Random effects, Mixtures and NPMLE

John Hinde
and
Jochen Einbeck
National University of Ireland, Galway

email: john.hinde@nuigalway.ie

Research supported by sfi science foundation ireland fondúireacht eolaíochta éireann

Normal Models

$$\mathbf{y} = \boldsymbol{\beta}^T\mathbf{x} + \boldsymbol{\epsilon}$$

- single error term includes
  - individual observation/measurement error
  - *experimental* unit variability
  - unobserved covariates
- for simplest data structures/designs use ***normal linear model***
- more complex situations
  - structure in *experimental* unit variability
  - repeated measures/longitudinal observations
  - ...

Normal Mixed Model

$$\mathbf{y} = \boldsymbol{\beta}^T\mathbf{x} + \boldsymbol{\gamma}^T\mathbf{z} + \boldsymbol{\epsilon}$$

- $\mathbf{z}$ unobserved ***random effects***
- shared random effects
  - multi-level/variance components models
  - longitudinal observations
  - spatial structure
- $\mathbf{z}$ normal
  - normal model with ***structured covariance matrix***
- standard mixed model analyses – ML, REML
- widely available in standard software

Generalized Linear Models

Models for counts, proportions, times, ...

$$\mathbf{y} \sim F(\boldsymbol{\mu}) \qquad g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \boldsymbol{\beta}^T\mathbf{x}$$

- distributional assumption relates to the observation/measurement process
- how does this model incorporate
  - experimental/individual unit variability?
  - unobserved covariates?

It doesn't!

hence overdispersion, etc

Orange Tissue-culture Experiment

Tissue culture experiment on the orange variety *Caipira* to study effect of carbohydrate sources on stimulation of somatic embryos from callus cultures.

- six carbohydrate sources: maltose, glucose, galactose, lactose, sucrose and glycerol;
- completely randomized block design with sugars at 5 different dose levels and 5 replicates of each treatment;
- response – **number of embryos** observed after four weeks

Aim: to study dose-response relationship

Poisson model shows clear ***overdispersion***.

Germination of Orobanche

Germination of two species of Orobanche seeds grown on plates of extract of either bean or cucumber root.

O. aegyptiaca 75		*O. aegyptiaca 73*	
Bean	Cucumber	Bean	Cucumber
10/39	5/6	8/16	3/12
23/62	53/74	10/30	22/41
23/81	55/72	8/28	15/30
26/51	32/51	23/45	32/51
17/39	46/79	0/4	3/7
	10/13		

- Binomial logit model – ***significant interaction; overdispersion***

Random Effect Models

Include random effect(s) in the linear predictor

$$\boldsymbol{\eta} = \boldsymbol{\beta}^T \mathbf{x} + \boldsymbol{\gamma}^T \mathbf{z}$$

- single conjugate random effect at individual level – standard overdispersion models
  - negative binomial for count data
  - beta-binomial for proportions
- $\mathbf{z}$ normal $\longrightarrow$ ***generalized linear mixed models***
- $\mathbf{z}$ unspecified $\longrightarrow$ ***nonparametric maximum likelihood***

Normal Random Effect

$$L(\boldsymbol{\beta}, \sigma) = \prod_{i=1}^{n} \int f(y_i|\boldsymbol{\beta}, \sigma, z_i)\phi(z_i)dz_i$$

where $\phi(z)$ is the normal density, and f the response density.

No analytic form for integral – approximate using K-point Gaussian Quadrature (mass points z_k with weights π_k)

$$L(\boldsymbol{\beta}, \sigma) \approx \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f(y_i|\boldsymbol{\beta}, \sigma, z_k)$$

Likelihood for K component mixture of response distribution with linear predictor for k-th component

$$\eta_{ik} = \boldsymbol{\beta}^T \mathbf{x}_i + \sigma z_k$$

EM Algorithm

Estimation for finite mixture conveniently viewed as EM algorithm.

