Communications in Statistics - Simulation and Computation
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lssp20

The Use of Working Variables in the Bayesian Modeling of Mean and Dispersion Parameters in Generalized Nonlinear Models with Random Effects

Andres Gutierrez[a]
[a] Universidad Santo Tomas, Bogota, Colombia

Accepted author version posted online: 13 Nov 2013.Published online: 28 Jul 2014.

To cite this article: Andres Gutierrez (2015) The Use of Working Variables in the Bayesian Modeling of Mean and Dispersion Parameters in Generalized Nonlinear Models with Random Effects, Communications in Statistics - Simulation and Computation, 44:1, 168-195, DOI: 10.1080/03610918.2013.770529

To link to this article: http://dx.doi.org/10.1080/03610918.2013.770529

Communications in Statistics—Simulation and Computation®, 44: 168–195, 2015

ISSN: 0361-0918 print / 1532-4141 online
DOI: 10.1080/03610918.2013.770529

The Use of Working Variables in the Bayesian Modeling of Mean and Dispersion Parameters in Generalized Nonlinear Models with Random Effects

ANDRES GUTIERREZ

Universidad Santo Tomas, Bogota, Colombia

This article is aimed at reviewing a novel Bayesian approach to handle inference and estimation in the class of generalized nonlinear models. These models include some of the main techniques of statistical methodology, namely generalized linear models and parametric nonlinear regression. In addition, this proposal extends to methods for the systematic treatment of variation that is not explicitly predicted within the model, through the inclusion of random effects, and takes into account the modeling of dispersion parameters in the class of two-parameter exponential family. The methodology is based on the implementation of a two-stage algorithm that induces a hybrid approach based on numerical methods for approximating the likelihood to a normal density using a Taylor linearization around the values of current parameters in an MCMC routine.

Keywords Bayesian analysis; Generalized linear models; Heteroscedasticity; MCMC; Mixed effects models; Nonlinear regression.

Mathematics Subject Classification 62F15; 62J02.

1. Introduction

Let us consider the nonlinear regression model

$$y_i = f(\mathbf{x}_i, \boldsymbol{\beta}) + e_i, \quad i = 1, \ldots, n \tag{1}$$

where y_i is the response variable, $\mathbf{x}_i = (x_{i1}, \ldots, x_{iP})'$ is a vector of known regression constants related to the interest (response) variable through a vector of fixed unknown parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_P)'$, $f(\cdot, \cdot)$ is a nonlinear differentiable function of $\boldsymbol{\beta}$, and the errors e_i are assumed to be independent with common mean zero, but not necessarily identically distributed. At the same time, we assume that the distribution of y_i is in the exponential family with likelihood given by the following general expression:

$$p(y_i \mid \theta_i) = a(y_i) \exp\{d(\theta_i)T(y_i) + c(\theta_i)\},$$

where θ_i is the parameter of interest, and the functional forms of $a(\cdot)$, $d(\cdot)$, $T(\cdot)$ and $c(\cdot)$ are supposed to be known. If y_i is a continuous random variable, then p is assumed to be a

Received June 9, 2011; Accepted January 21, 2013
Address correspondence to Andres Gutierrez, Universidad Santo Tomas, Bogota, Colombia; E-mail: hugogutierrez@usantotomas.edu.co

density with respect to the Lebesgue measure; if y_i is discrete-type, then p is a density with respect to the counting measure. Consequently, the mean of y_i is then given by

$$E(y_i) = \mu_i = h^{-1}(f(\mathbf{x}_i, \boldsymbol{\beta})), \tag{2}$$

where $h(\cdot)$ is a link function (McCullagh and Nelder, 1996). Under the former assumptions, we say that y_i follows a generalized nonlinear model. Note that if we let $f(\mathbf{x}_i, \boldsymbol{\beta}) = \mathbf{x}_i'\boldsymbol{\beta}$, then y_i follows the so-called generalized linear model (GLM). However, typical situations involve overdispersed data and the extra variation unaccounted for in the above model can be accommodated with the inclusion of some random effects (Zhao et al., 2006). This way, the structural form of the mean model should include some other terms involving parameters for those random effects and their associated covariates, if any. On the other hand, as claimed by Dey et al. (1997, p. 98), another way to force overdispersion is by considering a more general case of the exponential family and modeling the dispersion parameter in terms of random components that affect the variability of the response variable in the model. This article explores a Bayesian approach to estimate the parameters of the mean for random and fixed effects, while taking into account the particular variation of the responses, by jointly modeling the dispersion parameter of the density.

In this direction, a classical approach in normal regression analysis was proposed by Aitkin (1987); later a Bayesian methodology was propounded by Cepeda and Gammerman (2001) using working variables as in Gammerman (1997). Recently, Cepeda and Gammerman (2005) have reported some work based on a Bayesian methodology for modeling parameters in the two-parameter exponential family involving linearity of the mean. That linearity assumption was relaxed by Cepeda and Achcar (2010) in a paper dealing with heteroscedastic nonlinear normal models. The Bayesian counterpart of those approaches may be seen as particular cases of the results of this research, because the resulting algorithms are implicit in the general MCMC methodology proposed here.

The generalized linear mixed model (GLMM) combines the GLM and linear mixed model, and the Bayesian fitting procedure is based on the EM algorithm and MCMC methods as described by McCulloch and Searle (2001). Smyth (2002) modeled the variance parameter in heteroscedastic linear models, proposing an algorithm for restricted maximum likelihood estimation. The author uses a Levenberg–Marquardt modification in order to ensure that the likelihood increases iteration by iteration.

Gijbels et al. (2005) dealt with extended double exponential family models in the context of nonparametric techniques. Specifically, the authors use the P-spline approach to estimate the dispersion function and the mean function. This approach is quite flexible in the sense that it can handle both overdispersion and underdispersion phenomena and even a mixture of both. Rigby and Stasinopoulos (2005) presented the generalized additive model for location, scale, and shape (GAMLSS), where the assumption of an exponential family is relaxed and more general distributions could be considered. The authors present two algorithms based on the Newton–Raphson and Fisher scoring algorithm for estimation.

The recent work of Neykov et al. (2012) presents a robust modification of the classical maximum likelihood and extended quasi-likelihood for joint modeling of the mean and the dispersion. Specifically, the authors use the maximum trimmed likelihood estimator and the maximum extended trimmed quasi-likelihood (ETQL) estimators. In large datasets, the problem of joint modeling of the mean and dispersion functions is handled by the neural network modeling of mean and dispersion. These models generally have good predictive performance because of their flexibility (Hastie et al., 2001). Charalambous et al. (2010) presented a variable selection method for joint modeling of the mean and variance functions

in hierarchical generalized linear models, based on the penalized likelihood, the SCAD penalty.

This article provides a unifying approach to Bayesian inference in nonlinear mixed effect models, which is an extension of the methodology proposed by Gammerman (1997). This article adopts a computational strategy that is aimed at approximating the first stage likelihood to a normal distribution in order to facilitate an MCMC implementation based on the Metropolis–Hastings algorithm. The simulations and the empirical application provide an illustration for the method. Finally, one novel aspect is the modeling of dispersion using the parameterization proposed by Dey et al. (1997). Thus, we reduce the problem to the case when the mean and the dispersion parameters are orthogonal. We note that there are rather few research papers that address this issue in the Bayesian setup, although it is a standard component in the h-likelihood approach of Lee and Nelder (1996) and Lee et al. (2006).

After a brief introduction, Section 2 introduces the proposed approach under the uniparametric exponential family. That section provides the basis for the developments and proposed algorithms throughout the article. Section 3 deals with the problem of overdispersion in a two-parameter exponential family in two ways: first, we consider random effects in the mean model, where a standard generalized nonlinear model is employed, but for each unit a random effect is added to the fixed term resulting in a generalized nonlinear mixed model. Second, and not necessarily independent of the previous approach, we model the dispersion parameter in the two-parameter exponential family to create an overdispersed generalized nonlinear (mixed) model. One important practical issue is the monotonicity of the function that links covariates with mean and dispersion parameters. Section 4 shows the results of some empirical simulations carried out to check the performance of the methodology for the estimation of the parameters involved in the joint modeling of mean and dispersion. In Section 5, we undertake an analysis of a dataset recently reported, involving schooling rates in Colombia, by using some models considering a nonlinear function of the mean and different link functions. In that section, we report some comparisons of the proposed algorithm with other existing MCMC methods considered in dealing with this problem. Finally, Section 6 is devoted to a discussion of the results, some possible extensions, and further research in more intricate models. The proofs of the results and some suitable computational codes are given in the Appendix.

2. The Proposed Approach to Model Mean Parameters

As is usual in Bayesian statistics, the specification of any model requires assigning prior distributions for the parameters of interest. In particular, we assign the following prior density $p(\boldsymbol{\beta})$ to the parameters of the mean:

$$\boldsymbol{\beta} \sim N(\mathbf{b}, \mathbf{B}).$$

Then, following the Bayes rule, the posterior density of $\boldsymbol{\beta}$ is given by

$$\begin{aligned} p(\boldsymbol{\beta} \mid \mathbf{y}) &\propto \prod_{i=1}^{n} p(y_i \mid \theta_i) p(\boldsymbol{\beta}) \\ &\propto \exp\left\{ \sum_{i=1}^{n} [d(\theta_i) T(y_i) + c(\theta_i)] - \frac{1}{2} (\boldsymbol{\beta} - \mathbf{b})' \mathbf{B}^{-1} (\boldsymbol{\beta} - \mathbf{b}) \right\}. \end{aligned} \tag{3}$$

This distribution is not easy to sample from, because it does not have a known form and generally it is not log-concave. Given the generality of the nonlinear generalized model, we propose a two-stage hybrid algorithm based on the methodology developed by Gammerman (1997) and the iteratively re-weighted least squares (IRLS) method. At this stage, we are able to propose a suitable working variable which approximates the likelihood to a normal distribution. In the second stage, and given the properties of this proposal, we implement a Taylor linearization in the new pseudo-likelihood so that the hybrid algorithm can fully perform. The results of this two-stage technique can be seen as a generalization of many recent methods that perform Bayesian inference for this kind of model. The reader will note that the functional form of this approach looks entirely standard and can be considered as an extension of the method used in the seminal paper by Dellaportas and Smith (1993) on MCMC for generalized linear models.

2.1. *First Stage: Normal Approximation to the Likelihood*

Result 2.1 (Adaptation of IRLS algorithm)**.** *Assuming that the current value of $\boldsymbol{\beta}$ is $\boldsymbol{\beta}^{(c)}$, a suitable working variable, resulting from the Fisher scoring technique, that approximates the likelihood to a normal distribution is given by*

$$\dot{y}_i = f(\mathbf{x}_i, \boldsymbol{\beta}) + h'(h^{-1}[f(\mathbf{x}_i, \boldsymbol{\beta})])(y_i - h^{-1}[f(\mathbf{x}_i, \boldsymbol{\beta})]), \tag{4}$$

for $i = 1, \ldots, n$. This working variable leads to a Gaussian distribution with mean $f(\mathbf{x}_i, \boldsymbol{\beta})$ and variance $\tilde{V}_i = (h'(h^{-1}(\mathbf{x}_i'\boldsymbol{\beta})))^2 Var(y_i)$.

Following the path of a Bayesian analysis, the next step would be to use the Bayes rule in order to find an approximation to the full posterior distribution. This approximation is given by the following expression:

$$\begin{aligned} p(\boldsymbol{\beta} \mid \mathbf{y}) &\propto p(\mathbf{y} \mid \boldsymbol{\beta}) p(\boldsymbol{\beta}) \\ &\approx p(\dot{\mathbf{y}} \mid \boldsymbol{\beta}) p(\boldsymbol{\beta}) \\ &= \exp\left\{ \frac{-1}{2} [(\dot{\mathbf{y}} - f(\mathbf{x}_i, \boldsymbol{\beta}))' \tilde{\mathbf{V}}^{-1} (\dot{\mathbf{y}} - f(\mathbf{x}_i, \boldsymbol{\beta})) + (\boldsymbol{\beta} - \mathbf{b})' \mathbf{B}^{-1} (\boldsymbol{\beta} - \mathbf{b})] \right\}. \end{aligned}$$

The latter expression does not have closed form and it is not easy to take samples from. However, this is the first step of the algorithm and, observing in detail this approximated posterior density, one could propose for something to be done with the mean of the working variable. The next section deals in detail with this problem and its solution.

