A Solution to Separation for Clustered Binary Data

José Cortiñas Abrahantes and Marc Aerts

Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-BioStat), Center for Statistics, Hasselt University. Agoralaan 1, B-3590 Diepenbeek, Belgium

Summary. The presence of one or more covariates which perfectly or almost perfectly predict the outcome of interest (which is referred to as complete or quasi-complete separation, the latter denoting the case when such perfect prediction occurs only for a subset of observations in the data) has been extensively studied in the last four decades. Since 1984, when Albert and Anderson (1984) differentiated between complete and quasi-complete separation, several authors have studied this phenomenon and tried to provide answers or ways of identifying the problem (Lesaffre and Albert, 1989; Firth, 1993; Christmann and Rousseeuw, 2001; Rousseeuw and Christmann, 2003; Allison, 2004; Zorn, 2005; Heinze, 2006). From an estimation perspective, separation leads to infinite coefficients and standard errors, which makes the algorithm collapse or give inappropriate results. As a practical matter, separation forces the analyst to choose from a number of problematic alternatives for dealing with the problem, and in the past the elimination of such problematic variables were common practice to deal with such situations. In the last decade solutions using penalized likelihood have been proposed, but always dealing with independent binary data. Here we will propose a Bayesian solution to the problem when we deal with clustered binary data using informative priors that are supported by the data and compare it with an alternative procedure proposed by Gelman et al. (2008).

1. Introduction

In the setting of binary-response data, logistic regression is one of the most common tools used in real life situations. In general, the standard estimation method is based on the maximum likelihood principle. Maximum likelihood estimates, however, do not always produce finite estimates (Silvapulle, 1981). Whenever the likelihood does not have a maximum, the numerical algorithms implementing the search for this maximum are bound to encounter serious problems or to produce erroneous and non-sensible results. This non existence of finite estimated values in logistic regression is commonly known as the separation problem.

The name separation is due to the fact that the likelihood will not have a maximum whenever one of the independent variables used as a the linear predictor is able to perfectly separate or classify the observations into the respective groups of the response variable. The issue is known at least since Day and Kerridge (1967) and is not unique to modeling binary data. When modeling times-to-event, the problem is known as "monotone likelihood" (Bryson and Johnson, 1981). Later Albert and Anderson (1984) introduced a useful distinction between complete separation and quasi-separation. Although proposed for multinomial logit models, the distinction applies equally well to the binomial situation.

When we deal with logistic regression problems the existence of maximum likelihood estimates depends on the configuration of the sample points in the observation space, which we can subdivide in three mutually exclusive and exhaustive categories: complete separation, quasi-complete separation, and overlap. Complete separation refers to the case for which we can find a vector $\mathbf{b}$ that, if multiplied by the covariate value, correctly allocates all observations in the data to their corresponding response group, given for example, the sign of the product. In the particular case of a binary covariate X, complete separation corresponds to "empty cells" in the off diagonal of the implied 2×2 table formed by the two variables (covariate versus response). Basically we can say that the response Y can be perfectly predicted by the covariate X for all the observations in the data. The consequence is that there is no variance left to be explained in Y by the other covariates in the model and the likelihood will be flat, the diagonal elements of the inverse of the information matrix will be infinite in size, thus yielding infinite standard error estimates. Quasi-complete separation is found if the data are not completely separated and there exists a vector $\mathbf{b}$ such that we can

subdivide the space $(\mathbf{b} \cdot \mathbf{X})$ into separating spaces corresponding to the response groups, but there might exist members of several groups of the response that falls into this separating hyperplane. In the particular case of a binary covariate X, quasi-complete separation occurs when only one cell of the off-diagonal of the implied 2×2 table is "empty". Under quasi-complete separation, the parameter estimate for variable X and its standard errors will also be infinite, but the effect of other covariates may remain relatively unaffected. Whenever data do not exhibit either complete or quasi-complete separation, the data are said to overlap. For overlapping data the maximum likelihood estimates are proven to exist and to be unique (Silvapulle, 1981).

The phenomenon of separation in a dataset was studied via simulation by Heinze and Schemper (2002). According to their investigations, separation depends on sample size, on the number of dichotomous covariates involved in the analysis, the magnitude of the coefficients and how unbalanced the response is. They showed as well that separation can happen even when the absolute value of the model parameters is low. In view of these findings it is important to be able to detect such problems. Several authors have published on the issue of such detection. The detection methods range from very simple ad-hoc methods to optimization solutions that are proven to result in correct detection of the phenomenon in the dataset. A first simple way to deal with the problem is to cross-classify the independent variables one by one with the dependent variable and search for 0 values in the cells preliminary to the analysis. Alternatively, one can use very large coefficients with huge standard errors as indices for numerical problems. A third approach is to simply inspect how the coefficients evolve during the iterations of the fitting algorithm. If the estimates for a particular parameter do not stabilize but 'jump around' in the solution space, separation might well be the culprit. All of these checks, however, do not give certainty on the origin of the problem.

For this reason, Albert and Anderson (1984) propose a linear programming method that always leads to detection. Santner and Duffy (1986) expand on Albert and Anderson (1984) and provide a mixed linear programming algorithm that can be used to determine whether there is complete separation, quasi-complete separation or overlap. Another interesting complement to these methods was offered by researchers in robust statistics. Christmann and Rousseeuw (2001) developed a way not to find separation, but quantify the degree of overlap in a dataset. A software implementation of this approach is offered by the R package noverlap.

However, detecting the problem, is just the first step, solution to such problems are of practical importance, specially when dealing with data analysis presenting such issues. Several strategies have been developed to deal with both complete or quasi-complete separation. The most widely-used "solution" is simply to omit the problematic covariate or covariates from the analysis. Of course, this alternative is particularly unattractive, since omitting a covariate that clearly bears a strong relationship to the phenomenon of interest is nothing more than deliberately introducing specification bias. Another approach suggested by Clogg et al. (1991) modifies the data in order to eliminate the separation, supplementing additional "artificial" data across the various patterns of (categorical) covariates, and then conduct the analysis in the usual fashion on the augmented data. Unfortunately, it has been demonstrated in a simulation study (Heinze and Schemper, 2002) that it does not perform as well as other alternative solutions. Another solution could be for the case in which the problematic variable is of a nominal type with more than two categories, to group categories, which can help for quasi-complete separation (but not for complete separation). Exact logistic regression, already suggested by Cox (1970), could be an alternative approach, but the method can be computer-intensive thus practically less feasible.

Firth explored a general method of bias reduction using penalized likelihood in a 1992 paper and finalized his ideas in 1993 (Firth, 1993). The intuition of Firth's approach is to introduce a bias term into the standard likelihood function which itself goes to zero as sample size increases. Although the initial purpose of Firth was to reduce the bias of the maximum likelihood estimates, the method has many advantageous characteristics when data display separation (Heinze, 2006). Heinze and Schemper (2002) advocate using penalized profile likelihood confidence intervals instead of the Wald intervals used by Firth. Gao and Chen (2007) develop the concept further and introduce a second penalizing term to deal with ill-behaving due to multicollinearity. In contexts where multicollinearity is likely to arise, it might indeed be important to clearly deal with the two numerical problems separately.

Another alternative is to work in a Bayesian paradigm and use MCMC methods to fit such models, but according to Allison (2004) uninformative priors in general lead to convergence problems. And the use of informative priors might lead to results depending rather heavily on the mean and variance of the prior distribution. Thus, using Bayesian models when dealing with logistic regression models which suffers from separation issues should be done with care. In a recent paper, however, Gelman et al. (2008) develop a method that addresses this criticism. First, all nonbinary variables are scaled to have mean 0 and standard deviation 0.5, and then independent Student-t prior distributions are put on the coefficients. It is stated that the method always converges (even when complete separation is present), due to the use of what the authors called weakly informative priors.

The robust statistics community also contributed to the issue of finding solutions to separation, when they introduced the concept of hidden logistic regression (Rousseeuw and Christmann, 2003). These authors assume unobservable true responses and propose a maximum estimated likelihood estimator. The estimator is robust against separation and always exists. A weighted version is proposed as well to render the estimator robust against outliers.

In general several alternatives have been proposed in the past to deal with separation issues in logistic regression problems, but all, with exception of Gelman et al. (2008) have been extensively discussed in the context of independent observations. Here in this article we will propose an alternative approach when dealing with clustered binary data, in a Bayesian framework and using penalized likelihood to obtain plausible prior distributions for the regression coefficients corresponding to the covariates having separation issues. We will also compare this approach with the approach proposed by Gelman et al. (2008) in a simulation study, in terms of bias, variance and mean squared error of the estimates.

First a motivating case study is presented in Section 2. In Section 3, the model setting used throughout the paper is described. The proposed method, which combines results obtained from Firth's penalized likelihood approach with the use of appropriate priors in a Bayesian framework is described in Section 4. The simulation setting to evaluate and compare the performance of this alternative with the existing method of Gelman et al. (2008) is presented in Section 5. Section 6 present the results of case study. We close the paper with some discussion and recommendations in Section 7.

