An approximate Bayesian estimation of a Poisson Markov random field model for crash data

Ignacio Alvarez-Castro   Kristian Schmith   Jarad Niemi   Alicia Carriquiry

SIESTA - 28 de Mayo de 2019

Traffic accidents in Iowa:

- About 400 annual fatalities, 1 billion dollars yearly cost (McDonald, 2012).
- DOT wants to identify *hot spots*, i.e. the sites with potential risk.
- Data: Number of crash accidents on intersections from 3 towns

Winsorized Poisson Markov random fields (WPMRF)

- Model crashes at the intersection level, allowing spatial correlation among intersections.
- WPMRF for spatially correlated count data (Kaiser, 2002; Kaiser and Cressie, 1997).
- WPMRF are *doubly* intractable, then ABC for inference.

Model areal-referenced count data:

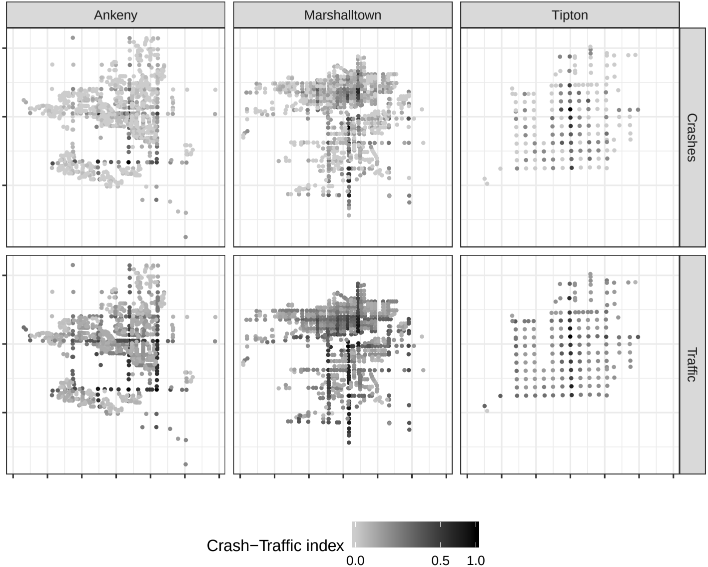
- response: discrete variable, available in a set of locations.
- neighborhood structure for spatial dependence
- continuous covariate

DOT provides:

- intersection crashes and traffic volume
- connectivity information

Neighbors: *link* between intersections.
EW neighbors: Distance in latitude is the smallest.

Crash accidents on intersections



Crash accidents on intersections

Table: Summary statistics

Town	Intersection		Crashes				Total traffic	
	Total	No Crash	Mean	Q90	I_{EW}	I_{NS}	I_{EW}	I_{NS}
Ankeny	893	0.6	3.54	20	0.19	0.28	0.42	0.57
Marshalltown	764	0.43	3.27	12	0.18	0.46	0.16	0.78
Tipton	159	0.61	0.7	3	0.09	0.53	0.09	0.79

Bayesian inference is the process of fitting a probability model to a set of data and summarizing the result by a probability distribution on relevant quantities (unobserved) conditional on observed data.

Main step:

- Obtain the conditional probability distribution of θ given the data y.

$$
\begin{aligned}
p(\theta|y) &= p(y|\theta)\ p(\theta)\,\tfrac{1}{p(y)} \\
&\propto p(y|\theta)\ p(\theta)
\end{aligned}
$$

What is ABC ?

- ABC: obtain posterior WITHOUT use likelihood.

- Main idea: accept θ^* when synthetic data from $p(y|\theta^*)$ are *similar* to observed data

- Useful when
  - likelihood is analytically or computationally intractable
  - is posible simulate data from the data model

$$\begin{array}{ll} \text{Data Model:} & y \sim Bin(20, p) \\ \text{Prior:} & p \sim Beta(1, 1) \end{array}$$

