A JOINT ANALYSIS OF COUNTS AND SEVERITY WITH ZERO-INFLATED LONGITUDINAL DATA

Giovani L. Silva
CEAUL & DMIST - Technical University of Lisbon
gsilva@math.ist.utl.pt

Elizabeth Juarez-Colunga
CSPH - University of Colorado Denver
elizabeth.juarez-colunga@ucdenver.edu

Charmaine Dean
DSAS/FS - University of Western Ontario
scidean@uwo.ca

This article presents a joint approach for analysis of longitudinal data, as the severity of the event of interest is jointly observed with its occurrence, motivated by a clinical trial involving participants who were healthy menstruating women prior to hysterectomy/ovariectomy for benign disease (Prior et al., 2007). The proposed joint modelling framework allows mixtures of discrete (counting of hotflush events) and categorical (severity of the events) response variables over time. Each response is related to individual-specific random effects, which may be correlated over time, through a generalized linear model. Because there is many zero counts in the motivating dataset, we adapted the proposed model to zero-inflated modelling by using different counting distributions. For Bayesian approach, Markov chain Monte Carlo methods are used for computing some posterior quantities of interest.

Keywords: Bayesian hierarchical models, Zero-inflated data, Joint analysis, Smoothing.

1 INTRODUCTION

In longitudinal studies, one can sometimes observed the frequency and intensity of episodes in order to explain the recurrent occurrence of an event of interest. For example, i) earthquake sequences in a certain region combined with their magnitude levels may indicate long or short free-event periods studying the joint distribution of count (number of earthquakes) and severity (earthquake magnitude level); ii) Prior to occurrence of a health event, such as onset of clinical disease or hospitalization, repeated subclinical events may point out that the individual is at risk.

As there are different measures of severity, *e.g.*, duration or intensity of the event, we need to model mixtures of count, categorical, and continuous response variables. Some authors have studied the joint analysis of different kinds of outcomes, including count and severity. Dunson (2000, 2003) suggested a Bayesian approach for analysis of multidimensional longitudinal data, motivated by studies using an item response battery, to measure traits of an individual repeatedly over time. Henrring and Yang (2007) presented a joint modelling of multiple episode occurrence and severity with a terminating event for assessing the relationship between vaginal bleeding during pregnancy and the gestational age at delivery.

This work is motivated by a real data set involving participants who were healthy menstruating women prior to hysterectomy/ovariectomy for benign disease (Prior *et al.*, 2007). Daily data provided the number of night and day hot flushes/night sweats compared by two treatments, medroxyprogesterone acetate (MPA) and conjugated equine oestrogen (CEE), to reduce the event of interest (flushes) and daily mean severity of flushes. The investigators are namely interested to conclude whether MPA and CEE are equivalent in the control of the number of hot flushes/night sweats and associated severity immediately following premenopausal ovariectomy.

When severity and count modelling is based on marked point processes, investigators usually focus on one of the two at a time: counting processes for recurring process and repeated measures approaches for the marks process. Recently, few papers have drawn their attention to approaches that aim to model both simultaneously. French and Heagerty (2009) developed Generalized Estimating Equations (GEE) methods for marks by assuming that the association between the recurring process and the marks can be explained through an exposure covariate which is measured over time. Cai *et al.* (2010) proposed an estimating equations approach based on a proportional means model for the marks; they obtained the estimates regarding the recurrent events process, and plug in those into the estimation for the marks model. In both approaches there is no direct way of testing the association between the two processes. Our proposal consists in modelling both the recurrent events process and the marks given linking random effects. This allows us to test for the existence of such association between the two processes.

On the other hand, in modelling of longitudinal count data, there are usually a relatively large number of zeros, so-called zero-inflated data (Ridout *et al.*, 2001; Ghosh *et al.*, 2006). The commonly used models are Poisson or Negative Binomial distributions for modelling discrete data with many zeros. We propose a Bayesian joint analysis of counts and severity with zero-inflated longitudinal data to formulate multivariate correlated models for a combination of binary, ordinal, discrete and continuous outcomes measuring the same underlying trend over time. Our models fall within the general framework of generalized linear latent and mixed models (Breslow and Clayton, 1993). However, the number of random effects makes this approach computationally very intensive. Following a Bayesian approach to inference, Markov chain Monte Carlo methods are used for computing some posterior quantities of interest. Ovariectomy data set is analyzed to illustrate the practicability of the proposed method, easily implemented using OpenBUGS (Spiegelhalter *et al.*, 2007).