2. Motivating Case Study: Analysis of *Salmonella* Data in Pigs

Since January 2005 blood samples of growing and fattening pigs are collected from all (fattening) pig herds in Belgium within the Aujeszky disease monitoring program and it is combined with an indirect commercial enzyme linked immunosorbent assay (ELISA, Idexx LaboratoriesR) for the detection of Salmonella-specific antibodies. Within this monitoring program, blood samples of 12 or more animals (depending on the herd size) have to be collected in each herd in Belgium every 4 months in order to determine the Aujeszky status. The presence of antibodies against Salmonella in each sample was determined by relating the optical density (OD) values to the mean positive kit control by calculating the sample to positive ratio corrected with negative background values of the kit (SP-ratio = Optical Density (ODsample - ODNeg Kit control) / (ODPos Kit control - ODNeg-kit control)). SP-ratio's are generally recorded between 0 and 4, and in general a dichotomized version is used to classify *Salmonella* infected animals(SP-ratio > 0.35). Beside SP-ratio's the final dataset contains other covariates as well (herd id number, sampling date, cleaning (if cleaning is performed after a batch is leaving the farm, before a new batch is coming), vermin (presence or absence during the stay of a batch),...). Other variables were also collected, but we will just focus on those previously mentioned to illustrate the problem and the solutions proposed. Variables cleaning and vermin were dichotomous and sampling time was considered as a continuous variable (number of days, calculated as the sampling date minus the starting date of the national Salmonella program (=01/01/2005)).

The dataset contain information from 83 herds, with a total of 4179 observations. The average number of pigs sampled per herd is around 50. We selected from the complete datasets (containing 314 herds) those herds with more than 30 observations. The selection is based on the fact that the monitoring program aims

Table 1. Contingency tables between dichotomous covariates and binary response.

Salmonella	Vermin		Cleaning	
	Absence	Presence	Absence	Presence
Absence	187	3386	296	3277
Presence	17	589	0	606

Table 2. Parameter estimates of logistic regression model (2.1) without random effects.

Coefficients	Estimate	Std. Error	Z value	P-value
Intercept	-18.2944	229.8633	-0.08	0.9366
SamplingTime	-0.0661	0.0436	-1.52	0.1296
Vermin	0.7331	0.2574	2.85	0.0044
Cleaning	15.9028	229.8632	0.07	0.9448

at identifying possible problematic pig herds, with high infestation of Salmonella. The criteria applied by the Belgian Federal Agency for the Safety of the Food Chain (FASFC) to assign a pig herd as "high risk" are: a pig herd with mean S/P-ratio (mean value calculated from maximum 12 samples per sampling) equal or higher than 0.6 during 3 consecutive samplings. It implies that only herds with at least 3 consecutive samplings will be considered in the analysis. More details on the data can be found in Bollaerts et al. (2007) and Cortiñas Abrahantes et al. (2009).

The model to be used is the following:

$$\Pr(Y_{ij} = 1|X_{1ij}, X_{2ij}, X_{3ij}) \equiv \pi_{ij} = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 \cdot X_{1ij} + \beta_2 \cdot X_{2ij} + \beta_3 \cdot X_{3ij} + b_i)}, \tag{2.1}$$

where binary variables X_{1ij} and X_{3ij} represent presence or absence of 'vermin' and 'cleaning' respectively and variable X_{2ij} refers to the continuous variable 'sampling time'. The term b_i is a herd specific random intercept accounting for possible correlation at the cluster (herd) level.

The exploratory analysis of the data reveals that variable 'vermin' and 'sampling time' do not seem to suffer from separation issues. However, if we do include 'cleaning', quasi-complete separation is introduced as shown in Table 1. Note that the cross-tabulation is not provided as a definitive proof of separation, but as an exploratory tool to distinguish possible issues. Other methods such as those of Albert and Anderson (1984); Santner and Duffy (1986); Christmann and Rousseeuw (2001) could be used for this purpose. In a first stage we have fitted a logistic regression model considering the three covariates previously mentioned and ignoring intra-cluster correlation. The results are shown in Table 2. It is important to note that variable 'sampling time' was standardized in order to avoid computational issues due to scale differences between both covariates. Also here separation issues can be observed through the large values obtained for parameter estimates and standard errors.

If we use the penalized likelihood approach (see Section 3.2), Table 3 now shows much smaller parameter estimates and standard errors are very much reduced as well, providing a different interpretation of the parameters in the model (which previously indicated no significant effect of 'cleaning', for instance).

If we consider the model with a herd specific random intercept to account for intra-herd correlation (see Section 3.1; results shown in Table 4), then we also observe large standard errors for variable 'cleaning' and for the intercept (parameter estimates are comparable to the one obtained in Table 2). If inference would be based on this model, there is no significant effect of any of the covariates. It is clear that separation is preventing us from getting valid inference, thus it should be accounted for when dealing with clustered binary data. Another interesting point is the fact that standard error estimates almost doubled as compared to the results of logistic regression (assuming independence), especially for the intercept and the parameter

Table 3. Parameter estimates of penalized likelihood approach.

Coefficients	Estimate	Std. Error	Z value	P-value
Intercept	-7.0876	0.8470	-8.37	7.9e-17
SamplingTime	-0.0661	0.0436	-1.52	0.12962
Vermin	0.7073	0.2546	2.78	0.00546
Cleaning	4.7232	0.8341	5.66	1.6e-08

Table 4. Parameter estimates of logistic regression model (2.1) with a random intercept.

Coefficients	Estimate	Std. Error	Z value	P-value
Intercept	-18.5465	446.4685	-0.042	0.967
SamplingTime	-0.0799	0.0480	-1.665	0.096
Vermin	0.6578	0.8075	0.815	0.415
Cleaning	15.6507	446.4678	0.035	0.972
$\sigma^2_{b_i}$	1.9949	1.4124		

associated to 'cleaning'. Standard errors are expected to become larger in case of a positive intra-cluster correlation, as in this case. In general, the impact depends on the magnitude of the variance of the random effect.

3. Separation and Clustered Binary Data

This section summarizes relevant basic concepts of GLMM's (Generalized Linear Mixed Models) to model clustered binary data, Firth's penalized likelihood approach to deal with the separation issue, and it discusses the difficulties with extending Firth's methods to GLMM's, to the combined setting where separation appears while using clustered data.

3.1. Generalized Linear Mixed Models

The GLMM (Breslow and Clayton, 1993; Engel and Keen, 1994; Wolfinger and O'Connell, 1993) is the most frequently used random-effects model in the context of non-Gaussian clustered or repeated data. It is a relatively straightforward extension of the generalized linear model for independent data to the context of hierarchical data on the one hand and the linear mixed model (Verbeke and Molenberghs, 2000) on the other hand. Let us first introduce some notation. Let $\mathbf{Y} = (\mathbf{Y_1}, \ldots, \mathbf{Y_N})$ be the vector of binary responses, where the vector $\mathbf{Y_i}$ is grouping the n_i measurements Y_{ij}, β is a vector of fixed-effects parameters, $\mathbf{X_{ij}}$ represents the covariates, and $\mathbf{b}$ is a vector of random effects associated with a vector of covariates $\mathbf{Z_{ij}}$, where $i = 1, 2, \ldots, N$ refers to the different clusters, and $j = 1, 2, \ldots, n_i$ the observations collected within each cluster. The random effects are assumed to be randomly distributed with mean $\mathbf{0}$ and variance-covariance matrix $D = D(\theta)$, which depends on a $d + 1$ dimensional vector of parameters $\theta = (\theta_0, \theta_1, \ldots, \theta_d)$. Note also that a general random effects structure could be used, and as such $\mathbf{b_i}$ could be representing a vector. The density function of $\mathbf{b_i}$ will be denoted by $f(\mathbf{b_i})$. Note that we do not need to specify the nature of the distribution, meaning that no explicit assumption about the distribution of the random effects is needed. Then the logistic model can be written as follow:

$$\Pr(Y_{ij} = 1|\mathbf{X_{ij}}) \equiv \pi_{ij} = \frac{1}{1 + \exp(-\mathbf{X_{ij}} \cdot \beta - \mathbf{Z_{ij}} \cdot \mathbf{b_i})}. \tag{3.1}$$

For this particular scenario several likelihood functions exist: conditional, augmented or marginal likelihood.

