Biostatistics (2015), **0**, 0, *pp.* 1–11
doi:10.1093/biostatistics/Supplementary09

Supplementary Material to Bayesian two-part spatial models for semicontinuous data with application to emergency department expenditures

BRIAN NEELON*

Department of Public Health Sciences, Medical University of South Carolina, 135 Cannon Street Suite 303 MSC 835, Charleston, SC 29425 USA

neelon@musc.edu

LI ZHU

Department of Biostatistics, University of Pittsburgh, 130 De Soto Street, Pittsburgh, Pennsylvania 15261, USA

SARA E. BENJAMIN NEELON

Department of Community and Family Medicine, Duke University School of Medicine, 2200 W. Main Street, Durham, North Carolina 27705, USA

*To whom correspondence should be addressed.

Table S1: *Empirical Kullback-Leibler divergence estimates comparing true and model-estimated densities for four selected block groups in the illustrative example.*

Block Group	LN	LSN	3-df LST	16-df LST	DPLN-ω_5
Average	1.518	0.100	0.076	0.093	0.056
3	1.695	0.115	0.077	0.106	0.065
35	2.172	0.215	0.132	0.223	0.085
148	1.685	0.075	0.059	0.073	0.043

Table S2: *Summary statistics for DSR data (N=44722).*

Variable	n	%
ED users	11083	25
Male gender	16425	36
Race		
Non-Hispanic white	22986	51
Non-Hispanic black	19432	44
Hispanic	2304	5
Insurance		
Private	24044	54
Medicare	8669	19
Medicaid	3909	9
Uninsured	8100	18
	Median	Range
Age (years)	47	(18, 105)
Block group sample size	267	(1, 803)
Block group median household income	45640	(5934, 157300)

Table S3: *DPLN-estimated* $\Pr(Y > 0)$ *and* $\mathrm{E}(Y|Y > 0)$ *for the four block groups highlighted in Figure 3. 95% credible intervals are given in parentheses.*

Block Group	$\Pr(Y > 0)$	$\mathrm{E}(Y|Y > 0)$
84	0.082 (0.061, 0.104)	3182 (2720, 3809)
86	0.081 (0.057, 0.108)	3713 (3082, 4315)
122	0.037 (0.025, 0.054)	3175 (2482, 3935)
24	0.028 (0.017, 0.046)	3910 (3052, 4905)

Table S4: *DPLN- and LST-estimated median and 90th quantiles* (Q_{90}) *for block groups 24 and 122. 95% credible intervals are given in parentheses.*

Model	Block Group	Median	Q_{90}
DPLN ($\omega = 1$)	24	2760 (2160, 3720)	8300 (6320, 10962)
	122	2200 (1760, 2821)	6662 (5040, 8642)
LST ($\nu = 16$)	24	3557 (2927, 4272)	10393 (8555, 12514)
	122	2819 (2324, 3436)	8238 (6786, 10067)