2.2. *Second Stage: Linear Approximation of the Mean*

Result 2.2. *If $\boldsymbol{\beta}^{(c)}$ is the current value of $\boldsymbol{\beta}$, the function $f(\cdot, \cdot)$ can be approximated by means of the following expression:*

$$f(\mathbf{x}_i, \boldsymbol{\beta},) \cong f\left(\mathbf{x}_i, \boldsymbol{\beta}^{(c)}\right) + \tilde{\mathbf{x}}_i \left(\boldsymbol{\beta} - \boldsymbol{\beta}^{(c)}\right), \tag{5}$$

where

$$\tilde{\mathbf{x}}_i = \nabla f\left(\mathbf{x}_i, \boldsymbol{\beta}^{(c)}\right) = \left[\frac{\partial f}{\partial \beta_1}, \ldots, \frac{\partial f}{\partial \beta_P} \right]\Bigg|_{\boldsymbol{\beta} = \boldsymbol{\beta}^{(c)}}. \tag{6}$$

Once we have found a convenient approximation to the nonlinear function f, we could propose another working variable so that its distribution has a closed linear normal form. In this manner, the new working variable may be combined with the prior distribution of $\boldsymbol{\beta}$ in order to provide a closed proposal for the posterior distribution which can be used in a Metropolis-type algorithm in order to draw samples from the original posterior distribution.

Result 2.3. Combining the prior distribution of the vector of parameters $\boldsymbol{\beta}$ with the normal approximation induced by the following working variable:

$$\tilde{y}_i = \tilde{\mathbf{x}}_i \boldsymbol{\beta}^{(c)} + h'\big(h^{-1}\big[f\big(\mathbf{x}_i, \boldsymbol{\beta}^{(c)}\big)\big]\big)\big(y_i - h^{-1}\big[f\big(\mathbf{x}_i, \boldsymbol{\beta}^{(c)}\big)\big]\big), \tag{7}$$

which is normally distributed with mean $\tilde{\mathbf{x}}_i \boldsymbol{\beta}^{(c)}$ and variance $\tilde{V}_i$, a suitable proposal distribution for the full posterior (3) with Gaussian kernel $q_{\boldsymbol{\beta}}$ is given by

$$q_{\boldsymbol{\beta}}\big(\boldsymbol{\beta}^{(c)}\big) = N(\mathbf{b}^*, \mathbf{B}^*), \tag{8}$$

where,

$$\mathbf{B}^* = (\mathbf{B}^{-1} + \tilde{\mathbf{X}}'\tilde{\mathbf{V}}^{-1}\tilde{\mathbf{X}})^{-1},$$
$$\mathbf{b}^* = \mathbf{B}^*(\mathbf{B}^{-1}\mathbf{b} + \tilde{\mathbf{X}}'\tilde{\mathbf{V}}^{-1}\tilde{\mathbf{y}}),$$

and $\tilde{\mathbf{X}} = (\tilde{\mathbf{x}}_1', \ldots, \tilde{\mathbf{x}}_n')$, $\tilde{\mathbf{y}} = (\tilde{y}_1, \ldots, \tilde{y}_n)'$.

2.3. *Some Useful Models*

2.3.1. *Normal nonlinear regression.* If y_i follows a normal distribution with unknown mean but known variance, then the working variable is given by $\tilde{y}_i = \tilde{\mathbf{x}}_i \boldsymbol{\beta}^{(c)} + y_i - f(\mathbf{x}_i, \boldsymbol{\beta}^{(c)})$ and $\tilde{V}_i = \mathrm{Var}(y_i) = \sigma^2$.

2.3.2. *Binomial nonlinear regression.* Suppose that $y_i \sim \mathrm{Binomial}(n_i, \pi_i)$, but our interest is focused on modeling the success probabilities $p_i = y_i/n_i$. Thus $\mu_i = \pi_i$ and, if we consider the natural link function $h(\mu_i) = logit(\mu_i) = logit(\pi_i)$, its inverse is given by $h^{-1}(v) = e^v/(1+e^v)$, and its first derivative $h'(v) = 1/(v(1-v))$, then the working variable is given by

$$\tilde{p}_i = \tilde{\mathbf{x}}_i' \boldsymbol{\beta}^{(c)} + \frac{\big(1 + \exp\big(f\big(\mathbf{x}_i', \boldsymbol{\beta}^{(c)}\big)\big)\big)^2}{\exp\big(f\big(\mathbf{x}_i', \boldsymbol{\beta}^{(c)}\big)\big)} p_i - \exp\big(f\big(\mathbf{x}_i', \boldsymbol{\beta}^{(c)}\big)\big) - 1.$$

In addition, noting that $\mathrm{Var}(p_i) = \pi_i(1-\pi_i)/n_i = \mu_i(1-\mu_i)/n_i$, then $\tilde{V}_i$ will be

$$\begin{aligned}\tilde{V}_i &= \big[h'\big(h^{-1}\big(f\big(\mathbf{x}_i', \boldsymbol{\beta}^{(c)}\big)\big)\big)\big]^2 \mathrm{Var}(p_i) \\ &= [h'(\mu_i)]^2 \frac{\mu_i(1-\mu_i)}{n_i} = \frac{1}{n_i \mu_i (1-\mu_i)} = \frac{\big(1 + \exp\big(f\big(\mathbf{x}_i', \boldsymbol{\beta}^{(c)}\big)\big)\big)^2}{n_i \exp\big(f\big(\mathbf{x}_i', \boldsymbol{\beta}^{(c)}\big)\big)}.\end{aligned}$$

2.3.3. *Poisson nonlinear regression.* If $y_i \sim \mathrm{Poisson}(a_i)$, then $\mu_i = a_i$, and suppose we use the natural link function $h(\mu_i) = \log(\mu_i)$; then the working variable is given by

$$\tilde{y}_i = \tilde{\mathbf{x}}_i' \boldsymbol{\beta}^{(c)} + \frac{y_i}{\exp\big(f\big(\mathbf{x}_i', \boldsymbol{\beta}^{(c)}\big)\big)} - 1.$$

Also, noting that $\text{Var}(y_i) = a_i = \mu_i = h^{-1}(f(\mathbf{x}_i', \boldsymbol{\beta}^{(c)}))$, we conclude that $\tilde{V}_i$ takes the following form:

$$\begin{aligned}\tilde{V}_i &= \left[h'\left(h^{-1}\left(f\left(\mathbf{x}_i', \boldsymbol{\beta}^{(c)}\right)\right)\right)\right]^2 Var(y_i) \\ &= \left(\frac{1}{\exp\left(f\left(\mathbf{x}_i', \boldsymbol{\beta}^{(c)}\right)\right)}\right)^2 \exp\left(f\left(\mathbf{x}_i', \boldsymbol{\beta}^{(c)}\right)\right) = \frac{1}{\exp\left(f\left(\mathbf{x}_i', \boldsymbol{\beta}^{(c)}\right)\right)}.\end{aligned}$$

2.4. *The MCMC Algorithm for Mean Parameters Estimation*

Based on the previous proposal, a Metropolis–Hastings algorithm (Hastings, 1970; Metropolis et al., 1953) could be carried out in order to simulate values from the full posterior distribution (3) and from the sampling of (8). This algorithm is described below.

1. Set the iteration counter chain to $j = 1$.
2. Set initial values $\boldsymbol{\beta}$ given by $\boldsymbol{\beta}^{(j-1)}$.
3. Propose a new value, $\boldsymbol{\phi}_{\boldsymbol{\beta}}$, generated from the jumping proposal distribution (8).
4. Compute the acceptance probability of the movement. If it is accepted, then $\boldsymbol{\beta}^{(j)} = \boldsymbol{\phi}_{\boldsymbol{\beta}}$; otherwise $\boldsymbol{\beta}^{(j)} = \boldsymbol{\beta}^{(j-1)}$.
5. Update the counter of the chain from j to $j + 1$.
6. Back to step 2 and repeat the procedure until the chain reaches the desired convergence.

3. Allowing for Intraclass Variation and Heteroscedasticity

Without loss of generality, let us assume that the distribution of y_i is in the exponential family with likelihood given by the following general but not unique expression:

$$p(y_i \mid \theta_i, \tau_i) = a(y_i) \exp\{d_1(\theta_i, \tau_i)T_1(y_i) + d_2(\theta_i, \tau_i)T_2(y_i) + b(\theta_i, \tau_i)\},$$

where θ_i and τ_i are the parameters of interest, and the functional forms of $a(\cdot)$, $d_1(\cdot, \cdot)$, $d_2(\cdot, \cdot)$, $T_1(\cdot)$, $T_2(\cdot)$, and $b(\cdot, \cdot)$ are assumed to be known (McCullagh and Nelder, 1996). Under this formulation, we suppose that the sample space of the natural parameter contains a bidimensional rectangle that contains the point $\tau = 0$. We recall that this particular parameterization induces the uniparametric exponential family when $\tau = 0$.

3.1. *Considering Random Effects*

A common way of considering overdispersion in a model is by means of the introduction of random effects in the mean. In this manner, we suppose that the expectation of y_i is given by

$$E(y_i \mid \theta_i, \tau_i) = \mu_i = h^{-1}(f(\mathbf{x}_i, \boldsymbol{\beta}, \mathbf{w}_i, \boldsymbol{\lambda}_i)), \tag{9}$$

where $h(\cdot)$ is the link function, $\mathbf{w}_i = (w_{i1}, \ldots, w_{iR})'$ is the vector containing auxiliary information for the ith unit and is related to the response variable through a vector of random effects $\boldsymbol{\lambda}_i = (\lambda_1, \ldots, \lambda_R)'$. Under the former assumptions, we say that y_i follows a generalized nonlinear mixed model. Note that if we let $f(\mathbf{x}_i, \boldsymbol{\beta}, \mathbf{w}_i, \boldsymbol{\lambda}_i) = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda}_i$, then y_i follows a generalized linear mixed model (McCulloch and Searle, 2001). In this manner,

if h is the identity function and the response variable is Gaussian, then we have the classical linear mixed regression model (Laird and Ware, 1982). Following the Bayesian approach, this general model is completed after assigning prior distributions for parameters $\boldsymbol{\beta}$, $\boldsymbol{\lambda}$, $\mathbf{L}$ given by

$$\begin{aligned} \boldsymbol{\beta} &\sim N(\mathbf{b}, \mathbf{B}), \\ \boldsymbol{\lambda}_i \mid \mathbf{L} &\sim N(\mathbf{0}, \mathbf{L}), \\ \mathbf{L} &\sim I - W(v, \mathbf{S}^{-1}), \end{aligned}$$

where the notation $I - W(v, \mathbf{S}^{-1})$ applies to an inverse Wishart distribution with v degrees of freedom and inverse scale matrix $\mathbf{S}$. Consequently, assuming prior independence, the joint prior distribution is $p(\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{L}) = p(\boldsymbol{\lambda} \mid \mathbf{L})p(\mathbf{L})p(\boldsymbol{\beta})$. Next, as a consequence of the Bayes rule, the full posterior distribution for the parameters of interest takes the following form:

$$p(\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{L} \mid \mathbf{y}) \propto \prod_{i=1}^{n} p(y_i \mid \theta_i, \tau_i) p(\boldsymbol{\lambda}_i \mid \mathbf{L}) p(\mathbf{L}) p(\boldsymbol{\beta}). \tag{10}$$

Thus, the conditional posterior density for $\boldsymbol{\beta}$ is given by

$$p(\boldsymbol{\beta} \mid \boldsymbol{\lambda}, \mathbf{L}, \mathbf{y}) \propto \exp\left\{ \frac{-1}{2} (\boldsymbol{\beta} - \mathbf{b})' \mathbf{B}^{-1} (\boldsymbol{\beta} - \mathbf{b}) + \sum_{i=1}^{n} [d_1(\theta_i, \tau_i) T_1(y_i) + d_2(\theta_i, \tau_i) T_2(y_i) + b(\theta_i, \tau_i)] \right\}. \tag{11}$$

And the conditional posterior density for $\boldsymbol{\lambda}$ is given by

$$p(\boldsymbol{\lambda} \mid \boldsymbol{\beta}, \mathbf{L}, \mathbf{y}) \propto \exp\left\{ \frac{-1}{2} \boldsymbol{\lambda}' \mathbf{L}^{-1} \boldsymbol{\lambda} + \sum_{i=1}^{n} [d_1(\theta_i, \tau_i) T_1(y_i) + d_2(\theta_i, \tau_i) T_2(y_i) + b(\theta_i, \tau_i)] \right\}. \tag{12}$$

Note that the above posteriors are in the same form as (3), and the same methodology with the two-stage approach of the latter section can be used in order to estimate the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$. The following result allows us to handle such situations as random effects in the mean of the model.