The conditional log likelihood ($l^{\mathrm{C}}(\beta, \mathbf{b})$)is given by the expression:

$$\begin{aligned} l^{\mathrm{C}}(\beta, \mathbf{b}) &= \sum_{i=1}^{N}\sum_{j=1}^{n_i} l_{ij}^{\mathrm{C}}(\beta, \mathbf{b}) \\ &= \sum_{i=1}^{N}\sum_{j=1}^{n_i} \left\{(Y_{ij}-1)\cdot(\mathbf{X_{ij}}\cdot\beta + \mathbf{Z_{ij}}\cdot\mathbf{b_i}) - ln\left[1+\exp(-\mathbf{X_{ij}}\cdot\beta - \mathbf{Z_{ij}}\cdot\mathbf{b_i})\right]\right\}. \end{aligned} \tag{3.2}$$

The augmented log likelihood ($l^{\mathrm{A}}(\beta, \mathbf{b})$) then would be:

$$\begin{aligned} l^{\mathrm{A}}(\beta, \mathbf{b}) &= \sum_{i=1}^{N}\sum_{j=1}^{n_i} l^{\mathrm{C}}(\beta, \mathbf{b})_{ij} + lnf(\mathbf{b_i}) \\ &= \sum_{i=1}^{N}\sum_{j=1}^{n_i} \left\{(Y_{ij}-1)\cdot(\mathbf{X_{ij}}\cdot\beta + \mathbf{Z_{ij}}\cdot\mathbf{b_i}) - ln\left[1+\exp(-\mathbf{X_{ij}}\cdot\beta - \mathbf{Z_{ij}}\cdot\mathbf{b_i})\right] + lnf(\mathbf{b_i})\right\}. \end{aligned} \tag{3.3}$$

In general, the marginal log likelihood is then used to estimate the parameters of the model β and $\mathbf{b}$, and is then calculated as:

$$l^{\mathrm{M}}(\beta, \mathbf{b}) = \sum_{i=1}^{N} ln \int_{-\infty}^{\infty} \exp\left(\sum_{j=1}^{n_i} l^{\mathrm{C}}(\beta, \mathbf{b})_{ij} + lnf(\mathbf{b_i})\right) d\mathbf{b_i}. \tag{3.4}$$

This likelihood does not always have a closed form, as it depends on the distribution assumed for the random effects. To obtain the maximum likelihood estimates for β and θ in those cases, it is necessary to use numerical integration or approximation procedures for the integral. Another drawback is that distributional assumptions of the random effects $\mathbf{b}$ are difficult to verify. Given the complexity of the problem, obtaining the score equations and information matrix is not a straightforward procedure, since integrations are taking part of the likelihood function. Fitting such models is not always a straightforward task, depending on the specific assumptions on the random effects. It is relatively straightforward to fit GLMM' with normally distributed random effects (e.g. using PROC NLMIXED in SAS).

For the analysis of the *Salmonella* data as well as for the simulations, we will take a random intercept GLMM with normally distributed random effect.

3.2. Firth's Penalized Likelihood Approach

The penalized likelihood approach was proposed by Firth (1993), whose central idea was to develop a method to eliminate the small–sample bias in maximum likelihood estimation. The intuition behind his approach is to introduce a bias term into the standard likelihood function which on the one hand can be considered negligible when the sample size increases, but on the other hand eliminates the bias for small sample sizes. The result of his method is the so-called "penalized" likelihood:

$$\mathbf{L^P}(\beta) = \mathbf{L}(\beta)\cdot|\mathbf{I}(\beta)|^{\frac{1}{2}}. \tag{3.5}$$

The loglikelihood then would be:

$$ln\mathbf{L}^{\mathbf{P}}(\beta) = ln\mathbf{L}(\beta) + \frac{1}{2} \cdot ln|\mathbf{I}(\beta)|, \tag{3.6}$$

where $\mathbf{I}(\beta)$ is the information matrix written as:

$$[\mathbf{I}(\beta)]^{-1} = \left\{ \mathrm{E}\left[\frac{\partial^2 ln\mathbf{L}(\beta)}{\partial\beta\partial\beta'} \right] \right\}^{-1} = \mathbf{X}' \cdot \mathbf{W} \cdot \mathbf{X} = \mathbf{X}' \cdot \mathrm{diag}\left[\pi_i \cdot (1-\pi_i)\right] \cdot \mathbf{X}. \tag{3.7}$$

Firth (1993) demonstrates that this modified likelihood is asymptotically consistent, while eliminating the usual small–sample bias. Its estimates exist in situations where standard likelihood estimates do not, in particular when we face complete or quasi complete separation in binary response models. Using this penalized log-likelihood, we obtain the penalized score equations:

$$\begin{aligned} \mathbf{U}^{\mathbf{P}}(\beta_t) &= \frac{\partial \left\{ ln\mathbf{L}(\beta) + \frac{1}{2} \cdot ln|\mathbf{I}(\beta)| \right\}}{\partial \beta_p} \\ &= \mathbf{U}(\beta_p) + \frac{1}{2} \cdot \mathrm{trace}\left\{ [\mathbf{I}(\beta)]^{-1} \cdot \frac{\partial \mathbf{I}(\beta)|}{\partial \beta_p} \right\}. \end{aligned} \tag{3.8}$$

Using the score equations and the information matrix (3.7), the penalized score equations for logistic regression are:

$$\mathbf{U}^{\mathbf{P}}(\beta_p) = \sum_{i=1}^{N} \left[Y_i - \frac{1-h_i}{1+\exp(-\mathbf{X_i} \cdot \beta)} + \frac{h_i}{2} \right] \cdot \mathbf{X_i} = 0, \tag{3.9}$$

where h_i is the i-th diagonal element of the penalized likelihood version of the "hat" matrix H:

$$H = \mathbf{W}^{\frac{1}{2}} \cdot \mathbf{X} \cdot \mathbf{X}' \cdot \mathbf{W} \cdot \mathbf{X} \cdot \mathbf{X}' \cdot \mathbf{W}^{\frac{1}{2}},$$

and $\mathbf{W}$ is defined as in equation (3.7).

Estimation of β in (3.9) can be done using standard Newton-Raphson or quasi-Newton methods to the modified score function, and standard mechanisms can be used to obtain standard error estimates, such as the square roots of the diagonal elements of $\left[\mathbf{I}^{\mathrm{P}}(\beta)\right]^{-1} = \left\{ \mathrm{E}\left[\frac{\partial^2 ln\mathbf{L}^{\mathbf{P}}(\beta)}{\partial\beta\partial\beta'} \right] \right\}^{-1}$.

In case of a binary logit model with a single dichotomous covariate, the penalized likelihood correction has an especially simple interpretation. It basically means that we need to add 0.5 to each cell of the 2×2 table. In general, it can be seen as an approach that "splits each original observation i into two new observations having response values Y_i and $1 - Y_i$ with iteratively updated weights $1 + h_i/2$ and $h_i/2$, respectively" (Heinze and Schemper, 2002). However, profile likelihood confidence intervals based on penalized likelihood estimates are in general asymmetric, this is due to the fact that the estimates are close to boundary of the eligible parameter space. This means that inferences based on Wald–type statistics can be misleading. We can say that the penalized likelihood approach offers a way out to the problem of separation in logit, probit, and other binary response generalized linear models. In fact this method provides a nice alternative to either omiting important covariates from the models or applying post-hoc data manipulation for those covariates.

3.3. Extensing Firth's Penalized Likelihood Approach to Clustered Binary Response Data

In fact this method could be extended to the case of clustered binary data, since it just involves correcting the likelihood by multiplying it by the squared root of the determinant of the information matrix. The issue in

the case of random effects logistic regression is that the likelihood and information matrix involve integration of the random effects, and no analytic solutions are available in this scenario. As we already discussed in Section 3.1, obtaining "ML–estimates" in this context, even when no separation problems are presented is not straightforward, thus adding this extra complication to the likelihood is going to make the maximization problem even more complicated. A possibility would be to use an approximation to the integral such as the Laplace approximation. It has been shown however that this approach could suffer from severe bias. An alternative would be to use numerical methods to maximize the penalized likelihood, but since it involves partial differentiation and integration it is neither an appealing option. Here we look for solutions using the Bayesian paradigm.

4. Bayesian Solutions to Separation with Clustered Binary Data

As it was already mentioned in the introduction, the Bayesian framework offers an alternative to deal with separation issues, but informative priors or weakly informative priors are needed in order to reach convergence. In the next subsections, we briefly describe the two-step approach proposed by Gelman et al. (2008) and an alternative procedure, also consisting of two stages. The two methods only differ in the first step with the selection of the priors.

4.1. Gelman's Approach

There is a vast literature covering noninformative, default, and reference prior distributions (Jeffreys, 1961; Hartigan, 1964; Bernardo, 1979; Spiegelhalter and Smith, 1982; Berger and Bernardo, 1989; Tibshirani, 1989; Ye and Berger, 1991; Berger and Bernardo, 1992a,b,c; Yang and Berger, 1994; Kass and Wasserman, 1996). Gelman et al. (2008) proposed a new prior distribution. Their approach differs from other work on the topic in the sense that they actually aim to include some prior information, with the objective to regularize the extreme inferences that are obtained using maximum likelihood or completely noninformative priors. Prior to their approach other authors have provided either fully informative prior distributions for which they used application-specific information, or noninformative priors, in general motivated by invariance principles. Their purpose was to propose what they called a weakly informative prior distribution that could be used in a wide range of applications. In previous work Gelman (2008) discussed a standardization procedure for input variables in order to directly compare and interpret parameter estimates. In their paper they proposed a two step procedure, the first step proposed by Gelman et al. (2008) consists of the standardization of the input variables, a procedure which has been previously applied to Bayesian generalized linear models by Raftery (1996). They propose that the standardization should be done as follow: in the case of binary inputs, the coding is such that they are shifted to have mean 0 and to differ by 1 in their lower and upper conditions. In other cases, inputs are shifted to have mean 0 and scaled to have a standard deviation of 0.5, assuring that continuous variables are on the same scale as symmetric binary inputs. The authors made a distinction between regression inputs and predictors (see Gelman (2008) for more details). Then prior distributions for the coefficients of the predictors are defined, assuming independence of the coefficients. Then they propose to use a Student-t prior distribution with mean 0, degrees-of-freedom ν and scale s, where ν and s are chosen in such a way that prior information constrains the coefficients to lie in a reasonable range. They argue in their paper that this prior distribution allows easy and stable computation for logistic regression. Gelman et al. (2008) discussed the use of a Cauchy prior distribution, given that it outperforms the normal, on average, due to the fact that it allows for large coefficients and at the same time performs a reasonable amount of shrinkage for coefficients near zero. They based the choice of the parameters to be used on the fact that if we consider the case of one-half of a success and one-half of a failure for a single binomial trial with probability p, which is just a logistic regression with only a constant term, then the likelihood is close to a t density function with 7 degrees of freedom (t_7) and scale 2.5 (Liu, 2004). In order to be more conservative and leaving the possibility to obtain very large values, they propose to use the Cauchy, or t_1, distribution, again with a scale of 2.5. The idea is then to assign independent Cauchy prior distributions with center 0 and scale 2.5 to each of the coefficients in the logistic regression except the constant term, for which they propose to use a Cauchy