In this way, this paper is organized as follows. Section 2 describes the motivating study for analyzing counts and severity jointly, while the associated methods of that joint analysis are introduced in Section 3. Section 4 deals with the results of the Bayesian analysis for the joint modelling of different kinds of outcomes, as well as the use of Markov chain Monte Carlo methods for estimating quantities of interest. In section 5, some concluding remarks are presented based on the joint analysis of that dataset.

2 Motivating study

Counts of number of hot flushes in a day along with their severities were recorded in a clinical trial that compares oestrogen therapy (conjugated equine oestrogen: denoted as CEE), the gold standard for the treatment of hot flushes with medroxyprogesterone acetate (referred to as MPA). The participants were healthy menstruating women prior to hysterectomy/ovariectomy for benign disease (Prior *et al.*, 2007) between the ages of 32 and 53 years; their body mass index was between 17.69 and 32.62 units. There were 20 women in the MPA group and 18 in the CEE group; one of the women from the MPA group was eliminated because of extreme number of hot flushes in a day reported, so in here we analyze 19 individuals in the MPA group. A primary goal was to investigate the effect of the treatments in reducing the event of interest (flushes) and their severity, and whether it was necessary to consider both outcomes counts and severities simultaneously, or the counts would give a good enough indication of the effectiveness of the treatment effect.

Figure 1 displays the weekly data for the presence of zero for count and severity by treatments: MPA and CEE. Some individuals have many zero count and severity and other little suggesting a great deal of zero-inflated values in the data. Visually, one can observe less presence of counts over time for the MPA group, and also less severe events for this group mostly all across time. See Figure 2 that shows plots of observed count and severity means over time by treatments: MPA and CEE. Investigating the no-zero counts and severity, we can notice that some individuals have many counts and other little suggesting a great deal of heterogeneity in the data.

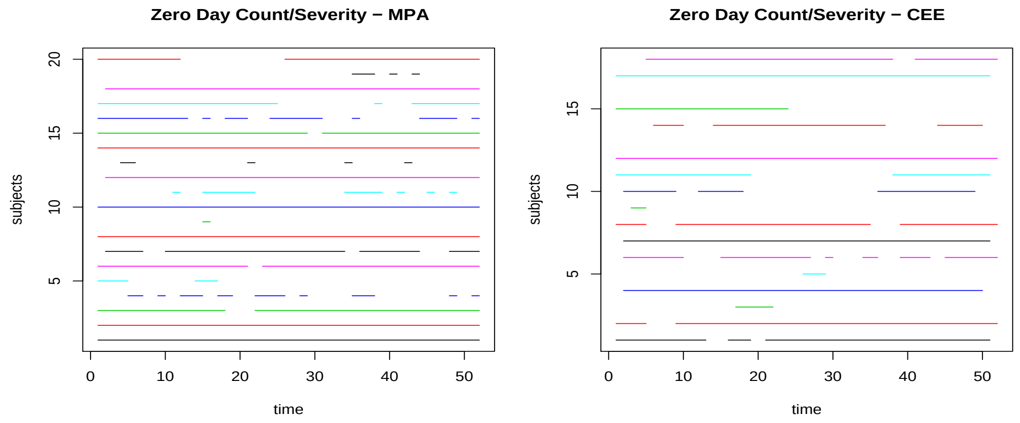

Figure 1: Plots of observed zero count and severity over time by treatments: MPA (left) and CEE (right).