Result 3.1. *Suppose that $\boldsymbol{\beta}^{(c)}$ and $\boldsymbol{\lambda}_i^{(c)}$ are the current values of $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}_i$, respectively, combining the prior distribution of the vector of parameters $\boldsymbol{\beta}$ with the normal approximation induced by the following working variable:*

$$\tilde{y}_i^b = \tilde{\mathbf{x}}_i \boldsymbol{\beta}^{(c)} + h'\big(h^{-1}\big[f\big(\mathbf{x}_i, \boldsymbol{\beta}^{(c)}, \mathbf{w}_i, \boldsymbol{\lambda}_i^{(c)}\big)\big]\big)\big(y_i - h^{-1}\big[f\big(\mathbf{x}_i, \boldsymbol{\beta}^{(c)}, \mathbf{w}_i, \boldsymbol{\lambda}_i^{(c)}\big)\big]\big), \tag{13}$$

which is normally distributed with mean $\tilde{\mathbf{x}}_i \boldsymbol{\beta}^{(c)}$ and variance $\tilde{V}_i$. Then, a suitable proposal distribution for the full posterior (11) with Gaussian kernel $q_{\boldsymbol{\beta}}$ is given by

$$q_{\boldsymbol{\beta}}\big(\boldsymbol{\beta}^{(c)}\big) = N(\mathbf{b}^*, \mathbf{B}^*), \tag{14}$$

where,

$$\mathbf{B}^* = (\mathbf{B}^{-1} + \tilde{\mathbf{X}}'\tilde{\mathbf{V}}^{-1}\tilde{\mathbf{X}})^{-1},$$
$$\mathbf{b}^* = \mathbf{B}^*(\mathbf{B}^{-1}\mathbf{b} + \tilde{\mathbf{X}}'\tilde{\mathbf{V}}^{-1}\tilde{\mathbf{y}}_b),$$

and $\tilde{\mathbf{y}}_b = (\tilde{y}_1^b, \ldots, \tilde{y}_n^b)'$. *The components of* $\tilde{\mathbf{X}}$ *are given by*

$$\tilde{\mathbf{x}}_i = \nabla_{\boldsymbol{\beta}} f\left(\mathbf{x}_i, \boldsymbol{\beta}^{(c)}, \mathbf{w}_i, \boldsymbol{\lambda}_i^{(c)}\right) = \left[\frac{\partial f}{\partial \beta_1}, \ldots, \frac{\partial f}{\partial \beta_P}\right]\Bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(c)}, \boldsymbol{\lambda}_i=\boldsymbol{\lambda}_i^{(c)}}.$$

Analogously, combining the prior distribution of the vector of parameters $\boldsymbol{\lambda}_i$ *with the normal approximation induced by the following working variable,*

$$\tilde{y}_i^l = \tilde{\mathbf{w}}_i\boldsymbol{\lambda}_i^{(c)} + h'\left(h^{-1}\left[f\left(\mathbf{x}_i, \boldsymbol{\beta}^{(c)}, \mathbf{w}_i, \boldsymbol{\lambda}_i^{(c)}\right)\right]\right)\left(y_i - h^{-1}\left[f\left(\mathbf{x}_i, \boldsymbol{\beta}^{(c)}, \mathbf{w}_i, \boldsymbol{\lambda}_i^{(c)}\right)\right]\right), \quad (15)$$

which is normally distributed with mean $\tilde{\mathbf{w}}_i\boldsymbol{\lambda}_i^{(c)}$ *and variance* $\tilde{V}_i$, *a suitable proposal distribution for the full posterior (12) with Gaussian kernel* $q_{\boldsymbol{\lambda}}$ *is given by*

$$q_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}^{(c)}) = N(\mathbf{l}^*, \mathbf{L}^*), \quad (16)$$

where,

$$\mathbf{L}^* = (\mathbf{L}^{-1} + \tilde{\mathbf{W}}'\tilde{\mathbf{V}}^{-1}\tilde{\mathbf{W}})^{-1},$$
$$\mathbf{l}^* = \mathbf{L}^*(\tilde{\mathbf{W}}'\tilde{\mathbf{V}}^{-1}\tilde{\mathbf{y}}_l),$$

with $\tilde{\mathbf{y}}_l = (\tilde{y}_1^l, \ldots, \tilde{y}_n^l)'$, *and* $\tilde{\mathbf{W}} = (\tilde{\mathbf{w}}_1', \ldots, \tilde{\mathbf{w}}_n')$, *whose components are given by*

$$\tilde{\mathbf{w}}_i = \nabla_{\boldsymbol{\lambda}_i} f\left(\mathbf{x}_i, \boldsymbol{\beta}^{(c)}, \mathbf{w}_i, \boldsymbol{\lambda}_i^{(c)}\right) = \left[\frac{\partial f}{\partial \lambda_1}, \ldots, \frac{\partial f}{\partial \lambda_R}\right]\Bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(c)}, \boldsymbol{\lambda}_i=\boldsymbol{\lambda}_i^{(c)}}.$$

Finally, it is straightforward to show that the conditional posterior density for $\mathbf{L}$ is

$$p(\mathbf{L} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{y}) \propto |\mathbf{L}|^{-(v+n)/2} \exp\left\{\frac{-1}{2} tr(\mathbf{L}^{-1}[\mathbf{S} + \mathbf{S}_{\boldsymbol{\lambda}}])\right\}, \quad (17)$$

which follows an $I - W(v + n, \mathbf{S} + \mathbf{S}_{\boldsymbol{\lambda}})$ density, with $\mathbf{S}_{\boldsymbol{\lambda}} = \sum_{i=1}^n \boldsymbol{\lambda}_i\boldsymbol{\lambda}_i'$. As mentioned by Gammerman (1997), in the case of a scalar random effect with variance σ_λ^2, the conditional posterior density for σ_λ^2 follows an inverse gamma (I-G) density.

Based on the previous approaches, the algorithm of Section 2.4 may be supplemented in order to allow the simulation of values from the full posterior distribution (10). We only have to modify step 3 by proposing a new value, $\boldsymbol{\phi}_{\boldsymbol{\beta}}$, generated from the jumping proposal (16). Next, we have to add two more steps: (a) propose a new value, $\boldsymbol{\phi}_{\boldsymbol{\lambda}}$, generated from the jumping proposal $q_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}^{(c)})$ and compute its acceptance rate and, (b) propose a new value for $\mathbf{L}$ directly generated from (17). By iterating this process, the chains will reach convergence.

3.2. *Modeling the Dispersion Parameter*

With the above procedure, it is possible to model the mean of the distribution when proposing values for the fixed and random parameters from the jumping proposals. However, in order to model the dispersion parameter τ_i, it is necessary to perform a similar methodology

for its inference and estimation. Dey et al. (1997) have reparameterized the formulation of the generalized linear model in the same way as Gelfand and Dalal (1990) do. Under this parameterization, it is shown that, assuming regularity conditions, μ is orthogonal to τ in the sense of Barndorf-Nielsen (1978) and Cox and Reid (1987). This leads to a classical generalized linear mixed model, and it is possible to fit a generalized linear mixed model with overdispersion (Jorgersen, 1987) when it is further assumed that the responses y_i are associated with some covariates $\mathbf{z}_i$ through a vector of parameters $\boldsymbol{\gamma}$ and a function $g(\cdot)$, such that

$$g(\tau_i) = \mathbf{z}_i'\boldsymbol{\gamma}, \tag{18}$$

where g is a strictly monotonic differentiable function. So, to force the dispersion in the model to be positive ($\tau_i > 0$), this function could be taken to be, for example, $\tau_i = \exp(\mathbf{z}_i'\boldsymbol{\gamma})$. Thus, if a normal prior distribution is considered for $\boldsymbol{\gamma}$, with prior mean $\mathbf{g}$ and prior variance $\mathbf{G}$, we conclude that its conditional posterior distribution is given by

$$p(\boldsymbol{\gamma} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{L}, \mathbf{y}) \propto \exp\left\{\frac{-1}{2}(\boldsymbol{\gamma} - \mathbf{g})'\mathbf{G}^{-1}(\boldsymbol{\gamma} - \mathbf{g}) + \sum_{i=1}^{n}[d_1(\theta_i, \tau_i)T_1(y_i) + d_2(\theta_i, \tau_i)T_2(y_i) + b(\theta_i, \tau_i)]\right\}. \tag{19}$$

As discussed before, since the above expression has no closed form and is not log-concave, the likelihood should be approximated to a normal density in order to combine it with the normal prior distribution of $\boldsymbol{\gamma}$. We will appeal again to the adaptation of the iterative weighted least-squares technique and the creation of another working variable. The following result shows the way to achieve this.

Result 3.2 *Assuming that there exists a variable t_i such that $E(t_i) = \tau_i = g^{-1}(\mathbf{z}_i'\boldsymbol{\gamma})$, the following working variable is defined (induced by the first-order Taylor approximation of $g(\cdot)$ in some neighborhood of $E(t_i)$) as*

$$\tilde{t}_i := \mathbf{z}_i'\boldsymbol{\gamma} + g'(g^{-1}(\mathbf{z}'\boldsymbol{\gamma}))(t_i - g^{-1}(\mathbf{z}'\boldsymbol{\gamma})). \tag{20}$$

The distribution of $\tilde{t}_i$ is normal with expectation $E(\tilde{t}_i) = \mathbf{z}_i'\boldsymbol{\gamma}$ and variance $\tilde{T}_i = Var(\tilde{t}_i) = [g'(g^{-1}(\mathbf{z}'\boldsymbol{\gamma}))]^2 Var(t_i)$. Then, if $\boldsymbol{\gamma}^{(c)}$ is the current parameter of $\boldsymbol{\gamma}$, a suitable proposal jumping distribution with Gaussian kernel $q_{\boldsymbol{\gamma}}$ is obtained as

$$q_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}^{(c)}) = \text{Normal}(\mathbf{g}^*, \mathbf{G}^*), \tag{21}$$

where

$$\mathbf{G}^* = (\mathbf{G}^{-1} + \mathbf{Z}'\tilde{\mathbf{T}}^{-1}\mathbf{Z})^{-1},$$
$$\mathbf{g}^* = \mathbf{G}^*(\mathbf{G}^{-1}\mathbf{g} + \mathbf{Z}'\tilde{\mathbf{T}}^{-1}\tilde{\mathbf{t}}),$$

with $\tilde{\mathbf{T}}$ as the diagonal matrix with entries $\tilde{T}_i$, and $\tilde{\mathbf{t}} = (\tilde{t}_1, \ldots, \tilde{t}_n)'$.

3.3. *Some Useful Models*

3.3.1. *Normal nonlinear regression with random effects.* We consider the mixed regression model $y_i = f(\mathbf{x}_i, \boldsymbol{\beta}, \mathbf{w}_i, \boldsymbol{\lambda}_i) + \varepsilon_i$ with $i = 1, \ldots, n$ and $\varepsilon_i \sim \text{Normal}(0, \sigma_i^2)$. Under this context, $h(\mu_i) = \mu_i = f(\mathbf{x}_i, \boldsymbol{\beta}, \mathbf{w}_i, \boldsymbol{\lambda}_i)$. Thus, the working variable for the vector of fixed effects is

$$\tilde{y}_i^b = y_i - f(\mathbf{x}_i, \boldsymbol{\beta}, \mathbf{w}_i, \boldsymbol{\lambda}_i) + \tilde{\mathbf{x}}_i' \boldsymbol{\beta}.$$

The working variable for the vector of random effects is

$$\tilde{y}_i^l = y_i - f(\mathbf{x}_i, \boldsymbol{\beta}, \mathbf{w}_i, \boldsymbol{\lambda}_i) + \tilde{\mathbf{w}}_i' \boldsymbol{\lambda},$$

with associated working variance $\tilde{V}_i = \text{Var}(y_i) = \sigma_i^2$. Finally, as $\tau_i = \sigma_i^2$ and considering $g(\sigma_i^2) = \log(\sigma_i^2) = \mathbf{z}_i' \boldsymbol{\gamma}$, we define $t_i = (y_i - f(\mathbf{x}_i, \boldsymbol{\beta}, \mathbf{w}_i, \boldsymbol{\lambda}_i))^2$. As $E(t_i/\sigma_i^2) \sim \chi_1^2$, then $E(t_i) = \sigma_i^2$ and $\text{Var}(t_i) = 2\sigma_i^4$. The working variable for the vector of variance effects is given by

$$\tilde{t}_i = \mathbf{z}_i' \boldsymbol{\gamma} + \frac{(y_i - f(\mathbf{x}_i', \boldsymbol{\beta}, \mathbf{w}_i', \boldsymbol{\lambda}))^2}{\exp(\mathbf{z}_i' \boldsymbol{\gamma})} - 1,$$

with associated working variance $\tilde{T}_i = (\frac{1}{\exp(\mathbf{z}_i' \boldsymbol{\gamma})})^2 (2\sigma_i^4) = 2$.