with center 0 and scale 10 (implying that the success probability for an average case should be between 10^{-9} and $1 - 10^{-9}$). About the computation of such an approach the authors state that, in principle, logistic regression with these prior distributions can be computed using the Gibbs and Metropolis algorithms. In this article we will follow this path, and apply their method in a simulation exercise using Gibbs sampling methodology. This method can easily be extended to the case of clustered binary data. It merely introduces another term in the model (usually called random effect) and t_1 prior distributions with scales as mentioned before for the so called "fixed coefficients" in the model. We left the prior distribution for the variance associated to the random effect to be non-informative (we have used inverse-gamma($\epsilon = 0.001, \epsilon = 0.001$)). It is worth noting that in this particular setting when the recommended uniform distribution (Gelman, 2006) on a wide range, $U(0, 100)$ is used as prior, results obtained are comparable. Gelman (2006) states that if variance of the random effect is estimated to be near zero, the resulting inferences will be sensitive to ϵ. But clearly in this case the variance of the random effect is not near zero, thus resulting inference are comparable for both prior distributions.

4.2. A New Approach Combining Penalized Likelihood with Bayesian Methodology

First of all, it is necessary to identify the regression inputs that could present separation issues. As it was discussed earlier in the introduction, several methods have been proposed for this purpose (Albert and Anderson, 1984; Santner and Duffy, 1986; Christmann and Rousseeuw, 2001). Next, we propose to use an alternative two stage approach: in the first stage we use Firth's penalized likelihood approach without considering the clustered nature of the data in our model, for the purpose of extracting information about the prior distributions that we should use in the second stage. In the first stage, we fit the following model:

$$\Pr(Y_{ij} = 1|\mathbf{X_{ij}}) \equiv \pi_i = \frac{1}{1 + \exp(-\mathbf{X_{ij}} \cdot \beta)}. \tag{4.1}$$

In order to simplify notation, we will assume that only one covariate or regression input X_{wij} is presenting separation problems. Then in the second stage, in line with the method proposed by Gelman et al. (2008), we propose to use a t_1 prior distribution, centered around the estimated value ($\hat{\beta}_w^p$) obtained from the penalized likelihood approach for the variable in question (X_{wij}) and scale equal to 2 times the estimated variance ($2 \cdot \hat{\sigma}^2_{\hat{\beta}_w^p}$), in order to be conservative and at the same time allowing the possibility to obtain extreme values around the center. Note that in case that more covariates present separation issues we just extract from the penalized model the estimated value of the coefficient associated to the covariates, as well as the estimated variances and follow the procedure mentioned above for each coefficient associated to the covariate in question. For the other regression inputs which do not present separation problems, we will just confine ourself to use non-informative priors (flat priors, normally distributed centered around 0 with variance equal to 10^5) . This method borrows strengths from the penalized likelihood approach to deal with the fact that informative priors are needed (which allows extreme values) for the variable suffering from separation issues. Another feature of this approach, different from Gelman's approach, is that prior distributions for all other regression inputs which do not present separation problems can remain non-informative.

4.3. Data Splitting

One might raise a concern on using the data twice: to define the prior distribution as well as for fitting the model. This concern holds for Gelman's approach as well as for the new proposed approach. Gelman et al. (2008) (p. 1363) states that "a prior distribution on standardized variables depends on the data, but this is not necessarily a bad idea". A way to deal with this issue is to split the data in two parts, similar to the splitting in a training and a test set as used in for instance random forests and other so-called data mining or machine-learing techniques. We applied this data splitting as follows: prior to the two steps, a training-test setting is adopted, using in the first step the training subset of the full data, to obtain the prior distribution while in the second step the test subset is used to fit the final model. We applied a 70%-30% splitting of the

sample for both stages (as commonly applied for random forests). Of course, other splitting ratios could be chosen. The use of all data for both stages as well as a 70%-30% splitting for both stages is illustrated in the analysis of *Salmonella* data in pigs (Section 6). For computational reasons, it was not feasable to fully apply the splitting procedure in the simulations (Section 5).

5. Simulation Study

Gelman's and the new approach will be applied in a simulation study and the performance of both methods will be evaluated by examining the bias, the variance and the mean squared error of all estimators. For both approaches estimates are calculated using 20000 iterations after an initial burn-in sample of 50000. The posterior inference about the model parameters is obtained from the marginal posterior distribution of each parameter. Convergence of the chains was assessed using the Gelman-Rubin convergence statistic, as modified by Brooks and Gelman (1998) as well as Geweke's test (Geweke, 1992), which compares the first part of each chain (after burn-in) to the last part of the chain.

5.1. Simulation Setting

In our simulation study we generate a binary response y_{ij}, where i is representing the cluster (N in total) and j the observation within the cluster (n_i). The probability of success for y_{ij} depends on 3 covariates (X_{1ij},X_{2ij} and X_{3ij}) and on a cluster specific effect (b_i) as follow:

$$\Pr(Y_{ij} = 1|X_{1ij}, X_{2ij}, X_{3ij}) \equiv \pi_{ij} = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 \cdot X_{1ij} + \beta_2 \cdot X_{2ij} + \beta_3 \cdot X_{3ij} + b_i)}. \tag{5.1}$$

We have considered two binary covariates (X_{1ij} and X_{3ij}) and a continuous covariate X_{2ij}, normally distributed with mean equal to 1 and variance 2.

One of the binary covariates equally divide our data at hand (X_{1ij}), while the other covariate (X_{3ij}) is separating the data into 70% in one of the categories and 30 % into the other one. The values of the parameters used in our simulations are $\beta_0 = -1$, $\beta_1 = 15$, $\beta_2 = -1$ and $\beta_3 = -1$, inducing a separation problem for regression input X_{1ij}. The cluster specific effects were considered normally distributed with mean 0 and variance σ_b^2. Several scenarios were considered, with varying values for the variance of the cluster specific effect ($\sigma_b^2 = 0.1$, 0.5 and 1); and varying values for the number of clusters ($N = 5$, 20 and 100). The number of observations within a cluster (n_i) was considered constant and equal to 100. We then generate 100 samples of size 500, 2000 or 10000 depending on the scenario. In the simulation exercise we did not split the data, but rather used the average of the estimated values and the estimated standard errors obtained from the penalized models for each simulated data without considering clustering in order to obtain a plausible prior distribution for the regression input presenting separation issues. This prior distribution is then used throughout all simulation runs. As the new two-stage procedure requires iterative fitting at both stages (Firth's logistic regression at stage 1 and GLMM at stage 2), it was not considered feasible to split the data in each run throughout all simulation scenarios.

5.2. Simulation Results

The results obtained for each scenario are displayed in Table 5 and 6. It can be seen that for the smaller number of clusters, a smaller bias (defined as the difference between true and estimated value) is obtained by Gelman's approach for the parameter associated to the variable presenting the separation issue, while for the variance of the random effects, rather the opposite is observed. It can also be noted that in the case that the number of clusters is large ($N = 100$), the two stage approach generally shows a smaller bias. In general, the variability of the estimates is smaller for the two stage approach, even for parameters β_0, β_2 and β_3 for

which we specify a totally non-informative prior, while Gelman's approach uses a weakly informative prior. This can be clearly seen in Figure 1 in which we have plotted the average over the 100 runs for each setting, together with the median and the credible intervals for each approach, for the setting in which the variance of the random effect is largest. For other settings results are similar and therefore not shown. It can also be highlighted that if N increases, then both approaches tend to provide very similar results, in terms of bias as well as variance.

It is also worth mentioning that posterior distributions for parameter with separation are skewed (figure not shown), thus Wald type of statistics might provide misleading inference. This can also be observed from the lower and upper percentiles displayed in Table 5 and 6.

If we compare the two approaches in term of mean squared error (bias squared plus variance) (Table 7), the two stage approach is performing better than the approach proposed by Gelman when we focus on β_0, β_1 and σ^2. For the other parameters Gelman's approach produces somewhat smaller but still very comparable mean squared errors.

