Fitting exponential family mixed models

Juni Palmgren[1,2] and Samuli Ripatti[1,3]

April 25, 2001

[1] Mathematical Statistics, Stockholm University, S-10691 Stockholm, Sweden. Email: juni@matematik.su.se

[2] Medical Epidemiology, Karolinska Institutet, Sweden

[3] Rolf Nevanlinna Institute, University of Helsinki, Finland

Abstract

The seminal papers by Nelder and Wedderburn (Generalized Linear Models, JRSS A 1972) and Cox (Regression models and life tables, JRSS B 1972) both rely on the assumption that conditionally on covariate information (including time) the observations are independent. The difficulty in identifying and measuring all relevant covariates has pushed for methods that can handle both mean and covariance structures jointly. There has been a parallel development of (i) marginal models and (ii) random effects models as multivariate extensions of the generalized linear model and the multiplicative hazard model, respectively. After a brief review of this development we focus on estimation and computational aspects of fitting random effects models. We discuss the use of penalized likelihood, Monte Carlo EM and MCMC methods using examples involving censored survival time responses and Poisson responses.

Keywords: Frailty; Generalized linear mixed model; Markov chain Monte Carlo; Monte Carlo EM; Penalized likelihood; Random effects.

1 Introduction

Soon after the introduction in 1972 Nelder's and Wedderburn's generalized linear model (Nelder and Wedderburn, 1972) was recognized as a useful conceptual framework for a wide class of regression models used in biomedical research. Cox's semi-parametric regression model for censored failure time data (Cox, 1972), also published in 1972, has had an equally profound influence on the statistical methodology used in the medical field. Since the 1970's there has been escalating efforts to extend these two families of non-normal non-linear models to allow for between-cluster heterogeneity and within-cluster dependence. Study designs such as group randomization, litter based toxicology studies, longitudinal studies, studies on spatial variation, as well as family studies have pushed for this development.

In Section 2 we account for some of the milestones in the development of multivariate generalized linear models and multivariate hazard regression models. In Section 3 we present the random effects model, and a battery of estimation and inference approaches are outlined in Section 4. Section 5 describes two data analyses examples: a hiearchical survival data problem where the lifetime of roses is assessed, and a Poisson random effects model for spatial smoothing of alcohol related mortality in Finland. We conclude in Section 6 by discussing pros and cons of the different estimation and inference procedures, and we argue that a new layer of unification is emerging for handling the multivariate generalized linear models and multivariate hazard regression models.

2 From univariate to multivariate models

2.1 The generalized linear model

The generalized linear model (GLM) is specified through the probability distribution for the observations, and the link function relating the regression parameters to the means. Conditionally on the means the observations are assumed statistically independent. The standard linear regression model is a GLM with normal distribution and identity link. The log linear model for count data is a GLM with Poisson distribution and logarithmic link. The logistic regression model is a GLM with binomial distribution and logit link. These three special cases are useful standard GLM's with attractive theoretical properties (McCullagh and Nelder, 1989).

2.2 Cox regression

Censored failure time data arise in many areas of biomedical research. Early methodology was confined to descriptive life table techniques and to the mathematical formulation of the survival experience over time. Cox's regression model changed the focus to partial likelihood inference for the relative hazard as function of covariate values, while the baseline time dynamics were treated as a secondary feature and modelled non-parametrically. Partial likelihood estimation of relative risk parameters may be viewed as a stratified analysis, in which time is controlled for by matching on the risk set at each time point when a failure occurs. Counting process and martingale theory provide the theoretical basis for the Cox regression model (Andersen, Borgan, Gill and Keiding, 1993). An important feature of the model, which nicely bridges the gap to the generalized linear model, is that conditional on the past the counting process behaves like a Poisson process, with independent increments and time varying rate function.

2.3 Multivariate responses

Logistic regression, Poisson regression and Cox regression can all be viewed as univariate probability models for a series of binary events (Clayton, 1994). They all share the property that conditional on measured exposures and covariates the responses are assumed statistically independent across individuals, with a constant event probability. For the Poisson and Cox models this conditional event probability is 'small' and the 'risk sets' are large. However, incomplete covariate information is often a reality, rendering the standard model specification too simplistic.

When no information is available on sources of unobserved heterogeneity, then one single overdispersion parameter may capture the additional component of variation. Compound distributions such as the beta-binomial or gamma-Poisson may be used, or an extra parameter may be multiplied to the Binomial or Poisson variance expressions (Williams, 1982; Breslow, 1984). When the data involve identified clusters, e.g. repeated measurements on the same individual or clusters of individuals in families, then a structured model can be specified for the between-cluster heterogeneity and the within-cluster dependence.