3.3.2. *Gamma nonlinear regression with random effects.* If $y_i \sim \text{Gamma}(a_i, b_i)$, then $\mu_i = a_i/b_i$ and by using the canonical[1] link function $h(\mu_i) = 1/\mu_i = f(\mathbf{x}_i', \boldsymbol{\beta}, \mathbf{w}_i', \boldsymbol{\lambda})$, the working variable for the vector of fixed effects is given by

$$\tilde{y}_i^b = \tilde{\mathbf{x}}_i' \boldsymbol{\beta} + f(\mathbf{x}_i', \boldsymbol{\beta}, \mathbf{w}_i', \boldsymbol{\lambda}) - (f(\mathbf{x}_i', \boldsymbol{\beta}, \mathbf{w}_i', \boldsymbol{\lambda}))^2 y_i.$$

The working variable for the vector of random effects is

$$\tilde{y}_i^l = \tilde{\mathbf{w}}_i' \boldsymbol{\lambda} + f(\mathbf{x}_i', \boldsymbol{\beta}, \mathbf{w}_i', \boldsymbol{\lambda}) - (f(\mathbf{x}_i', \boldsymbol{\beta}, \mathbf{w}_i', \boldsymbol{\lambda}))^2 y_i.$$

By considering that $a_i = \mu_i b_i = \tau_i$, $b_i = a_i/\mu_i = \tau_i/\mu_i$, $Var(y_i) = a_i/b_i^2 = \mu_i^2/\tau_i$ and modeling the dispersion parameter as $\tau_i = \exp(\mathbf{z}_i' \boldsymbol{\gamma})$, the algorithm could be implemented after defining

$$\begin{aligned} \tilde{V}_i &= [h'(h^{-1}(f(\mathbf{x}_i', \boldsymbol{\beta}, \mathbf{w}_i', \boldsymbol{\lambda})))]^2 Var(y_i) \\ &= f(\mathbf{x}_i', \boldsymbol{\beta}, \mathbf{w}_i', \boldsymbol{\lambda})^4 \mu_i^2/\tau_i \\ &= \frac{f(\mathbf{x}_i', \boldsymbol{\beta}, \mathbf{w}_i', \boldsymbol{\lambda})^2}{\exp(\mathbf{z}_i' \boldsymbol{\gamma})}. \end{aligned}$$

As considered, $\tau_i = a_i$. Then, $g(a_i) = \log(a_i) = \mathbf{z}_i' \boldsymbol{\gamma}$ and we define $t_i = b_i y_i$ because $E(t_i) = a_i$. Thus, taking into account that

$$b_i = \frac{a_i}{\mu_i} = \frac{\exp(\mathbf{z}_i' \boldsymbol{\gamma})}{\mu_i} = \frac{\exp(\mathbf{z}_i' \boldsymbol{\gamma})}{h^{-1}(f(\mathbf{x}_i', \boldsymbol{\beta}, \mathbf{w}_i', \boldsymbol{\lambda}))} = f(\mathbf{x}_i', \boldsymbol{\beta}, \mathbf{w}_i', \boldsymbol{\lambda}) \exp(\mathbf{z}_i' \boldsymbol{\gamma}),$$

[1]Ntzoufras (2009) stated that because the gamma distribution parameters are positive, then one must make sure that, using data from the response and covariates, the canonical link function yields strictly nonnegative values. It is suggested to use such functions as logarithmic link to force this kind of result.

the working variable for the vector of variance effects is given by the following expression:

$$\begin{aligned}\tilde{t}_i &= \mathbf{z}_i'\boldsymbol{\gamma} + \frac{1}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})}(b_i y_i - \exp(\mathbf{z}_i'\boldsymbol{\gamma})) \\ &= \mathbf{z}_i'\boldsymbol{\gamma} + \frac{1}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})}(f(\mathbf{x}_i', \boldsymbol{\beta}, \mathbf{w}_i', \boldsymbol{\lambda})\exp(\mathbf{z}_i'\boldsymbol{\gamma})y_i - \exp(\mathbf{z}_i'\boldsymbol{\gamma})) \\ &= \mathbf{z}_i'\boldsymbol{\gamma} + f(\mathbf{x}_i', \boldsymbol{\beta}, \mathbf{w}_i', \boldsymbol{\lambda})y_i - 1,\end{aligned}$$

with working variance given by

$$\tilde{T}_i = [g'(g^{-1}(\mathbf{z}_i'\boldsymbol{\gamma}))]^2 \mathrm{Var}(t_i) = \frac{1}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})}.$$

Thus, the model is completely specified by

$$\tilde{y}_i \sim \mathrm{Gamma}(\tau_i, \tau_i/\mu_i).$$

3.3.3. *Beta nonlinear regression with random effects.* On the other hand, let us assume that $y_i \sim Beta(a_i, b_i)$. Note that $\mu_i = a_i/(a_i + b_i)$ and suppose we use the *logit* link function. Therefore, $h(\mu_i) = logit(\mu_i) = \log(\mu_i/(1-\mu_i))$, and the working variable for the vector of fixed effects is

$$\tilde{y}_i^b = \tilde{\mathbf{x}}_i'\boldsymbol{\beta} + \frac{(1+\exp(f(\mathbf{x}_i', \boldsymbol{\beta}, \mathbf{w}_i', \boldsymbol{\lambda})))^2}{\exp(f(\mathbf{x}_i', \boldsymbol{\beta}, \mathbf{w}_i', \boldsymbol{\lambda}))} y_i - \exp(f(\mathbf{x}_i', \boldsymbol{\beta}, \mathbf{w}_i', \boldsymbol{\lambda})) - 1.$$

Analogously, the working variable for the vector of random effects is

$$\tilde{y}_i^l = \tilde{\mathbf{w}}_i'\boldsymbol{\lambda} + \frac{(1+\exp(f(\mathbf{x}_i', \boldsymbol{\beta}, \mathbf{w}_i', \boldsymbol{\lambda})))^2}{\exp(f(\mathbf{x}_i', \boldsymbol{\beta}, \mathbf{w}_i', \boldsymbol{\lambda}))} y_i - \exp(f(\mathbf{x}_i', \boldsymbol{\beta}, \mathbf{w}_i', \boldsymbol{\lambda})) - 1.$$

By considering that $\tau_i = a_i + b_i$ and modeling τ_i such that $g(\tau_i) = \log(a_i + b_i) = \mathbf{z}_i'\boldsymbol{\gamma}$ we observe that $a_i = \mu_i\tau_i$, $b_i = \tau_i - \mu_i\tau_i$ and $\mathrm{Var}(y_i) = a_i b_i/[(a_i + b_i)^2(a_i + b_i + 1)] = \mu_i(1-\mu_i)/(\tau_i + 1)$. Then, the associated variance for the above working variable is as follows:

$$\tilde{V}_i = [h'(h^{-1}(f(\mathbf{x}_i', \boldsymbol{\beta}, \mathbf{w}_i', \boldsymbol{\lambda})))]^2 Var(y_i) = \frac{1}{\exp(f(\mathbf{x}_i', \boldsymbol{\beta}, \mathbf{w}_i', \boldsymbol{\lambda}))(\exp(\mathbf{z}_i'\boldsymbol{\gamma}) + 1)}.$$

As considered before, if $\tau_i = a_i + b_i$, then we define $t_i = \frac{(a_i+b_i)^2}{a_i} y_i$ because $E(t_i) = a_i + b_i = \tau_i$. So then, taking into account that

$$\begin{aligned}t_i = \frac{(a_i+b_i)^2}{a_i} y_i = \frac{\tau_i}{\mu_i} y_i &= \frac{g^{-1}(\mathbf{z}_i'\boldsymbol{\gamma})}{h^{-1}(f(\mathbf{x}_i', \boldsymbol{\beta}, \mathbf{w}_i', \boldsymbol{\lambda}))} y_i \\ &= \frac{\exp(\mathbf{z}_i'\boldsymbol{\gamma})(1+\exp(f(\mathbf{x}_i', \boldsymbol{\beta}, \mathbf{w}_i', \boldsymbol{\lambda})))}{\exp(f(\mathbf{x}_i', \boldsymbol{\beta}, \mathbf{w}_i', \boldsymbol{\lambda}))} y_i,\end{aligned}$$

the working variable for the vector of variance effects is given by

$$\tilde{t}_i = \mathbf{z}_i'\boldsymbol{\gamma} + \frac{1}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})}(t_i - \exp(\mathbf{z}_i'\boldsymbol{\gamma}))$$
$$= \mathbf{z}_i'\boldsymbol{\gamma} + \frac{1 + \exp(f(\mathbf{x}_i', \boldsymbol{\beta}, \mathbf{w}_i', \boldsymbol{\lambda}))}{\exp(f(\mathbf{x}_i', \boldsymbol{\beta}, \mathbf{w}_i', \boldsymbol{\lambda})))} y_i - 1,$$

with associated variance

$$\tilde{T}_i = [g'(g^{-1}(\mathbf{z}_i'\boldsymbol{\gamma}))]^2 Var(t_i) = \frac{(1 + \exp(f(\mathbf{x}_i', \boldsymbol{\beta}, \mathbf{w}_i', \boldsymbol{\lambda})))^2}{\exp(f(\mathbf{x}_i', \boldsymbol{\beta}, \mathbf{w}_i', \boldsymbol{\lambda}))(\exp(\mathbf{z}_i'\boldsymbol{\gamma}) + 1)}.$$

In this way, the model is completely specified as

$$\tilde{y}_i \sim Beta(\mu_i\tau_i, \tau_i - \mu_i\tau_i).$$

3.4. *The MCMC Algorithm for Mean and Dispersion Parameters Estimation*

With the latter assumption, it is possible to implement a Metropolis-type hybrid algorithm (Robert and Casella, 2009, p. 230), completed with the inclusion of the joint modeling of mean and dispersion parameters by means of the following steps:

1. Set the iteration counter chain to $j = 1$.
2. Set initial values $\boldsymbol{\beta}^{(j-1)}$, $\boldsymbol{\lambda}^{(j-1)}$, $\boldsymbol{\gamma}^{(j-1)}$ and $\mathbf{L}^{(j-1)}$.
3. Propose a new value, $\boldsymbol{\phi}_{\boldsymbol{\beta}}$, generated from the jumping proposal (14).
4. Compute the acceptance probability of the movement. If it is accepted, then $\boldsymbol{\beta}^{(j)} = \boldsymbol{\phi}_{\boldsymbol{\beta}}$; otherwise $\boldsymbol{\beta}^{(j)} = \boldsymbol{\beta}^{(j-1)}$.
5. Propose a new value, $\boldsymbol{\phi}_{\boldsymbol{\lambda}}$, generated from the jumping proposal (16).
6. Compute the acceptance probability of the movement. If it is accepted, then $\boldsymbol{\lambda}^{(j)} = \boldsymbol{\phi}_{\boldsymbol{\lambda}}$; otherwise $\boldsymbol{\lambda}^{(j)} = \boldsymbol{\lambda}^{(j-1)}$.
7. Propose a new value, $\boldsymbol{\phi}_{\boldsymbol{\gamma}}$, generated from the jumping proposal (21).
8. Compute the acceptance probability of the movement. If it is accepted, then $\boldsymbol{\gamma}^{(j)} = \boldsymbol{\phi}_{\boldsymbol{\gamma}}$; otherwise $\boldsymbol{\gamma}^{(j)} = \boldsymbol{\gamma}^{(j-1)}$.
9. Propose a new value, $\mathbf{L}^{(j)}$ directly generated from (17).
10. Update the counter of the chain from j to $j + 1$.
11. Back to step 2 and repeat the procedure until the chain reaches the desired convergence.

4. Simulations

4.1. *Modeling the Mean Parameters*

In this section, we introduce six simulations carried out in order to empirically demonstrate the potential of the proposed Bayesian technique[2]. In these simulations, we wanted to study the performance of the algorithm in the presence of some of the most important techniques of the statistical framework; namely linear regression, nonlinear regression, generalized linear

[2]All of the simulations and the empirical application were carried out using the statistical software R (R Development Core Team, 2009). The codes are available upon request to the author.

models, and generalized nonlinear models. Also, the algorithm is applied for both discrete and continuous response variables. The results are very satisfactory and they show that this technique performs well in many scenarios. We want to stress the high computational efficiency showed in all of the simulated chains; the convergence in the four scenarios, with different initial values, is almost immediate. Also, the acceptance rates of the Metropolis-type algorithm are high. For all of the simulations, the results reported are based on a sample of 1000 draws after a burn-in stage of 5000. Also, the prior distributions used were flat for every parameter involved in the modeling of the mean; i.e., $\beta \sim N(0, 10^3)$.

Case 1: Normal linear model. As a first simulation, we conducted a study to test this Bayesian approach for the most simple case. We have already observed that if the errors of the model yield a normal distribution with null mean and variance σ^2 and the function $f(\mathbf{x}_i, \boldsymbol{\beta}) = \mathbf{x}_i'\boldsymbol{\beta}$, then the proposed algorithm becomes a Gibbs sampler type algorithm. Although there is nothing new behind a Gibbs sampler algorithm in normal linear models, we include this case in order to show that it is indeed a particular case of the proposed methodology, and the acceptance rate is equal to unity. Initially, we simulated $n = 50$ values of one explanatory variable x_1 generated from a continuous uniform distribution in the interval (1, 1.5). The values of the response variable y_i ($i = 1, \ldots, n$) were simulated from a normal distribution with mean

$$\mu_i = 13 + 25x_{1i},$$

and constant variance $\sigma^2 = 1$.

