package preserve the multivariate Gaussian unconditional distribution of the random effects,  $\mathcal{B}$  (eqn. ??). Because most families used for the conditional distribution,  $\mathcal{Y}|\mathcal{B} = \mathbf{b}$ , do not incorporate a separate scale factor,  $\sigma$ , we remove it from the definition of  $\Sigma$  and from the distribution of the spherical random effects,  $\mathcal{U}$ . That is,

$$\mathcal{U} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_q) \tag{1}$$

and

$$\Sigma_\theta = \Lambda_\theta \Lambda_\theta'.$$
 (2)

The conditional distributions,  $\mathcal{Y}|\mathcal{B} = \mathbf{b}$  and  $\mathcal{Y}|\mathcal{U} = \mathbf{u}$ , preserve the properties that the components of  $\mathcal{Y}$  are conditionally independent and that the mean,  $\mu_{\mathcal{Y}|\mathcal{U}=\mathbf{u}}$ , depends on  $\mathbf{u}$  only through the linear predictor,

$$\boldsymbol{\eta} = \mathbf{Z}\Lambda_\theta\mathbf{u} + \mathbf{X}\boldsymbol{\beta}.$$
 (3)

The mapping from  $\mu_{\mathcal{Y}|\mathcal{U}=\mathbf{u}}$  to  $\boldsymbol{\eta}$ , which is called the *link function* and written

$$\mathbf{Z}\Lambda_\theta\mathbf{u} + \mathbf{X}\boldsymbol{\beta} = \boldsymbol{\eta} = \mathbf{g}(\mu_{\mathcal{Y}|\mathcal{U}=\mathbf{u}}),$$
 (4)

is a *diagonal mapping* in the sense that there is a scalar function,  $g$ , such that the  $i$ th component of  $\boldsymbol{\eta}$  is  $g$  applied to the  $i$ th component of  $\mu_{\mathcal{Y}|\mathcal{U}=\mathbf{u}}$ . (The name “diagonal” reflects the fact that the Jacobian matrix,  $\frac{d\boldsymbol{\eta}}{d\boldsymbol{\mu}'}$ , of such a mapping will be diagonal.)

The scalar link function must be invertible over its range. The vector-valued *inverse link* function,  $\mathbf{g}^{-1}$ , will be the scalar inverse link,  $g^{-1}$ , applied component-wise to  $\boldsymbol{\eta}$ .

Common forms of the conditional distribution are Bernoulli, for binary responses, binomial for binary responses that are recorded as the number of trials and the number of successes, and Poisson, for count data. The combination of a distributional form and a link function is called a *family*. For distributional forms in the exponential family there is a *canonical link*. For Bernoulli or binomial forms the canonical link is the *logit* link function

$$\eta_i = \log\left(\frac{\mu_i}{1 - \mu_i}\right);$$
 (5)

for the Poisson distribution the canonical link is the natural logarithm.

The form of the distribution determines the conditional variance,  $\text{Var}(\mathcal{Y}|\mathcal{U} = \mathbf{u})$ , as a function of the conditional mean and, possibly, a separate scale factor. (In the most common cases the conditional variance is completely determined by the conditional mean.)

The likelihood of the parameters, given the observed data, is now

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{y}_{\text{obs}}) = \int_{\mathbb{R}^q} f_{\mathcal{Y},\mathcal{U}}(\mathbf{y}_{\text{obs}}, \mathbf{u}) d\mathbf{u}$$
 (6)

where, as in the case of linear mixed models,  $f_{\mathcal{Y},\mathcal{U}}(\mathbf{y}_{\text{obs}}, \mathbf{u})$  is the unscaled conditional density of  $\mathcal{U}$  given  $\mathcal{Y} = \mathbf{y}_{\text{obs}}$ . The notation here is a bit blurred because, although the joint distribution of  $\mathcal{Y}$  and  $\mathcal{U}$  is always continuous with respect to  $\mathcal{U}$ , it can be (and often is) discrete with respect to  $\mathcal{Y}$ . However, when we condition on the observed value  $\mathcal{Y} = \mathbf{y}_{\text{obs}}$ , the resulting function is continuous with respect to  $\mathbf{u}$  so the unscaled conditional density is indeed well-defined as a density, up to a scale factor.