Multivariate models for the mean and dependence structures for responses measured on a wide variety of scales has been the focus of escalating methodological interest. Two main routes have emerged: (i) the marginal models and (ii) the random effects models. We briefly touch on (i) here and discuss (ii) in detail in Section 3. In 1986 Liang and Zeger proposed a general procedure for multivariate generalized linear models (Liang and Zeger, 1986). Their focus was on the estimation of regression parameters that linked covariate effects to population averages. The within-cluster dependence was treated as a nuisance, needing to be accounted for since it affects the power of tests and the precision of regression estimates. Zhao and Prentice (1989) extended the Liang and Zeger procedure by setting up two sets of estimating equations jointly, one for the mean parameters and one for the dependence parameters. Wei, Lin and Weissfeld (1989) considered semi-parametric regression models in which two or more distinct failure times are recorded on each individual. Each marginal failure time is modelled by a semi-parametric Cox model, and the dependence is accounted for when estimating the parameter uncertainty. The Wei, Lin and Weissfeld model is a multivariate failure time analogue to the Liang and Zeger generalized estimating equation (GEE) approach for the generalized linear model.

3 Random effects models

While the marginal models focus on inference for the fixed regression parameters, the random effects models jointly describe the mean and dependence using fixed and random regression parameters. Stiratelli, Laird and Ware (1984) elegantly extend to the multivariate binary setting the Laird and Ware (1982) mixed model for normally distributed repeated measures. Breslow and Clayton (1993) give a thorough account of random effects generalized linear models, and call them generalized linear mixed models. For right censored failure time data the random effects are referred to as frailties (Vaupel, Manton and Stallard, 1979). There is a rather extensive literature on so called shared frailty models with a simple covariance structure (e.g. Klein, 1992; Hougaard, 1991; Andersen et al., 1993). Below we present the generalized linear mixed model and the frailty model in general terms, and proceed to discuss estimation and inference.

3.1 The model

The generalized linear mixed model: Let Y_i, for $i = 1, \ldots, n$, denote the observation on unit i. Let $\boldsymbol{\beta}$ denote a p -vector of unknown fixed effect parameters, with an associated known design vector $\boldsymbol{X}_i$ for unit i. Let $\boldsymbol{b}_i$ denote a q -vector of unknown random effect parameters, with associated known design vector $\boldsymbol{Z}_i$. For given $\boldsymbol{b} = (b_1 \ldots b_n)$, the conditional distribution for Y_i is of exponential family form $p(Y_i \mid \gamma_i) = c_i(y_i) \exp(\gamma_i y_i - a(\gamma_i))$, with γ_i the canonical parameter, $a(.)$ a known monotone differentiable function, $E(Y_i \mid \gamma_i) = \mu_i = a'(\gamma_i)$ the mean parameter and $var(Y_i \mid \gamma_i) = v(\mu_i) = a''(\gamma_i)$ the variance function. Following (Breslow and Clayton, 1993; McCullagh and Nelder, 1989) we write the generalized linear model for unit i in the form

$$\begin{aligned} p(Y_i \mid \mu_i) &= \exp\left[\int_{y_i}^{\mu_i} \frac{y_i - u}{v(u)} du\right] \\ h(\mu_i) &= \boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{Z}_i \boldsymbol{b}, \qquad (1) \end{aligned}$$

with $h(.)$ a known, monotone, differentiable function linking the regression parameters to the mean. Conditionally on $\boldsymbol{b}$ the observations are assumed independent. At the second stage a distribution is imposed on $\boldsymbol{b}$, capturing the structure for between cluster heterogeneity and within cluster dependence as defined through the design vectors $\boldsymbol{Z}_i$. We assume that jointly $\boldsymbol{b} \sim p(\boldsymbol{b} \mid D(\boldsymbol{\theta}))$, with $\boldsymbol{\theta}$ a vector of unknown parameters which vary independently of $\boldsymbol{\beta}$.

The frailty model: Let T_i, for $i = 1, \ldots, n$, denote the event time, C_i the censoring time, $U_i = \min(T_i, C_i)$ and $\delta_i = I_{\{T_i \leq C_i\}}$. Given the random effects, or frailties $\boldsymbol{b} = (b_1 \ldots b_n)$, the event times are assumed independent and the conditional hazard function $\lambda_i(t)$ for unit i has the form

$$\lambda_i(t) = \lambda_0(t) \exp(\boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{Z}_i \boldsymbol{b}), \qquad (2)$$

with $\lambda_0(t)$ the baseline hazard and $b \sim p(\boldsymbol{b} \mid D(\boldsymbol{\theta}))$ as before.

For models (1) and (2) the random effects b may be viewed as a set of latent observations, and the model may be characterized as an incomplete data model. Besides making inferences about the regression parameters $\boldsymbol{\beta}$ and the variance component parameters θ, the purpose of the modelling is often to make predictions for the random effects $\boldsymbol{b}$.

3.2 The likelihood

Following the missing data terminology, the complete data are $(\boldsymbol{Y}, \boldsymbol{b})$, but only $\boldsymbol{Y}$ are observed. The observed data likelihood takes the form

$$p(\boldsymbol{Y} \mid \boldsymbol{\beta}, \boldsymbol{\theta}) = \int p(\boldsymbol{Y}, \boldsymbol{b} \mid \boldsymbol{\beta}, \boldsymbol{\theta}) d\boldsymbol{b} = \int p(\boldsymbol{Y} \mid \boldsymbol{\beta}, \boldsymbol{b})\, p(\boldsymbol{b} \mid \boldsymbol{\theta}) d\boldsymbol{b}, \tag{3}$$

and we write for model (1)

$$\log p(\boldsymbol{Y} \mid \boldsymbol{\beta}, \boldsymbol{b}) = \sum_{i=1}^{n} \left[\frac{y_i - \mu_i}{v(\mu_i)} \right], \tag{4}$$

with

$$h(\mu_i) = \boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{Z}_i \boldsymbol{b}.$$