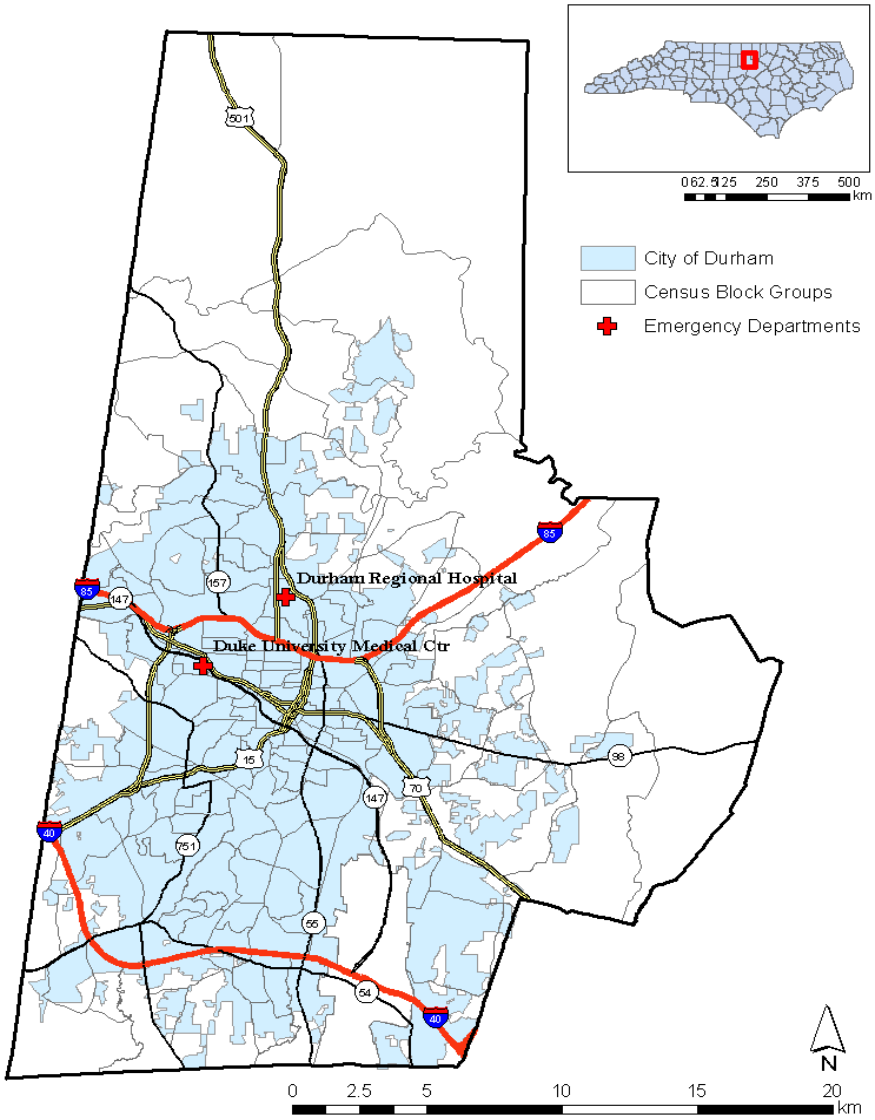

Figure S1: Map of Durham County block groups, city limits, and emergency department locations.