We also compared coverage probabilities for 95% credible intervals for both approaches (see Table 7), and it can be seen that for the scenario in which $N = 5$ the two approaches show undercoverage for parameter β_2, with Gelman's approach being more affected, while the rest of the parameters present comparable results. For the scenario with $N = 20$, coverages are around 95% for the new stage approach except for σ^2 (around 90%) when the variance of the cluster specific effect equals 0.5. For Gelman's approach, the coverage is around 90% not only for σ^2, but also for β_0 when $\sigma^2 = 0.1$. For the scenario with $N = 100$, we see that the β_1 and β_3 coverage values are affected for both approaches, being 83% for the scenario in which $\sigma^2 = 1$.

Another way of evaluating the methods is by examining the length of the credible intervals for all 100 simulated data. Figure 2 displays the average length from the 100 simulated datasets for the scenario in which we consider the largest variability for the random effects, together with 2.5 and 97.5 percentiles from the 100 credible interval lengths within this scenario. For other scenarios similar conclusions could be drawn (figures not shown). It is clear that in general Gelman's approach produced wider intervals than the new two stage approach, specially for parameter β_1 (coefficient associated to the regression input presenting separation issues), and this is probably due to heavier tails for the weakly informative prior proposed by Gelman, thus allowing for more extreme values. It is also consistent with the higher standard deviations and higher MSE values for Gelman's approach. However, for the other parameters in general if N increases, they tend to be of the same magnitude.

In order to illustrate how convergence was evaluated, we will present some of the plots for one of the scenarios considered (referred to as Scenario A), in particular the scenario in which $N = 5$, $n_i = 100$ and with variance of the cluster specific random intercept (σ_b^2) is equal to 1. In Figure 3, plots proposed by Brooks and Gelman (1998) are shown for each of the parameters in model 5.1. It can be seen that convergence is reached, since the 97.5 quantile is very close to the value 1, which indicates no convergence problem. We have also used the plots proposed by Geweke (1992) (results not shown) to evaluate convergence for all five parameter in the model. Also here, the Z-scores are within allowable ranges, and the diagnostics do not detect any problems with lack of convergence of the posterior samples. Also autocorrelations were visually inspected and the graph for all parameters (figures not shown) shows that the lag autocorrelation values decreases to values close to zero very quickly. In order to get an overall evaluation of the 100 simulated data for Scenario A, we depicted a histogram based on the 97.5 quantile values for each parameter obtained for the 100 simulated data (figure not shown). From this also we can see that most of the upper quantile values are around 1, in particular for β_2 and β_3, and a somewhat wider range for the other 3 parameters, but even for them, upper quantiles are below 1.3, with around 90 % of the values below 1.1. This illustrates that convergence was reached and good mixing was observed.

Table 5. Simulated Data for the different scenarios. Results obtained for the new two stage approach. Average of mean, median, standard deviation (StDev), 2.5 and 97.5 percentiles of the posterior distribution over the 100 simulated datasets. True values are $\beta_0 = -1$, $\beta_1 = 15$, $\beta_2 = -1$ and $\beta_3 = -1$.

Setting when number of cluster is 5 ($N = 5$)															
	$\sigma_b^2 = 0.1$					$\sigma_b^2 = 0.5$					$\sigma_b^2 = 1$				
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\sigma}_b^2$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\sigma}_b^2$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\sigma}_b^2$
Mean	-1.058	12.748	-1.077	-1.074	0.486	-1.034	12.607	-1.060	-1.034	1.239	-0.999	12.571	-1.042	-1.035	2.197
StDev	0.470	3.574	0.204	0.479	1.198	0.592	3.374	0.201	0.473	2.883	0.711	3.373	0.200	0.472	4.638
Percentile 2.5	-2.016	9.145	-1.493	-2.017	0.012	-2.299	9.050	-1.474	-1.965	0.063	-2.482	9.017	-1.452	-1.966	0.176
Median	-1.042	12.068	-1.071	-1.073	0.200	-1.013	11.930	-1.053	-1.033	0.570	-0.980	11.886	-1.035	-1.033	1.139
Percentile 97.5	-0.173	20.355	-0.700	-0.136	2.686	0.062	20.299	-0.688	-0.103	6.739	0.394	20.295	-0.672	-0.112	10.633
Setting when number of cluster is 20 ($N = 20$)															
	$\sigma_b^2 = 0.1$					$\sigma_b^2 = 0.5$					$\sigma_b^2 = 1$				
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\sigma}_b^2$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\sigma}_b^2$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\sigma}_b^2$
Mean	-1.002	13.722	-1.014	-1.023	0.138	-0.985	13.678	-1.013	-1.014	0.572	-0.987	13.673	-1.007	-1.008	1.165
StDev	0.191	3.066	0.097	0.228	0.122	0.242	3.140	0.098	0.228	0.325	0.298	3.170	0.097	0.228	0.569
Percentile 2.5	-1.384	10.387	-1.211	-1.473	0.014	-1.473	10.411	-1.210	-1.463	0.163	-1.587	10.446	-1.203	-1.458	0.437
Median	-0.999	13.125	-1.012	-1.022	0.104	-0.981	13.042	-1.011	-1.013	0.502	-0.981	13.018	-1.005	-1.006	1.045
Percentile 97.5	-0.632	20.667	-0.83	-0.579	0.459	-0.519	20.756	-0.827	-0.570	1.397	-0.411	20.827	-0.822	-0.564	2.599
Setting when number of cluster is 100 ($N = 100$)															
	$\sigma_b^2 = 0.1$					$\sigma_b^2 = 0.5$					$\sigma_b^2 = 1$				
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\sigma}_b^2$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\sigma}_b^2$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\sigma}_b^2$
Mean	-0.988	15.129	-1.004	-1.012	0.094	-0.987	15.143	-1.005	-1.007	0.498	-0.992	15.084	-1.003	-1.001	1.022
StDev	0.083	2.904	0.044	0.101	0.047	0.105	2.937	0.044	0.101	0.116	0.129	2.733	0.043	0.101	0.203
Percentile 2.5	-1.151	11.916	-1.090	-1.210	0.023	-1.196	12.001	-1.091	-1.205	0.307	-1.248	12.086	-1.089	-1.199	0.685
Median	-0.987	14.551	-1.003	-1.012	0.088	-0.986	14.53	-1.004	-1.006	0.486	-0.991	14.469	-1.003	-1.001	1.002
Percentile 97.5	-0.828	21.834	-0.919	-0.814	0.204	-0.783	21.988	-0.920	-0.809	0.759	-0.743	21.879	-0.919	-0.804	1.476

Table 6. Simulated Data for the different scenarios. Results obtained for Gelman's approach. Average of mean, median, standard deviation (StDev), 2.5 and 97.5 percentiles of the posterior distribution over the 100 simulated datasets. True values are $\beta_0 = -1$, $\beta_1 = 15$, $\beta_2 = -1$ and $\beta_3 = -1$.

Setting when number of cluster is 5 ($N = 5$)															
	$\sigma_b^2 = 0.1$					$\sigma_b^2 = 0.5$					$\sigma_b^2 = 1$				
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\sigma}_b^2$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\sigma}_b^2$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\sigma}_b^2$
Mean	-1.141	13.596	-1.027	-0.958	0.551	-1.171	13.349	-1.017	-0.925	1.343	-1.199	13.547	-1.001	-0.931	2.471
StDev	0.489	4.675	0.206	0.458	1.655	0.625	4.139	0.205	0.455	3.316	0.762	4.56	0.203	0.452	5.230
Percentile 2.5	-2.176	8.435	-1.454	-1.871	0.013	-2.582	8.427	-1.438	-1.829	0.065	-2.932	8.463	-1.421	-1.833	0.182
Median	-1.117	12.106	-1.019	-0.952	0.206	-1.123	12.187	-1.009	-0.920	0.602	-1.132	12.211	-0.993	-0.926	1.234
Percentile 97.5	-0.251	26.164	-0.647	-0.078	3.080	-0.075	24.356	-0.639	-0.048	7.122	0.112	26.295	-0.626	-0.061	12.100
Setting when number of cluster is 20 ($N = 20$)															
	$\sigma_b^2 = 0.1$					$\sigma_b^2 = 0.5$					$\sigma_b^2 = 1$				
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\sigma}_b^2$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\sigma}_b^2$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\sigma}_b^2$
Mean	-1.015	14.857	-1.013	-0.99	0.137	-1.008	14.721	-1.013	-0.984	0.569	-1.023	15.110	-1.010	-0.979	1.169
StDev	0.192	4.077	0.099	0.227	0.122	0.243	3.810	0.099	0.225	0.327	0.302	3.941	0.099	0.226	0.578
Percentile 2.5	-1.399	9.972	-1.211	-1.436	0.014	-1.498	10.066	-1.214	-1.429	0.160	-1.638	10.262	-1.209	-1.424	0.434
Median	-1.013	13.667	-1.011	-0.989	0.102	-1.003	13.623	-1.011	-0.983	0.498	-1.017	14.072	-1.007	-0.979	1.045
Percentile 97.5	-0.648	25.185	-0.825	-0.547	0.456	-0.544	24.05	-0.824	-0.545	1.395	-0.446	24.888	-0.822	-0.538	2.619
Setting when number of cluster is 100 ($N = 100$)															
	$\sigma_b^2 = 0.1$					$\sigma_b^2 = 0.5$					$\sigma_b^2 = 1$				
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\sigma}_b^2$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\sigma}_b^2$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\sigma}_b^2$
Mean	-0.982	15.296	-1.011	-1.006	0.094	-0.984	15.362	-1.012	-1.000	0.498	-0.993	15.097	-1.003	-1.001	1.022
StDev	0.082	2.606	0.044	0.101	0.047	0.106	2.455	0.044	0.101	0.116	0.129	2.777	0.044	0.101	0.204
Percentile 2.5	-1.146	11.574	-1.098	-1.203	0.023	-1.193	11.760	-1.100	-1.198	0.306	-1.246	12.084	-1.090	-1.197	0.685
Median	-0.982	14.783	-1.010	-1.005	0.087	-0.983	14.956	-1.012	-1.000	0.486	-0.992	14.486	-1.003	-1.001	1.001
Percentile 97.5	-0.822	21.027	-0.926	-0.809	0.203	-0.780	20.807	-0.927	-0.803	0.757	-0.743	21.852	-0.918	-0.805	1.478