For model (2) we denote the data by $\boldsymbol{Y} = (\boldsymbol{U}, \boldsymbol{\delta})$ and write

$$\log p(\boldsymbol{Y} \mid \lambda_0(t), \boldsymbol{\beta}, \boldsymbol{b}) = \sum_{i=1}^{n} \delta_i [\log \lambda_i(t)] - \exp[\Lambda_i(t)], \tag{5}$$

with

$$\begin{aligned} \lambda_i(t) &= \lambda_0(t) \exp(\boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{Z}_i \boldsymbol{b}) \\ \Lambda_i(t) &= \int_0^t \lambda_i(s) ds. \end{aligned}$$

4 Estimation and inference

For given b, the complete data log likelihood in (4) or (5) is easy to maximize, suggesting that the EM-algorithm is a natural choice for computing maximum likelihood estimates based on (3).

4.1 The EM algorithm

The EM algorithm finds the maximum of the observed data likelihood (3) by alternates between finding the expectation of the unobserved part of the data, given the observed data (E-step), and maximizing the complete data likelihood as if the non-observables were observed (M-step) (Dempster, Laird and Rubin, 1974). The random effects b are treated as unobserved data and they are imputed in the E-step. More precisely, for $\boldsymbol{\psi} = (\boldsymbol{\beta}, \boldsymbol{\theta})$ in model (4) and $\boldsymbol{\psi} = (\lambda_0(t), \boldsymbol{\beta}, \boldsymbol{\theta})$ in model (5), the E-step in iteration (r) involves the evaluation of

$$Q(\boldsymbol{\psi}, \boldsymbol{\psi}^{(r)}) = E[\log(p(\boldsymbol{Y}, \boldsymbol{b} \mid \boldsymbol{\psi})) \mid \boldsymbol{Y}, \boldsymbol{\psi}^{(r)}]$$

$$= \int \log(p(\boldsymbol{Y}, \boldsymbol{b} \mid \boldsymbol{\psi}))p(\boldsymbol{b} \mid \boldsymbol{Y}, \boldsymbol{\psi}^{(r)})d\boldsymbol{b}. \tag{6}$$

In the M-step the Q function is maximized with respect to $\boldsymbol{\psi}$ to obtain $\boldsymbol{\psi}^{(r+1)}$. The M-step equals maximization of the complete data log-likelihood (4) or (5), and standard software for the generalized linear model or the Cox model can be used, treating $Z_i b$ as an offset term. However, the elegance of the simple M-step is shadowed by the fact that the E-step in (6) involves an integral of the same dimension as in the observed data likelihood (3). A computatational problem thus remains, to which several solutions have been suggested, including penalized likelihood methods based on the Laplace approximation to the integral, and simulation based Monte Carlo EM and Markov chain Monte Carlo procedures.

4.2 Penalized likelihood

Breslow and Clayton (1993) derive a penalized likelihood solution for the generalized linear mixed model (4) assuming Gaussian random effects. We recapture their argument and present a parallel approximation for the semi-parametric frailty model (5) (Ripatti and Palmgren, 2000). For Gaussian random effects we have $p(\boldsymbol{b} \mid \boldsymbol{\theta}) \propto |D(\boldsymbol{\theta})|^{-\frac{1}{2}} \exp[-\frac{1}{2}\boldsymbol{b}'\boldsymbol{D}(\boldsymbol{\theta})^{-1}\boldsymbol{b}]$, and we write (3) in the form

$$c\,|\boldsymbol{D}|^{-\frac{1}{2}} \int \exp[-\kappa(\boldsymbol{b})]d\boldsymbol{b}.$$

with

$$\boldsymbol{\kappa}(\boldsymbol{b}) = \log p(\boldsymbol{Y} \mid \boldsymbol{\beta}, \boldsymbol{b}) - \frac{1}{2}\boldsymbol{b}'\boldsymbol{D}^{-1}\boldsymbol{b}. \tag{7}$$

Let $\boldsymbol{\kappa}'$ and $\boldsymbol{\kappa}''$ denote the q-vector and the $q \times q$ matrix of first- and second order partial derivatives of κ with respect to b. Ignoring the multiplicative constant c, the Laplace approximation to the marginal log likelihood takes the form

$$l(\boldsymbol{\beta}, \boldsymbol{\theta}) \approx -\frac{1}{2}\log|\boldsymbol{D}(\boldsymbol{\theta})| - \frac{1}{2}\log\left|\boldsymbol{\kappa}''(\tilde{\boldsymbol{b}})\right| - \boldsymbol{\kappa}(\tilde{\boldsymbol{b}},) \tag{8}$$

with $\tilde{\boldsymbol{b}} = \tilde{\boldsymbol{b}}(\boldsymbol{\beta}, \boldsymbol{\theta})$ the solution to $\boldsymbol{\kappa}'(\tilde{\boldsymbol{b}}) = 0$.