E-Step: Calculate component weights w_{ik} – the posterior probability that observation y_i comes from component k:

$$w_{ik} = \frac{\pi_k f_{ik}}{\sum_\ell \pi_\ell f_{i\ell}}$$

M-step: Estimate $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}$ by

- fitting response model to expanded data (K-copies)
- y-variate $\mathbf{y}^* = (\mathbf{y}', \mathbf{y}', \ldots, \mathbf{y}')'$
- explanatory variables for y^*_{ik}: $\mathbf{x}_i$ and z_k
- weights w_{ik}

Multi-Centre Beta-blocker Trial

- trial of beta-blockers to reduce mortality after myocardial infarction
- 22 centres
- single treatment – treatment and control groups
- patients *within* centres
- response r deaths out of n for each group

- centres very different sizes – 38 to 1916

require generalizability

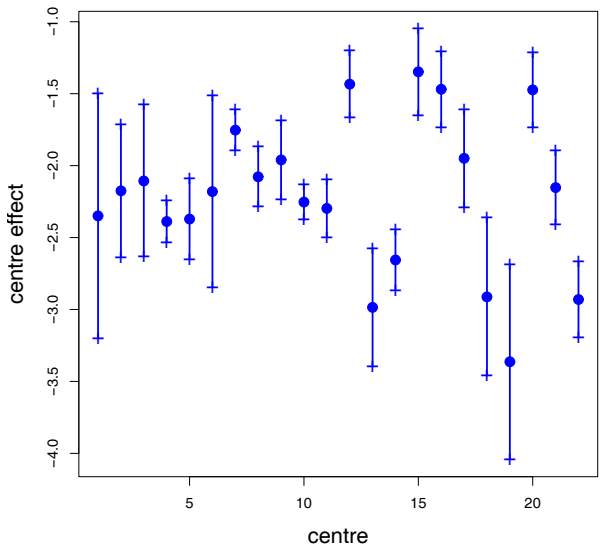
- Simple fixed effect model, ignoring among-centre variation

  Using binomial logit model, on log-odds scale

  $$\text{treatment effect} = -0.257 \quad (\text{s.e.} = 0.049)$$

  Residual deviance: 305.76 on 42 df

- considerable among-centre variation . . .



Normal Random Effect – Variance Component

$$\text{logit}(p_{ij}) = \alpha + \beta\,\text{treat}_{ij} + \sigma\, Z_i \quad ; \quad Z_i \sim N(0,1)$$

Gaussian quadrature for betablockers

K	α	se	β	se	σ	$-2\log L$
1	−2.197	0.034	−0.257	0.049	0.000	523.2
2	−2.034	0.035	−0.257	0.050	0.366	365.5
3	−2.239	0.034	−0.258	0.050	0.360	321.0
5	−2.238	0.034	−0.258	0.050	0.455	319.8
10	−2.087	0.034	−0.262	0.050	0.454	318.6
20	−2.180	0.034	−0.261	0.050	0.432	316.7

Arbitrary Random Effects

$$L(\boldsymbol{\beta}, g) = \prod_{i=1}^{n} \int f(y_i|\boldsymbol{\beta}, z_i) g(z_i) dz_i$$

where $g(z)$ is the unspecified mixing distribution.

Use non-parametric maximum likelihood (NPML) estimate of g

a finite discrete distribution on K mass points

$$\begin{pmatrix} z_1, & z_2, & \cdots, & z_K \\ \pi_1, & \pi_2, & \cdots, & \pi_K \end{pmatrix}$$

The joint likelihood is

$$L(\boldsymbol{\beta}, K, \pi_1, ..., \pi_{K-1}, z_1, ...z_K) = \prod_{i=1}^{n} \left\{ \sum_{k=1}^{K} f(y_i|z_k, \boldsymbol{\beta}) \pi_k \right\}$$

Non-parametric Maximum Likelihood

EM technique is easily extended to incorporate estimation of a discrete mixing distribution for z with K mass points.

E-step: as before using current estimate of mixing distribution in place of quadrature points and weights.