Case 2: Normal nonlinear model. A second simulation study was carried out for $n = 30$ units for a model with independent normal errors with null mean and constant variance $\sigma^2 = 1$. The response variable y_i follows a normal distribution with mean function given by

$$\mu_i = 10 \times \exp(2x_{i1}),$$

where x_1 was generated as before.

Case 3: Poisson linear model. The third simulation is related to a nonnormal and non-continuous response variable. We conducted a study for $n = 150$ values of one explanatory variable x_1 generated as before. The values of the response variable y_i were generated from a Poisson distribution such that

$$\log(\mu_i) = 2 + 1.2x_{1i}.$$

Case 4: Poisson nonlinear model. This simulation was based on a more intricate model. We simulated $n = 30$ units from a nonlinear model when the distribution of response variable was considered as a Poisson density. Then, the mean of the response variable y_i was such that

$$\log(\mu_i) = 0.5 \times \exp(1.4x_{1i}),$$

where x_1 was generated as before.

Case 5: Binomial nonlinear regression. For this exercise, a sample of $n = 70$ binomial trials was drawn. For each draw $i = 1, \ldots, 70$, a particular number of trials n_i (between 40 and 60) was simulated. The probability of success on each trial was modeled in the following way:

$$logit(\pi_i) = 0.5 \times \exp(0.1x_{1i}),$$

Table 1
Summary of the simulations: parameter estimates, corresponding standard deviations, and acceptance rates (AR) for mean models given in cases 1–6

Case	β_1	β_2	$\hat{\beta}_1$ (sd)	$\hat{\beta}_2$ (sd)	AR
1	13	25	13.042 (1.195)	24.979 (0.926)	1.000
2	10	2	10.072 (0.319)	1.997 (0.023)	0.884
3	2	1.2	1.813 (0.280)	1.342 (0.226)	0.959
4	0.5	1.4	0.488 (0.079)	1.411 (0.106)	0.525
5	0.5	0.1	0.450 (0.042)	0.108 (0.006)	0.859
6	15	−2	14.55 (0.541)	−1.930 (0.072)	0.972

where x_1 was generated from a continuous uniform distribution in the interval (1, 20). Doing this, $y_i \sim \text{Binomial}(n_i, \pi_i)$.

Case 6: Binomial linear regression. The simulated values for the number of trials were generated just as before. The probability of success on each trial was modeled in the following way:

$$logit(\pi_i) = 15 - 2x_{1i},$$

where x_1 was generated from a continuous uniform distribution in the interval (3, 10).

Table 1 exhibits the efficiency of the proposed methodology for the estimation of mean fixed parameters in both linear and nonlinear models. The first column indicates the simulated model, Columns 2 and 3 show the real values for the parameters of the mean model, the next two columns show the estimated values and the corresponding standard deviations for the above parameters. The last column indicates the acceptance rate of the algorithm.

4.2. *Joint Modeling of Mean and Dispersion Parameters*

In this section, we introduce another six simulations carried out for some densities belonging to the two-parameter exponential family. For all of the simulations, the results reported are based on a sample of 2000 draws every 10 iterations after a burn-in stage of 5000. For each simulated exercise, the models involved one random parameter λ_i per unit $i = 1, \ldots, n$. Also, the prior distributions used were flat for every parameter involved in the modeling of the mean; i.e., $\beta \sim N(0, 10^3)$, $\gamma \sim N(0, 10^3)$, $\lambda_i \sim N(0, \sigma_\lambda^2)$, and $p(\sigma_\lambda^2) \propto 1$. Doing this, the full conditional posterior for σ_λ^2 is $I - G(n/2, \sum_{i=1}^{n} \lambda_i^2)$.

Case 7: Normal nonlinear mixed model. We simulated $n = 200$ values from an explanatory variable x_1 generated from a continuous uniform distribution in the interval (10, 15) and from an explanatory variable z_1 generated from a continuous uniform distribution in the interval (5, 10). The values of the response variable y_i (i=1,\ldots ,n) were simulated from a normal distribution with mean

$$\mu_i = 10 \times \exp(0.5x_{1i}) + \lambda_i,$$

and variance

$$\sigma_i^2 = \exp(-5 + 2z_{1i}),$$

where $\lambda_i \sim N(0, \sqrt{0.5})$.

Case 8: Normal linear mixed model. This simulation study was carried out for $n = 200$ units from a Gaussian model. The explanatory variable x_1 was generated as before, while the explanatory variable z_1 was generated from a continuous uniform distribution in the interval (20, 40). The response variable was simulated from a normal distribution with mean

$$\mu_i = 30 + 2.5x_{1i} + \lambda_i,$$

and variance

$$\sigma_i^2 = \exp(-5 + 0.2z_{1i}),$$

where $\lambda_i \sim N(0, \sqrt{1})$.

Case 9: Gamma nonlinear mixed model. We conducted this study for $n = 50$ values from an explanatory variable x_1 generated from a continuous uniform distribution in the interval (0, 10) and from an explanatory variable z_1 generated from a continuous uniform distribution in the interval (0.5, 0.7). The values of the response variable y_i were simulated from a Gamma distribution with

$$\mu_i = \frac{1}{0.3 \times \exp(0.8x_{1i}) + \lambda_i},$$

and

$$\tau_i = \exp(5 + 5z_{1i}),$$

where $\lambda_i \sim N(0, \sqrt{0.1})$.

Case 10: Gamma linear mixed model. For this exercise, we simulated $n = 50$ values from explanatory variables x_1 and z_1 generated as in case 9. The values of the response variable y_i were simulated from a Gamma distribution with

$$\mu_i = \frac{1}{3 + 0.2x_{1i} + \lambda_i},$$

and

$$\tau_i = \exp(12 + 10z_{1i}),$$

where $\lambda_i \sim N(0, \sqrt{0.1})$.

Case 11: Beta nonlinear mixed model. This simulation is related to the Beta distribution, which restricts its realizations to the continuous interval (0,1). We generated $n = 100$ values from an explanatory variable x_1 generated from a continuous uniform distribution in the interval (10, 20) and from an explanatory variable z_1 generated from a continuous uniform distribution in the interval (15, 20). The values of the response variable y_i were simulated from a Beta distribution where

$$logit(\mu_i) = 0.2 \times \exp(0.1x_{1i}) + \lambda_i,$$

and

$$\tau_i = \exp(0.5 + 1z_{1i}),$$

with $\lambda_i \sim N(0, \sqrt{0.15})$.

Case 12: Beta linear mixed model. Finally, we simulated $n = 100$ values from explanatory variables x_1 and z_1 generated as in case 11. The values of the response variable y_i were simulated from a Beta distribution with

$$logit(\mu_i) = 3 - 0.2x_{1i} + \lambda_i,$$

and

$$\tau_i = \exp(0.4 + 0.8z_{1i}),$$

where $\lambda_i \sim N(0, \sqrt{0.1})$. Table 2 shows the performance of the proposed methodology in estimating the mean fixed parameters, the variance of the random effects, and the dispersion model parameters. The first column indicates the simulated model, Columns 2 and 3 show the real values for the parameters of the mean model, Columns 4 and 5 show the real values for the parameters of the dispersion model, and Column 5 displays the real value of the variance for the random parameter. The remaining columns show the estimated values and the corresponding standard deviations for the above parameters.

5. Application: The Schooling Rate Data Revisited

In this section, we consider the data reported and analyzed by Cepeda and Achcar (2010, p. 414). These data report the schooling rate in Colombia during the years 1991–2003, that was obtained by dividing the number of students enrolled in school by the number of potential students during this period for the ages 5–19 years. This information was obtained from Colombia's National Statistics Department and its Ministry of Education. This rate shows the general level of participation in a given level of education (from basic primary, to basic secondary and additional 10th and 11th years), and indicates the capacity of the education system to enroll students of a particular age group. The rate data are: 51.25, 56.79, 59.32, 63.09, 68.82, 67.84, 69.14, 74.18, 74.76, 76.71, 75.18, 76.23, and 79.15, for years 1991, 1992, ..., 2003, respectively.

As the variable of interest is the schooling rate and supposing a slightly nonconstant variance that may depends on the values of the explanatory variable, defined as the year of the corresponding reported rate, we use three approaches to model these data. First, we review the basic model described in Cepeda and Achcar (2010), which is a normal nonlinear model assuming that the variance model is given by a quadratic function of time. Then, we use a beta distribution to fit a nonlinear model, assuming that the variance model is given by a linear function of time. Finally, we add a random effect to the former nonlinear normal model, and the variance is assumed to remain linear. Note that this kind of data makes possible the comparison of the proposed approach with other similar methodologies.

For the implementation of the method, we followed the advice of Gelman and Shirley (2010) by simulating several parallel chains with dispersed initial values. After a burn-in stage, where we discarded the first half of the simulated values, we checked the convergence using within-chain analysis to monitor stationarity and we used a suitable thinning constant. Then, we mixed all of the simulations into a large chain of size 2000 and we summarized the posterior target distributions. We used the `R` statistical software in order to perform our methodology and the MCMC algorithm, which yielded close estimates to those found in Cepeda and Achcar (2010) with the software BUGS.

Table 2

Summary of the simulations: parameter estimates and corresponding standard deviations for mean and dispersion mixed models given in cases 7–12

Case	β_1	β_2	γ_1	γ_2	σ_λ^2	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\sigma}_\lambda^2$
7	10	0.5	−5	2	0.5	10.04 (0.063)	0.49 (0.004)	−4.34 (0.596)	1.81 (0.078)	0.505 (0.124)
8	30	2.5	−5	0.2	1	29.97 (0.926)	2.50 (0.083)	−6.87 (1.054)	0.25 (0.142)	0.996 (0.142)
9	0.3	0.8	5	5	0.1	0.31 (0.004)	0.79 (0.002)	5.15 (0.186)	4.71 (0.311)	0.15 (0.027)
10	3	0.2	12	10	0.1	3.01 (0.002)	0.19 (0.004)	12.00 (0.005)	9.99 (0.005)	0.099 (0.021)
11	0.2	0.1	0.5	1	0.15	0.20 (0.002)	0.09 (0.006)	0.50 (0.003)	0.99 (0.002)	0.157 (0.023)
12	3	−0.2	0.4	0.8	0.1	3.04 (0.027)	−0.20 (0.001)	0.40 (0.002)	0.79 (0.001)	0.104 (0.015)

5.1. *The Basic Modeling*

As claimed by Cepeda and Achcar (2010), it is reasonable to fit a normal model to the data. In this way, they fitted[3] the following nonlinear model for the mean:

$$\mu_i = \frac{\beta_1}{1 + \beta_2 \exp(\beta_3 x_i)}, \tag{22}$$

along with a quadratic function for the variance

$$\log(\sigma_i^2) = \gamma_1 + \gamma_2 x_i + \gamma_3 x_i^2, \tag{23}$$

where $x_i = i$ and $i = 1, \ldots, 13$. Note that the variance is larger from years 5 to 8 and we can conclude that variance depends on the explanatory variable through the latter function. The mean and dispersion models were fitted assuming vague prior distributions for each parameter. We found that the posterior mean of β_1 is 80.03 with a standard deviation of (2.086); the posterior mean of β_2 is 0.723 (0.06173); and, the posterior mean of β_3 is -0.2582 (0.04024). For the dispersion fitting, the posterior mean of γ_1 is 0.2363 (2.132); the posterior mean of γ_2 is -0.3338 (0.6032); and the posterior mean of γ_3 is 0.01927 (0.03919). For this model, the deviance information criterion (DIC) (Spiegelhalter et al., 2002) is 52.96.