For the generalized linear mixed model Breslow and Clayton argue that if the variance function $v(\mu)$ varies slowly (or not at all) as a function of the mean μ, then the first two terms in (8) may be ignored. An approximate solution to the likelihood in (4) is thus obtained by maximizing $\boldsymbol{\kappa}(\boldsymbol{b})$ in (7), which corresponds to Green's penalized log likelihood (Green, 1987). Following the same rationale, Ripatti and Palmgren (2000) derive expressions for $\boldsymbol{\kappa}(\boldsymbol{b})$,$\boldsymbol{\kappa}'(\boldsymbol{b})$ and $\boldsymbol{\kappa}''(\boldsymbol{b})$ for the frailty model. They further show that for fixed θ the values $\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}), \hat{\boldsymbol{b}}(\boldsymbol{\theta})$, which maximize the penalized log likelihood (7) based on $\log p(\boldsymbol{Y} \mid \lambda_0(t), \boldsymbol{\beta}, \boldsymbol{b})$ in (5) also maximize the penalized partial log likelihood

$$\sum_{i=1}^{n} \delta_i \left((\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}) - \log \sum_{j \in R(t_i)} \exp(\boldsymbol{X}_j\boldsymbol{\beta} + \boldsymbol{Z}_j\boldsymbol{b}) \right) - \frac{1}{2}\boldsymbol{b}'\boldsymbol{D}(\boldsymbol{\theta})^{-1}\boldsymbol{b}. \tag{9}$$

For given $\boldsymbol{\theta}$, the estimating equations for $\boldsymbol{\beta}(\boldsymbol{\theta})$, $\boldsymbol{b}(\boldsymbol{\theta})$, based on the first partial derivatives of the penalized log likelihood (7) derived from the generalized linear mixed model (4) are of the form

$$\sum_{i=1}^{n} [y_i - \mu_i]\, \boldsymbol{X}_i = 0 \tag{10}$$

$$\sum_{i=1}^{n} [y_i - \mu_i]\, \boldsymbol{Z}_i - \boldsymbol{D}^{-1}\boldsymbol{b} = 0, \tag{11}$$

with $h(\mu_i) = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}$. The corresponding estimating equations for $\boldsymbol{\beta}(\boldsymbol{\theta})$, $\boldsymbol{b}(\boldsymbol{\theta})$ for the frailty model derived from (9) are

$$\sum_{i=1}^{n} \delta_i\, [1 - \nu_i]\, \boldsymbol{X}_i = 0 \tag{12}$$

$$\sum_{i=1}^{n} \delta_i\, [1 - \nu_i]\, \boldsymbol{Z}_i - \boldsymbol{D}^{-1}\boldsymbol{b} = 0, \tag{13}$$

with

$$\nu_i = \frac{\exp(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b})}{\sum_{j \in R(t_i)} \exp(\boldsymbol{X}_j\boldsymbol{\beta} + \boldsymbol{Z}_j\boldsymbol{b})}.$$

We find $\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}), \hat{\boldsymbol{b}}(\boldsymbol{\theta})$ by alternating between solving the equations (10) and (11) for the generalized linear mixed model, and between solving (12) and (13) for the frailty model. Note that solving (10) or (12) corresponds to the M-step in the EM-algorithm for β, and can be done with standard software for the generalized linear model or the Cox regression model, using estimated values of the random effects in an offset term. Maximizing the penalized likelihood (7) rather than the marginal likelihood (3) has replaced the awkward integral in the E-step with estimating equations (11) and (13), respectively.

Once $\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}), \hat{\boldsymbol{b}}(\boldsymbol{\theta})$ are computed, we update $\boldsymbol{\theta}$ in $\boldsymbol{D}(\boldsymbol{\theta})$ by maximizing the approximate profile likelihood derived from (8)

$$l(\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}), \boldsymbol{\theta}) \approx -\frac{1}{2}\log|\boldsymbol{D}(\boldsymbol{\theta})| - \frac{1}{2}\log\left|\boldsymbol{\kappa}''(\hat{\boldsymbol{b}})\right| - \frac{1}{2}\hat{\boldsymbol{b}}'\boldsymbol{D}(\boldsymbol{\theta})^{-1}\hat{\boldsymbol{b}}. \tag{14}$$

For the generalized linear mixed model Breslow and Clayton compute $\boldsymbol{\kappa}''$ in (14) from the likelihood (4), both with and without a REML adjustment for the degrees of freedom. For the frailty model

Ripatti and Palmgren (2000) compute $\boldsymbol{\kappa}''$ in (14) from the penalized partial likelihood (9) rather than from the full likelihood (5). The choice is motivated by the former performing better in simulations, and it is obtained as a side product from the previous iteration step.

4.3 Monte Carlo EM

Instead of alternating between (10) – (11) or (12) – (13) and (14) to obtain an approximate solution to (3), samples may be drawn from the predictive distribution $p(\boldsymbol{b} \mid \boldsymbol{Y}, \boldsymbol{\psi}^{(r)})$ in (6), and the sample mean computed instead of the expectation in the E-step of the EM-algorithm. The distribution $p(\boldsymbol{b} \mid \boldsymbol{Y}, \boldsymbol{\psi}^{(r)})$ is not a standard multivariate distribution, but rejection or importance sampling may be used (e.g. Gilks, Richardson and Spiegelhalter, 1996; Gelman, Carlin, Stern and Rubin, 1995). If enough samples are drawn, then the Monte Carlo EM-iterations converge to the maximum of the marginal likelihood (3). Booth and Hobert (1999) and Ripatti, Larsen and Palmgren (2001) suggest procedures where the number of samples is automatically increased when approaching the target, thus gaining absolute convergence for the MCEM algorithm.