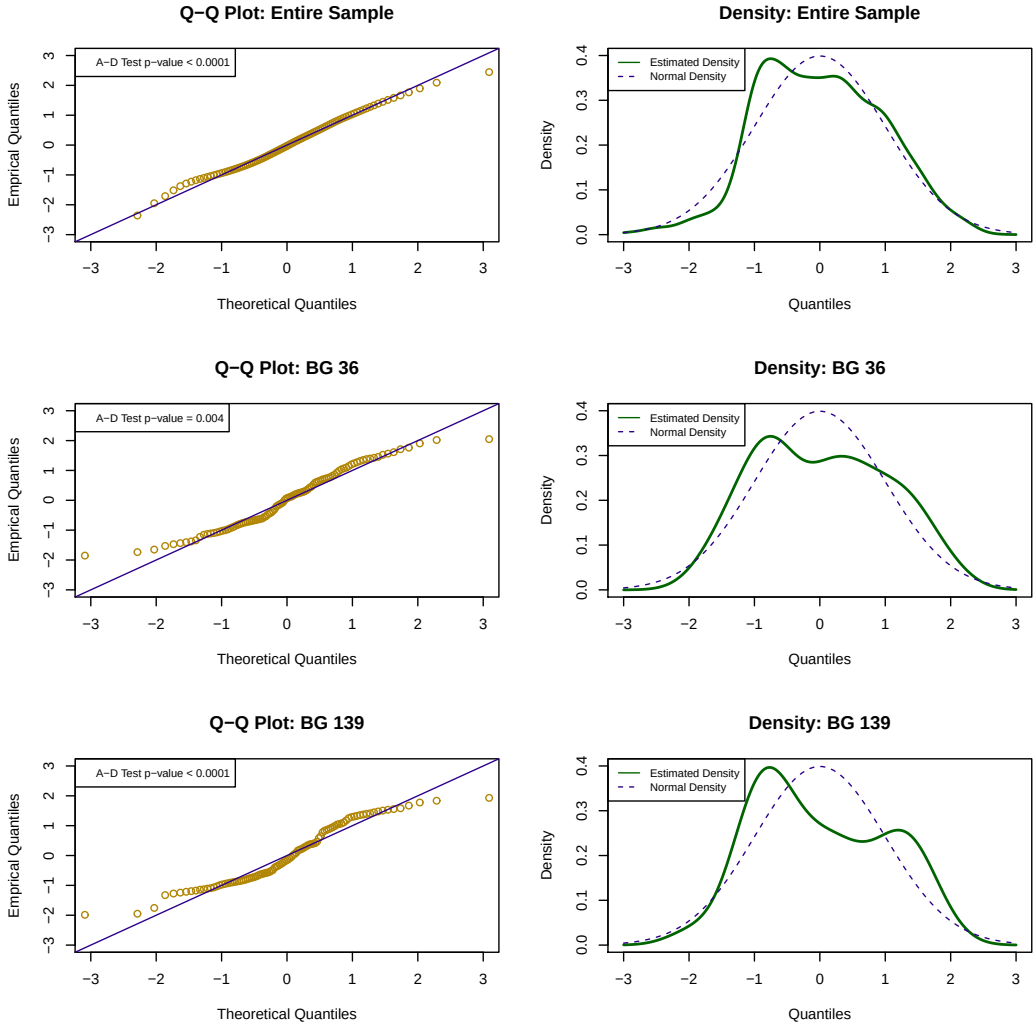

Figure S2: Residual diagnostics on log scale from an LN spatial model fitted to the nonzero expenditures in the ED application. Left panels: Q-Q plots of standardized residuals for the entire sample and for two select block groups. Right panels: Estimated residual densities with standard normal densities provided as reference. P-values are from Anderson-Darling (A-D) tests of normality. BG: Block Group.

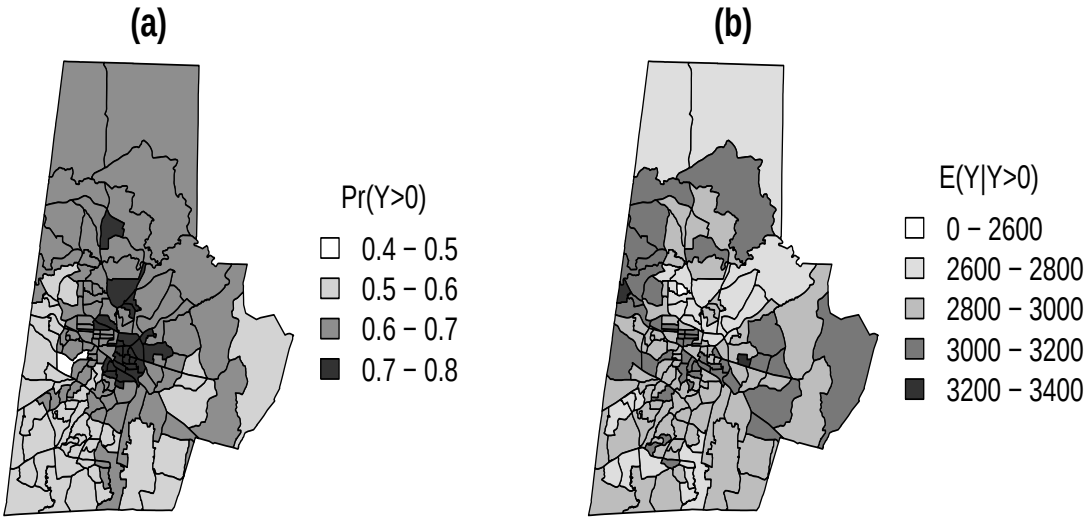

Figure S3: DPLN-estimated (a) $\Pr(Y > 0)$ and (b) $\mathrm{E}(Y|Y > 0)$ for patient group 2 from the ED expenditure analysis.