To evaluate the integrand in (6) we use the value of the `dev.resids` function in the GLM family. This vector,  $\mathbf{d}(\mathbf{y}_{\text{obs}}, \mathbf{u})$ , with elements,  $d_i(\mathbf{y}_{\text{obs}}, \mathbf{u})$ ,  $i = 1, \dots, n$ , provides the deviance of a generalized linear model as

$$\sum_{i=1}^n d_i(\mathbf{y}_{\text{obs}}, \mathbf{u}).$$

(We should note that there some confusion in R (and in its predecessor, S) about what exactly the deviance residuals for a family are. As indicated above, we will use this name for the

value of the `dev.resids` function in the family. The signed square root of this vector, using the signs of  $\mathbf{y}_{\text{obs}} - \mu$ , is returned from `residuals` applied to a fitted model of class "glm" when `type="deviance"`, the default, is specified. Both are called “deviance residuals” in the documentation but, although they are related, they are not the same.)

The likelihood can now be expressed as

$$L(\boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{y}_{\text{obs}}) = \int_{\mathbb{R}^q} \exp \left( - \frac{\sum_{i=1}^n d_i(\mathbf{y}_{\text{obs}}, \mathbf{u}) + \|\mathbf{u}\|^2}{2} \right) (2\pi)^{-q/2} d\mathbf{u} \quad (7)$$

As for linear mixed models, we simplify evaluation of the integral (6) by determining the value,  $\tilde{\mathbf{u}}_{\beta, \theta}$ , that maximizes the integrand. When the conditional density,  $\mathcal{U} | \mathcal{Y} = \mathbf{y}_{\text{obs}}$ , is multivariate Gaussian, this conditional mode will also be the conditional mean. However, for most families used in GLMMs, the mode and the mean need not coincide so we use a more general term and call  $\tilde{\mathbf{u}}_{\beta, \theta}$  the *conditional mode*. We first describe the numerical methods for determining the conditional mode using the Penalized Iteratively Reweighted Least Squares (PIRLS) algorithm then return to the question of evaluating the integral (6).

## 2.1. Determining the conditional mode

The iteratively reweighted least squares (IRLS) algorithm is an efficient method of determining the maximum likelihood estimates of the coefficients in a generalized linear model. We extend it to a *penalized iteratively reweighted least squares* (PIRLS) algorithm for determining the conditional mode,  $\tilde{\mathbf{u}}_{\beta, \theta}$ . This algorithm has the form

1. Given parameter values,  $\boldsymbol{\beta}$  and  $\boldsymbol{\theta}$ , and starting estimates,  $\mathbf{u}_0$ , evaluate the linear predictor,  $\boldsymbol{\eta}$ , the corresponding conditional mean,  $\boldsymbol{\mu}_{\mathcal{Y}|\mathcal{U}=\mathbf{u}}$ , and the conditional variance. Establish the weights as the inverse of the variance. We write these weights in the form of a diagonal weight matrix,  $\mathbf{W}$ , although they are stored and manipulated as a vector.
2. Solve the penalized, weighted, nonlinear least squares problem

$$\arg \min_{\mathbf{u}} \left( \left\| \mathbf{W}^{1/2} (\mathbf{y}_{\text{obs}} - \boldsymbol{\mu}_{\mathcal{Y}|\mathcal{U}=\mathbf{u}}) \right\|^2 + \|\mathbf{u}\|^2 \right) \quad (8)$$

3. Update the weights,  $\mathbf{W}$ , and check for convergence. If not converged, go to step 2.

We use a Gauss-Newton algorithm with an orthogonality convergence criterion (Bates and Watts 1988, §2.2.3) to solve the penalized, weighted, nonlinear least squares problem in step 2. At the  $i$ th iteration we determine an increment,  $\boldsymbol{\delta}_i$ , as the solution to the penalized, weighted, linear least squares problem