4.4 Covariances for $\hat{\boldsymbol{\psi}}$

For $\boldsymbol{\psi} = (\boldsymbol{\beta}, \boldsymbol{\theta})$ in model (4) and $\boldsymbol{\psi} = (\lambda_0(t), \boldsymbol{\beta}, \boldsymbol{\theta})$ in model (5), we write the Louis' Louis's (1982) observed information

$$I(\boldsymbol{\psi}) = E\left(-\frac{\partial^2 \log(p(\boldsymbol{Y}, \boldsymbol{b} \mid \boldsymbol{\psi}))}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}'} \mid \boldsymbol{Y}, \hat{\boldsymbol{\psi}}\right) - \mathrm{var}\left(\frac{\partial \log(p(\boldsymbol{Y}, \boldsymbol{b} \mid \boldsymbol{\psi}))}{\partial \boldsymbol{\psi}} \mid \boldsymbol{Y}, \hat{\boldsymbol{\psi}}\right), \tag{15}$$

with $\mathrm{cov}(\hat{\boldsymbol{\psi}}) = I^{-1}(\boldsymbol{\psi})$, evaluated at $\boldsymbol{\psi} = \hat{\boldsymbol{\psi}}$. A discretized baseline hazard with jumps at distinct event times is used for $\lambda_0(t)$. Note that the conveniant procedure of computing the covariance matrix for $\hat{\beta}(\theta)$ from the estimating equations (10) or (12) neglects the additional variation stemming from the uncertainty in the estimated $\hat{\boldsymbol{\theta}}$. This additional variation is captured in the second term in the information matrix (15), and needs to be computed separately when using the penalized likelihood estimating equations. When using Monte Carlo EM both terms in (15) are obtained as a side product from the samples in the last iteration.

4.5 Posterior inference and MCMC

We make a conceptual shift and treat $\boldsymbol{\psi} = (\boldsymbol{\beta}, \boldsymbol{\theta})$ in model (1) and $\boldsymbol{\psi} = (\lambda_0(t), \boldsymbol{\beta}, \boldsymbol{\theta})$ in model (2) as random, and the data Y as fixed. We write the observed data posterior

$$p(\boldsymbol{\psi} \mid \boldsymbol{Y}) = \int p(\boldsymbol{\psi} \mid \boldsymbol{Y}, \boldsymbol{b}) p(\boldsymbol{b} \mid \boldsymbol{Y}) d\boldsymbol{b}. \tag{16}$$

Using Bayes' theorem the complete data posterior $p(\boldsymbol{\psi} \mid \boldsymbol{Y}, \boldsymbol{b}) p(\boldsymbol{b} \mid \boldsymbol{Y})$ inside the integral (16) is proportional to the product of a prior distribution $p(\boldsymbol{\psi})$ and the complete data likelihood $p(\boldsymbol{Y}, \boldsymbol{b} \mid \boldsymbol{\psi})$ in (3). If samples from $p(\boldsymbol{b} \mid \boldsymbol{Y})$ could be drawn easily, then it would be straight forward to evaluate the observed data posterior (16) as a Monte Carlo mean. We write $p(\boldsymbol{b} \mid \boldsymbol{Y})$ as

$$p(\boldsymbol{b} \mid \boldsymbol{Y}) = \int p(\boldsymbol{b} \mid \boldsymbol{Y}, \boldsymbol{\psi}) p(\boldsymbol{\psi} \mid \boldsymbol{Y}) d\boldsymbol{\psi}. \tag{17}$$

From the symmetry of the expressions in (16) and (17) an iterative two-step algorithm is suggested, involving an imputation step (I-step), with draws $\boldsymbol{b}^{(r)}$ from the conditional predictive distribution $p(\boldsymbol{b} \mid \boldsymbol{Y}, \boldsymbol{\psi}^{(r)})$ in (17), and a posterior step (P-step), with draws $\boldsymbol{\psi}^{(r+1)}$ from the conditional posterior distribution $p(\boldsymbol{\psi} \mid \boldsymbol{Y}, \boldsymbol{b}^{(r)})$ in (16). Under broad regularity conditions the sequence $\{\boldsymbol{\psi}^{(r)}, \boldsymbol{b}^{(r)}, r = 1, 2, \ldots\}$ converges to the joint posterior $p(\boldsymbol{\psi}, \boldsymbol{b} \mid \boldsymbol{Y})$, and the sequences of the components to their respective marginal posteriors $p(\boldsymbol{\beta} \mid \boldsymbol{Y})$, $p(\boldsymbol{b} \mid \boldsymbol{Y})$ and $p(\boldsymbol{\theta} \mid \boldsymbol{Y})$ (Gilks, Richardson and Spiegelhalter, 1996). For the hazard model Clayton (1991) discusses how to sample from the non-parametric distribution for the conditional baseline hazard $\lambda_0(t)$. Note that sampling in the P-step (I-step) depends on the previous I-step (P-step), but given the previous I-step (P-step) is conditionally independent of the previous P-step (I-step). This motivates the terminology Markov chain Monte Carlo. The I-step and P-step may be seen as stochastic counterparts to the E-step and M-step of the EM-algorithm. For large samples the likelihood will overrule the prior, and the mode and the curvature of the posterior (16) will coincide with the mode and the curvature of the likelihood (3). Note that per definition the credible intervals for $p(\boldsymbol{\beta} \mid \boldsymbol{Y})$ and $p(\boldsymbol{b} \mid \boldsymbol{Y})$ include the uncertainty in $\boldsymbol{\theta}$. For specific problems there is an extensive literature on clever choices of conditional distributions that are easy to sample from, and on computational tricks to speed up the sampling process and to ensure that all parts of the parameter space are covered (Gelman and Rubin, 1992).