6. Analysis of *Salmonella* Data in Pigs

In this section we deal with the separation issue previously demonstrated in combination with the possibility of introducing random effects to deal with non-independent observations for the *Salmonella* data introduced in Section 2. We present the results of the two methods previously discussed (Gelman's methods and the proposed two stage approaches). The results are summarized in Table 8. This table shows results when using all data for both steps (upper part), and when using a 70%-30% data splitting as discussed in Section 4.3 (lower part). For both situations the conclusions are very similar and in line with the simulations. It can be observed that both methods produce comparable posterior means and medians. The estimates for the intercept and the effect of 'cleaning' ($\hat{\beta}_3$) for the new two stage approach are somewhat smaller, compared the estimates of Gelman's approach. But in general the proposed approach produces smaller posterior standard deviation. Table

In order to get some more insights, we plotted the prior densities for both approaches (Figure 4). We have centered both densities at 0 in order to compare both density functions more easily. It is clear that the tails of the density for Gelman's approach are much heavier, thus giving more chance to extreme values, which is then in turn reflected in wider credible intervals. The prior distributions proposed by Gelman et al. (2008) might be too conservative as also reflected by the larger posterior standard deviation.

7. Concluding Remarks

Several approaches have been proposed to deal with separation in the case of independent binary responses. But in practice in many situations, the assumption of independence between observations is not satisfied, as for example, when we deal with clustered data, regional data or meta-analysis, just to mention a few. In such situations, separation problems can appear, and solutions in this setting are also needed. Extensions of the penalized likelihood method in this particular scenarios are not straightforward, since likelihood functions involved integrating out the random effects. The Bayesian framework offers an attractive alternative, but concessions need to be made, in term of information in our prior distributions. Here we evaluate two different strategies, the first one proposed by Gelman et al. (2008) that first standardizes the regression inputs and then uses weakly informative priors for the parameters in the model, and another new strategy which also consists of two stages: a penalized likelihood approach assuming independence to construct weakly informative priors for the variables presenting separation issues (and non-informative priors for the other regression inputs), followed by a GLLM in the second stage.

The results of the simulations show that if the number of clusters is large, both approaches produce comparable results in terms of point estimates and credible intervals. In case the number of clusters is small, the new two stage approach provides credible intervals which are narrower than the approach proposed by Gelman et al. (2008). This was also observed in the Salmonella case study. It was also confirmed that the posterior distribution of parameter associated to regression inputs presenting separation problems are not symmetric, implying that Wald–type inference should not be used.

Acknowledgments

The authors gratefully acknowledge support from the fund of Scientific Research (FWO, Research Grant G.0151.05) and Belgian IUAP/PAI network P6/03 "Statistical Techniques and Modeling for Complex Substantive Questions with Complex Data" of the Belgian Government (Belgian Science Policy). We would also like to thank the Yves van der Stede (CODA, from Belgium) for providing the data which motivated this research.

Table 7. Average of Mean Squared Error (coverage) over 100 simulated datasets for both approach for the scenarios considered.

		$\hat{\beta}_0(\beta_0)$	$\hat{\beta}_1(\beta_1)$	$\hat{\beta}_2(\beta_2)$	$\hat{\beta}_3(\beta_3)$	$\hat{\sigma}_b^2(\sigma_b^2)$
Two Stage Approach						
$N = 5$	$\sigma^2 = 0.1$	0.224(0.99)	17.846(0.97)	0.047(0.90)	0.235(0.94)	1.585(0.98)
	$\sigma^2 = 0.5$	0.351(0.97)	17.108(0.95)	0.044(0.90)	0.225(0.94)	8.859(0.98)
	$\sigma^2 = 1.0$	0.506(0.94)	17.281(0.95)	0.042(0.93)	0.224(0.96)	22.942(0.98)
$N = 20$	$\sigma^2 = 0.1$	0.037(0.93)	11.036(0.97)	0.010(0.97)	0.053(0.93)	0.016(0.98)
	$\sigma^2 = 0.5$	0.059(0.92)	11.608(0.95)	0.010(0.98)	0.052(0.94)	0.111(0.89)
	$\sigma^2 = 1.0$	0.089(0.94)	11.810(0.93)	0.010(0.96)	0.052(0.96)	0.350(0.94)
$N = 100$	$\sigma^2 = 0.1$	0.007(0.97)	7.330(0.87)	0.002(0.94)	0.010(0.91)	0.002(0.92)
	$\sigma^2 = 0.5$	0.011(0.97)	7.511(0.87)	0.002(0.95)	0.010(0.91)	0.013(0.95)
	$\sigma^2 = 1.0$	0.017(0.96)	7.476(0.83)	0.002(0.96)	0.010(0.92)	0.042(0.96)
Gelman's Approach						
$N = 5$	$\sigma^2 = 0.1$	0.259(0.94)	23.826(0.97)	0.043(0.86)	0.212(0.94)	2.943(0.98)
	$\sigma^2 = 0.5$	0.420(0.93)	19.854(0.95)	0.042(0.90)	0.212(0.97)	11.705(0.98)
	$\sigma^2 = 1.0$	0.619(0.93)	22.902(0.95)	0.041(0.91)	0.209(0.98)	29.514(0.98)
$N = 20$	$\sigma^2 = 0.1$	0.037(0.89)	16.640(0.97)	0.010(0.96)	0.051(0.95)	0.016(0.98)
	$\sigma^2 = 0.5$	0.059(0.93)	14.595(0.95)	0.010(0.97)	0.051(0.95)	0.112(0.90)
	$\sigma^2 = 1.0$	0.091(0.93)	15.540(0.93)	0.010(0.96)	0.051(0.96)	0.362(0.93)
$N = 100$	$\sigma^2 = 0.1$	0.007(0.95)	6.877(0.91)	0.002(0.93)	0.010(0.92)	0.002(0.93)
	$\sigma^2 = 0.5$	0.011(0.96)	6.159(0.93)	0.002(0.96)	0.010(0.91)	0.013(0.96)
	$\sigma^2 = 1.0$	0.017(0.97)	7.724(0.83)	0.002(0.96)	0.010(0.91)	0.042(0.95)

Table 8. Salmonella in Pigs. Estimates obtained using Gelman's and two stage approach, considering a herd specific random intercept.

Full Dataset Results										
	Gelman's Approach					Two Stage Approach				
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\sigma}_b^2$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\sigma}_b^2$
Mean	-8.720	0.545	-0.081	5.849	2.172	-8.150	0.640	-0.080	5.240	2.180
StDev	3.234	0.741	0.048	3.104	0.493	1.236	0.870	0.048	0.909	0.483
Percentile 2.5	-16.990	-0.910	-0.180	2.330	1.400	-10.770	-1.010	-0.170	3.740	1.410
Median	-8.050	0.530	-0.081	5.159	2.113	-8.120	0.620	-0.080	5.170	2.120
Percentile 97.5	-4.780	2.000	0.013	13.850	3.343	-5.872	2.406	0.017	7.587	3.314
Test Dataset Results										
	Gelman's Approach					Two Stage Approach				
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\sigma}_b^2$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\sigma}_b^2$
Mean	-7.196	0.462	0.346	4.586	2.159	-7.060	0.706	0.348	4.252	2.138
StDev	3.482	0.928	0.096	3.356	1.047	1.293	1.174	0.093	0.652	0.997
Percentile 2.5	-14.52	-1.386	0.160	1.021	0.869	-9.667	-1.665	0.163	3.034	0.876
Median	-6.653	0.470	0.345	3.965	1.921	-7.060	0.739	0.348	4.214	1.915
Percentile 97.5	-3.113	2.294	0.533	11.830	4.758	-4.507	3.016	0.532	5.649	4.691

References

Albert, A., and Anderson, J. A. (1984). On the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika*, **71**, 1-10.

Allison, P. (2004). Convergence problems in logistic regression, in: Altman, M., Gill, J., McDonald, M.P., *Numerical issues in statistical computing for the social scientist*, New York: Wiley, 238–252.

Berger, J. O., and Bernardo, J. M. (1989). Estimating a product of means: Bayesian analysis with reference priors. *Journal of the American Statistical Association*, **84**, 200–207.

Berger, J. O., and Bernardo, J. M. (1992a). Reference priors in a variance components problem. In *Bayesian Analysis in Statistics and Econometrics*, P. Goel and N. Iyengar, Eds. New York: Springer Verlag.

