# SPATIAL SCAN STATISTICS FOR MODELS WITH OVERDISPERSION AND INFLATED ZEROS

Max S. de Lima[a], Luiz H. Duczmal[b], José C. Neto[a] and Letícia P. Pinto[b]

[a]*Federal University of Amazonas,* [b]*Federal University of Minas Gerais, Brazil*

**Supplementary Material**

## S1 Maximum likelihood estimators for $H_0$ and $H_1$

The logarithm of the likelihood ratio for the ZIDP model under $H_1$ is

$$\begin{aligned} l_Z^a(p,\theta_1,\theta_2,\phi;\mathbf{y},\mathbf{u}) &= \sum_{i=1}^{L}(u_i\log p+(1-u_i)\log(1-p)) \\ &+ \sum_{s_i\in Z}(1-u_i)\log f_{DP}(y_i|\theta_1 n_i,\phi) \\ &+ \sum_{s_i\notin Z}(1-u_i)\log f_{DP}(y_i|\theta_2 n_i,\phi) \\ &= l_Z^a(p;\mathbf{u})+l_Z^a(\theta_1,\phi;\mathbf{y},\mathbf{u})+l_Z^a(\theta_2,\phi;\mathbf{y},\mathbf{u}). \end{aligned} \tag{S1.1}$$

Then, in step **E**:

$$\begin{aligned} \mathbb{E}\left\{l_Z^a(p,\theta_1,\theta_2,\phi;\mathbf{y},\mathbf{u})|\mathbf{y},\boldsymbol{\gamma}^{(k)}\right\} &= \sum_{i=1}^{L}\left(u_i^{(k)}\log p+(1-u_i^{(k)})\log(1-p)\right) \\ &+ \sum_{s_i\in Z}(1-u_i^{(k)})\log f_{DP}(y_i|\theta_1 n_i,\phi) \\ &+ \sum_{s_i\notin Z}(1-u_i^{(k)})\log f_{DP}(y_i|\theta_2 n_i,\phi). \end{aligned}$$

- Step M for $p$:

$$\frac{\partial}{\partial p}\mathbb{E}\left\{l_Z^a(p,\theta_1,\theta_2,\phi;\mathbf{y},\mathbf{u})|\mathbf{y},\boldsymbol{\gamma}^{(k)}\right\}=\frac{\sum_{s_i\in\mathcal{S}}(u_i^{(k)})}{p}+\frac{\sum_{s_i\in\mathcal{S}}(1-u_i^{(k)})}{1-p}. \tag{S1.2}$$

Equating to zero, we obtain

$$\hat{p}^{(k+1)} = \frac{\sum_{i=1}^{L} u_i^k}{L}.$$

- Step M for $\theta_1$:

$$\log f_{DP}(y_i|\theta_1 n_i, \phi) \propto \frac{1}{2}\log\phi - \theta_1\phi n_i + \phi y_i \log\theta_1.$$

Then

$$\frac{\partial}{\partial\theta_1}\mathbb{E}\left\{l_Z^a(p,\theta_1,\theta_2,\phi;\mathbf{y},\mathbf{u})|\mathbf{y},\boldsymbol{\gamma}^{(k)}\right\} = \sum_{s_i\in Z}(1-u_i^{(k)})\left[-n_i\phi + y_i\phi/\theta_1\right]. \quad \text{(S1.3)}$$

Equating to zero, we obtain

$$\hat{\theta}_1^{(k+1)} = \frac{\sum_{s_i\in Z}(1-u_i^{(k)})y_i}{\sum_{s_i\in Z}(1-u_i^{(k)})n_i}.$$

- Step M for $\theta_2$:

$$\frac{\partial}{\partial\theta_2}\mathbb{E}\left\{l_Z^a(p,\theta_1,\theta_2,\phi;\mathbf{y},\mathbf{u})|\mathbf{y},\boldsymbol{\gamma}^{(k)}\right\} = \sum_{s_i\notin Z}(1-u_i^{(k)})\left[-n_i\phi + y_i\phi/\theta_2\right], \quad \text{(S1.4)}$$

and

$$\hat{\theta}_2^{(k+1)} = \frac{\sum_{s_i\notin Z}(1-u_i^{(k)})y_i}{\sum_{s_i\notin Z}(1-u_i^{(k)})n_i}.$$