Table 1: Parameter estimates and standard errors for two models for the rose survival data based on the MCEM algorithm and penalized partial likelihood (PPL) estimating equations.

	Model 1		Model 2	
Parameter	MCEM	PPL	MCEM	PPL
Treatment A	0.23(0.16)	0.24(0.16)	0.26(0.16)	0.28(0.17)
Treatment B	0.38(0.15)	0.39(0.16)	0.40(0.16)	0.41(0.17)
Treatment D	0.25(0.16)	0.25(0.16)	0.27(0.16)	0.22(0.17)
Block 1	-0.49(0.17)	-0.48(0.17)	-0.51(0.16)	-0.50(0.18)
Block 2	-0.50(0.15)	-0.51(0.16)	-0.52(0.15)	-0.51(0.16)
Block 3	-0.27(0.15)	-0.28(0.15)	-0.29(0.16)	-0.29(0.15)
$\hat{\theta}$	0.18(0.07)	0.22(0.06)		
$\hat{\theta}_A$			0.12(0.10)	0.30(0.13)
$\hat{\theta}_B$			0.15(0.11)	0.27(0.11)
$\hat{\theta}_C$			0.32(0.13)	0.21(0.12)
$\hat{\theta}_D$			0.11(0.10)	0.12(0.09)

5 Data analyses

5.1 Lifetime of roses

In the first example, we study data from a greenhouse experiment on vase lifetimes of cv. Frisco rose cuts. This is an incomplete randomized block design with four blocks, three plots in each block and eight plants per plot. From each plant, several rose cuts were picked and put to a vase, and for each cut the lifetime in the vase was recorded. There were total of 716 cuts with 3 censored lifetimes because of bent rose necks. There were four different lighting treatments randomized within blocks, and the primary interest is to study the effects of treatments on the average vase life as well as on the between plant variation. The details of the experiment are reported in Särkkä, Rita and Ripatti (2000).

We fit two different models to these data. The first is a shared frailty model

$$\lambda_{ij}(t) = \lambda_0(t)\exp(X_{1ij}\beta_1 + X_{2ij}\beta_2 + b_i), \tag{18}$$

where $i = 1, \ldots, 224$ for plant i and $j = 1, \ldots, n_i; 1 \leq n_i \leq 10$ for cut j within plant i, X_{1ij} is a vector indicating which of the four blocks the plant belongs to and X_{2ij} which of the four treatments is allocated to cut j in plant i. The random effects are assumed to be independent realizations from a normal distribution, i.e. $b_i \sim N(0, \theta)$. The second model allows the frailty variances to differ between the four treatments, i.e. the covariance matrix for $b = (b_1, \ldots, b_{224})$ is diagonal with variances $\theta_A, \theta_B, \theta_C, \theta_D$ depending on the treatment allocation for the respective cut.

Table 1 shows estimates and standard errors for the parameters in the two models based on the MCEM algorithm and on the penalized partial likelihood estimating equations. For both models and estimation methods treatment B gives the shortest lifetime and treatment C the longest. When the model allows for differential variability, then the MCEM fit indicates that the lifetimes of roses treated with C vary the most. The difference between the variance component estimates is not, however, significant, and differential variability does not show in the penalized likelihood fit. The rose data are discussed in more detail in Ripatti, Larsen and Palmgren (2001).

5.2 Alcohol related mortality in Finland

In the second example we smooth alcohol related mortality rates in 452 Finnish municipalities, using Bayesian GLMM (for details of the study, see Mäkelä, Ripatti and Valkonen (2001)). The observed number of deaths O_i in municipality i are assumed to follow a Poisson distribution with expectation $\mu_i, i = 1, \ldots, 452$. Each μ_i is assumed to depend log-linearly on the logarithm of the expected number of deaths E_i and a municipality specific random effect b_i

$$\log(\mu_i) = \log(\mathrm{E}_i) + b_i. \tag{19}$$

The expected mortality E_i is computed based on the size and structure of the population in the municipality, and it is treated as fixed. Conditionally on E_i and b_i, the observed counts O_i, for $i = 1, \ldots, 452$, are assumed independent. For the random effects b_i, a Markov random field prior (Besag, York, and Mollié, 1991) is specified, with mean equal to the average of the effects from municipalities immediately adjacent to municipality i. The variance function for the random effects b_i is set to θ/k_i,