$$\boldsymbol{\delta}_i = \arg \min_{\boldsymbol{\delta}} \left\| \begin{bmatrix} \mathbf{W}^{1/2} (\mathbf{y}_{\text{obs}} - \boldsymbol{\mu}_i) \\ \mathbf{u}_i \end{bmatrix} - \begin{bmatrix} \mathbf{W}^{1/2} \mathbf{M}_i \mathbf{Z} \boldsymbol{\Lambda}_{\boldsymbol{\theta}} \\ \mathbf{I}_q \end{bmatrix} \mathbf{u} \right\|^2 \quad (9)$$

where  $\mathbf{u}_i$  is current value of  $\mathbf{u}$ ,  $\boldsymbol{\mu}_i$  is the corresponding conditional mean of  $\mathcal{Y} | \mathcal{U} = \mathbf{u}_i$  and  $\mathbf{M}_i$  is the Jacobian matrix of the vector-valued inverse link, evaluated at  $\boldsymbol{\mu}_i$ . That is

$$\mathbf{M}_i = \left. \frac{d\boldsymbol{\mu}}{d\boldsymbol{\eta}'} \right|_{\boldsymbol{\eta}_i}, \quad (10)$$

which will be a diagonal matrix so, as for the weights, we store and manipulate the Jacobian as a vector.

The minimizer,  $\delta_i$ , of (9) satisfies

$$\mathbf{P} (\mathbf{\Lambda}'_{\theta} \mathbf{Z}' \mathbf{M}_i \mathbf{W} \mathbf{M}_i \mathbf{Z} \mathbf{\Lambda}_{\theta} + \mathbf{I}_q) \mathbf{P}' \delta_i = \mathbf{\Lambda}'_{\theta} \mathbf{Z}' \mathbf{M}_i \mathbf{W} (\mathbf{y}_{\text{obs}} - \boldsymbol{\mu}_i) - \mathbf{u}_i \quad (11)$$

which we solve using the sparse Cholesky factor. At convergence, the factor,  $\mathbf{L}_{\beta, \theta}$ , satisfies

$$\mathbf{L}_{\beta, \theta} \mathbf{L}'_{\beta, \theta} = \mathbf{P} (\mathbf{\Lambda}'_{\theta} \mathbf{Z}' \mathbf{M} \mathbf{W} \mathbf{M} \mathbf{Z} \mathbf{\Lambda}_{\theta} + \mathbf{I}_q) \mathbf{P}' \quad (12)$$

## 2.2. Evaluating the likelihood for GLMMs using the Laplace approximation

A second-order Taylor series approximation to  $-2 \log[f_{\mathcal{Y}, \mathcal{U}}(\mathbf{y}_{\text{obs}}, \mathbf{u})]$  based at  $\tilde{\mathbf{u}}$  provides an approximation of unscaled conditional density as a multiple of the density for the multivariate Gaussian  $\mathcal{N}(\tilde{\mathbf{u}}, \mathbf{L}\mathbf{L}')$ . The change of variable

$$\mathbf{u} = \tilde{\mathbf{u}} + \mathbf{L}\mathbf{z} \quad (13)$$

provides

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{y}_{\text{obs}}) &= \int_{\mathbb{R}^q} f_{\mathcal{Y}, \mathcal{U}}(\mathbf{y}_{\text{obs}}, \mathbf{u}) d\mathbf{u} \\ &\approx \tilde{f} |\mathbf{L}| \int_{\mathbb{R}^q} e^{-\|\mathbf{z}\|^2/2} (2\pi)^{-q/2} d\mathbf{z} \\ &= \tilde{f} \text{abs}(|\mathbf{L}|) \end{aligned} \quad (14)$$

or, on the deviance scale,

$$-2\ell(\boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{y}_{\text{obs}}) \approx \sum_{i=1}^n d_i(\mathbf{y}_{\text{obs}}, \tilde{\mathbf{u}}) + \|\tilde{\mathbf{u}}\|^2 + \log(|\mathbf{L}|^2) + \frac{q}{2} \log(2\pi) \quad (15)$$

### *Decomposing the deviance for simple models*

A special, but not uncommon, case is that of scalar random effects associated with levels of a single grouping factor,  $\mathbf{h}$ . In this case the dimension,  $q$ , of the random effects is the number of levels of  $\mathbf{h}$  — i.e. there is exactly one random effect associated with each level of  $\mathbf{h}$ . We will write the vector of variance-covariance parameters, which is one-dimensional, as a scalar,  $\theta$ . The matrix  $\mathbf{\Lambda}_{\theta}$  is a multiple of the identity,  $\theta \mathbf{I}_q$ , and  $\mathbf{Z}$  is the  $n \times q$  matrix of indicators of the levels of  $\mathbf{f}$ . The permutation matrix,  $\mathbf{P}$ , can be set to the identity and  $\mathbf{L}$  is diagonal, but not necessarily a multiple of the identity.