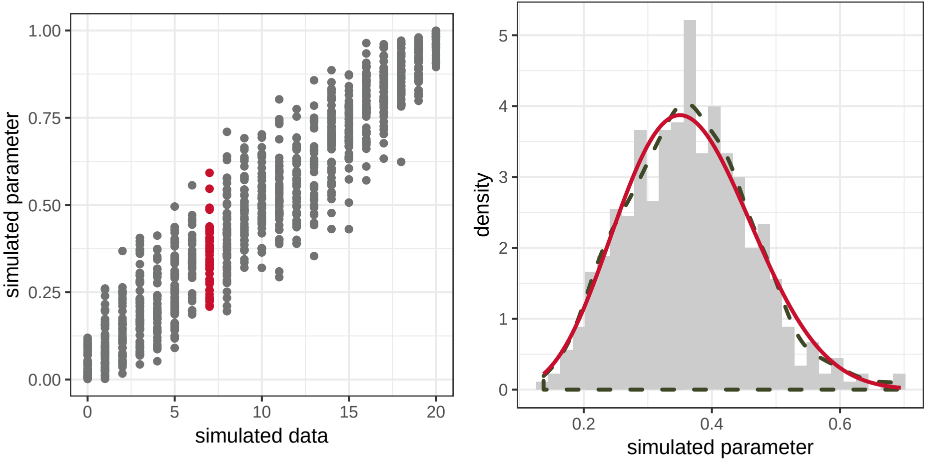
asume we observe $y_0 = 7$, then:

$$p|y_0 \sim Beta(7 + 1, 20 - 7 + 1)$$

We could obtain the posterior by doing:

```
N = 1e4; y = 7
the  <- runif(N)
ysim <- rbinom(N, 20, the)
post <- data.frame(the = the[ ysim == y ] )
```

A (very) small example



- 501 points where $y_{sim} = y_0 = 7$ ($\approx 5\%$ of simulations)
- Red line is the true posterior
- Black line is a KDE estimation

In practice:

Distance $y_{sim} \neq y_0$, so we accept *ysim similar*, $\rho(ysim, y_0) \leq \epsilon$

Summary statistic $\rho(ysim, y_0)$ is bad, then $\rho(T(ysim), T(y_0))$

Improvement
- better sampling scheme
- post-processing: model can be used to improve approximation

Rejection ABC

Algorithm 1 ABC-Rejection sampler

1. Compute $t_0 = T(y_{obs})$
2. For $(i \in 1: S)$
   - generate $\theta_i \sim \pi(\theta)$, and $y_i^* \sim f(y|\theta_i)$
   - compute $t_i = T(y_i^*)$ and $d_i = \rho(t_i, t_0)$
3. Return $\{\theta_i : d_i < d_{(k_S)}\}$, the k_S-nearest neighbors of t_0

Output:

$$\{(\theta_i, t_i)\}_{i=1}^{k_S} \sim p(\theta, T(y)|d_i < d_{(k_S)})$$

a random sample from $(\theta, T(y))$ joint density restricted to a neighborhood of $t_0 = T(y_{obs})$ (Biau et al., 2015).

Consider a second example

Data Model: $y_i \sim Bin(20, p) \; i = 1, \ldots, 15$

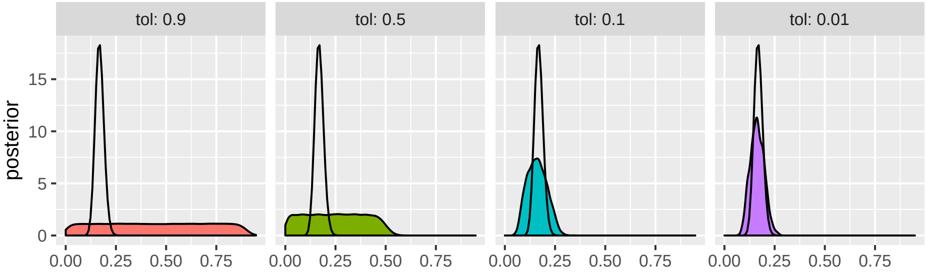
Prior: $p \sim Beta(1, 1)$

Data $y_{obs} = (3, 6, 5, 3, 3, 2, 4, 6, 5, 2, 5, 8, 4, 3, 2)$

Posterior $p|y \sim Beta(61 + 1, 239 + 1)$

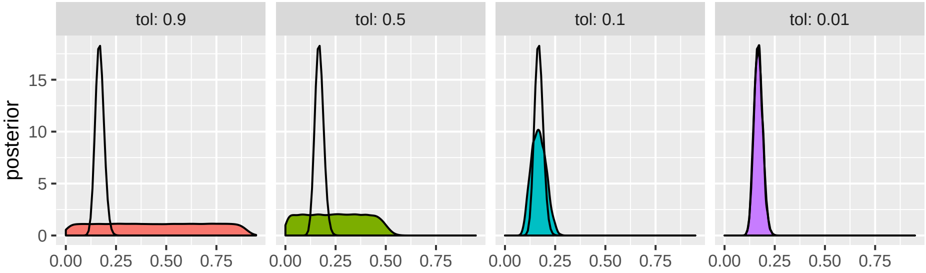
- Simultate $1e5$ datasets, $y^k = y$ 0 times !
- Use proportion as summary $T(y) = \sum y/(20 \times 15)$

- Smaler ϵ improve approximation.