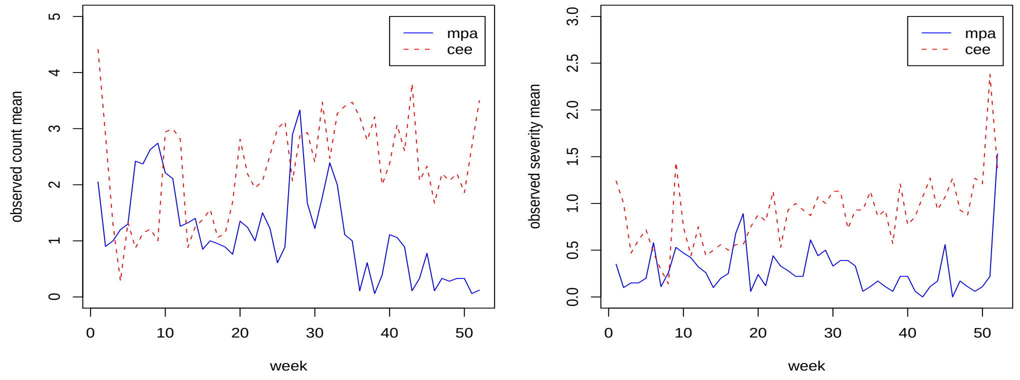

Figure 2: Plots of observed count (left) and severity (right) means by treatments.

3 Bayesian inference on joint modelling

Suppose that n subjects may experience a type of (recurrent) event during a period of time. Let $\mathbf{y}_{ij}$ be the outcome vector for individual i and time j, $j = 1, \ldots, J_i \leq J$, $i = 1, \ldots, n$. In the particular case of the motivating example, the outcomes are observed daily for the ith subject on days $t_{i1} \leq t_{i2} \leq \cdots$; hence they can also be denoted as $y_k(t_{i1}), y_k(t_{i2}), \ldots$, $k = 1, 2$, with $y_1(t) \equiv N(t)$ the number of events on day t and $y_2(t) \equiv m(t)$ the corresponding severity associated with day t.

Assuming that y_{ijk} has a distribution in the exponential family with canonical parameter μ_{ijk}, we define the following model:

$$h_k(\mu_{ijk}) = \eta_{ijk} = \mathbf{z}'_{ijk}\boldsymbol{\Omega}_k + S_{ik}(j) + \xi_{ij} + \varphi_{ik}, \tag{1}$$

where $h_k(\cdot)$ is a link function for the kth outcome, η_{ijk} is the linear predictor for outcome k and subject i at time j, $\mathbf{z}_{ijk}$ is a vector of observed covariates, $\boldsymbol{\Omega}_k$ the regression coefficient vector for outcome k, $S_{ik}(j)$ represents the subject and outcome specific trend taking the flexible form of a spline, ξ_{ij} is a time-dependent random effect that takes account of the dependency among the outcomes measured for a subject at a given time, and φ_{ik} is a measurement error component.

For the hot flush dataset context, the outcome vector is $\mathbf{y}_{ij} = (y_{ij1}, y_{ij2})$, where y_{ij1} is the weekly number of hot flushes and y_{ij2} is the number of days rated as high severity per week for each of the 38 patients during one year, $i = 1, \ldots, 38$, $j = 1, \ldots, J_i \leq 52$. Assuming Poisson and binomial distributions for y_{ij1} and y_{ij2}, respectively, with parameters $\mu_{ij1} = \lambda_{ij}$ (mean of counts) and $\mu_{ij2} = \theta_{ij}$ (probability of high severity), a simple and flexible form for model (1) is expressed by

$$\begin{aligned} \log(\lambda_{ij}) &= \alpha_0 + \mathbf{z}'_{ij}\boldsymbol{\alpha} + S_{i1}(j) + \xi_{ij} + \varphi_{i1} & (2)\\ \text{logit}(\theta_{ij}) &= \beta_0 + \mathbf{z}'_{ij}\boldsymbol{\beta} + S_{i2}(j) + \xi_{ij} + \varphi_{i2}. & (3) \end{aligned}$$

For simplicity, we consider $S_{ik}(j) = S_k(j)$, $\forall\, i$, and here we have fixed covariates so $\mathbf{z}_{ij} = \mathbf{z}_i$, $\forall\, j$. In addition, a more general structure for the random effects would use in (3):

$$\begin{aligned} \xi_{ij} &= \omega_\xi \xi_{ij} + \nu_{ij} & (4)\\ \varphi_{i2} &= \omega_\varphi \varphi_{i1} + \varphi_i. & (5) \end{aligned}$$

Null values for ω_ξ and ω_φ indicate no dependency between counts and severities. Here, $\mathbf{z}_i = (z_{1i}, z_{2i}, z_{3i})$, where z_{1i} indicates treatment (MPA=1, CEE=0), z_{2i} represents age, and z_{3i} is body mass index (BMI).

