Estimating random effect in Negative Binomial Regression

There are various ways of presenting the negative binomial distribution, and a detailed explanation can be found in Hilbe(2007). It can be viewed as a Poisson-Gamma mixture, which means we assume that Y's are Poisson distributed with mean μ following a Gamma distribution. The density function is then expressed as

$$f(y;k,\mu) = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)} \times \left(\frac{k}{\mu+k}\right)^k \times \left(1-\frac{k}{\mu+k}\right)^y$$

where $\Gamma(y+1) = (y+1)!$.
The mean and variance of Y are given by

$$E(Y) = \mu \qquad \text{Var}(Y) = \mu + \frac{\mu^2}{k}$$

Note that by letting $\phi = 1/k$ where ϕ is usually recognized as dispersion parameter, we can reexpress the variance as $\text{Var}(Y) = \mu + \phi\mu^2$

Basic Setup of NB Regression
The NB regression is specified with three steps

1. Y_i is negative binomial distributed with mean μ_i and dispersion parameter ϕ.

2. the predictor is given by $\eta = X\alpha + Zb$, where α and b are fixed and random effects, respectively. Suppose $b \sim N(0,\sigma^2)$

3. there is a link between the mean Y and predictor $\eta = g(\mu)$, by default glmmadmb uses ....

For SAS `PROC NLMIXED` , the theory behind this procedure

1. Assumptions and Notations

The observed data vector $\boldsymbol{y}_i$ for each I subjects, $i = 1, \ldots, s$ and $\boldsymbol{y}_i$ are assumed to be independent across i, but within-subject covariance is likely to exist because each of the elements of $\boldsymbol{y}_i$ is measured on the same subject(that implies it can deal with repeated measurement). The joint probability density function

$$p(\boldsymbol{y}_i|\boldsymbol{X}_i, \boldsymbol{\phi}, \boldsymbol{u}_i) q(\boldsymbol{u}_i|\boldsymbol{\xi})$$

where $\boldsymbol{X}_i$ is a matrix of observed explanatory variables and $\boldsymbol{\phi}, \boldsymbol{\xi}$ are vectors of unknown parameters.

Let $\boldsymbol{\theta} = [\boldsymbol{\phi}, \boldsymbol{\xi})]$ is of dimension n. Then we can make inference about $\boldsymbol{\theta}$ by the marginal likelihood funciton

$$m(\boldsymbol{\theta}) = \prod_{i=1}^{s} \int p(\boldsymbol{y}_i|\boldsymbol{X}_i, \boldsymbol{\phi}, \boldsymbol{u}_i) q(\boldsymbol{u}_i|\boldsymbol{\xi}) d\boldsymbol{u}_i$$

. Essentially we obtain $\hat{\boldsymbol{\theta}}$ by minimizing

$$f(\boldsymbol{\theta}) = -\log[m(\boldsymbol{\theta})]$$

Next we apply the NB regression setup here. For a particular gene k (index suppressed here)

$$y_{ij} \sim NB(\mu_i, \phi), \;\; \eta_i = \log(\mu_i) = log(R_j N_j) + \beta + e_i$$

where $e_i \sim N(0, \sigma^2)$ and i, j index jth sample in ith group. not sure if normalization is needed yet, in which case we might begin with $\log(\mu_i) = \beta + \log(N_i R_i) + e_i$. The data set I analyze consists of 4 labs of arabidopsis experiment with 2 or 3 samples in each lab. That being said, I am assuming the means for different labs vary only in terms of random effect e_i. Note: for now we just assume ϕ is a constant within different samples for a particular gene. Note that $\boldsymbol{\theta} = (\beta, \sigma^2)$, and for SAS `PROC NLMIXED` the NB $Y \sim \text{negbin}(n, p)$ log-likelihood is

$$l(n, p; y) = \log[\Gamma(n + y)] - \log[\Gamma(n)] - \log[\Gamma(y + 1)] + n\log(p) + y\log(1 - p)$$

$$E[Y] = kP = k\left(\frac{1-p}{p}\right), \text{Var}[Y] = kP(1 - P) = k\left(\frac{1-p}{p}\right)\frac{1}{p}$$

with $n \geq 0, 0 < p < 1$. That is equivalent to $\mu = kP, k = 1/\phi$ under NB2 parameterization. Therefore the $p(\cdot)$ can be written as

$$p(y_i, p, k) = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)} \times p^k (1-p)^y$$