5.2. *Deviating from Normality*

We may suppose that the response variable divided by 100 may be observed values of a beta random variable. Thus, the logit function is proposed to be the link function of the mean; i.e, $h(\mu_i) = logit(\mu_i) = f(\mathbf{x}_i', \boldsymbol{\beta}) = \beta_1/(1 + \beta_2 \exp(\beta_3 x_i))$. Also, we found that a better model for the dispersion parameter is the one given by $\log(\tau_i) = \gamma_1 + \gamma_2 x_i$, supposing that the variance slightly increases as time does. Then, based on our former approach, the working variable for the vector of fixed effects is given by

$$\tilde{y}_i^b = \tilde{\mathbf{x}}_i' \boldsymbol{\beta} + \frac{(1 + \exp(\beta_1/(1 + \beta_2 \exp(\beta_3 x_i))))^2}{\exp(\beta_1/(1 + \beta_2 \exp(\beta_3 x_i)))} y_i - \exp(\beta_1/(1 + \beta_2 \exp(\beta_3 x_i))) - 1,$$

where $\boldsymbol{\beta}' = (\beta_1, \beta_2, \beta_3)$, and

$$\tilde{\mathbf{x}}_i' = \left(\frac{1}{1 + \beta_2 \exp(\beta_3 x_i)}, -\frac{\beta_1 \exp(\beta_3 x_i)}{(1 + \beta_2 \exp(\beta_3 x_i))^2}, -\frac{x_i \beta_1 \beta_2 \exp(\beta_3 x_i)}{(1 + \beta_2 \exp(\beta_3 x_i))^2} \right).$$

The associated variance of $\tilde{y}_i^b$ is

$$\tilde{V}_i = \frac{1}{\exp(\beta_1/(1 + \beta_2 \exp(\beta_3 x_i)))(\exp(\gamma_1 + \gamma_2 x_i) + 1)}.$$

On the other hand, the working variable for the vector of fixed effects is given by

$$\tilde{t}_i = \gamma_1 + \gamma_2 x_i + \frac{1 + \exp(\beta_1/(1 + \beta_2 \exp(\beta_3 x_i)))}{\exp(\beta_1/(1 + \beta_2 \exp(\beta_3 x_i)))} y_i - 1,$$

[3]Although they developed a theoretical methodology for the Bayesian estimation of normal nonlinear models, they used an empirically computational approach to fit the above model based on BUGS software (Spiegelhalter et al., 2004). However, their paper lacks a valid methodology for the estimation of data that deviates from normality. Instead, we want to use this dataset in order to show the flexibility of our approach by fitting the same models and comparing the results obtained.

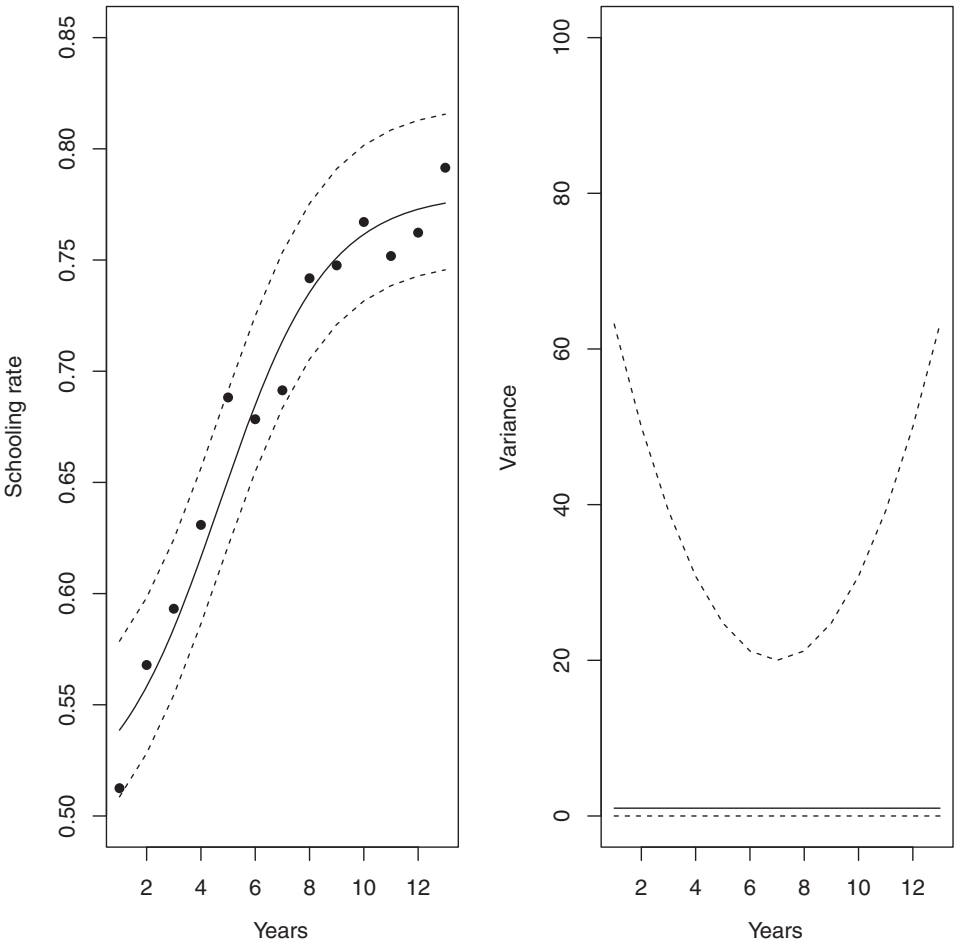

Figure 1. Beta nonlinear model. (a) Prediction: fitted posterior mean (solid line) and 95% prediction interval (dotted lines). (b) Posterior variance with uncertainty limits.

with variance

$$\tilde{T}_i = \frac{(1 + \exp(\beta_1/(1 + \beta_2 \exp(\beta_3 x_i))))^2}{\exp(\beta_1/(1 + \beta_2 \exp(\beta_3 x_i)))(\exp(\gamma_1 + \gamma_2 x_i) + 1)}.$$

For this model, we estimated that the posterior mean for β_1 was 1.2662, with a standard deviation of (0.0794); the posterior mean for β_2 was 11.6801 (5.5752); and, the posterior mean for β_3 was -0.4863 (0.0992). Finally, the posterior mean for γ_1 was 5.6565 (0.0899), and the posterior mean for γ_2 was 0.0899 (0.0072). For this model, the DIC was -59.207, showing the good performance of our approach. Figure 1 shows the behavior of the variance and the prediction through the fitted model. The variance is linear over time and small throughout the years.

5.3. *Modeling with Random Effects*

Moreover, if the context of the problem indicates that it is possible to consider a random effect in the nonlinear model then, based on the results of this research, it would be plausible to perform such inferences by carrying out the proper algorithms. For example, we can perform another analysis on the schooling rate data by considering normality of the data with the following mean:

$$\mu_i = f(\mathbf{x}_i', \boldsymbol{\beta}, w_i, \lambda_i) = \frac{\beta_1 + \lambda_i}{1 + \beta_2 \exp(\beta_3 x_i)}, \tag{24}$$

and by forcing overdispersion through

$$\log\left(\sigma_i^2\right) = \gamma_1 + \gamma_2 x_i. \tag{25}$$

It is not hard to find the suitable working variables for the mean and dispersion parameters along with their corresponding variances, in order to implement the MCMC algorithm proposed in this article. Using this approach, the working variable for the vector of fixed effects is given by

$$\tilde{y}_i^b = y_i - \frac{\beta_1 + \lambda_i}{1 + \beta_2 \exp(\beta_3 x_i)} + \tilde{\mathbf{x}}_i' \boldsymbol{\beta}.$$

The working variable for the random effect coefficient is

$$\tilde{y}_i^l = y_i - \frac{\beta_1 + \lambda_i}{1 + \beta_2 \exp(\beta_3 x_i)} + \tilde{w}_i \lambda,$$

with associated working variance $\tilde{V}_i = Var(y_i) = \sigma_i^2$. Note that $\tilde{w}_i = \frac{1}{1+\beta_2 \exp(\beta_3 x_i)}$ and

$$\tilde{\mathbf{x}}_i' = \left(\frac{1}{1 + \beta_2 \exp(\beta_3 x_i)}, -\frac{(\beta_1 + \lambda_i)\exp(\beta_3 x_i)}{(1 + \beta_2 \exp(\beta_3 x_i))^2}, -\frac{x_i(\beta_1 + \lambda_i)\beta_2 \exp(\beta_3 x_i)}{(1 + \beta_2 \exp(\beta_3 x_i))^2} \right).$$

The working variable for the vector of variance effects is given by

$$\tilde{t}_i = \gamma_1 + \gamma_2 x_i + \frac{\left(y_i - \frac{\beta_1+\lambda_i}{1+\beta_2 \exp(\beta_3 x_i)}\right)^2}{\exp(\gamma_1 + \gamma_2 x_i)} - 1,$$

with associated working variance $\tilde{T}_i = (\frac{1}{\exp(\gamma_1+\gamma_2 x_i)})^2 (2\sigma_i^4) = 2$.

For this mixed model, we estimated that the posterior mean for β_1 was 80.3, with a standard deviation of (0.01029); the posterior mean for β_2 was 0.6712 (0.00018); and, the posterior mean for β_3 was -0.2408 (0.00035). The posterior mean of the random effect is null and its posterior variance is 0.4164 (0.1616). Finally, the posterior mean for γ_1 was 52.6 (13.35), and the posterior mean for γ_2 was -2.885 (1.613). For this model, the DIC was -3135 and we can claim that this model is improved when adding a random effect. Figure 2 shows the behavior of the variance and the prediction through the fitted model. We can see that this model fits almost perfectly. This is due to the inclusion of the random effect. The variance[4] is linear over time and smaller than the previous models considered.

[4]Note that the variance upper bound, close to year 12, is larger compared with other years. This feature agrees with the prediction interval of the mean, that is wider at year 12. However, when comparing the limits of the prediction interval for the model with random effects (Fig. 2) with the

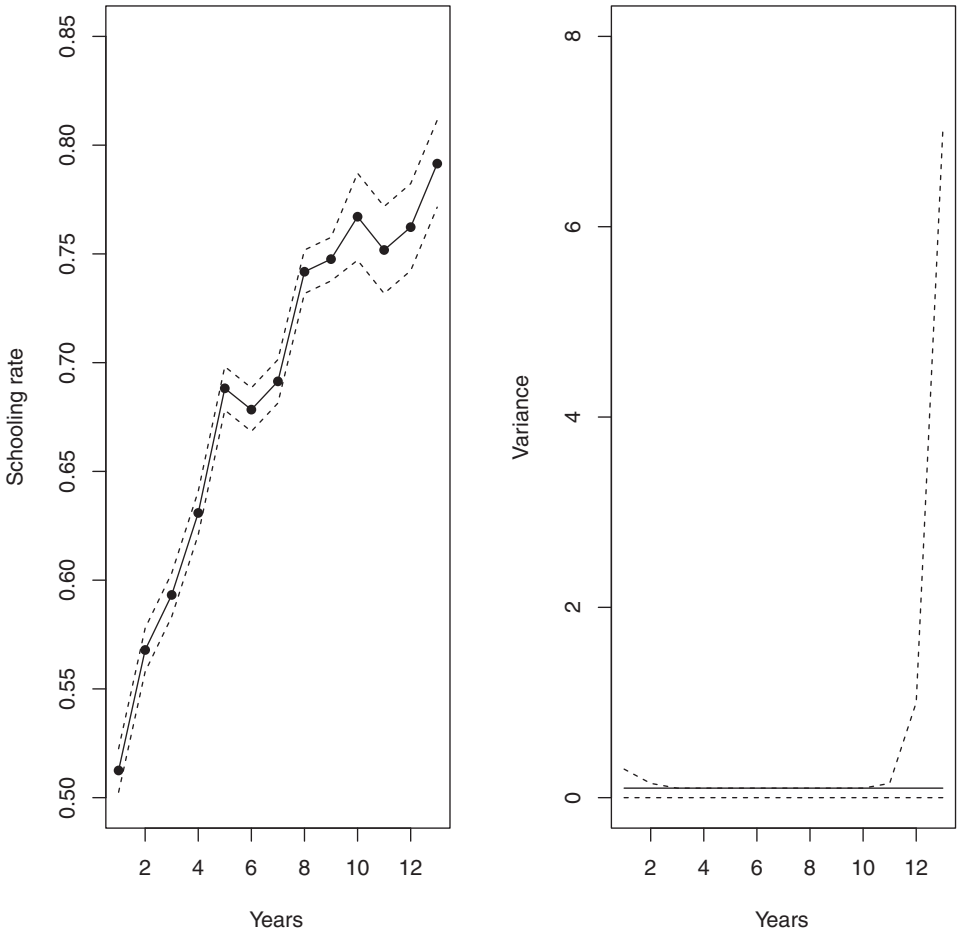

Figure 2. Normal nonlinear model with a random effect. (a) Prediction: fitted posterior mean (solid line) and 95% prediction interval (dotted lines). (b) Posterior variance with uncertainty limits.

6. Conclusions and Further Research

This research was based on the development of a systematic Bayesian methodology in order to perform MCMC algorithms to compute posterior estimates of mean and dispersion parameters in nonlinear models, even when considering random effects. The resulting techniques are general cases of some recent approaches and are capable of performing estimations in a vast range of statistical areas, such as, generalized linear and nonlinear models, mixed linear and nonlinear models or even classical regression.