In order to accommodate zero-inflation, we construct a mixture of a Poisson-binomial distribution and a distribution that is degenerate at $(0,0)$. Note that the conditional probability $P(y_{ij2}=0|y_{ij1}=0)=1$. Conditional on random effects, the probability distribution of the (bivariate) zero-inflated Poisson-binomial is $f(y_{ij1}, y_{ij2})$,

$$\begin{array}{rcl} f(0,0) & = & (1-p_{ij}) + p_{ij} \times \mathrm{e}^{-\lambda_{i,j}} \times (1-\theta_{i,j})^7, \\ f(r,s) & = & p_{ij} \times \frac{\mathrm{e}^{-\lambda_{i,j}} {\lambda_{i,j}}^{y_{ij1}}}{y_{ij1}!} \times \binom{7}{y_{ij2}} {\theta_{i,j}}^{y_{ij2}} (1-\theta_{i,j})^{7-y_{ij2}} \end{array} \tag{6}$$

for $r=1,2,\ldots;\ s=0,1,\ldots,7$, where p_{ij} represents the probability of an excess zero for subject i at time j and is modelled as

$$\mathrm{logit}(p_{ij}) = \gamma_0 + \mathbf{z}'_{ij}\boldsymbol{\gamma}. \tag{7}$$

One typically assumes independent normal prior distributions for the random effects, *i.e.*,

$$\varphi_{ik} \sim N(0, \sigma^2_{\varphi_k}), \tag{8}$$

$$\xi_{ij} \sim N(0, \sigma^2_{\xi}), \tag{9}$$

$k=1,2,$, as well as for the intercept $(\alpha_0, \beta_0, \gamma_0)$, regression $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ and spline $(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2)$ coefficients – note that

$$S_k(j) = \sum_{l=1}^{L} \delta_{kl} B_l(j), \tag{10}$$

where $B_l(\cdot)$, $l=1,\ldots,L$, are the cubic B-spline basis functions, $k=1,2$, $j=1,\ldots,J$. However, spline models (2) and (3) offer reasonable flexibility in modelling these temporal effects, and is parsimonious especially in terms of its interpretation, being particularly useful for exploratory analysis of trends (vide *e.g.* MacNab and Dean, 2001).

For the variance component hyperparameters, one usually assigns an inverse gamma prior, *i.e.*,

$$\sigma^2_{\varphi_1} \sim IG(c_1, d_1), \tag{11}$$

$$\sigma^2_{\varphi_2} \sim IG(c_2, d_2), \tag{12}$$

$$\sigma^2_{\xi} \sim IG(c_3, d_3), \tag{13}$$

whose kernel density is equal to $x^{-(c+1)}\exp(-d/x)$, $x>0$. Similar distributions are assumed for the variances of the coefficients mentioned above. In fact, these inverse gamma priors, as well as the normal priors for those coefficients, are usually assigned highly dispersed, but proper, priors.

4 Results of the joint data analysis

For the hot flush dataset context, the outcome vector is $\mathbf{y}_{ij} = (y_{ij1}, y_{ij2})$, where y_{ij1} is the weekly number of hot flushes and y_{ij2} is the number of days rated as high severity per week for each of the 38 patients during one year, $i=1,\ldots,38$, $j=1,\ldots,J_i \leq 52$.

Using 20,000 iterations and 10,000 of burn-in (thin=40) for MCMC methods in OpenBuUGS (Spiegelhalter *et al.*, 2007), some longitudinal zero-inflated Poisson-binomial models (1) were fitted and compared by Deviance (D). For details about D and other comparison model measures, see for instance section 3.3 in Silva *et al.* (2008). Some interesting examples are in Table 1. Model M_3 provides a reasonable fit, with posterior means, standard deviations and 95% highest posterior density (HPD) credible intervals (CI) for model parameters.