Because each element of  $\boldsymbol{\mu}$  depends on only one element of  $\mathbf{u}$  and the elements of  $\mathcal{Y}$  are conditionally independent, given  $\mathcal{U} = \mathbf{u}$ , the conditional densities of the  $u_j, j = 1, \dots, q$  given  $\mathcal{Y} = \mathbf{y}_{\text{obs}}$  are independent. We partition the indices  $1, \dots, n$  as  $\mathbb{I}_j, j = 1, \dots, q$  according to the levels of  $\mathbf{h}$ . That is, the index  $i$  is in  $\mathbb{I}_j$  if  $h_i = j$ . This partitioning also applies to the deviance residuals in that the  $i$ th deviance residual depends only on  $u_j$  when  $i \in \mathbb{I}_j$ .

Writing the univariate conditional densities as

$$f_j(\mathbf{y}_{\text{obs}}, u_j) = \exp \left( -\frac{\sum_{i \in \mathbb{I}_j} d_i(\mathbf{y}_{\text{obs}}, u_j) + u_j^2}{2} \right) (2\pi)^{-1/2} \quad (16)$$

we have

$$f_{\mathcal{Y},\mathcal{U}}(\mathbf{y}_{\text{obs}}, \mathbf{u}) = \prod_{j=1}^q f_j(\mathbf{y}_{\text{obs}}, u_j) \quad (17)$$

and

$$L(\boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{y}_{\text{obs}}) = \prod_{j=1}^q \int_{\mathbb{R}} f_j(\mathbf{y}_{\text{obs}}, u) du \quad (18)$$

We consider this special case both because it occurs frequently and because, for some software, it is the only type of GLMM that can be fit. Also, in this particular case we can graphically assess the quality of the Laplace approximation by comparing the actual integrand to its approximation.

Consider the `cbpp` data on contagious bovine pleuropneumonia incidence according to season and herd, available in the `lme4` package.

```
> str(cbpp)

'data.frame': 56 obs. of 4 variables:
 $ herd      : Factor w/ 15 levels "1","2","3","4",...: 1 1 1 1 2 2 2..
 $ incidence: num  2 3 4 0 3 1 1 8 2 0 ...
 $ size      : num  14 12 9 5 22 18 21 22 16 16 ...
 $ period    : Factor w/ 4 levels "1","2","3","4": 1 2 3 4 1 2 3 1 2..
```

and the model

```
> print(m1 <- glmer(cbind(incidence, size-incidence) ~ period + (1|herd),
+                   cbpp, binomial), corr=FALSE)

Generalized linear mixed model fit by maximum likelihood (Laplace
Approximation) [glmerMod]
Family: binomial ( logit )
Formula: cbind(incidence, size - incidence) ~ period + (1 | herd)
Data: cbpp
      AIC      BIC    logLik deviance df.resid
194.0531 204.1799 -92.0266 184.0531      51
Random effects:
Groups Name      Std.Dev.
herd (Intercept) 0.6421
Number of obs: 56, groups: herd, 15
Fixed Effects:
(Intercept)      period2      period3      period4
      -1.3983      -0.9919      -1.1282      -1.5797
```

This model has been fit by minimizing the Laplace approximation to the deviance. We can assess the quality of this approximation by evaluating the unscaled conditional density at  $u_j(z) = \tilde{u}_j + z/\mathbf{L}_{j,j}$  and comparing the ratio,  $f_j(\mathbf{y}_{\text{obs}}, u)/(\tilde{f}_j\sqrt{2\pi})$ , to the standard normal