Berger, J. O., and Bernardo, J. M. (1992b). Ordered group reference priors with application to a multinomial problem. *Biometrika*, **79**, 25–37.

Berger, J. O., and Bernardo, J. M. (1992c). On the development of the reference prior method. In *Bayesian Analysis*, J.M.Bernardo, J.O.Berger, D.V.Lindley, and A.F.M.Smith, Eds., vol. 4. London: Oxford University Press.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference (with discussion). *Journal of the Royal Statistical Society Series B*, **41**, 113-147.

Bollaerts, K., Aerts, M., Ribbens, S., Van der Stede, Y., Boone, I., Mintiens, K. (2007). Identification of Salmonella Risk Herds in Belgium using semiparametric quantile regression. *Journal of the Royal Statistical Society: Series A*, **171**, 449-464.

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

Brooks, S. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, **7**, 434–455.

Bryson, M.C. and Johnson, M.E. (1981). The incidence of monotone likelihood in the Cox model. *Technometrics*, **23**, 381–383.

Christmann, A. and Rousseeuw, P.J. (2001). Measuring overlap in logistic regression. *Computational Statistics and Data Analysis*, **37**, 65–75.

Clogg, Clifford C., D. B. Rubin, N. Schenker, B. Schultz and L. Weidman. 1991. Multiple Imputation of Industry and Occupation Codes in Census PublicUse Samples Using Bayesian Logistic Regression. *Journal of the American Statistical Association*, **86**, 68-78.

Cortiñas Abrahantes J., Bollaerts K., Aerts M., Ogunsanya V., Van der Stede Y. (2009). Salmonella serosurveillance: Different statistical methods to categorise pig herds based on serological data. *Preventive Veterinary Medicine*, **89**, 59–66.

Cox, David R. 1970. *Analysis of Binary Data.* New York: Wiley.

Day, N.E. and Kerridge, D.F. (1967). A general maximum likelihood discriminant, *Biometrics*, **23**, 313–323.

Demidenko, E. (2001). Computation aspects of the probit model. *Mathematical Communications*, **6**, 233–247.

Engel, B. and Keen, A. (1994). A simple approach for the analysis of generalized linear mixed models. *Statistica Neerlandica*, **48**, 1-22.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, **80**, 27–38.

Gao, S. and Chen, J. (2007). Asymptotic properties of a double penalized maximum likelihood estimator in logistic regression. *Statistics and Probability Letters*, **77**, 925–930.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, **1**, 515-533.

Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, **27**, 2865–2873.

Gelman, A., Jakulin, A., Pittau,M.G. and Su, T.S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics*, **2**, 1360–1383.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In *Bayesian Statistics 4* (ed JM Bernado, JO Berger, AP Dawid and AFM Smith). Clarendon Press, Oxford, UK. 169-193.

Hartigan, J. (1964). Invariant prior distributions. *Annals of Mathematical Statistics*, **35**, 836-845.

Heinze, G. (2006). A comparative investigation of methods for logistic regression with separated or nearly separated data. *Statistics in Medicine*, **25**, 4216–4226.

Heinze, G. and Ploner M. (2003). Fixing the nonconvergence bug in logistic regression with SPLUS and SAS. *Computer Methods and Programs in Biomedicine*, **71**, 181–187.

Heinze, G. and Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine*, **21**, 2409–2419.

Jeffreys, H. (1961). *Theory of Probability*, 3rd ed. Oxford Univ. Press.

Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, **91**, 1343-1370.

King, E. N., and Ryan, T. P. (2002). A Preliminary Investigation of Maximum Likelihood Logistic Regression Versus Exact Logistic Regression. *The American Statistician*, **56**, 163–170.

Lesaffre, E. and Albert, A. (1989). Partial Separation in Logistic Discrimination. *Journal of the Royal Statistical Society, Series B*, **51**, 109-116.

Lesaffre, E. and Kaufmann, H. (1992). Existence and uniqueness of the maximum likelihood estimator for a multivariate probit model. *Journal of the American Statistical Association*, **87**, 805–811.

Liu, C. (2004). Robit regression: A simple robust alternative to logistic and probit regression. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives* (A. Gelman and X. L. Meng, eds.) 227-238. Wiley, London.

Raftery, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika*, **83**, 251-266.

Rousseeuw, P.J. and Christmann, A. (2003). Robustness against separation and outliers in logistic regression. *Computational Statistics and Data Analysis*, **43**, 315–332.

Santner, T.J. and Duffy, D.E. (1986). A note on Albert and J.A Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, **73**, 755–758.

Silvapulle, M.J. (1981). On the existence of maximum likelihood estimators for the binomial response models. *Journal of the Royal Statistical Society B*, **43**, 310–313.

Spiegelhalter, D. J. and Smith, A. F. M. (1982). Bayes factors for linear and log-linear models with vague prior information. *Journal of the Royal Statistical Society Series B*, **44**, 377-387.

Stokes, H.H. (2004). On the advantage of using two or more econometric software systems to solve the same problem. *Journal of Economic and Social Measurement*, **29**, 307–320.

Tibshirani, R. (1989). Noninformative priors for one parameter of many. *Biometrika*, **76**, 604–608.

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

Wolfinger, R. and O'Connell, M. (1993). Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, **48**, 233243.

Yang, R. and Berger, J. O. (1994). Estimation of a covariance matrix using reference prior. *Annals of Statistics*, **22**, 1195-1211.

Ye, K. and Berger, J. O. (1991). Noninformative priors for inferences in exponential regression models. *Biometrika*, **78** , 645–656.

Zorn, C. (2005). A solution to separation in binary response models. *Political Analysis*, **13**, 157–170.

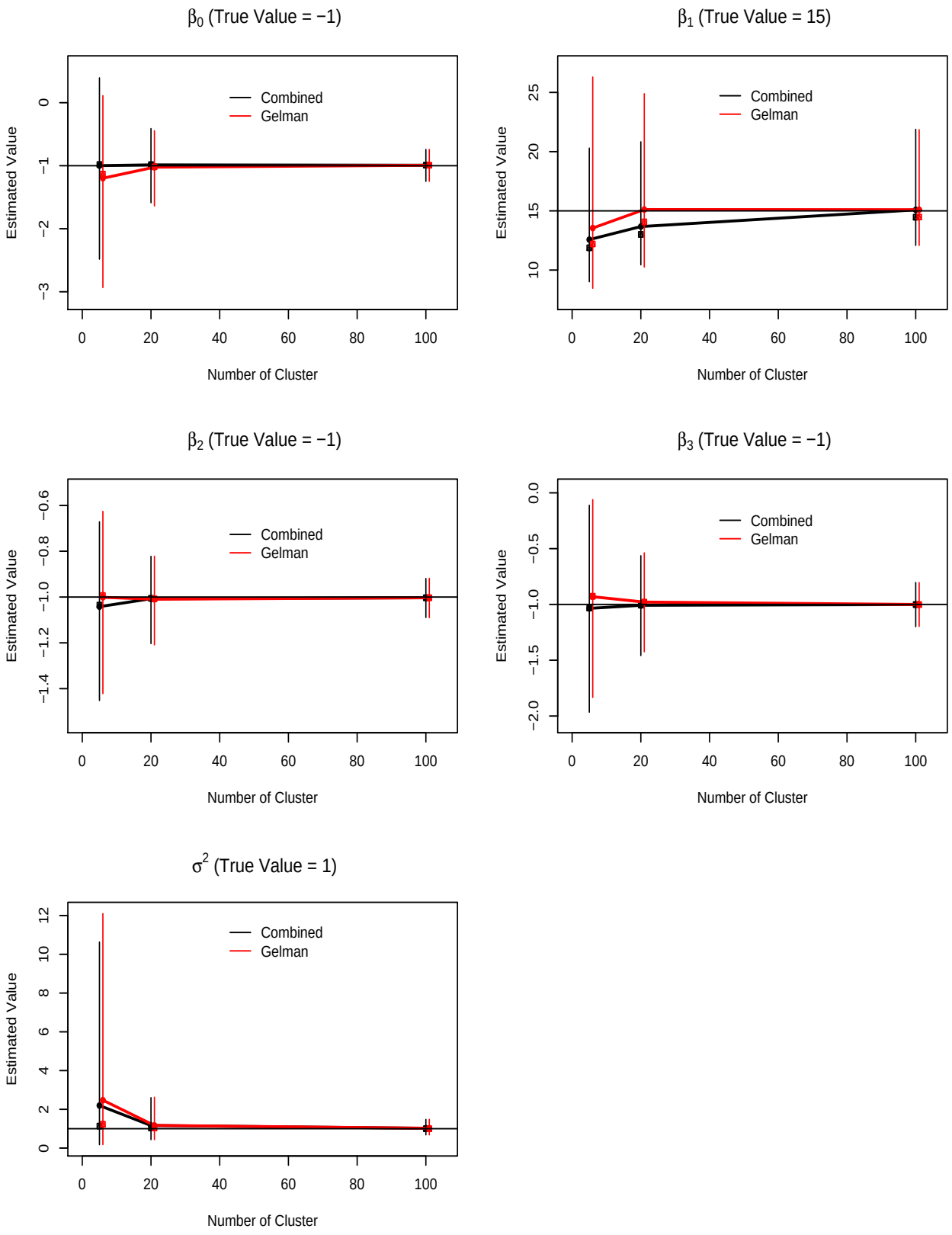

Fig. 1. *Average of mean, median,* 2.5 *and* 97.5 *percentiles of the posterior distribution over the 100 simulated datasets for all parameters when* $\sigma^2 = 1$.