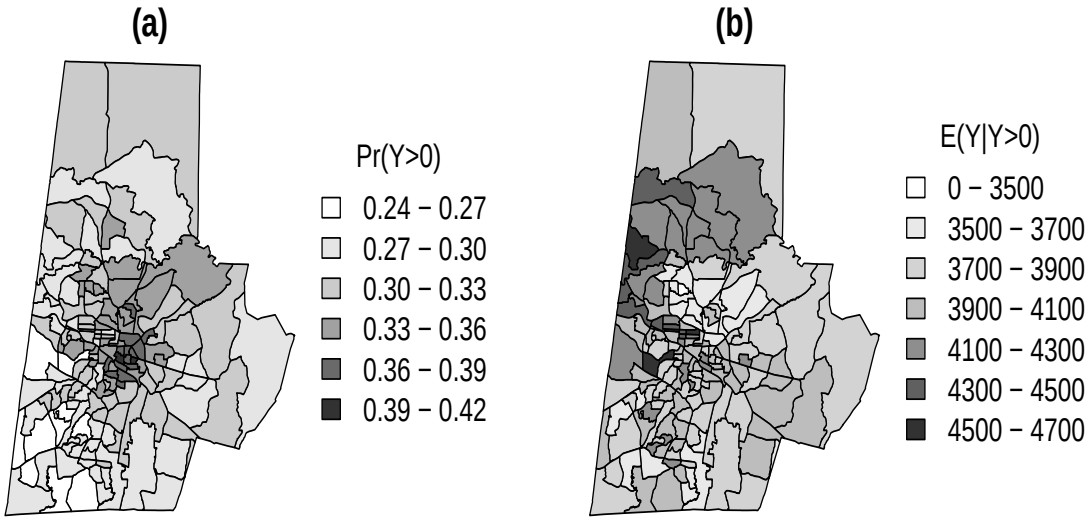

Figure S4: DPLN-estimated (a) $\Pr(Y > 0)$ and (b) $\mathrm{E}(Y|Y > 0)$ for patient group 3 from the ED expenditure analysis.

MCMC Algorithms

MCMC Updates for Two-Part Lognormal Model

1. Update U_{ij}. Assuming a probit link for the binary part in equation (2.2), the full conditional for the latent variable U_{ij} $(i = 1, \ldots, n;\ j = 1, \ldots, n_i)$ is

$$U_{ij}|- \sim \mathrm{TN}_{[0,+\infty)}(\boldsymbol{x}'_{ij}\boldsymbol{\gamma} + s_1(v_{ij}) + b_{1i}, 1), \quad \text{if } y_{ij} > 0$$

$$U_{ij}|- \sim \mathrm{TN}_{(-\infty,0]}(\boldsymbol{x}'_{ij}\boldsymbol{\gamma} + s_1(v_{ij}) + b_{1i}, 1), \quad \text{if } y_{ij} = 0.$$

2. Update $\boldsymbol{\gamma}$ and the spline coefficients associated with $s_1(v_{ij})$. For brevity, let $\boldsymbol{\gamma}$ denote a $p \times 1$ vector of fixed effect *and* spline coefficients, where p denotes the total number of coefficients. Assuming a joint $\mathrm{N}_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ prior, the joint full conditional for $\boldsymbol{\gamma}$ is given by $\mathrm{N}(\boldsymbol{\mu}_{\boldsymbol{\gamma}}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}})$, where

$$\boldsymbol{\Sigma}_{\boldsymbol{\gamma}} = \left(\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{X}'\boldsymbol{X}\right)^{-1} \quad \text{and}$$

$$\boldsymbol{\mu}_{\boldsymbol{\gamma}} = \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\left\{\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{X}'(\boldsymbol{u} - \boldsymbol{D}\boldsymbol{b}_1)\right\}.$$

   Here, $\boldsymbol{u}$ is an $N \times 1$ vector of sampled values, u_{ij}, from the previous step, $\boldsymbol{X}$ is an $N \times p$ design matrix combining fixed effects covariates and spline basis functions, and $\boldsymbol{D}$ is a $N \times n$ random effects design matrix for the $n \times 1$ vector $\boldsymbol{b}_1$.

3. Update $\boldsymbol{b}_1$. Given the conditional prior specification in equation (3.3), the full conditional for $\boldsymbol{b}_1$ is $\mathrm{N}(\boldsymbol{\mu}_{\boldsymbol{b}_1}, \boldsymbol{\Sigma}_{\boldsymbol{b}_1})$, where

$$\boldsymbol{\Sigma}_{\boldsymbol{b}_1} = \left\{\frac{\boldsymbol{Q}}{(1-\rho^2)\Lambda_{11}} + \boldsymbol{D}'\boldsymbol{D}\right\}^{-1} \quad \text{and}$$

$$\boldsymbol{\mu}_{\boldsymbol{b}_1} = \boldsymbol{\Sigma}_{\boldsymbol{b}_1}\left\{\frac{\rho\boldsymbol{Q}\boldsymbol{b}_2}{(1-\rho^2)\sqrt{\Lambda_{11}\Lambda_{22}}} + \boldsymbol{D}'\left(\boldsymbol{u} - \boldsymbol{X}\boldsymbol{\gamma}\right)\right\}.$$

   Note that this update relies on data ($\boldsymbol{u}$ and $\boldsymbol{X}$) from all N observations.