Models defined in terms of the linear predictors		D
M_1:	$\eta_{ij1}=\alpha_0+\mathbf{z}_i'\boldsymbol{\alpha}+S_1(j)+\varphi_{i1}$	4821
	$\eta_{ij2}=\beta_0+\mathbf{z}_i'\boldsymbol{\beta}+S_1(j)+S_2(j)+\varphi_{i1}+\varphi_{i2}$	
M_2:	$\eta_{ij1}=\alpha_0+\mathbf{z}_i'\boldsymbol{\alpha}+\xi_{ij}$	3873
	$\eta_{ij2}=\beta_0+\mathbf{z}_i'\boldsymbol{\beta}+\xi_{ij}$	
M_3:	$\eta_{ij1}=\alpha_0+\mathbf{z}_i'\boldsymbol{\alpha}+S_1(j)+\xi_{ij}+\varphi_{i1}$	3003
	$\eta_{ij2}=\beta_0+\mathbf{z}_i'\boldsymbol{\beta}+S_2(j)+\omega_\xi\xi_{ij}+\nu_{ij}+\omega_\varphi\varphi_{i1}+\varphi_{i2}$	

Table 1: Model comparison based on posterior mean of deviance (D).

Table 2 displays the posterior quantities of interest (parameters) for selected model M_3: mean, standard deviation (s.d.), and 95% highest posterior density (HPD) credible interval. Based on these results, i) there is covariate influence only from body mass index (z_3) for probability of excess zero component (see γ_3 estimates); ii) the measurement error component for counts (φ_{i1}) is lightly shared with that component for severities (see ω_φ estimates), whereas the time-dependent random effect (ξ_{ij}) for counts is higher shared with that component for severities (see ω_ξ estimates); iii) there are some unobserved heterogeneity of the subject-count (φ_{i1}), subject-severity (φ_{i2}), shared subject-time (ξ_{ij}) and severity subject-time (ν_{ij}) components (see $\sigma^2_{\varphi_1}$, $\sigma^2_{\varphi_2}$, σ^2_ξ and σ^2_ν estimates).

Based on both Table 2 and Figure 3, posterior distributions of ω_φ and ω_ξ for selected model M_3 are both significantly different from zero, indicating dependence between the subject and subject-time random effects for count and severity. From Figure 4, the posterior distributions of the variance

parameter	mean	s.d.	95% HPD credible interval
α_0	-0.69	0.693	(-2.17,0.63)
α_1	-0.52	1.103	(-2.73,1.66)
α_2	0.06	0.127	(-0.19,0.32)
α_3	-0.27	0.191	(-0.63,0.12)
β_0	-4.08	0.881	(-5.80,-2.34)
β_1	-1.09	1.183	(-3.41,1.32)
β_2	-0.04	0.145	(-0.33,0.25)
β_3	-0.17	0.201	(-0.56,0.23)
γ_0	0.79	0.150	(0.50,1.09)
γ_1	-0.89	0.227	(-1.33,-0.45)
γ_2	0.04	0.026	(-0.01,0.09)
γ_3	0.46	0.059	(0.35,0.58)
ω_φ	0.78	0.179	(0.44,1.14)
ω_ξ	2.09	0.19	(1.72,2.46)
$\sigma^2_{\varphi_1}$	10.67	3.73	(4.79,17.9)
$\sigma^2_{\varphi_2}$	5.80	2.07	(2.34,9.84)
σ^2_ξ	0.91	0.122	(0.67,1.14)
σ^2_ν	2.43	0.484	(1.53,3.38)

Table 2: Posterior estimates for selected model parameters (Model M_3).

components σ^2_ν, σ^2_ξ, $\sigma^2_{\varphi_1}$ and $\sigma^2_{\varphi_2}$ for selected model M_3, one can suggest a higher unobserved heterogeneity on subject random effects than that kind of heterogeneity on subject-time random effects. Finally, the posterior mean of the mass probability at (0,0) over the full observation period for selected model M_3 is 0.73 (overall), 0.77 (MPA) and 0.68 (CEE). This varies over time by treatments (see Figure 5).

5 Conclusions

In conclusion, this joint modelling enables association between related outcomes and provides better power in identifying effects. It is also useful for providing an understanding of the mechanisms generating the outcomes. In the example, there was no difference in the treatment effects in the mean counts and severities; this was an important scientific observation. However, the proportion of zeros differed in the treatment arms. Future work in this area will include the incorporation of autoregressive random effects and the study of the gains in efficiency through this sort of joint modelling.

