Aster Models with Random Effects

Charles J. Geyer

School of Statistics
University of Minnesota

March, 15, 2013

Co-authors: Ruth Shaw (Minnesota), Julie Etterson (Minnesota Duluth), Caroline Ridley (EPA), and Robert Latta (Dalhousie)

Outline

- **Aster models**. A kind of *generalized* generalized linear model. Allows dependence among components of response vector (simple graphical model). Allows different components of response to have different families (some Bernoulli, some Poisson, etc.). For life history analysis.
- **Generalized linear mixed models (GLMM)**. Are there any good algorithms? Are they ever in asymptopia? Can inference for them ever be trusted?
- **Combination of the Two**.

Aster Models

Aster models (named after the flowers) are "generalized generalized linear models."

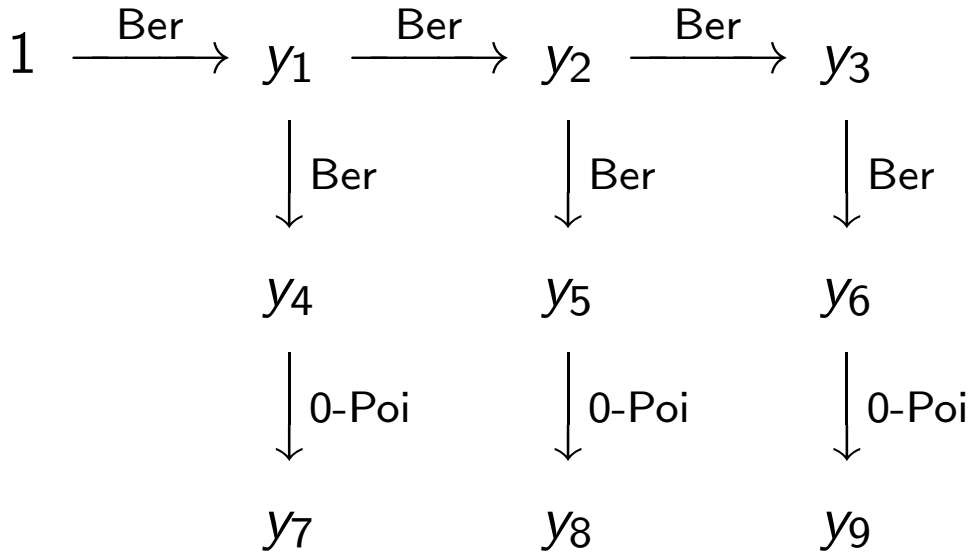
Allow different components of response vector to have different distributions (some Bernoulli, some Poisson, some zero-truncated Poisson, some normal).

Allow different components of response vector to have dependence specified by simple graphical model.

Are exponential family models.

Are for "life history analysis" when there is survival but interest is not in mere survival but in what happens after survival.

An Aster Graph



y_i are components of response vector for one individual (all individuals have isomorphic graphs). 1 is the constant 1.

Arrows indicate conditional distributions of variable at head of arrow (successor) given variable at tail of arrow (predecessor). Ber = Bernoulli, 0-Poi = zero-truncated Poisson.

Abstract Aster Graph

Not general graphical models.

Nodes (variables) have at most one predecessor, hence graph is disjoint union of trees.

Conditional distribution for each arrow is one-parameter exponential family with successor variable as canonical statistic and predecessor variable as sample size.

Aster Model Log Likelihood

Log likelihood

$$l(\theta) = \sum_{j \in J} \left[y_j \theta_j - y_{p(j)} c_j(\theta_j) \right]$$

This is recognizable as log likelihood for joint exponential family because log likelihood is linear in the y's.

Reparameterize to canonical parameters for joint exponential family (details in Geyer, et al., *Biometrika*, 2007).

Virtues of Aster Models

Has all the virtues of exponential families.

Canonical affine submodels of exponential families are themselves full exponential families.

Log likelihood is strictly concave and MLE are unique if they exist.

Submodel canonical statistics are sufficient statistics (sufficient dimension reduction).

Multivariate monotone relationship between canonical and mean-value parameters gives simple interpretation.

Reboot

End of advertisement for aster models.

For rest of talk have any exponential family with log likelihood

$$l(\varphi) = y^T \varphi - c(\varphi)$$

(canonical statistic vector y and canonical parameter vector φ).