M-step: Estimate $\boldsymbol{\beta}$ and $\{z_k\}$ by

- fitting response model to expanded data (K-copies)
- explanatory variables $\mathbf{x}_i$ and **a K-level factor**
- weights w_{ik}
- $\widehat{\pi}_k = \sum_{i=1}^{n} \frac{w_{ik}}{n}$

Fit models for different values of K until joint likelihood is maximized.

Beta-Blocker: NPML Variance Component

$$\text{logit}(p_{ij}) = \alpha + \beta\,\text{treat}_{ij} + Z_i \quad ; \quad Z_i \text{ unspecified}$$

K	β	se	σ	$-2\log L$	# Z-pars
1	−0.257	0.049	0.000	523.2	0
3	−0.258	0.050	*0.428*	318.7	4
5	−0.258	0.050	*0.488*	310.4	8
10	−0.258	0.050	*0.489*	308.6	18
Normal	−0.261	0.050	0.432	316.7	1

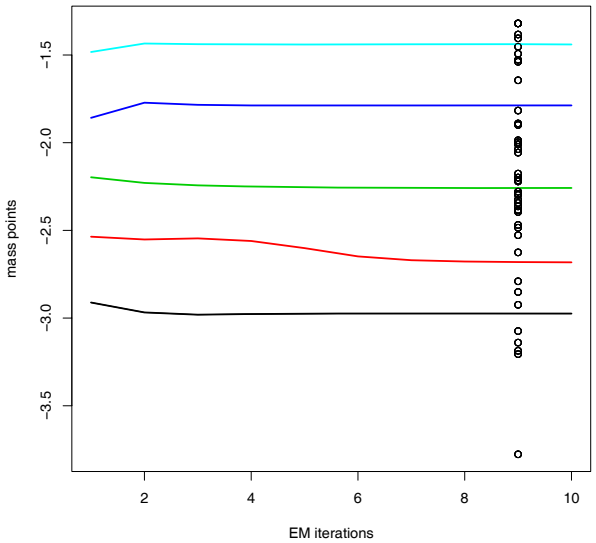
No evidence against normality

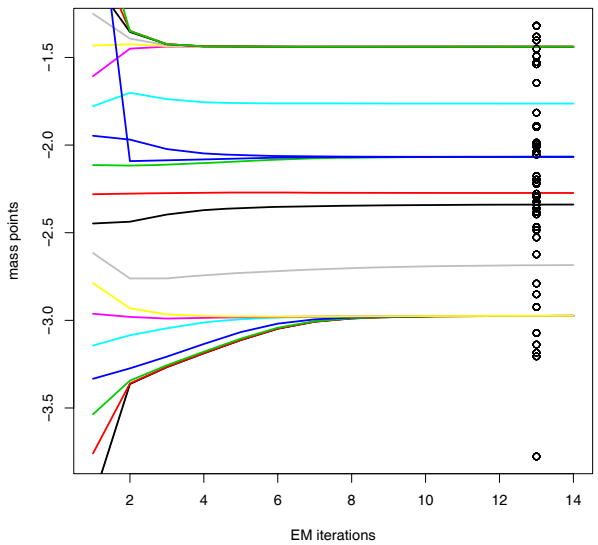
R-code and Computation

Simple model specification

```
allvc(cbind(r,(n-r))~treat,data=betablok,
        family=binomial,random=~1|center,
          k=5,random.distribution='np',tol=0.25))
```

- fixed and random models
- Gaussian or NPML random effects
- # of mass-points
- control over start for mass points – `tol`





Beta-blocker: Random Treatment Model

$\text{logit}(p_{ij}) = \alpha + \beta\,\text{treat}_{ij} + Z_i + U_i\,\text{treat}_{ij}$; Z_i, U_i unspecified