We observe that in particular, when $f(\mathbf{x}_i, \boldsymbol{\beta}) = \mathbf{x}_i'\boldsymbol{\beta}$, then $\tilde{\mathbf{X}} = \mathbf{X}$, the vector of working variables $\tilde{\mathbf{y}}$ becomes the vector $\dot{\mathbf{y}}$ and the algorithm has only one stage. Thus, the algorithm becomes a replica of that in Cepeda and Gammerman (2005), with constant variance. Besides, if f remains nonlinear, the link function is the identity and the errors of the model (1) follow a normal distribution with null mean and constant variance, then the

limits of the former model (Fig. 1), we found that the variance of the model with random effects is actually much smaller, and therefore it has a better model fit.

working variable becomes $\tilde{y}_i = \tilde{\mathbf{x}}_i \boldsymbol{\beta}^{(c)} + y_i - f(\mathbf{x}_i, \boldsymbol{\beta}^{(c)})$ and the algorithm is completed after two stages. In this case, the algorithm also becomes a replica of that in Cepeda and Achcar (2010), with constant variance. However, when $f(\mathbf{x}_i, \boldsymbol{\beta}) = \mathbf{x}_i' \boldsymbol{\beta}$, the link function is the identity and the errors of the model (1) follow a normal distribution with null mean and constant variance, we have that $\tilde{\mathbf{X}} = \mathbf{X}$; in this case, the vector of working variables $\tilde{\mathbf{y}}$ becomes the vector $\mathbf{y}$, $\tilde{\mathbf{V}} = \boldsymbol{\Sigma} = \text{Cov}(\mathbf{y})$, the proposal jumping distribution (8) fully coincides with the posterior distribution (3), and the acceptance probability would be one at each step of the algorithm. This is the case of a Gibbs sampler algorithm (Geman and Geman, 1984), widely known in the Bayesian framework (Gelman et al., 2003).

Although the proposed approach is new, to our knowledge, the idea of joint modeling of mean and dispersion parameters is not. Thus, the above general algorithm is related to other recent works in the context of joint modeling of mean and dispersion parameters under the Bayesian framework. So then, if we assume that $\text{Var}(\boldsymbol{\lambda}) = \mathbf{0}$; i.e, the model does not consider any random effect, then the algorithm omits steps 5, 6, and 9. Under this formulation, this methodology yields a new Bayesian technique to the joint modeling of mean and dispersion parameters in nonlinear generalized models. However, if the mean function is considered to be linear, then the algorithm becomes that of Cepeda and Gammerman (2005). Moreover, if the nonlinearity remains but the link function is the identity and the residuals of the model follow a Gaussian distribution, the algorithm becomes that of Cepeda and Achcar (2010). If we consider normality and linearity of the mean, then the algorithm becomes that of Cepeda and Gammerman (2001). On the other hand, if we keep the random effect assumption, but omit the nonlinearity and consider that the dispersion of the response variables is constant among the units, then we must omit steps 7 and 8, and the algorithm becomes that of Gammerman (1997). Finally, as an extreme case, if we consider normality, linearity, and constant variance and if we omit the random effects, then this technique yields a Gibbs-type algorithm and the acceptance rate will be unity.

Among others, and given the generality of the nonlinear mixed generalized model considered, we especially find this approach useful when modeling data with multilevel or hierarchical structures (Gelman and Hill, 2007), when the mean response takes a linear (Raudenbush and Bryk, 2002) or nonlinear (Verbeke and Molenberghs, 2000) form, and even when the distribution of the response deviates from normality (Barry et al., 2003). Another application of interest is given by small area estimation (Pfefferman, 2002; Rao, 2001), where inference is needed in subdomains of interest. Specifically, this approach could be suitable for both models Type A and Type B. The first one (Rao, 2003, p. 75) is related to the use of auxiliary information at the aggregate (area) level, and the second (Rao, 2003, p. 76) deals with auxiliary information at the unit level. Those models are particular cases of a general mixed model. Note that when the variable of interest is discrete, it is possible to propose a generalized model with random effects to model the parameters in the areas of interest just as Ghosh et al. (1998) did.

As a final note, we claim that a strong assumption of this methodology is the independence of the errors of the models; i.e., we consider that the errors are uncorrelated. However, in practice, it is common to deal with datasets where the units cannot be considered independent among themselves. This is the case of longitudinal studies, repeated measures, or spatial and temporal models. Thus, when independence fails, the proposed approach is no longer valid. At the moment, extensions allowing correlation between subjects are being considered by the authors following the approach of Cepeda and Nuez (2009) and antedependence models (Zimmerman and Nuez, 2009).

Appendix A: BUGS Codes

Here, we include the BUGS codes that should be used in order to perform the joint estimation of mean and dispersion parameters in a generalized nonlinear model, without random effects, when the link function of the mean is different from the identity, as stated before. These codes are a different type from those in Cepeda and Achcar (2010), where they seem to work only with the identity link function of the mean in order to fit the data to their so-called double generalized nonlinear models. In that paper, they seem to omit that one must be careful when using some BUGS codes that suggest using flat priors on the coefficients in the linear predictor, because it is well known that such a specification leads to improper posterior distributions, as Ibrahim and Laud (1991), Hobert and Casella (1996), and Gelfand and Sahu (1999) claim. These arguments extend to the modeling of dispersion.

We strongly emphasize that, when considering other link functions, omitting their inverses leads to wrong estimates of the parameters. Consequently, we should conclude that the approach given by Cepeda and Achcar (2010) remains valid only when assuming normality of the response variable or when the link function is the identity if considering other densities. Thus, in cases such as Beta or Gamma distributions with different link functions, we encourage the use of the approach presented in this article or the exploratory use of the following computational BUGS codes:

Normal nonlinear model:

```
model{
for(i in 1:N){
y[i] dnorm(mu[i], tau[i])
f[i]<-b1/(1+b2*exp(b3*x[i]))
mu[i]<-f[i]
sigma[i]<-exp(g1+g2*x[i]+g3*x[i]*x[i])
tau[i]<-1/sigma[i]
            }
b1 dflat()
b2 dflat()
b3 dflat()
g1 dflat()
g2 dflat()
g3 dflat()
}
```

Beta nonlinear model:

```
for(i in 1:N){
y[i] dbeta(a[i], b[i])
a[i]<-mu[i]*tau[i]
b[i]<-tau[i]-mu[i]*tau[i]
f[i]<-b1/(1+b2*exp(b3*x[i]))
logit(mu[i])<-f[i]
tau[i]<-exp(g1+g2*x[i])
y.pred[i] dbeta(a[i], b[i])
            }
b1 dflat()
b2 dflat()
```

```
b3 dflat()
g1 dflat()
g2 dflat()
}
```

Gamma nonlinear model:

```
model{
for(i in 1:N){
y[i] dgamma(a[i], b[i])
a[i]<-tau[i]
b[i]<-tau[i]/mu[i]
f[i]<-b1/(1+b2*exp(b3*x[i]))
mu[i]<-1/f[i]
tau[i]<-exp(g1+g2*x[i])
            }
b1 dflat()
b2 dflat()
b3 dflat()
g1 dflat()
g2 dflat()
}
```

Normal nonlinear model with random effects:

```
model{
for(i in 1:N){
y[i] dnorm(mu[i], tau[i])
mu[i]<-b1[i]/(1+b2*exp(b3*x[i]))
sigma[i]<-exp(g1+g2*x[i])
tau[i]<-1/sigma[i]

b1[i]<-a1+u[i]
u[i] dnorm(0,tau.a1)
            }
a1 dflat()
b2 dflat()
b3 dflat()
g1 dflat()
g2 dflat()
tau.a1 dgamma(0.0001,0.0001)
}
```


Appendix B: Proofs of Results

Proof of result 2.1. Let us consider y_i, which has an exponential distribution with mean $E(y_i) = \mu_i$ and variance $Var(y_i) = v_i$. Also, we define a working variable, $\dot{y}_i$, as the observation yielded by the Fisher scoring technique over the function $h(y_i)$ evaluated at the point $E(y_i)$ in some neighborhood of μ_i. In this manner, and taking into account that

$\mu_i = h^{-1}(f(\mathbf{x}_i, \boldsymbol{\beta}))$, we have that

$$\begin{aligned}\dot{y}_i &:= h(E(y_i)) + h'(E(y_i))(y_i - E(y_i)) \\ &= h(\mu_i) + h'(\mu_i)(y_i - \mu_i) \\ &= f(\mathbf{x}_i, \boldsymbol{\beta}) + h'\left(h^{-1}[f(\mathbf{x}_i, \boldsymbol{\beta})]\right)\left(y_i - h^{-1}[f(\mathbf{x}_i, \boldsymbol{\beta})]\right).\end{aligned}$$

Following the work of West (1985), $\dot{y}_i$ has an associated normal distribution with mean

$$\begin{aligned}E[\dot{y}_i] &= f(\mathbf{x}_i, \boldsymbol{\beta}) + h'\left(h^{-1}[f(\mathbf{x}_i, \boldsymbol{\beta})]\right) E\left(y_i - h^{-1}[f(\mathbf{x}_i, \boldsymbol{\beta})]\right) \\ &= f(\mathbf{x}_i, \boldsymbol{\beta}) + h'\left(h^{-1}[f(\mathbf{x}_i, \boldsymbol{\beta})]\right)\left(\mu_i - h^{-1}[f(\mathbf{x}_i, \boldsymbol{\beta})]\right) \\ &= f(\mathbf{x}_i, \boldsymbol{\beta}),\end{aligned}$$

and variance $\tilde{V}_i$ given by

$$\begin{aligned}\mathrm{Var}[\dot{y}_i] := \tilde{V}_i &= Var[f(\mathbf{x}_i, \boldsymbol{\beta}) + h'\left(h^{-1}[f(\mathbf{x}_i, \boldsymbol{\beta})]\right)\left(y_i - h^{-1}[f(\mathbf{x}_i, \boldsymbol{\beta})]\right)] \\ &= (h'\left(h^{-1}[f(\mathbf{x}_i, \boldsymbol{\beta})]\right))^2 \mathrm{Var}[y_i] \\ &= (h'\left(h^{-1}[f(\mathbf{x}_i, \boldsymbol{\beta})]\right))^2 v_i.\end{aligned}$$

On the other hand, the pseudo-likelihood of the vector of working variables $\dot{\mathbf{y}} = (\dot{y}_i, \ldots, \dot{y}_n)'$ is given by

$$p(\dot{\mathbf{y}} \mid \boldsymbol{\beta}) \propto \exp\left\{\frac{-1}{2}(\dot{\mathbf{y}} - f(\mathbf{x}_i, \boldsymbol{\beta}))'\tilde{\mathbf{V}}^{-1}(\dot{\mathbf{y}} - f(\mathbf{x}_i, \boldsymbol{\beta}))\right\}, \tag{26}$$

with $\tilde{\mathbf{V}}$ being the diagonal matrix with entries $\tilde{V}_i$.

Proof of result 2.2. The demonstration follows immediately by applying the first order Taylor linearization for a multivariate function evaluated at the point $\boldsymbol{\beta} = \boldsymbol{\beta}^{(c)}$.

Proof of result 2.3. First, and as a consequence of the previous results, the approximation of the pseudo-likelihood is given by replacing $f(\mathbf{x}_i, \boldsymbol{\beta})$ by $f(\mathbf{x}_i, \boldsymbol{\beta}^{(c)}) + \tilde{\mathbf{x}}_i(\boldsymbol{\beta} - \boldsymbol{\beta}^{(c)})$. Note that the first term of the quadratic form in (5) becomes

$$\begin{aligned}\dot{y}_i - f(\mathbf{x}_i, \boldsymbol{\beta}) &\cong \dot{y}_i - f(\mathbf{x}_i, \boldsymbol{\beta}^{(c)}) - \tilde{\mathbf{x}}_i(\boldsymbol{\beta} - \boldsymbol{\beta}^{(c)}) \\ &= \dot{y}_i - f(\mathbf{x}_i, \boldsymbol{\beta}^{(c)}) + \tilde{\mathbf{x}}_i\boldsymbol{\beta}^{(c)} - \tilde{\mathbf{x}}_i\boldsymbol{\beta} \\ &= \tilde{y}_i - \tilde{\mathbf{x}}_i\boldsymbol{\beta}.\end{aligned}$$

Note that

$$\begin{aligned}\tilde{y}_i &= \dot{y}_i - f(\mathbf{x}_i, \boldsymbol{\beta}^{(c)}) + \tilde{\mathbf{x}}_i\boldsymbol{\beta}^{(c)} \\ &= f(\mathbf{x}_i, \boldsymbol{\beta}^{(c)}) + h'\left(h^{-1}[f(\mathbf{x}_i, \boldsymbol{\beta}^{(c)})]\right)\left(y_i - h^{-1}[f(\mathbf{x}_i, \boldsymbol{\beta}^{(c)})]\right) - f(\mathbf{x}_i, \boldsymbol{\beta}^{(c)}) + \tilde{\mathbf{x}}_i\boldsymbol{\beta}^{(c)} \\ &= \tilde{\mathbf{x}}_i\boldsymbol{\beta}^{(c)} + h'\left(h^{-1}[f(\mathbf{x}_i, \boldsymbol{\beta}^{(c)})]\right)\left(y_i - h^{-1}[f(\mathbf{x}_i, \boldsymbol{\beta}^{(c)})]\right).\end{aligned}$$