- Step M for $\phi$:

$$\begin{aligned}\frac{\partial}{\partial\phi}\mathbb{E}\left\{l_Z^a(\hat{p},\hat{\theta}_1,\hat{\theta}_2,\phi;\mathbf{y},\mathbf{u})|\mathbf{y},\boldsymbol{\gamma}^{(k)}\right\} &= \frac{\partial}{\partial\phi}\left[\sum_{s_i\in Z}(1-u_i^{(k)})\log f_{DP}(y_i|\hat{\theta}_1 n_i,\phi)\right] \\ &+ \frac{\partial}{\partial\phi}\left[\sum_{s_i\notin Z}(1-u_i^{(k)})\log f_{DP}(y_i|\hat{\theta}_2 n_i,\phi)\right] \\ &= \frac{\sum_{i=1}^{L}(1-u_i^{(k)})}{2\phi} \\ &- \left\{\sum_{s_i\in Z}(1-u_i^{(k)})y_i\log(\theta_i/\hat{\theta}_1) + \sum_{s_i\notin Z}(1-u_i^{(k)})y_i\log(\theta_i/\hat{\theta}_2)\right\}.\end{aligned}$$

Equating to zero, we obtain

$$\phi_1^{(k+1)} = \frac{\sum_{i=1}^{L}(1-u_i^{(k)})}{2\left\{\sum_{s_i\in Z}(1-u_i^{(k)})y_i\log(\theta_i/\theta_1^{(k+1)}) + \sum_{s_i\notin Z}(1-u_i^{(k)})y_i\log(\theta_i/\theta_2^{(k+1)})\right\}}.$$

The estimators for $H_0$ are similarly obtained by maximizing $l_0^a(\hat{\theta_0},\phi;\mathbf{y},\mathbf{u})$.

# S2 Convergence of the EM Algorithm

The convergence of the ZIDP EM algorithm is studied through simulations. A proof of the convergence is also given in the end of this section.

A simple and effective initialization procedure of the ZIDP EM algorithm could be stated as follows.

- Suppose that initially $H_0$ is true and all zeros are structural, such that $u_i = 1$ when $y_i = 0$ and zero if $y_i > 0$. Initialize the EM algorithm with

$$p_1^{(0)} = p_0^{(0)} = \frac{\#(y_i = 0)}{L}, \quad \theta_0^{(0)} = \theta_1^{(0)} = \theta_2^{(0)} = \frac{\sum_{i:y_i>0} y_i}{\sum_{i:y_i>0} n_i}, \tag{S2.5}$$

$$\phi_1^{(0)} = \phi_0^{(0)} = \frac{L - \#(y_i = 0)}{2\left\{\sum_{i:y_i>0} y_i \log(\theta_i/\theta_0^{(0)})\right\}}, \tag{S2.6}$$

where $p_1^{(0)} = p_0^{(0)}$ is the proportion of zeros, $\theta_0^{(0)} = \theta_1^{(0)} = \theta_2^{(0)}$ is the global case rate, and $\phi_1^{(0)} = \phi_0^{(0)}$ is the positive cases overdispersion.

A corresponding simulation is shown, employing the 12 **ScanZIOP** models described in Table 2 to verify the convergence. For each set of initial values given in (S2.5) and (S2.6), 100 simulations were run. The results of Table 1 show the fast convergence for all the studied models.

We show that the ZIDP EM algorithm converges for every initial value $\gamma^{(0)}$. The proof is based on the general convergence theory of EM algorithms (Vaida( 2005)). A proof is given for the numerator of expression (3.1). The corresponding proof for the denominator is similar. Let $\boldsymbol{\gamma} = (p, \theta_1, \theta_2, \phi) \in \Theta = (0,1) \times (0,1) \times (0,1) \times (0,1]$, and

$$Q(\boldsymbol{\gamma}|\widetilde{\boldsymbol{\gamma}}) = \mathbb{E}\left\{l_Z^a(p, \theta_1, \theta_2, \phi; \mathbf{y}, \mathbf{u})|\mathbf{y}, \widetilde{\boldsymbol{\gamma}}\right\}$$