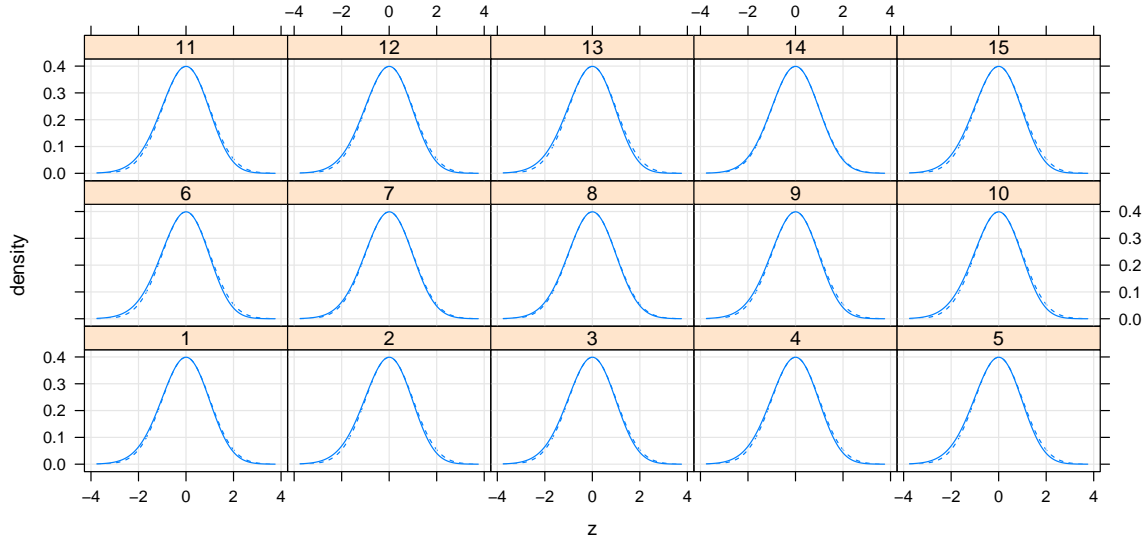


Figure 1: Comparison of univariate integrands (solid line) and standard normal density function (dashed line)

density,  $\phi(z) = e^{-z^2/2}/\sqrt{2\pi}$ , as shown in Figure 1. *consider Q-Q plots to emphasize deviations from normality?* As we can see from this figure, the univariate integrands are very close to the standard normal density, indicating that the Laplace approximation to the deviance is a good approximation in this case.

### 3. Adaptive Gauss-Hermite quadrature for GLMMs

When the integral (6) can be expressed as a product of low-dimensional integrals, we can use Gauss-Hermite quadrature to provide a closer approximation to the integral. Univariate Gauss-Hermite quadrature evaluates the integral of a function that is multiplied by a “kernel” where the kernel is a multiple of  $e^{-z^2}$  or  $e^{-z^2/2}$ . For statisticians the natural candidate is the standard normal density,  $\phi(z) = e^{-z^2/2}/\sqrt{(2\pi)}$ . A  $k$ th-order Gauss-Hermite formula provides knots,  $z_i, i = 1, \dots, k$ , and weights,  $w_i, i = 1, \dots, k$ , such that

$$\int_{\mathbb{R}} t(z)\phi(z) dz \approx \sum_{i=1}^k w_i t(z_i)$$

The function `GHrule` in **lme4** (based on code in the **SparseGrid** package) provides knots and weights relative to the standard normal kernel for orders  $k$  from 1 to 25. For example,

```
> GHrule(5)

      z      w    ldnorm
[1,] -2.856970e+00 0.01125741 -5.0000774
[2,] -1.355626e+00 0.22207592 -1.8377997
```

