Random Effect Modelling for Regression Models with Gamma - Distributed Response

Jochen Einbeck[1] and John Hinde[1]

[1] Department of Mathematics, National University of Ireland, Galway, Ireland. E-mails: {jochen.einbeck, john.hinde}@nuigalway.ie

Keywords: EM algorithm, random effect models, nonparametric maximum likelihood, overdispersion, gamma distribution, generalized linear model.

1 Introduction

Assume there is given a set of explanatory vectors $x_1, \ldots, x_n$ and a set of observations $y_1, \ldots, y_n$ sampled from an exponential family distribution $f(y_i|\beta, \phi_i)$ with dispersion parameter ϕ_i. In a generalized linear model (GLM), predictors and response are assumed to be related through a link function h,

$$\mu_i \equiv E(y_i|\beta, \phi_i) = h(\eta_i) \equiv h(x_i'\beta).$$

The variance $\sigma_i^2 = \mathrm{Var}(y_i|\beta, \phi_i) = \phi_i v(\mu_i)$ depends on a function $v(\mu_i)$ which is entirely determined by the choice of the particular exponential family. However, often the actual variance in the data is larger than the variance according to this strict mean-variance relationship. In order to account for this *overdispersion*, a random effect z_i with density $g(z)$ is included into the linear predictor $\eta_i = \beta' x_i + z_i$. The marginal likelihood can now be approximated by a finite mixture

$$L = \prod_{i=1}^{n} \int f(y_i|z_i, \beta, \phi_i) g(z_i)\, dz_i \approx \prod_{i=1}^{n} \left\{ \sum_{k=1}^{K} \pi_k f_{ik} \right\},$$

where $f_{ik} = f(y_i|z_k, \beta, \phi_k)$, z_k are the mass points and π_k their masses. The log-likelihood $l = \log L$ is then maximized using a standard EM algorithm. If the random effect distribution is not specified parametrically, one refers to this method as 'Nonparametric Maximum Likelihood' (NPML) estimation, introduced for overdispersed GLM's by Aitkin (1996) and adapted to variance component models, allowing for shared random effects, by Aitkin (1999).

2 Scope

Surprisingly, existing software packages for EM-based NPML estimation do not support the Gamma distribution $\Gamma(\nu, \nu/\mu)$, with shape ν and rate ν/μ, where the dispersion takes the form $\phi = 1/\nu$. GLIM4 (Aitkin & Francis, 1995) and C.A.MAN (Böhning et al., 1992) just support exponentially distributed response, i.e. $\nu = 1$. One reason for this may be general computational difficulties with NPML arising through highly fluctuating EM trajectories as demonstrated by Einbeck & Hinde (2005).
This gap is filled in Einbeck & Hinde (2006) using a damped version of the EM algorithm. The dispersion (shape) parameter is not only allowed to be constant or component-specific, but also to vary smoothly between components. The latter provides a convenient way forward when computation using component-specific shape parameters breaks down due to the occurrence of likelihood spikes.
NPML estimation of random effect models with gamma-distributed response is deemed to be particularly useful for the analysis of duration problems (e.g. models for hospital stay duration, as will be discussed in the presentation) and survival data. The method extends immediately to variance component models. The R code is available at `www.nuigalway.ie/maths/je/npml.html`.

References

Aitkin M (1996). A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing.* 6, 251–262.

Aitkin M (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics.* 55, 117–128.

Aitkin M and Francis B (1995). Fitting overdispersed generalized linear models by nonparametric maximum likelihood. *The GLIM Newsletter.* 25, 37–45.

Böhning D., Schlattmann P, and Lindsey B (1992). Computer-assisted analysis of mixtures (C.A.MAN): statistical algorithms. *Biometrics.* 48, 283–303.

Einbeck J and Hinde J (2005). Making the EM algorithm for NPML estimation less sensitive to tuning parameters. *CASI - 2005. Book of Abstracts*, 52–53.

Einbeck J and Hinde J (2006). A note on NPML estimation for exponential family regression models with unspecified dispersion parameter. *Austrian Journal of Statistics*, in press.