be the function to be maximized at step $M$ of the $EM$ algorithm. Then by Theorem 3 of Vaida(2005), for every initial value $\boldsymbol{\gamma}^{(0)}$ the EM algorithm converges when $Q(\boldsymbol{\gamma}|\widetilde{\boldsymbol{\gamma}})$ has an unique point of maximum. Combining (1.3) and (3.2), it follows that

$$\begin{aligned} Q(\boldsymbol{\gamma}|\widetilde{\boldsymbol{\gamma}}) &= \sum_{i=1}^{L} \{\widetilde{u}_i \log p + (1 - \widetilde{u}_i) \log(1-p)\} \\ &+ \sum_{s_i \in Z} (1 - \widetilde{u}_i) \left\{ \frac{1}{2}\phi - n_i \phi \theta_1 + \log\left(\frac{e^{-y_i} y_i^{y_i}}{y_i!}\right) + \phi y_i \log\left(\frac{e n_i \theta_1}{y_i}\right) \right\} \\ &+ \sum_{s_i \notin Z} (1 - \widetilde{u}_i) \left\{ \frac{1}{2}\phi - n_i \phi \theta_2 + \log\left(\frac{e^{-y_i} y_i^{y_i}}{y_i!}\right) + \phi y_i \log\left(\frac{e n_i \theta_2}{y_i}\right) \right\}. \end{aligned}$$

The partial derivatives and the critical point of $Q(.|\widetilde{\boldsymbol{\gamma}})$ are given in the Supplementary Materials section and the corresponding Hessian matrix is

$H(\boldsymbol{\gamma}|\widetilde{\boldsymbol{\gamma}}) =$

$$= \begin{bmatrix} \frac{\sum_{s_i \in \mathcal{S}} (\widetilde{u}_i)}{p^2} - \frac{\sum_{s_i \in \mathcal{S}} (1-\widetilde{u}_i)}{(1-p)^2} & 0 & 0 & 0 \\ 0 & -\frac{\phi}{\theta_1^2} \sum_{s_i \in Z} (1-\widetilde{u}_i) y_i & 0 & \sum_{s_i \in Z} (1-\widetilde{u}_i)\{-n_i + \frac{y_i}{\theta_1}\} \\ 0 & 0 & -\frac{\phi}{\theta_2^2} \sum_{s_i \in Z} (1-\widetilde{u}_i) y_i & \sum_{s_i \notin Z} (1-\widetilde{u}_i)\{-n_i + \frac{y_i}{\theta_2}\} \\ 0 & \sum_{s_i \in Z} (1-\widetilde{u}_i)\{-n_i + \frac{y_i}{\theta_1}\} & \sum_{s_i \notin Z} (1-\widetilde{u}_i)\{-n_i + \frac{y_i}{\theta_2}\} & -\frac{\sum_{i=1}^{L} (1-\widetilde{u}_i)}{2\phi^2} \end{bmatrix}$$

Note that $H(\boldsymbol{\gamma}|\widetilde{\boldsymbol{\gamma}})$ is symmetric and negative definite. Thus, $Q(\boldsymbol{\gamma}|\widetilde{\boldsymbol{\gamma}})$ is strictly concave. If $\widehat{\boldsymbol{\gamma}} = (\widehat{p}, \widehat{\theta}_2, \widehat{\theta}_2, \widehat{\phi})$ is the MLE obtained in the Supplementary Materials section, then $H(\widehat{\boldsymbol{\gamma}}|\widetilde{\boldsymbol{\gamma}})$ is a diagonal matrix whose non-null entries are the entries of the diagonal of $H(\boldsymbol{\gamma}|\widetilde{\boldsymbol{\gamma}})$ applied to $\widehat{\boldsymbol{\gamma}}$ such that $H(\widehat{\boldsymbol{\gamma}}|\widetilde{\boldsymbol{\gamma}})$ is also symmetric and negative definite. Thus $\widehat{\boldsymbol{\gamma}}$ is an unique global maximum of $Q(\boldsymbol{\gamma}|\widetilde{\boldsymbol{\gamma}})$.