Canonical Affine Submodel with Random Effects

$$\varphi = a + M\alpha + Zb,$$

where

- a is a known vector (*offset vector*),
- M and Z are known matrices (*model matrices* for fixed effects and random effects, respectively),
- α is vector of unknown parameters (*fixed effects*)
- b is a normal random vector (*random effects*) with mean vector zero and variance matrix D.

We assume the matrix D is diagonal. Its diagonal components are called *variance components*. Let ν denote vector of variance components.

The Right Thing (TRT)

Complete data log likelihood

$$l_c(\alpha, b, \nu) = l(a + M\alpha + Zb) - \tfrac{1}{2} b^T D^{-1} b - \tfrac{1}{2} \log \det(D)$$

Missing data log likelihood

$$l_m(\alpha, \nu) = \log \left(\int e^{l_c(\alpha, b, \nu)} \, db \right)$$

Cannot do integral except in very simple random effects models (only one random effect per individual). So cannot do The Right Thing (TRT), must do Some Wrong Thing (SWT).

Laplace Approximation

Breslow and Clayton (*JASA*, 1993) proposed to replace the undoable integral in TRT with its *Laplace approximation*: replace complete data log likelihood by its quadratic approximation (Taylor series up to quadratic terms), then integral has "e to a quadratic" form of normal integral and can be done analytically.

Laplace Approximation (cont.)

Where to expand around? Point b^* where first derivative is zero (considered as function of b for fixed α and ν).

Then approximate log integrated likelihood (SWT) is

$$q(\alpha, \nu) = l(a + M\alpha + Zb^*) - \tfrac{1}{2}(b^*)^T D^{-1} b^* \\ - \tfrac{1}{2} \log\det\left[Z^T W(a + M\alpha + Zb^*)ZD + I\right]$$

where

$$W(\varphi) = \nabla^2 c(\varphi).$$

Note: b^* is function of α and ν and D is function of ν.

Laplace Approximation (cont.)

Maximizers $\hat{\alpha}$ and $\hat{\nu}$ of SWT approximate maximum likelihood estimators.

$\hat{b} = b^*(\hat{\alpha}, \hat{\nu})$ is "estimator" of random effects vector (scare quotes because random effects are random variables not parameters).

Not actually what Breslow and Clayton (1993) recommended. Their scheme adds more complication to the Laplace approximation. Not clear (to me) that any software uses full Breslow and Clayton scheme. Not clear (to me) how to generalize their scheme from GLM to arbitrary exponential family.

Now What?

So much for point estimates, how about hypothesis tests and confidence intervals?

Negative of second derivative matrix of SWT should approximate observed Fisher information.

Problem: second partial derivatives of q involve second partial derivatives of W and fourth partial derivatives of original exponential family log likelihood.

No problem in theory — exponential family log likelihoods are infinitely differentiable — but big problem in practice, expecially for aster models (R package `aster` computes first and second derivatives, but not higher order derivatives).

Now What? (cont.)

Solution: (also taken from Breslow and Clayton, 1993) treat W as constant function of its arguments, equivalent to assuming that complete data log likelihood is exactly quadratic. Essentially, we are throwing away third and higher order derivatives of the original exponential family log likelihood.

Equivalent to changing q to

$$q(\alpha, \nu) = l(a + M\alpha + Zb^*) - \tfrac{1}{2}(b^*)^T D^{-1} b^* - \tfrac{1}{2} \log\det\left[Z^T \widehat{W} Z D + I\right]$$

where $\widehat{W}$ is a constant matrix, which should be chosen close to $W(a + M\hat{\alpha} + Z\hat{b})$. Notice that q is profile of

$$p(\alpha, b, \nu) = l(a + M\alpha + Zb) - \tfrac{1}{2} b^T D^{-1} b - \tfrac{1}{2} \log\det\left[Z^T \widehat{W} Z D + I\right]$$

Approximate Fisher Information

Now using implicit function theorem and chain rule (one of two technical innovations in Geyer, et al., submitted)

$$q_{\psi\psi}(\psi) = p_{\psi\psi}(\psi, b^*) - p_{\psi b}\big(\psi, b^*\big) p_{bb}\big(\psi, b^*\big)^{-1} p_{b\psi}\big(\psi, b^*\big)$$

where ψ is vector of all parameters (includes α and ν), subscripts indicate partial derivatives, and b^* is a function of ψ.