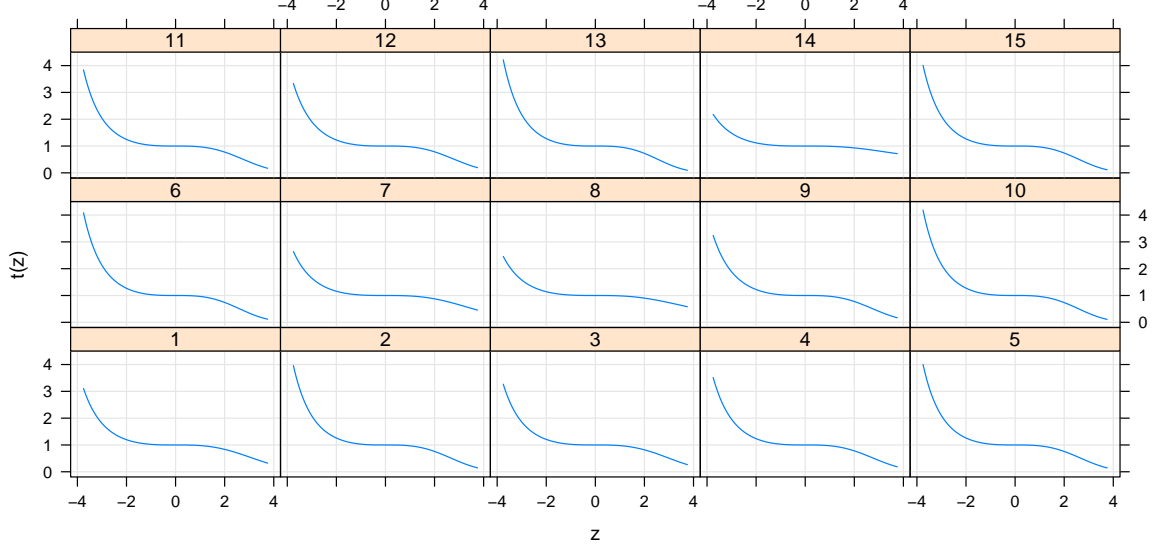


Figure 2: The function  $t(z)$ , which is the ratio of the normalized unscaled conditional density to the standard normal density, for each of the univariate integrals in the evaluation of the deviance for model `m1`. These functions should behave like low-order polynomials.

```
[3,] 3.865099e-17 0.53333333 -0.9189385
[4,] 1.355626e+00 0.22207592 -1.8377997
[5,] 2.856970e+00 0.01125741 -5.0000774
```

The choice of the value of  $k$  depends on the behavior of the function  $t(z)$ . If  $t(z)$  is a polynomial of degree  $k - 1$  then the Gauss-Hermite formula for orders  $k$  or greater provides an exact answer. The fact that we want  $t(z)$  to behave like a low-order polynomial is often neglected in the formulation of a Gauss-Hermite approximation to a quadrature. The quadrature knots on the  $u$  scale are chosen as

$$u_{i,j}(z) = \tilde{u}_j + z_i / \mathbf{L}_{j,j}, \quad i = 1, \dots, k; \quad j = 1, \dots, q \quad (19)$$

exactly so that the function  $t(z)$  should behave like a low-order polynomial over the region of interest, which is to say the region where quadrature knots with large weights are located. The term “adaptive Gauss-Hermite quadrature” reflects the fact that the approximating Gaussian density is scaled and shifted to provide a second order approximation to the logarithm of the unscaled conditional density.