Table 1: Average and standard deviation (sd) estimated for $(\theta_1, \theta_2, p, \phi)$ and the average number of steps to convergence, given initial values $\gamma^{(0)} = (p_1^{(0)}, \phi_0^{(1)}, \theta_1^{(0)}, \theta_2^{(0)})$ according to expressions (S2.5) and (S2.6), using 100 simulations for each set of initial values.

| | Estimated parameters and number of iterations (sd) | | | | |
|---|---|---|---|---|---|
| $(p, 1/\phi, \theta_1, \theta_2)$ | $p$ | $1/\phi$ | $\theta_1$ | $\theta_2$ | # iterations |
| (0.2,1.5,0.0015,0.001) | 0.20791 | 1.14280 | 0.00152 | 0.00097 | 5.06500 |
| | (0.05474) | (0.15122) | (0.00016) | (0.00005) | (0.50837) |
| (0.2,2.0,0.0015,0.001) | 0.20549 | 1.83715 | 0.00152 | 0.00100 | 5.02000 |
| | (0.05257) | (0.10727) | (0.00020) | (0.00003) | (0.37551) |
| (0.3,1.5,0.0015,0.001) | 0.29586 | 1.37003 | 0.00155 | 0.00100 | 5.06000 |
| | (0.06129) | (0.14800) | (0.00018) | (0.00003) | (0.37118) |
| (0.3,2.0,0.0015,0.001) | 0.30910 | 1.76217 | 0.00152 | 0.00099 | 5.12000 |
| | (0.05942) | (0.12095) | (0.00019) | (0.00004) | (0.38350) |
| (0.2,1.5,0.002,0.001) | 0.19744 | 1.38187 | 0.00200 | 0.00100 | 5.07000 |
| | (0.04842) | (0.13809) | (0.00016) | (0.00003) | (0.40837) |
| (0.2,2.0,0.002,0.001) | 0.20769 | 1.89586 | 0.00199 | 0.00100 | 5.03000 |
| | (0.05304) | (0.11352) | (0.00022) | (0.00003) | (0.43705) |
| (0.3,1.5,0.002,0.001) | 0.30102 | 1.35221 | 0.00197 | 0.00100 | 5.11000 |
| | (0.05476) | (0.14807) | (0.00024) | (0.00003) | (0.34510) |
| (0.3,2.0,0.002,0.001) | 0.29884 | 1.87121 | 0.00197 | 0.00100 | 5.09000 |
| | (0.05919) | (0.12110) | (0.00022) | (0.00004) | (0.32083) |
| (0.2,1.5,0.003,0.001) | 0.19467 | 1.40841 | 0.00300 | 0.00100 | 4.85000 |
| | (0.05014) | (0.14273) | (0.00022) | (0.00003) | (0.43519) |
| (0.2,2.0,0.003,0.001) | 0.19442 | 1.79782 | 0.00302 | 0.00100 | 4.84000 |
| | (0.04815) | (0.12181) | (0.00023) | (0.00003) | (0.48659) |
| (0.3,1.5,0.003,0.001) | 0.29579 | 1.39204 | 0.00302 | 0.00100 | 4.95000 |
| | (0.06080) | (0.14756) | (0.00027) | (0.00003) | (0.21904) |
| (0.3,2.0,0.003,0.001) | 0.29735 | 1.77302 | 0.00302 | 0.00100 | 4.92000 |
| | (0.05513) | (0.13040) | (0.00026) | (0.00003) | (0.33874) |

Table 2: Comparison of type I error probabilities for **ScanP**, **ScanZIP**, **ScanOP** and **ScanZIOP** with several values of $\phi$ and $p$.