4. Update $\boldsymbol{\beta}$ and the spline coefficients associated with $s_2(v_{ij})$. As with $\boldsymbol{\gamma}$ above, let $\boldsymbol{\beta}$ denote a $p \times 1$ vector of fixed effect and spline coefficients. Assuming an $\mathrm{N}_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ prior, the full

conditional for $\boldsymbol{\beta}$ is $\text{N}(\boldsymbol{\mu_\beta}, \boldsymbol{\Sigma_\beta})$, where

$$\boldsymbol{\Sigma_\beta} = \left(\boldsymbol{\Sigma}_0^{-1} + \frac{\boldsymbol{X}_1' \boldsymbol{X}_1}{\sigma^2} \right)^{-1} \quad \text{and}$$
$$\boldsymbol{\mu_\beta} = \boldsymbol{\Sigma_\beta} \left\{ \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 + \frac{\boldsymbol{X}_1' (\ln \boldsymbol{y}_1 - \boldsymbol{D}_1 \boldsymbol{b}_2)}{\sigma^2} \right\}.$$

Here, $\boldsymbol{y}_1$ is an $N_1 \times 1$ vector of nonzero expenditures, with N_1 denoting the number of nonzero observations; $\boldsymbol{X}_1$ is an $N_1 \times p$ design matrix containing fixed effects covariates and spline bases for the nonzero observations; and $\boldsymbol{D}_1$ is the corresponding $N_1 \times n$ random effects design matrix for the nonzero observations.

5. Update σ^2. Assuming an $\text{IG}(g, s)$ prior for σ^2, the full conditional is $\text{IG}(g^*, s^*)$, where $g^* = g + \frac{N_1}{2}$, and $s^* = s + \frac{1}{2} \{\ln \boldsymbol{y}_1 - \boldsymbol{X}_1 \boldsymbol{\beta} - \boldsymbol{D}_1 \boldsymbol{b}_2\}' \{\ln \boldsymbol{y}_1 - \boldsymbol{X}_1 \boldsymbol{\beta} - \boldsymbol{D}_1 \boldsymbol{b}_2\}$.

6. Update $\boldsymbol{b}_2$. Given the conditional prior in equation (3.4), update $\boldsymbol{b}_2$ from its $\text{N}(\boldsymbol{\mu_{b_2}}, \boldsymbol{\Sigma_{b_2}})$ full conditional, where

$$\boldsymbol{\Sigma_{b_2}} = \left\{ \frac{\boldsymbol{Q}}{(1-\rho^2)\Lambda_{22}} + \frac{\boldsymbol{D}_1' \boldsymbol{D}_1}{\sigma^2} \right\}^{-1} \quad \text{and}$$
$$\boldsymbol{\mu_{b_2}} = \Sigma_{\boldsymbol{b_2}} \left\{ \frac{\rho \boldsymbol{Q} \boldsymbol{b}_2}{(1-\rho^2)\sqrt{\Lambda_{11}\Lambda_{22}}} + \frac{\boldsymbol{D}_1' (\ln \boldsymbol{y}_1 - \boldsymbol{X}_1 \boldsymbol{\beta})}{\sigma^2} \right\}.$$

Note that this update relies only on data ($\boldsymbol{y}_1$ and $\boldsymbol{X}_1$) for the N_1 positive observations.

7. Update $\boldsymbol{\Lambda}$. Assuming an $\text{IW}(\kappa_0, \boldsymbol{S}_0)$ prior, the full conditional for $\boldsymbol{\Lambda}$ is $\text{IW}(\kappa^*, \boldsymbol{S}^*)$, where $\kappa^* = n + \kappa_0 - 1$, and $\boldsymbol{S}^* = \boldsymbol{S}_0 + \boldsymbol{b}^{*'} \boldsymbol{Q} \boldsymbol{b}^*$. Here, $\boldsymbol{b}^*$ denotes an $n \times 2$ matrix with first column equal to $\boldsymbol{b}_1$ and second column equal to $\boldsymbol{b}_2$. Use $\boldsymbol{\Lambda}$ to extract Λ_{11}, Λ_{22} and ρ.