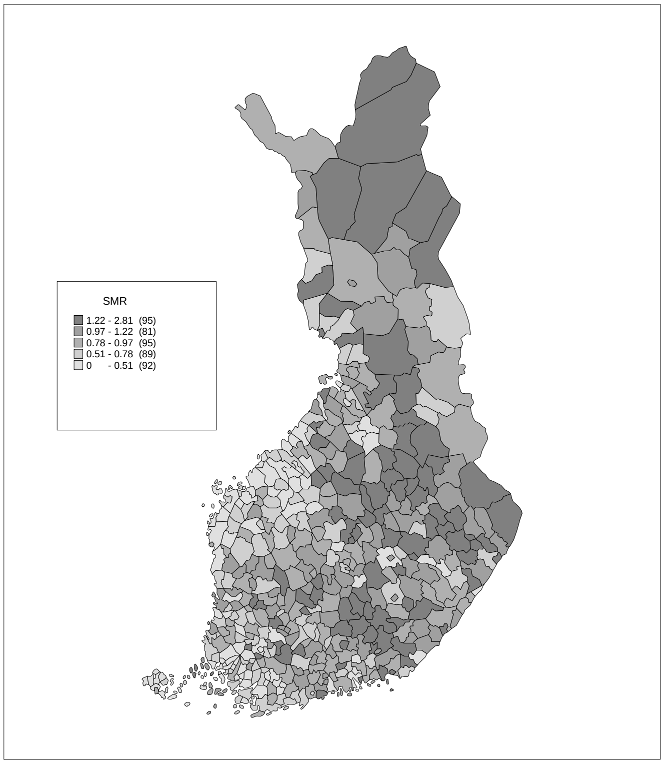

Figure 1: Raw standardized mortality ratios in Finnish municipalities.

where k_i is the number of municipalities adjacent to i, and θ is a random term, with $1/\theta$ following a gamma-distribution $\Gamma(1, 1)$. Gibb's sampling is used to draw from the posterior distribution $p(\theta, b \mid Y)$. Raw and smoothed standardized mortality ratios ($\mathrm{SMR}_i = \mu_i/\mathrm{E}_i$) based on posterior means are plotted in Figures 1 and 2, respectively. The more extreme SMR's in Figure 1 are smoothed in Figure 2, and a clear pattern of high mortality is shown in Northern and Eastern Finland, with lower rates in the West.

6 Discussion

We emphasise the parallel approaches to estimation and inference for the generalized linear mixed model and the frailty model. In our treatment of the frailty model the baseline hazard is profiled out. In penalized likelihood this is done following the profiling argument for the partial likelihood in the Cox model (Johansen, 1983). In the MCEM fit the complete data log likelihood is log $p(\boldsymbol{Y} \mid \lambda_0(t), \boldsymbol{\beta}, \boldsymbol{\theta})$ in (5), and the M-step involves the standard partial likelihood procedure together with the Breslow estimator (Breslow, 1974) for the cumulative hazard. Sampling in the E-step may be done using an

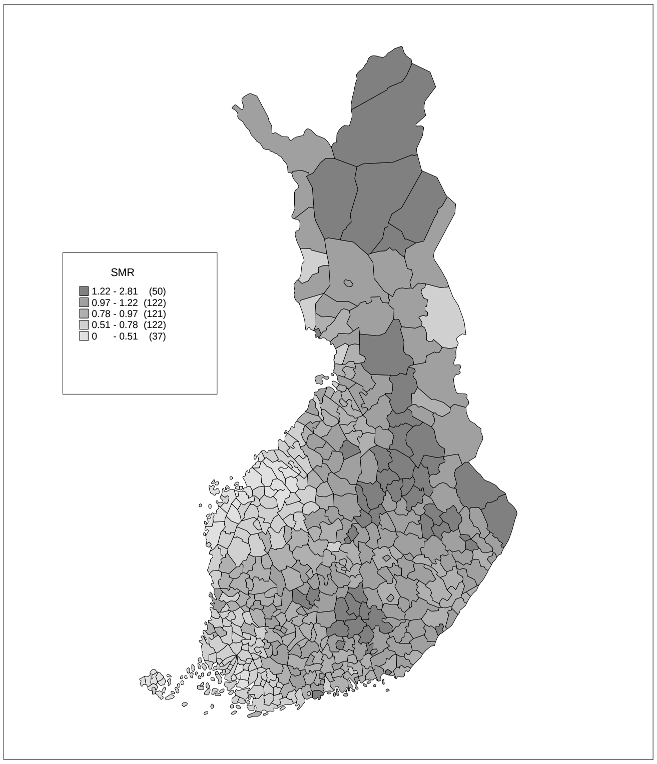

Figure 2: Posterior modes of the estimated standardized mortality ratios.