| | | Method | | | |
|---|---|---|---|---|---|
| | $1/\phi$ | **ScanP** | **ScanZIP** | **ScanOP** | **ScanZIOP** |
| $p = 0.3$ | 1.00 | 0.609 | 0.035 | 0.393 | 0.035 |
| | 1.50 | 0.813 | 0.253 | 0.136 | 0.090 |
| | 2.00 | 0.894 | 0.499 | 0.227 | 0.086 |
| | 3.00 | 0.974 | 0.833 | 0.394 | 0.085 |
| $p = 0.2$ | 1.00 | 0.390 | 0.038 | 0.371 | 0.040 |
| | 1.50 | 0.643 | 0.282 | 0.076 | 0.097 |
| | 2.00 | 0.773 | 0.520 | 0.096 | 0.093 |
| | 3.00 | 0.937 | 0.868 | 0.251 | 0.090 |
| $p = 0.1$ | 1.00 | 0.383 | 0.042 | 0.211 | 0.043 |
| | 1.50 | 0.594 | 0.263 | 0.026 | 0.091 |
| | 2.00 | 0.791 | 0.560 | 0.030 | 0.010 |
| | 3.00 | 0.894 | 0.888 | 0.080 | 0.090 |
| $p = 0.0$ | 1.00 | 0.026 | 0.048 | 0.056 | 0.047 |
| | 1.50 | 0.236 | 0.304 | 0.080 | 0.097 |
| | 2.00 | 0.382 | 0.608 | 0.066 | 0.095 |
| | 3.00 | 0.726 | 0.910 | 0.056 | 0.090 |

Table 3: Estimated values of Power for **ScanP**, **ScanZIP**, **ScanOP** and **ScanZIOP** with several values of ($\theta, p$ and $\phi$).

| | | Method | | | |
|---|---|---|---|---|---|
| | $(p, 1/\phi)$ | **ScanP** | **ScanZIP** | **ScanOP** | **ScanZIOP** |
| $\theta = 0.5$ | (0.2,1.5) | 0.994 | 0.800 | 0.195 | 0.518 |
| | (0.2,2.0) | 0.998 | 0.830 | 0.245 | 0.366 |
| | (0.3,1.5) | 0.998 | 0.716 | 0.323 | 0.442 |
| | (0.3,2.0) | 1.000 | 0.768 | 0.355 | 0.346 |
| $\theta = 1.0$ | (0.2,1.5) | 1.000 | 0.984 | 0.452 | 0.944 |
| | (0.2,2.0) | 1.000 | 0.982 | 0.493 | 0.898 |
| | (0.3,1.5) | 1.000 | 0.950 | 0.432 | 0.882 |
| | (0.3,2.0) | 1.000 | 0.956 | 0.478 | 0.830 |
| $\theta = 2.0$ | (0.2,1.5) | 1.000 | 1.000 | 0.894 | 1.000 |
| | (0.2,2.0) | 1.000 | 1.000 | 0.890 | 0.992 |
| | (0.3,1.5) | 1.000 | 0.998 | 0.830 | 1.000 |
| | (0.3,2.0) | 1.000 | 0.996 | 0.821 | 0.978 |

Table 4: Estimated values of Sensitivity (**SS**) and Positive Predicted Value (**PPV**) for **ScanP**, **ScanZIP**, **ScanOP** and **ScanZIOP** with several values of ($\theta, p$ and $\phi$).