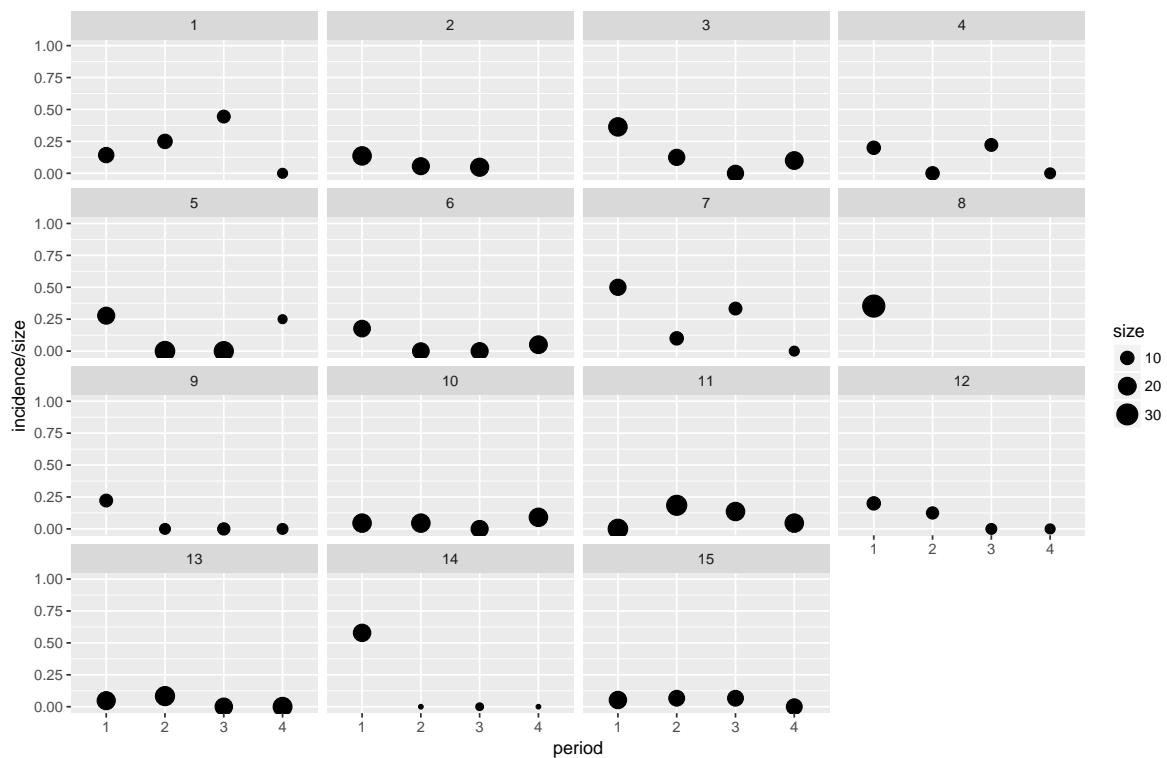
Figure 2 shows  $t(z)$  for each of the unidimensional integrals in the likelihood for the model `m1` at the parameter estimates.

## 4. Examples

```
> head(cbpp)
```

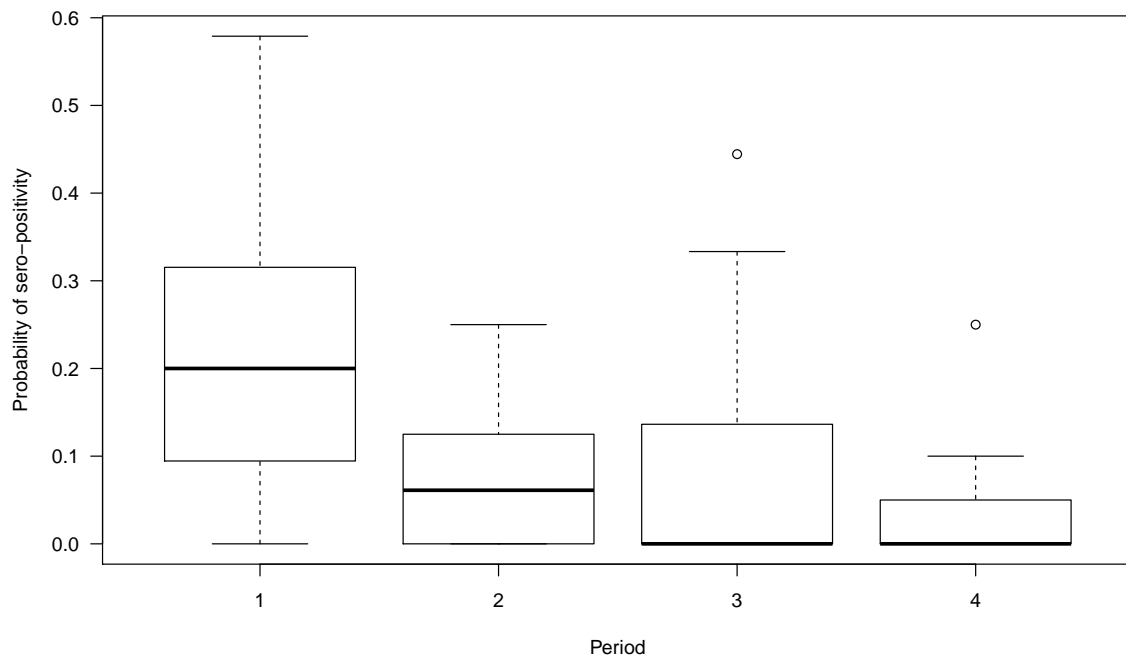
	herd	incidence	size	period
1	1	2	14	1
2	1	3	12	2
3	1	4	9	3
4	1	0	5	4
5	2	3	22	1
6	2	1	18	2

```
> library(ggplot2)
> ggplot(cbpp) + facet_wrap(~herd) +
+   geom_point(aes(period, incidence/size, size = size)) +
+   scale_y_continuous(limits = c(0, 1))
```



```
> boxplot(incidence/size ~ period, data = cbpp, las = 1,
+   xlab = 'Period', ylab = 'Probability of sero-positivity')
```