Form familiar from the conditional variance formula for normal distributions

$$\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

Guarantees $-q_{\psi\psi}(\psi)$ is positive definite whenever (ψ, b^*) is strong local maximum of p.

Zero Variance Components

All of preceeding makes no sense when some variance component is zero so D^{-1} is undefined. A well known trick is to change parameters from variance components ν to "standard deviation components" σ defined by $\nu_i = \sigma_i^2$ except that σ_i is allowed to be negative. Then log approximate integrated likelihood is defined and differentiable on neighborhoods of zero variance components.

But this causes other problems. Spurious zeros of first derivative whenever any $\sigma_i = 0$. So cannot use first derivative equal to zero as test for local maximum, *nor can rely on optimization software that uses such derivative tests*.

Zero Variance Components (cont.)

So (the other technical innovation in Geyer, et al., submitted) we do not use "standard deviation components". Instead use theory of constrained optimization to derive test based on directional derivatives

$$\bar{p}_{\nu_j}(\alpha, b, \nu) - \frac{1}{4} \sum_{\substack{i \in I \\ D_{ii} = \nu_j}} \bar{p}_{b_i}(\alpha, b, \nu)^2 \geq 0$$

where $\bar{p}$ is the part of p that does not contain D^{-1}.

Does it Work?

I won't present examples, because I would have to get you to understand some biology (Geyer, et al., submitted, has three examples with real data).

These examples prove (as all such examples prove) that if you put numbers into our code, then numbers come out.

Moreover, our analyses are re-analyses. The biologists had published the data using conventional normal-response-normal-random-effects models (even though the responses were highly non-normal), and our re-analyses roughly agree with theirs (the biological conclusions are not overturned).

When Should it Work?

The Laplace approximation assumes the complete data log likelihood is close to quadratic in the random effects b, and the second approximation (throwing away third and higher order derivatives) assumes it is also close to quadratic in the fixed effects.

This will happen if the model that treats random effects as fixed would be "in asymptopia" (close to quadratic in all effects), which happens if there are only a few effects of all kinds and a lot of data to estimate them.

When Should it Work? (cont.)

Suppose we could do exact maximum likelihood (by magic).

In typical examples, should we expect data "in asymptopia" so hypothesis tests and confidence intervals based on the "usual" asymptotics of maximum likelihood work?

Sung and Geyer (*Annals of Statistics*, 2007) looked at this and, although this is not the main point of the paper, concluded, in short, no! Examples we could find in the literature of GLMM data were very far from asymptopia.

When Should it Work? (cont.)

This should not be suprising. Multimodality and other bad behavior of log likelihoods is well known in the classical mixed model (normal random effects and normal response) literature.

It is only surprising to those who have bought into the feel good optimism of the GLMM literature.

A Formula Attributed to Lewis

For any missing data model with complete data density $f_\theta(x, y)$ with x missing and y observed and log likelihood

$$l(\theta) = \log \int f_\theta(x, y)\, dx$$

(a formula attributed to Lewis, 1982, in the EM literature, although also found in Sundberg, 1974)

$$\begin{aligned} \nabla^2 l(\theta) = {} & E_\theta\{\nabla^2 \log f_\theta(X, Y) \mid Y = y\} \\ & + \operatorname{var}_\theta\{\nabla \log f_\theta(X, Y) \mid Y = y\} \end{aligned}$$

Right hand side is difference of positive definite matrices. Result may be close to singular although neither term is.

One equation that says why missing data is hard: Fisher information much less than if missing data were observed.

When Does it Work?

For "bad" examples where second derivative matrix of log likelihood is nearly singular, it often will be singular or indefinite in computer arithmetic.

Then software (mine or anyone else's) will fail. Even if the error messages are nice and friendly, the problem does not get done.

When one bootstraps and hence simulates lots of random datasets, one often finds that some fail, and that makes bootstrap inference difficult or useless if there are enough failures.

Web Site

`http://www.stat.umn.edu/geyer/aster/`

has links to papers and tech reports. All tech reports done with `Sweave` so everything is exactly reproducible by anyone who has R.