Distance in summary statistic

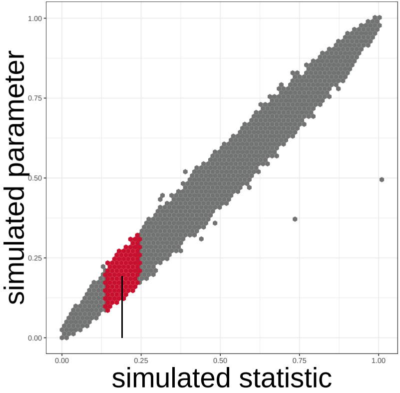
- Is better to meassure distance with the summary statistic.



Rejection estimate: acepted p_k (red points) are unchanged.

Consider a model:
$p_k = \beta_0 + \beta_1 s_k + e$

Regression model to correct discrepancy between s_k and s_0.



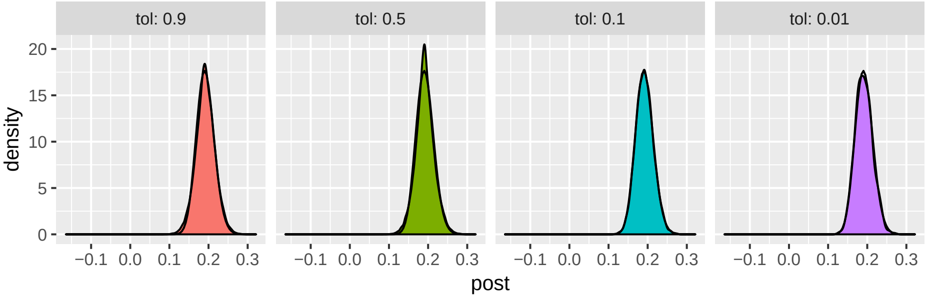
Post-processing simulations

$$\tilde{p}_k = p_k + (\hat{\beta}_0 + \hat{\beta}_1 s_0) - (\hat{\beta}_0 + \hat{\beta}_1 s_k)$$

- correct for discrepancy between s_k and s_0
- project p_k parallel to the model line.
- d_k as weight
- use more complex auxiliary models

Post-processing simulations

- regression reduce sensibility to ϵ
- Potentially: we could use all simulated p_k



Markov random field

$y(s_i)$ the response variable for location s_i
N_i the set of neighbors of s_i
$y(N_i) \equiv \{y(s_j) : j \in N_i\}$

Markov property

$$p(y(s_i)|\theta, y) = p(y(s_i)|\theta, y(N_i)).$$

Auto-models (Besag, 1974)

- like glm, with natural parameter as function of neighbors
- *neg-potential function* $Q(y|\theta) \equiv log\left[\frac{p(y|\theta)}{p(y_0|\theta)}\right]$

$$p(\theta|y) = \frac{e^{Q(y|\theta)}}{k(\theta)} p(\theta) \frac{1}{p(y)} \tag{1}$$

Posterior in (1) is Double intractable: $k(\theta)$, $p(y)$

- Poisson Auto-models does not allow positive correlation (Kaiser and Cressie, 2000)
- Winsorization: $Y \equiv \tilde{Y} \cdot I(\tilde{Y} \leq R) + R \cdot I(\tilde{Y} > R)$, where $R < \infty$, and $\tilde{Y} \sim Poi(\lambda)$

$$\begin{aligned} &y(s_i)|N_i \sim WP(\lambda_i, R) \\ &\lambda_i = \beta_0 + \beta_1 X_i + \sum_k \sum_{j \in N_{i,k}} \eta_k \left[y(s_j) - \beta_0 - \beta_1 X_j\right] \end{aligned} \tag{2}$$

- $y(s_i)$ is the number of crashes occurred at intersection s_i
- X_i represents the total traffic in the intersection s_i.
- Anisotropic dependence: $\eta_k \in (\eta_{NS}, \eta_{EW})$

Two stages:
Stage 1: Rejection ABC:

- Is not possible to simulate directly from WP-MRF model
- Run a **MCMC** chain to obtain 1 simulated data set