independent sampler (Ripatti, Larsen and Palmgren, 2001) or a dependent sampler (Vaida and Xu, 2000). For the Bayesian Markov chain Monte Carlo procedure an independent increments gamma-process may be used as the conditional distribution for the baseline hazard (Clayton, 1991). Note that although we derive the penalized likelihood estimating equations assuming Gaussian random effects, other distributions may be used for the MCEM and the MCMC procedures. All estimation and inference approaches are computer intensive. None can be singled out as universally best, but they all give acceptable accuracy over a wide range of conditions. The likelihood methods in sections 4.1 – 4.5 are justified by large sample arguments. The penalized likelihood estimating equations are computationally simpler than the other methods, and they have been shown to perform well in many situations. A separate routine is, however, needed to give standard errors for the estimates, whereas Monte Carlo sampling in the E-step of the EM-algorithm provides the Louis' observed information matrix as a side product. The posterior procedures are conceptually attractive, void of ad hoc fixes. If there are plentiful of data, then likelihood inference and posterior inference will give similar results. The Bayesian approach to inference is, however, valid also when data are sparse, and the possibility to include informative priors allows external information to be added to the model in a coherent way. The overruling difficulty with the posterior MCMC sampling is to assess convergence. In contrast, the likelihood is monotonically increasing in each iteration, and assessing convergence is a non-issue. This applies to the MCEM procedure provided the Monte Carlo error is small.

By adding layers of random effects to the linear predictor of the generalized linear model or the multiplicative hazard model, a large and flexible class of models is offered for empirical use. Complex hierarchical structures and missing data constitute natural parts of the model specification. Although estimation and inferences are not straight forward, a unified and reasonably well understood framework is emerging. When incorporated into the applied statisticians toolbox this large class of models allows increased freedom and flexibility to tailor the statistical framework to the applied problem at hand. Formal or informal procedures to assess the sensitive of results to the model structure and distributional assumptions should be part of the toolbox.

References

Andersen, P. K., Borgan, O., Gill, R. D. and Keiding, N. (1993) *Statistical models based on counting processes*. Berlin: Springer-Verlag.

Besag, J.E., York, J.C., and Mollié (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion). *Ann. Inst. Statist. Math.* **43**, 1–59.

Booth, J. G. and Hobert, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society B* **61**, 265–285.

Breslow, N.E. (1974). Covariance analysis of censored survival data. *Biometrics* **30**, 89 – 99.

Breslow, N.E. (1984). Extra-Poisson variation in log linear models. *Applied Statistics* **33**, 38 – 44.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear models. *Journal of the American Statististical Association* **88**, 9–25.

Clayton, D. G. (1991). A Monte Carlo method for Bayesian inference in frailty models. *Biometrics* **47**, 467–485.

Clayton, D. (1994). Some aproaches to the analysis of recurrent event data. *Statistical Methods in Medical Research* **3**, 244–262.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society B* **34**, 187–220.

Dempster, A. P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via EM algorithm (with discussion). *Journal of the Royal Statistical Society B* **39**, 1–38.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis.* London: Chapman & Hall.

Gelman, A and Rubin D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* **7**, 457–511.

Gilks, W.R., Richardson, S, and Spiegelhalter, D.J. (1996). *Markov chain Monte Carlo in practice.* London: Chapman and Hall.

Green, P. J. (1987). Penalized likelihood for general semi-parametric regression model. *International Statististical Review* **55**, 245–259.

Hougaard, P. (1991). Modelling heterogeneity in survival data. *Journal of Applied Probability* **28**, 695 – 701.

Johansen, S. (1983). An extension of Cox's regression model. *International Statistical Review* **51**, 158–262.

Klein, J. P. (1992). Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics* **48**, 795–806.

Laird N.M. and Ware J.H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963–74.

Liang, K-Y. and Zeger, SL. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 11–22.

Louis, T.A. (1982). Finding observed information using the EM algorithm. *Journal of the Royal Statistical Society B* **44**, 98 – 130.

Mäkelä, P., Ripatti, S, and Valkonen T. (2001). Alue-erot modesten alokoholikoulleisuudessa. *Suomen Lääkärilehti* in press.

McCullagh, P., and Nelder, J.A. (1989). *Generalized Linear Models* (2nd ed.). London: Chapman and Hall.

Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized Linear Models *Journal of the Royal Statistical Society A* **135**, 370–384.

Ripatti, S., Larsen K., and Palmgren J. (2001) Maximum likelihood inference for multivariate frailty models using a Monte Carlo EM Algorithm. *submitted*.

Ripatti, S. and Palmgren, J. (2000). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics* **56** 1016-1022.

Särkkä, L. E., Rita, H.J., and Ripatti, S. (2000). Cut rose flower longevity and its variation between plants of cv. Frisco grown in different lighting periods. *submitted*.

Stiratelli R., Laird N.M., Ware H. (1984). Random-effects models for serial observations with binary response. *Biometrics* **40**, 961–71.

Vaida, F. and Xu, R. (2000). Proportional hazards model with random effects. *Statistics in Medicine* **19**, 3309–3324.

Vaupel, J. W., Manton, K. G. and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* **16**, 439–454.

Wei, L. J., Lin, D. Y. and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modelling marginal distributions. *Journal of the American Statistical Association* **84**, 1065–1073.

Williams, D.A. (1982). Extra-binomial variation in logistic linear models. *Applied Statistics 1982* , 144–148.

Zhao, L.P. and Prentice, R.L. (1989). Correlated binary regession using a quadratic exponential model. *Biometrika* **77**, 642 –28.