8. Apply sum-to-zero constraints to $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$.

MCMC Updates for Two-Part Log Skew-Normal Model

For the two-part LSN model, the updates of the binary part are similar to those given for the two-part LN model. The continuous part is updated as follows:

1. Update the latent variable Z_{ij} in equation (2.4). If we assume a truncated normal $\mathrm{TN}_{[0,\infty)}(0,1)$ prior for all Z_{ij} such that $y_{ij} > 0$, the full conditional for Z_{ij} is $\mathrm{N}(\mu_{z_{ij}}, \sigma^2_{z_{ij}})$, where

$$\sigma^2_{z_{ij}} = \frac{\xi^2}{\psi^2 + \xi^2} \quad \text{and}$$
$$\mu_{z_{ij}} = \sigma^2_{z_{ij}} \psi \left\{ \ln y_{ij} - \boldsymbol{x}'_{ij}\boldsymbol{\beta} - b_{2i} \right\} / \xi^2.$$

   Concatenate the sampled values, z_{ij}, to form an $N_1 \times 1$ vector $\boldsymbol{z}_1$.

2. Update ψ, $\boldsymbol{\beta}$ and the spline coefficients associated with $s_2(v_{ij})$. Combine the $N_1 \times p$ matrix $\boldsymbol{X}_1$ defined earlier with the vector $\boldsymbol{z}_1$ columnwise to form an $N_1 \times (p+1)$ matrix $\boldsymbol{X}^*_1$ that now includes fixed effect covariates, spline basis functions, and the latent vector $\boldsymbol{z}_1$. Likewise, combine the vector $\boldsymbol{\beta}$ defined earlier and the parameter ψ defined in equation (2.4) to form a $(p+1) \times 1$ vector $\boldsymbol{\beta}^*$ that includes fixed and spline coefficients as well as ψ. Assigning a joint prior $\mathrm{N}_{p+1}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, the full conditional for $\boldsymbol{\beta}^*$ is $\mathrm{N}(\boldsymbol{\mu}_{\boldsymbol{\beta}^*}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}^*})$, where

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}^*} = \left(\boldsymbol{\Sigma}_0^{-1} + \frac{\boldsymbol{X}_1^{*'} \boldsymbol{X}_1^*}{\sigma^2} \right)^{-1} \quad \text{and}$$
$$\boldsymbol{\mu}_{\boldsymbol{\beta}^*} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}^*} \left\{ \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 + \frac{\boldsymbol{X}_1^{*'} (\ln \boldsymbol{y}_1 - \boldsymbol{D}_1 \boldsymbol{b}_2)}{\sigma^2} \right\}.$$

3. Update ξ^2, $\boldsymbol{b}_2$ and $\boldsymbol{\Lambda}$. Updates for ξ^2, $\boldsymbol{b}_2$ and $\boldsymbol{\Lambda}$ are identical to updating σ^2, $\boldsymbol{b}_2$ and $\boldsymbol{\Lambda}$ in the LN model if we replace $\boldsymbol{X}_1$ and $\boldsymbol{\beta}$ with $\boldsymbol{X}^*_1$ and $\boldsymbol{\beta}^*$, respectively. As before, apply sum-to-zero constraints to $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$.

MCMC Updates for Two-Part Log Skew-t Model

The updates of the binary part of the two-part LST are also similar to those given in two-part LN model. The updates for the continuous part are similar to those for the LSN, but with the following modifications:

1. Update Z_{ij} in equation (2.6). Given a truncated normal $\mathrm{TN}_{[0,\infty)}(0, 1/w_{ij})$ prior for all Z_{ij}

such that $y_{ij} > 0$, the full conditional of Z_{ij} is $\text{N}(\mu_{z_{ij}}, \sigma^2_{z_{ij}})$, where

$$\sigma^2_{z_{ij}} = \frac{\xi^2}{(\psi^2 + \xi^2)w_{ij}} \quad \text{and}$$
$$\mu_{z_{ij}} = \sigma^2_{z_{ij}} \psi w_{ij} \left\{ \ln y_{ij} - \boldsymbol{x}'_{ij}\boldsymbol{\beta} - b_{2i} \right\} / \xi^2.$$