By noting $\mu_i = e^{\beta + e_i} = k(1-p)/p$ we have

$$p = \frac{1}{k^{-1} e^{\beta + e_i} + 1}$$

Subsequently

$$p(y_i|\beta, e_i, k) = \frac{\Gamma(y_i+k)}{\Gamma(k)\Gamma(y_i+1)} \times \left(\frac{1}{k^{-1}e^{\beta+e_i}+1}\right)^k \left(1 - \frac{1}{k^{-1}e^{\beta+e_i}+1}\right)^{y_i}$$

Now that $e_i \sim N(0, \sigma^2)$ with $q(e_i|\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{e_i^2}{2\sigma^2}\right)$ gives

$$m(\beta, \sigma^2, k) = \prod_{i=1}^{n} \int p(y_i|\beta, e_i, k) q(e_i|\sigma^2) de_i$$

$$= \prod_{i=1}^{n} \int \frac{\Gamma(y_i+k)}{\Gamma(k)\Gamma(y_i+1)} \times \left(\frac{1}{k^{-1}e^{\beta+e_i}+1}\right)^k \left(1 - \frac{1}{k^{-1}e^{\beta+e_i}+1}\right)^{y_i} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{e_i^2}{2\sigma^2}\right) de_i$$

2. Integration Approximation

`PROC NLMIXED` uses adaptive Gaussian Quadrature (Pinheiro and Bates 1995) while R Package `glmmADMB` adopts Laplace Approximation (reference ???). According to SAS documentation, the latter is just a 1st order special case of the former. Let $p(\beta, e_i) = \frac{1}{k^{-1}e^{\beta+e_i}+1}$, then rewriting $m(\beta, \sigma^2, k)$ gives

$$m(\beta, \sigma^2, k) = \prod_{i=1}^{n} \frac{\Gamma(y_i+k)}{\Gamma(k)\Gamma(y_i+1)} \int p(\beta, e_i)^k (1-p(\beta, e_i))^{y_i} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{e_i^2}{2\sigma^2}\right) de_i$$

$$= \prod_{i=1}^{n} \frac{\Gamma(y_i+k)}{\Gamma(k)\Gamma(y_i+1)} \int e^{k\log[p(\beta,e_i)]} e^{y_i \log[1-p(\beta,e_i)]} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{e_i^2}{2\sigma^2}\right) de_i$$

$$= \prod_{i=1}^{n} \frac{\Gamma(y_i+k)}{\Gamma(k)\Gamma(y_i+1)} \frac{1}{\sqrt{2\pi\sigma^2}} \int \exp\left[k\log(p) + y_i \log(1-p) - \frac{e_i^2}{2\sigma^2}\right] de_i \qquad (*)$$

The integral is approximated by Gaussian Quadrature. Denote

$$l(e_i, y_i) = k\log(p) + y_i \log(1-p) - \frac{e_i^2}{2\sigma^2}$$

Let e_i^* maximizes $l(e_i, y_i)$., then $(*)$ can be approximated by

Gaussian Quadrature here

Can try

```
file.show(system.file("tpl","glmmadmb.tpl",package="glmmADMB"))"
```

to see how parameters are estimated in `glmmADMB`.
It seems the ADMB-RE package (implementing random effects in nonlinear models) is also adaptive to non-normally distributed random effects and C++ programs are available. SAS code for NB regression