To account for repeated measures

```
> (gm1 <- glmer(incidence/size ~ period + (1 | herd), family = binomial,
+               data = cbpp, weights = size))
```

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) [glmerMod]

Family: binomial ( logit )

Formula: incidence/size ~ period + (1 | herd)

Data: cbpp

Weights: size

	AIC	BIC	logLik	deviance	df.resid
	194.0531	204.1799	-92.0266	184.0531	51

Random effects:

Groups	Name	Std.Dev.
herd	(Intercept)	0.6421

Number of obs: 56, groups: herd, 15

Fixed Effects:

	(Intercept)	period2	period3	period4
	-1.3983	-0.9919	-1.1282	-1.5797

```
> summary(gm1)
```

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) [glmerMod]

```

Family: binomial ( logit )
Formula: incidence/size ~ period + (1 | herd)
Data: cbpp
Weights: size

      AIC      BIC   logLik deviance df.resid
194.1    204.2    -92.0    184.1      51

Scaled residuals:
      Min       1Q   Median       3Q      Max
-2.3816 -0.7889 -0.2026  0.5142  2.8791

Random effects:
Groups Name      Variance Std.Dev.
herd  (Intercept) 0.4123   0.6421
Number of obs: 56, groups: herd, 15

Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.3983     0.2312  -6.048 1.47e-09
period2       -0.9919     0.3032  -3.272 0.001068
period3       -1.1282     0.3228  -3.495 0.000474
period4       -1.5797     0.4220  -3.743 0.000182

Correlation of Fixed Effects:
      (Intr) perid2 perid3
period2 -0.363
period3 -0.340  0.280
period4 -0.260  0.213  0.198

```

```

> #profile.gm1 <- profile(gm1)
> #confint(gm1)
> #set.seed(1)
> #boot.gm1 <- bootMer(gm1, function(x) getME(x, 'theta'), nsim = 100)
> #quantile(boot.gm1$t, probs = c(0.025, 0.975))

```

## References

Bates DM, Watts DG (1988). *Nonlinear Regression Analysis and Its Applications*. Wiley, Hoboken, NJ. ISBN 0-471-81643-4.

**Affiliation:**

Steven C. Walker  
Department of Mathematics & Statistics  
McMaster University  
1280 Main Street W  
Hamilton, ON L8S 4K1, Canada  
E-mail: [scwalker@math.mcmaster.ca](mailto:scwalker@math.mcmaster.ca)

Rune Haubo Bojesen Christensen  
Technical University of Denmark  
Matematiktorvet  
Building 324, room 220  
2800 Kgs. Lyngby  
E-mail: [rhbc@dtu.dk](mailto:rhbc@dtu.dk)

Douglas Bates  
Department of Statistics, University of Wisconsin  
1300 University Ave.  
Madison, WI 53706, U.S.A.  
E-mail: [bates@stat.wisc.edu](mailto:bates@stat.wisc.edu)

Martin Mächler  
Seminar für Statistik, HG G 16  
ETH Zurich  
8092 Zurich, Switzerland  
E-mail: [maechler@stat.math.ethz.ch](mailto:maechler@stat.math.ethz.ch)

Benjamin M. Bolker  
Departments of Mathematics & Statistics and Biology  
McMaster University  
1280 Main Street W  
Hamilton, ON L8S 4K1, Canada  
E-mail: [bolker@mcmaster.ca](mailto:bolker@mcmaster.ca)