| | | | Method | | | |
|---|---|---|---|---|---|---|
| | | $(p, 1/\phi)$ | **ScanP** | **ScanZIP** | **ScanOP** | **ScanZIOP** |
| | | (0.2,1.5) | 0.723 | 0.600 | 0.176 | 0.415 |
| | $\theta = 0.5$ | (0.2,2.0) | 0.687 | 0.568 | 0.218 | 0.281 |
| | | (0.3,1.5) | 0.651 | 0.533 | 0.251 | 0.351 |
| | | (0.3,2.0) | 0.645 | 0.513 | 0.288 | 0.260 |
| | | (0.2,1.5) | 0.784 | 0.909 | 0.427 | 0.882 |
| **SS** | $\theta = 1.0$ | (0.2,2.0) | 0.781 | 0.846 | 0.454 | 0.794 |
| | | (0.3,1.5) | 0.742 | 0.838 | 0.377 | 0.798 |
| | | (0.3,2.0) | 0.729 | 0.791 | 0.415 | 0.719 |
| | | (0.2,1.5) | 0.405 | 0.950 | 0.782 | 0.951 |
| | $\theta = 2.0$ | (0.2,2.0) | 0.423 | 0.946 | 0.775 | 0.944 |
| | | (0.3,1.5) | 0.473 | 0.900 | 0.675 | 0.900 |
| | | (0.3,2.0) | 0.494 | 0.890 | 0.677 | 0.881 |
| | | (0.2,1.5) | 0.277 | 0.482 | 0.044 | 0.335 |
| | $\theta = 0.5$ | (0.2,2.0) | 0.279 | 0.395 | 0.049 | 0.199 |
| | | (0.3,1.5) | 0.177 | 0.424 | 0.053 | 0.272 |
| | | (0.3,2.0) | 0.188 | 0.344 | 0.058 | 0.164 |
| | | (0.2,1.5) | 0.581 | 0.823 | 0.287 | 0.787 |
| **PPV** | $\theta = 1.0$ | (0.2,2.0) | 0.557 | 0.749 | 0.257 | 0.715 |
| | | (0.3,1.5) | 0.375 | 0.781 | 0.154 | 0.732 |
| | | (0.3,2.0) | 0.387 | 0.680 | 0.171 | 0.620 |
| | | (0.2,1.5) | 0.262 | 0.977 | 0.723 | 0.979 |
| | $\theta = 2.0$ | (0.2,2.0) | 0.268 | 0.951 | 0.741 | 0.949 |
| | | (0.3,1.5) | 0.254 | 0.962 | 0.618 | 0.960 |
| | | (0.3,2.0) | 0.265 | 0.920 | 0.590 | 0.915 |

Table 5: Estimated values of Power, Sensitivity (**SS**) and Positive Predicted Value (**PPV**) for **ScanZIOP** in a cluster with low population size, for several values of ($\theta, p$ and $\phi$) .

| | | Method ScanZIOP | | |
|---|---|---|---|---|
| | $(p, 1/\phi)$ | **Power** | **SS** | **PPV** |
| | (0.2,1.5) | 0.480 | 0.380 | 0.292 |
| $\theta = 0.5$ | (0.2,2.0) | 0.324 | 0.296 | 0.261 |
| | (0.3,1.5) | 0.422 | 0.326 | 0.305 |
| | (0.3,2.0) | 0.316 | 0.238 | 0.214 |
| | (0.2,1.5) | 0.958 | 0.861 | 0.858 |
| $\theta = 1.0$ | (0.2,2.0) | 0.886 | 0.774 | 0.752 |
| | (0.3,1.5) | 0.902 | 0.794 | 0.823 |
| | (0.3,2.0) | 0.826 | 0.692 | 0.691 |
| | (0.2,1.5) | 0.996 | 0.923 | 0.979 |
| $\theta = 2.0$ | (0.2,2.0) | 1.000 | 0.916 | 0.964 |
| | (0.3,1.5) | 0.990 | 0.856 | 0.973 |
| | (0.3,2.0) | 0.994 | 0.856 | 0.952 |

Table 6: Estimated values of Power, Sensitivity (**SS**) and Positive Predicted Value (**PPV**) for **ScanZIOP** in a cluster with high population size, for several values of ($\theta, p$ and $\phi$) .

| | | Method ScanZIOP | | |
|---|---|---|---|---|
| | $(p, 1/\phi)$ | **Power** | **SS** | **PPV** |
| $\theta = 0.5$ | (0.2,1.5) | 0.518 | 0.415 | 0.335 |
| | (0.2,2.0) | 0.366 | 0.281 | 0.199 |
| | (0.3,1.5) | 0.442 | 0.351 | 0.272 |
| | (0.3,2.0) | 0.346 | 0.260 | 0.164 |
| $\theta = 1.0$ | (0.2,1.5) | 0.944 | 0.882 | 0.787 |
| | (0.2,2.0) | 0.898 | 0.794 | 0.715 |
| | (0.3,1.5) | 0.882 | 0.794 | 0.732 |
| | (0.3,2.0) | 0.830 | 0.719 | 0.620 |
| $\theta = 2.0$ | (0.2,1.5) | 1.000 | 0.951 | 0.979 |
| | (0.2,2.0) | 0.992 | 0.944 | 0.949 |
| | (0.3,1.5) | 1.000 | 0.900 | 0.960 |
| | (0.3,2.0) | 0.978 | 0.881 | 0.915 |