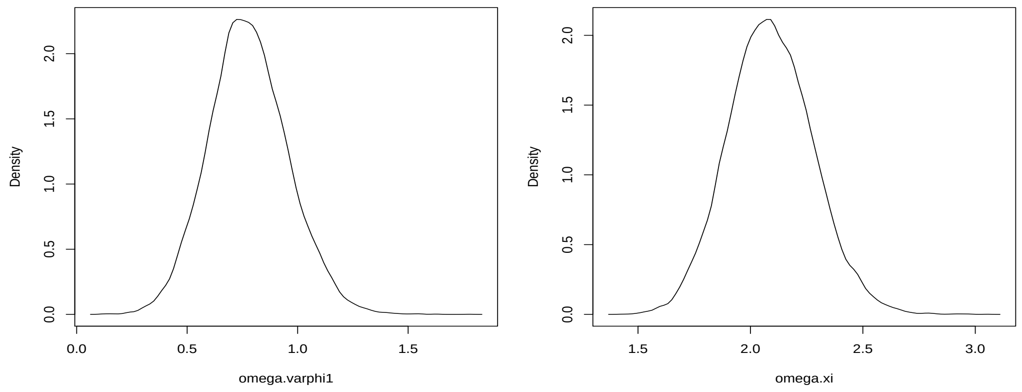

Figure 3: Posterior distributions of ω_φ (left) and ω_ξ (right) based on M_3.

Acknowledgements

This paper was partially supported by Pest-OE/MAT/UI0006/2011.

References

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

[2] Cai, J., Zeng, D., and Pan, W. (2010). Semiparametric proportional means model for marker data contingent on recurrent event. *Lifetime Data Analysis*, 16, 250-270.

[3] Dunson, D.B. (2000). Bayesian latent variable models for clustered mixed outcomes, *Journal of the Royal Statistical Society*, 62, 355-366.

[4] Dunson, D.B. (2003). Dynamic latent trait models for multidimensional longitudinal data, *Journal of the American Statistical Association*, 98, 555-563.

[5] French, B., and Heagerty, P.J. (2009). Marginal mark regression analysis of recurrent marked point process data. *Biometrics*, 65. 415-422.

[6] Henrring, A.H., and Yang, J. (2007) Bayesian modeling of multiple episode occurrence and severity with a terminating event, *Biometrics*, 63, 381-388.