Then, by conditioning in the point $\boldsymbol{\beta} = \boldsymbol{\beta}^{(c)}$, the new working variable $\tilde{y}_i$ has a Gaussian distribution with mean

$$\begin{aligned} E(\tilde{y}_i) &= E\left[\tilde{\mathbf{x}}_i\boldsymbol{\beta}^{(c)} + h'\left(h^{-1}[f(\mathbf{x}_i, \boldsymbol{\beta}^{(c)})]\right)\left(y_i - h^{-1}[f(\mathbf{x}_i, \boldsymbol{\beta}^{(c)})]\right)\right] \\ &= \tilde{\mathbf{x}}_i\boldsymbol{\beta}^{(c)} + h'\left(h^{-1}[f(\mathbf{x}_i, \boldsymbol{\beta}^{(c)})]\right)\left(E[y_i] - h^{-1}[f(\mathbf{x}_i, \boldsymbol{\beta}^{(c)})]\right) \\ &= \tilde{\mathbf{x}}_i\boldsymbol{\beta}^{(c)}, \end{aligned}$$

and variance

$$\begin{aligned} \mathrm{Var}(\tilde{y}_i) &= \mathrm{Var}\left[\tilde{\mathbf{x}}_i\boldsymbol{\beta}^{(c)} + h'\left(h^{-1}[f(\mathbf{x}_i, \boldsymbol{\beta}^{(c)})]\right)\left(y_i - h^{-1}[f(\mathbf{x}_i, \boldsymbol{\beta}^{(c)})]\right)\right] \\ &= (h'\left(h^{-1}[f(\mathbf{x}_i, \boldsymbol{\beta}^{(c)})]\right))^2\mathrm{Var}[y_i] \\ &= \tilde{V}_i. \end{aligned}$$

As a consequence, the new pseudo-likelihood will be given by

$$p(\tilde{\mathbf{y}} \mid \boldsymbol{\beta}) \propto \exp\left\{\frac{-1}{2}(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}'\boldsymbol{\beta})'\tilde{\mathbf{V}}^{-1}(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}'\boldsymbol{\beta})\right\}.$$

Combining the former distribution with the prior distribution of $\boldsymbol{\beta}$, we clearly see that the proper posterior distribution induced by this approach is given by (9).

Proof of result 3.1. The demonstration follows immediately by noting that

$$\begin{aligned} \nabla f(\mathbf{x}_i, \boldsymbol{\beta}^{(c)}, \mathbf{w}_i, \boldsymbol{\lambda}_i^{(c)})\begin{pmatrix} \boldsymbol{\beta} - \boldsymbol{\beta}^{(c)} \\ \boldsymbol{\lambda}_i - \boldsymbol{\lambda}_i^{(c)} \end{pmatrix} &= \nabla_{\boldsymbol{\beta}} f(\mathbf{x}_i, \boldsymbol{\beta}^{(c)}, \mathbf{w}_i, \boldsymbol{\lambda}_i^{(c)})(\boldsymbol{\beta} - \boldsymbol{\beta}^{(c)}) \\ &\quad + \nabla_{\boldsymbol{\lambda}_i} f(\mathbf{x}_i, \boldsymbol{\beta}^{(c)}, \mathbf{w}_i, \boldsymbol{\lambda}_i^{(c)})(\boldsymbol{\lambda}_i - \boldsymbol{\lambda}_i^{(c)}). \end{aligned}$$

Then, by using results 2.1 and 2.2, a first general working variable is given by

$$\tilde{y}_i := \tilde{\mathbf{x}}_i\boldsymbol{\beta}^{(c)} + \tilde{\mathbf{w}}_i\boldsymbol{\lambda}_i^{(c)} + h'\big(h^{-1}\big[f\big(\mathbf{x}_i, \boldsymbol{\beta}^{(c)}, \mathbf{w}_i, \boldsymbol{\lambda}_i^{(c)}\big)\big]\big)\big(y_i - h^{-1}\big[f\big(\mathbf{x}_i, \boldsymbol{\beta}^{(c)}, \mathbf{w}_i, \boldsymbol{\lambda}_i^{(c)}\big)\big]\big).$$

This working variable has expectation $\tilde{\mathbf{x}}_i'\boldsymbol{\beta} + \tilde{\mathbf{w}}_i'\boldsymbol{\lambda}_i$ and variance $\tilde{V}_i$. Then, for the vector of fixed effects $\boldsymbol{\beta}$, we define $\tilde{y}_i^b = \tilde{y}_i - \tilde{\mathbf{w}}_i'\boldsymbol{\lambda}_i$ with mean $\tilde{\mathbf{x}}_i'\boldsymbol{\beta}$ and variance $\tilde{V}_i$. The proof is completed by following a similar reasoning for the vector of random effects $\boldsymbol{\lambda}_i$.

Proof of result 3.2. After supposing the existence of the variable t_i with mean $E(t_i) = \tau_i$, we define the new working variable, according to the differentiability of g, as the first-order Taylor approximation of g, evaluated at the point $E(t_i)$, as follows:

$$\begin{aligned} \tilde{t}_i &= g(E(t_i)) + g'(E(t_i))(t_i - E(t_i)) \\ &= \mathbf{z}_i'\boldsymbol{\gamma} + g'(g^{-1}(\mathbf{z}'\boldsymbol{\gamma}))(t_i - g^{-1}(\mathbf{z}'\boldsymbol{\gamma})). \end{aligned}$$

Note that t_i has an associated Gaussian distribution with mean $E(\tilde{t}_i) = \mathbf{z}_i'\boldsymbol{\gamma}$ and variance $\tilde{T}_i = \mathrm{Var}(\tilde{t}_i) = [g'(g^{-1}(\mathbf{z}'\boldsymbol{\gamma}))]^2\mathrm{Var}(t_i)$. By assuming prior independence between t_i and t_j $(i \neq j)$, and defining $\tilde{\mathbf{t}} = (\tilde{t}_1, \ldots, \tilde{t}_i, \ldots, \tilde{t}_n)'$, the pseudo-likelihood of the vector of working variables is given by

$$p(\tilde{\mathbf{t}} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \mathbf{L}) \propto \exp\left\{\frac{-1}{2}(\tilde{\mathbf{t}} - \mathbf{Z}'\boldsymbol{\gamma})'\tilde{\mathbf{T}}^{-1}(\tilde{\mathbf{t}} - \mathbf{Z}'\boldsymbol{\gamma})\right\}.$$

Thus, after combining the previous density with the prior of $\boldsymbol{\gamma}$, we can easily obtain the result.

References

Aitkin, M. (1987). Modelling variance heterogeneity in normal regression using GLIM. *Applied Statistics* 36:332–339.

Barndorf-Nielsen, O. E. (1978). *Information and Exponential Families in Statistical Theory*. New York: Wiley.

Barry, S. C., Brooks, S. P., Catchpole, E. A., Morgan, B. J. T. (2003). The analysis of ring-recovery data using random effects. *Biometrics* 59:54–65.

Cepeda, E., Achcar, J. A. (2010). Heteroscedastic nonlinear regression models. *Comunnications in Statistics Simulation and Computation* 39:405–419.

Cepeda, E., Gammerman, D. (2001). Bayesian modeling of variance heterogeneity in normal regression models. *Brazilian Journal of Probability and Statistics* 14:207–221.

Cepeda, E., Gammerman, D. (2005). Bayesian methodology for modeling parameters in the two parameter exponential family. *Estadstica* 57:93–105.

Cepeda, E., Nuez, V. A. (2009). Bayesian modeling of the mean and covarince matrix in normal nonlinear models. *Journal of Statistical Computation and Simulation* 79:837–853.

Charalambous, C., Pan, J., Tranmer, M. (2010). *Variable Selection in Joint Modelling of the Mean and Variance via h-likelihood*. Research Report no. 14. Probability and Statistics Group, School of Mathematics, Manchester: The University of Manchester.

Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *Journal of the Royal Statistical Association B* 49:1–39.

Dellaportas, P., Smith, A. F. M. (1993). Bayesian inference for generalized linear and proportional hazards model via Gibbs sampling. *Applied Statistics* 42:443–460.

Dey, D. K., Gelfand, A. E., Peng, F. (1997). Overdispersed generalizaed linear models. *Journal of Statistical Planning and Inference* 64:93–107.

Gammerman, D. (1997). Efficient sampling from the posterior distributions in generalized linear mixed models. *Statistics and Computing* 7:57–68.

Gelfand, A. E., Dalal, S. R. (1990). A note on overdispersed exponential families. *Biometrika* 77:55–64.

Gelfand, A. E., Sahu, S. K. (1999). Identifiability, improper priors, and Gibbs sampling for generalized linear models. *Journal of the American Statistical Association* 91:247–253.

Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D. B. (2003). *Bayesian Data Analysis.* 2 ed. Boca Raton, FL: Chapman and Hall/CRC.

Gelman, A., Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Gelman, A., Shirley, K. (2010). Inference from simulations and monitoring convergence. *Handbook of Markov Chain Monte Carlo*. Boca Raton, FL: CRC.

Geman, S., Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6:721–741.

Ghosh, M., Natarajan, K., Stroud, T. W. F., Carlin, B. P. (1998). Generalized linear models for small area estimation. *Journal of American Statististical Association* 93:273–282.

Gijbels, I., Prosdocimi, I., Claeskens, G. (2010). Nonparametric estimation of mean and dispersion functions in extended generalized linear models. *TEST*. 19:580–608.

Hastie, T., Tibshirani, R., Friedman, J. (2001). *The Elements of Statistical Learning*. New York: Springer.

Hastings, W. K. (1970). Monte carlo sampling methods using Markov chains and their applications. *Biometrika* 57:97–109.

Hobert, J. P., Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association* 91:1461–1473.

Ibrahim, J. G., Laud, P. W. (1991). On Bayesian analysis of generalized linear models using Jeffreys's prior. *Journal of the American Statistical Association* 86:981–986.

Jorgersen, B. (1987). Exponential dispersion models (with discussion). *Journal of the Royal Statistical Association B* 49:150.

Laird, N. M., Ware, J. H. (1982). Random effects models for longitudinal data. *Biometrics* 38:963–974.

Lee, Y., Nelder, J. A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society, series B* 58:619–678.

Lee, Y., Nelder, J. A., Pawitan, Y. (2006). *Generalized Linear Models with Random Effects: A Unified Analysis via h-Likelihood*. Boca raton, FL: Chapman and Hall/CRC.

McCullagh, P., Nelder, J. A. (1996). *Generalized Linear Models*. Boca Raton, FL: Chapman and Hall.

McCulloch, C. E., Searle, S. R. (2001). *Generalized, Linear and Mixed Models*. New York: Wiley.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics* 21:1087–1092.

Neykov, N. M., Filzmoser, P., Neytchev, P. N. (2012). Robust joint modeling of mean and dispersion through trimming. *Computational Statistics and Data Analysis* 56(1): 34–48.

Ntzoufras, I. (2009). *Bayesian Modeling Using WinBUGS*. Hoboken, NJ: Wiley.

Pfefferman, D. (2002). Small area estimation: New developments and directions. *International Statistic Review* 70:125–143.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rao, J. N. K. (2001). EB and EBLUP in Small Area Estimation. *Empirical Bayes and Likelihood Inference*. New York: Springer. pp. 33–43.

Rao, J. N. K. (2003). *Small Area Estimation*. Hoboken, NJ: Wiley.

Raudenbush, S. W., Bryk, A. S. (2002). *Hierarchical Linear Models*. Thousand Oaks, CA: Sage.

Rigby, R. A., Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 54(3): 507–554.

Robert, C. P., Casella, G. (2009). *Introducing Monte Carlo Methods with R*. New York: Springer.

Smyth, G. K. (2002). An efficient algorithm for REML in heteroscedastic regression. *Journal of Graphical and Computational Statistics* 11:836–847.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., VanderLinde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society* B 64:583–639.

Spiegelhalter, D., Thomas, A., Best, N., Lunn, D. (2004). *WinBUGS User Manual*.

Verbeke, G., Molenberghs, G. (2000). *Linear Models for Longitudinal Data*. New York: Springer-Verlag.

West, M. (1985). Generalized Linear Models: Outlier Acommodation, Scale Parameters and Prior distributions (with discussion). In: *Bayesian Statistics 2*, Eds., Bernardo, J. M., De Groot, M. H., Lindley, D. V., and Smith, A. F. M. Oxford: Oxford University Press. pp. 461–484.

Zhao, Y., Staudenmayer, J., Coull, B. A., Wand, M. P. (2006). General design Bayesian generalized linear mixed models. *Statistical Science* 21:35–51.

Zimmerman, D. L., Nuez, V. A. (2009). *Antedependence Models For Longitudinal Data*. Boca Raton, FL: CRC Press.