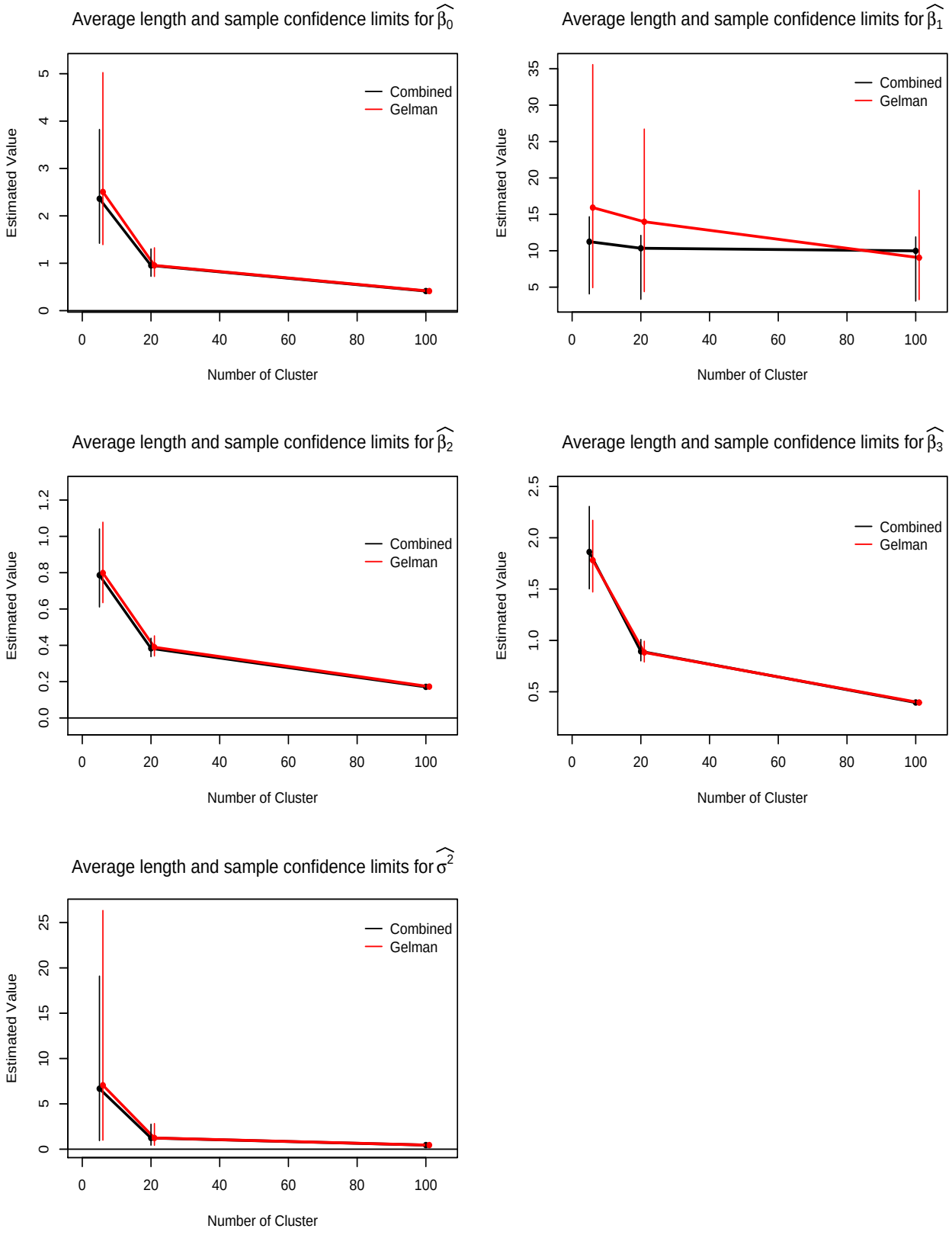

Fig. 2. *Average of length,* 2.5 *and* 97.5 *percentiles of the posterior distribution over the 100 simulated datasets for all estimated parameters when* $\sigma^2 = 1$.

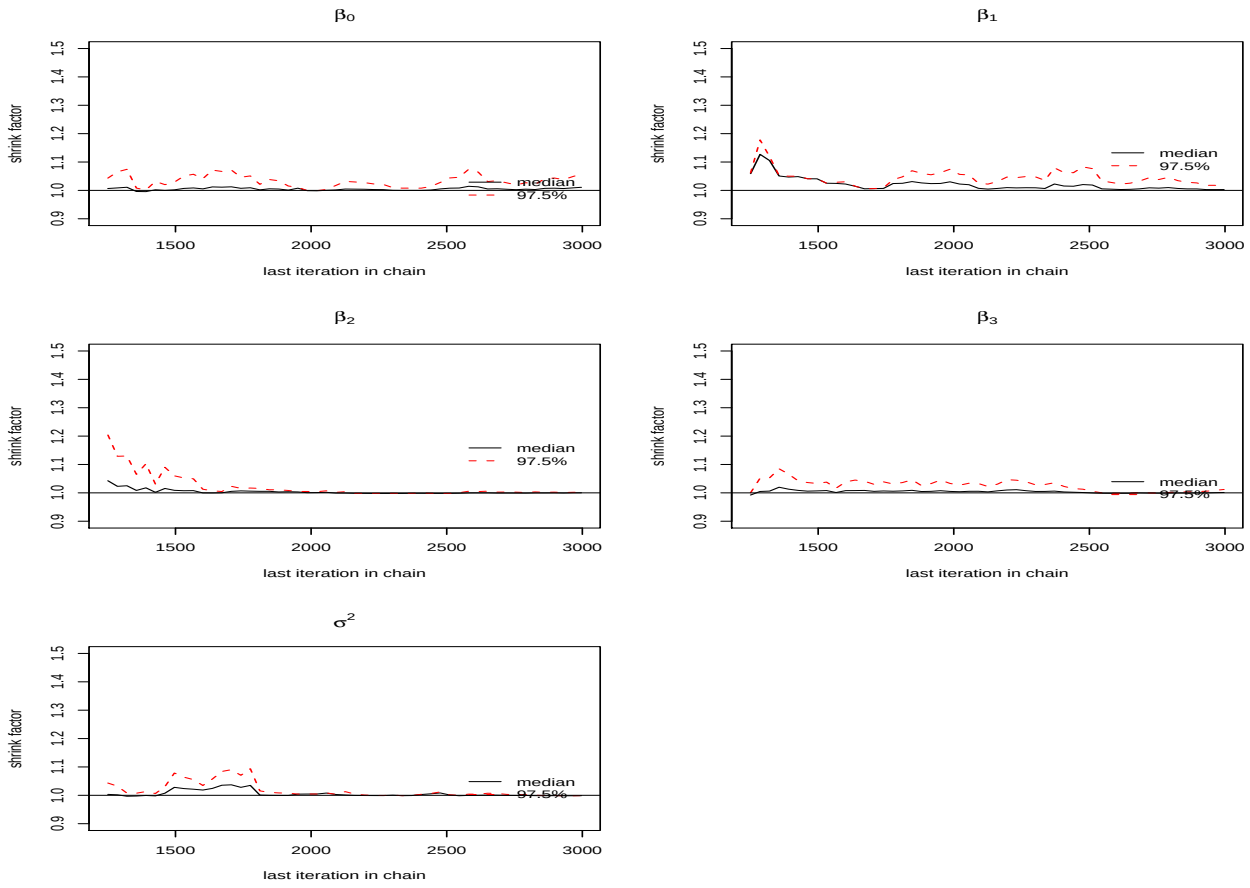

Fig. 3. *Gelman plot for all 5 parameters of the model for one specific dataset in Scenario A.*

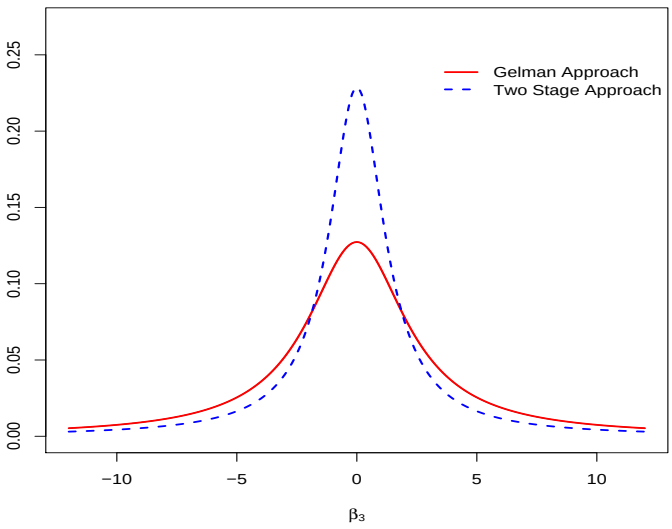

Fig. 4. *Prior densities for the "cleaning" parameter for both approaches.*