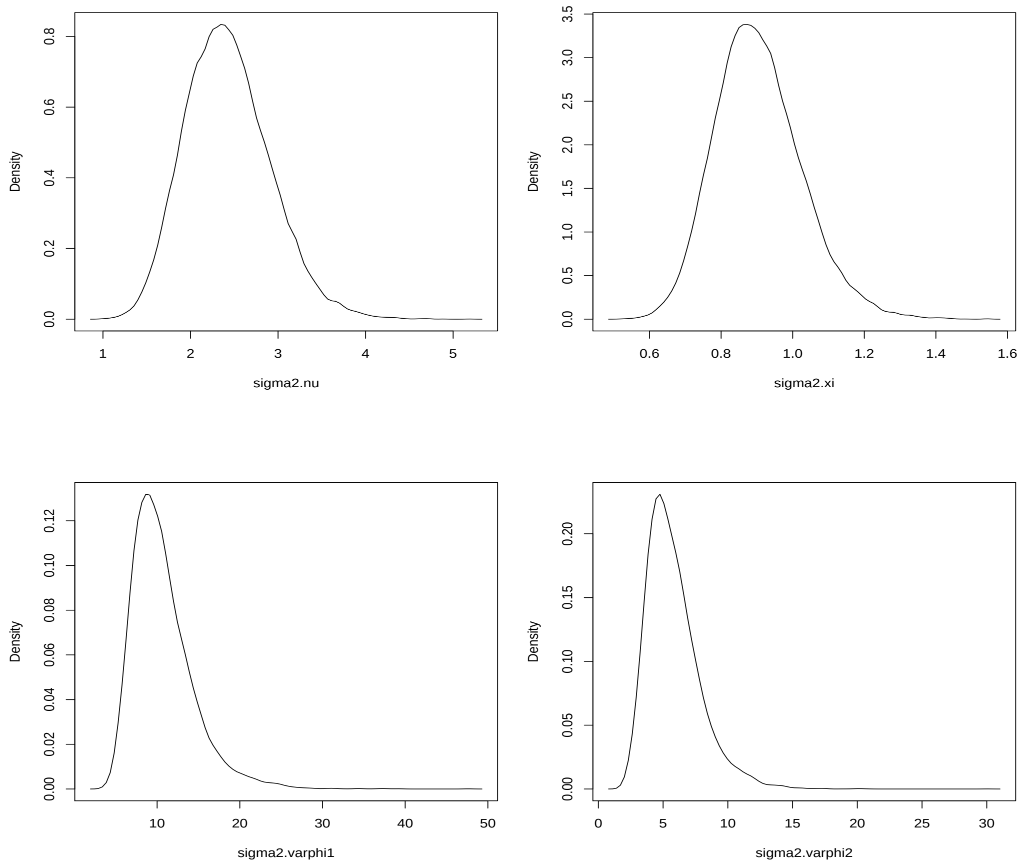

Figure 4: Posterior distributions of σ^2_ν (top-left), σ^2_ξ (top-right), $\sigma^2_{\varphi_1}$ (bottom-left) and $\sigma^2_{\varphi_2}$ (bottom-right) based on M_3.

[7] MacNab, Y.C., and Dean, C.B. (2001). Autoregressive spatial smoothing and temporal spline smoothing for mapping rates. *Biometrics*, 57, 949-956.

[8] Prior, J.C., Nielsen, J.D., Hitchcock, C.L., Williams, L.A., Vigna, Y.M., and Dean C.B. (2007). Medroxyprogesterone and conjugated oestrogen are equivalent for hot flushes: a 1-year randomized double-blind trial following premenopausal ovariectomy, *Clinical Science*, 112, 517-525.

[9] Ridout, M.S., Hinde, J.P., and Demetrio, C.G.B. (2001) A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics*, 57, 219-223.

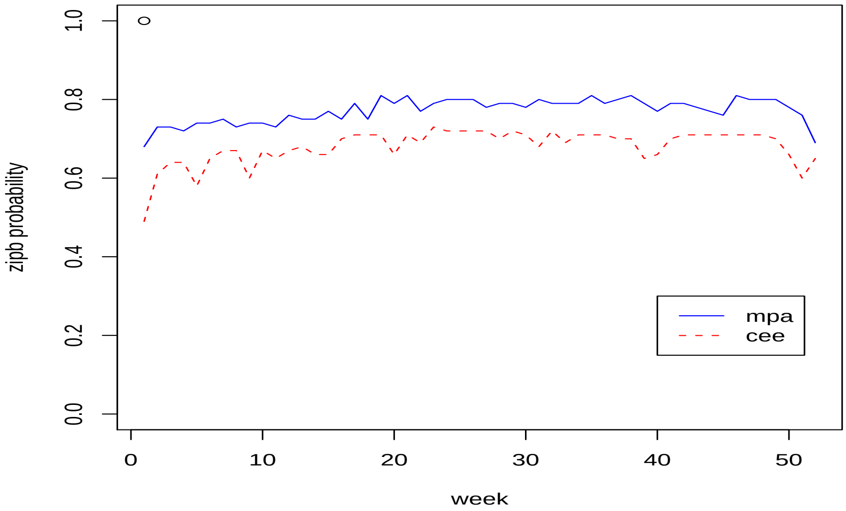

Figure 5: Posterior means of the mass probability at (0,0) by treatments MPA and CEE based on M_3.

[10] Ghosh, S.K., Mukhopadhyay, P., and Lu, J.-C. (2006). Bayesian analysis of zero-inflated regression models. *Journal of Statistical Planning and Inference*, 136, 1360-1375.

[11] Silva, G.L., Dean, C.B, Niyonsenga, T. and Vanasse, A. (2008). Hierarchical Bayesian spatiotemporal analysis of revascularization odds using smoothing splines. *Statistics in Medicine*, 27, 2381-2401.

[12] Spiegelhalter, D., Thomas, A., Best, N., and Lunn, D. (2007). *OpenBUGS User Manual* (version 3.0.2). Department of Epidemiology and Public Health, Imperial College, St Mary's Hospital London.