2. Update W_{ij}. Assuming a $\text{Ga}(\frac{\nu}{2}, \frac{\nu}{2})$ prior, for all $y_{ij} > 0$, the full conditional for W_{ij} is:

$$\text{Ga}\left(\frac{\nu}{2} + 1, \quad \frac{\nu}{2} + \frac{\left\{ \ln y_{ij} - \boldsymbol{x}'_{ij}\boldsymbol{\beta} - b_{2i} \right\}^2}{2\xi^2} + \frac{z^2_{ij}}{2} \right).$$

Concatenate the sampled values, w_{ij}, to form an $N_1 \times 1$ vector $\boldsymbol{w}_1$.

3. Update other parameters. Define $\widetilde{\boldsymbol{X}}_1 = \boldsymbol{X}^*_1 \sqrt{\boldsymbol{w}_1}$, $\widetilde{\boldsymbol{D}}_1 = \boldsymbol{D}_1 \sqrt{\boldsymbol{w}_1}$, and $\widetilde{\boldsymbol{y}}_1 = \ln(\boldsymbol{y}_1)\sqrt{\boldsymbol{w}_1}$, where $\boldsymbol{X}^*_1$ and $\boldsymbol{D}_1$ are defined earlier. Updates for $\boldsymbol{\beta}^*$, ξ^2, $\boldsymbol{b}_2$ and $\boldsymbol{\Lambda}$ are identical to updating $\boldsymbol{\beta}^*$, ξ^2, $\boldsymbol{b}_2$ and $\boldsymbol{\Lambda}$ in the LSN model if we replace $\boldsymbol{X}^*_1$, $\boldsymbol{D}_1$ and $\boldsymbol{y}_1$ with $\widetilde{\boldsymbol{X}}_1$, $\widetilde{\boldsymbol{D}}_1$ and $\widetilde{\boldsymbol{y}}_1$, respectively. Apply sum-to-zero constraints as before.

MCMC Updates for DP Mixture of Two-Part Lognormal Models

For the DPLN, we follow the stick-breaking construction outlined in Section 2.4 and iterate through following steps:

1. Update the component index C_{ij}. For all i and j, sample C_{ij} from its multinomial full conditional:

$$Pr(C_{ij} = k|-) = \frac{\left\{ \upsilon_k \prod_{l<k}(1 - \upsilon_l) \right\} f(y_{ij}|C_{ij} = k)}{\sum_{r=1}^{K} \left\{ \upsilon_r \prod_{s<r}(1 - \upsilon_s) \right\} f(y_{ij}|C_{ij} = r)},$$

where $f(y_{ij}|C_{ij} = k)$ is defined in equation (2.10) and K is the total number of mixture components after truncation.

2. Update the stick breaking weights υ_k. Assuming a $\text{Be}(1, \omega)$ prior, the full conditional for υ_k $(k = 1, \ldots, K)$ is:

$$\upsilon_k|- \sim \text{Be}\left(1 + \sum_{i=1}^{n}\sum_{j=1}^{n_i} \mathbb{1}(C_{ij} = k), \quad \omega + \sum_{i=1}^{n}\sum_{j=1}^{n_i} \mathbb{1}(C_{ij} > k) \right).$$

3. Conditional on $C_{ij} = k$, update the latent variable U_{ij} from its truncated normal full conditional, which is analogous the update for the LN model but with component-k likelihood and parameters.

4. For $k = 1, \ldots, K$, update the component-specific parameters $\boldsymbol{\gamma}_k$, $\boldsymbol{\beta}_k$, $\boldsymbol{b}_{1k}$, $\boldsymbol{b}_{2k}$, $\boldsymbol{\Lambda}_k$ and σ_k^2 following full conditionals similar to those for the LN model, but now with component-specific likelihood and parameters. For all k, apply sum-to-zero constraints to $\boldsymbol{b}_{1k}$ and $\boldsymbol{b}_{2k}$.