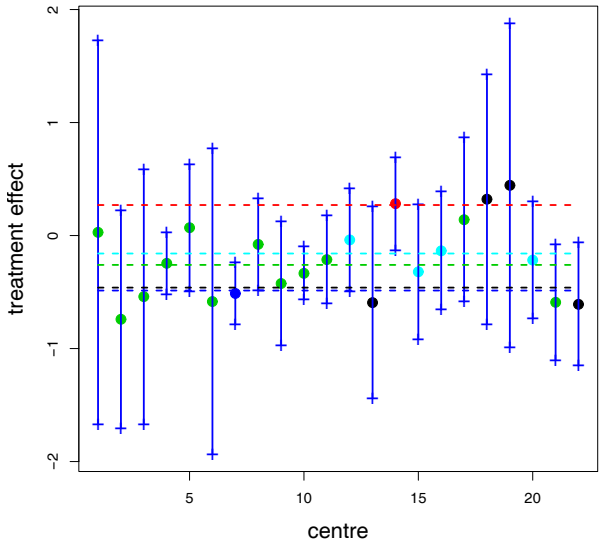
Finite mass-point distribution for (Z, U)

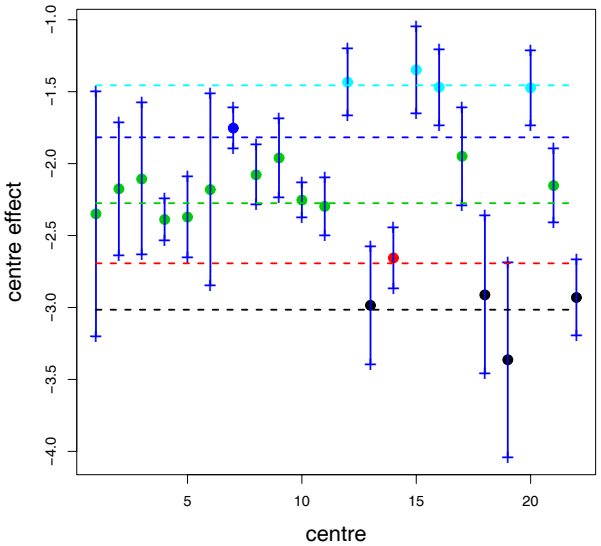
K	$-2\log L$	# (Z,U)-pars
1	523.2	0
3	316.6	6
5	299.0	12
VC-5	310.4	8
Normal	316.7	1

Regressions in each component

k	β_{0k}	β_{1k}	π_k
1	-1.486	-0.159	0.1816
2	-2.255	-0.260	0.4937
3	-2.895	-0.460	0.1581
4	-1.684	-0.487	0.0869
5	-2.937	$+0.270$	0.0798

- average treatment effect $= -0.250$
- significant treatment variation across centres
- component 5 — single large centre with ***increased*** death risk under treatment





Foetal Lamb Data

Numbers of movements of a foetal lamb in 240 5-second intervals ($\bar{x} = 0.36$, $s^2 = 0.66$) – many ***zero*** observations

Model	$-2 \log L$	# Z-pars
Poisson	348.5	0
Negative Binomial	319.7	1
Poisson-lognormal	318.8	1
ZIP	327.3	1
NPML	318.0	3

Mixing Distributions

Model	Distribution	Mean	Variance
Poisson	$\begin{pmatrix} 0.36 \\ 1 \end{pmatrix}$	0.36	0
Negative binomial	gamma	0.36	0.24
ZIP	$\begin{pmatrix} 0 & 0.85 \\ 0.58 & 0.42 \end{pmatrix}$	0.36	0.17
NPML	$\begin{pmatrix} 0 & 0.51 & 3.83 \\ 0.42 & 0.57 & 0.02 \end{pmatrix}$	0.36	0.31

Conclusions

- NPML provides a flexible framework for random effect models with minimal assumptions
- NPMLE provides a useful yardstick to assess random effect distribution
- NPML can accommodate and identify outliers
- the appropriate analysis of many glms requires random effects

References

Aitkin, M, Francis, B, Hinde, J (2005)
Statistical Modelling in Glim4, 2nd Edition, Oxford.

Software

Thanks to Ross Darnell and Brian Francis for initial work.

R-code and examples available from:

`http://www.nuigalway.ie/maths/je/npml.html`