Stage 2: Post-processing simulated values

- non-lineal regression model (Blum and François, 2010)
- accommodates non-constant variance, not affected by $S()$ dimension

Moran statistic:

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(z_i - \bar{z})(z_j - \bar{z})}{\sum_{i=1}^{n}(z_i - \bar{z})^2}$$

Summary statistic, computed over simulated data set:

- Overall mean
- Correlation between (simulated) response and covariate
- Directional Moran statistic, I_{NS}, I_{EW}

Results: parameter inference

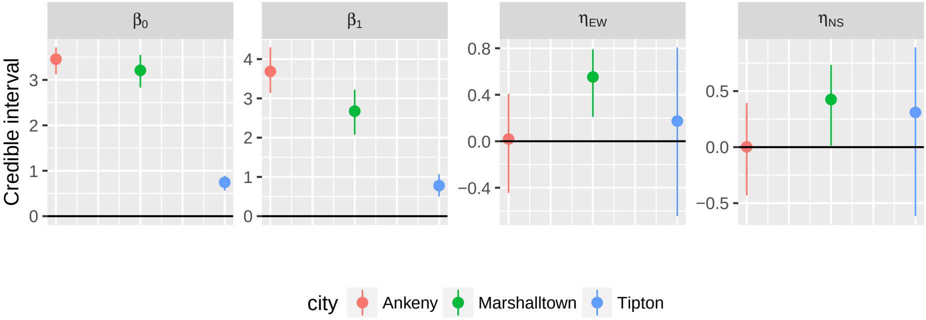


Figure: Parameter osterior credible intervals. Each facet corresponds to one of the four parameters in the model, the color of the points and lines represents the city.

Risk meassure

$$R_i = P(y(s_i) > y_{obs}(s_i) | y_{obs})$$

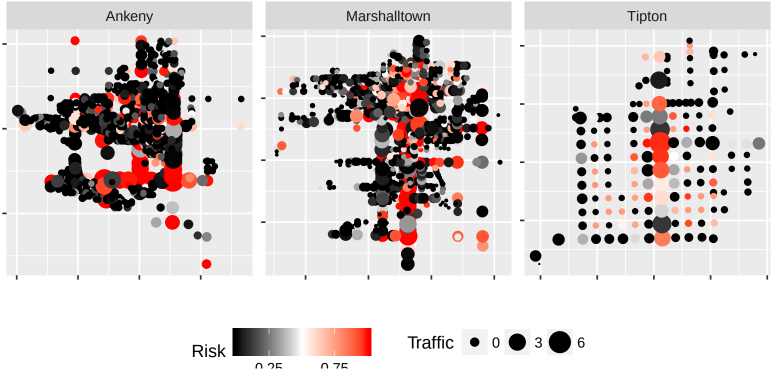


Figure: Intersection's risk at each intersection. The facets represent each of the three cities with data: Ankeny, Marshalltown and Tipton. Within each city (a facet) dots represent the intersections, color represents the risk measure of the intersection, and dot size represents the traffic volume.

Discussion

- Scope of the method: Areal spatial data model, both covariate and spatial correlation at the observation level.
- WP-MRF model that puts dependence structure directly on the crash numbers.
- ABC to make inference
  - ABC-rejection scheme, where each data simulation obtained via MCMC
  - Conditional density estimation with accepted parameter values
- Positive dependence in Marshalltown
- Medium risk intersections outside main roads

Besag, J. (1974), "Spatial interaction and the statistical analysis of lattice systems," *Journal of the Royal Statistical Society. Series B (Methodological)*, 192–236.

Biau, G., Cérou, F., Guyader, A., et al. (2015), "New insights into approximate bayesian computation," in *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, Institut Henri Poincaré, vol. 51, pp. 376–403.

Blum, M. G. and François, O. (2010), "Non-linear regression models for Approximate Bayesian Computation," *Statistics and Computing*, 20, 63–73.

Kaiser, M. S. (2002), "Markov random field models," in *Encyclopedia of Environmetrics*, Wiley John & Sons, pp. 1213–1225.

Kaiser, M. S. and Cressie, N. (1997), "Modeling Poisson variables with positive spatial dependence," *Statistics & Probability Letters*, 35, 423–432.

— (2000), "The construction of multivariate distributions from Markov random fields," *Journal of Multivariate Analysis*, 73, 199–220.

McDonald, T. (2012), "Traffic Safety Analysis for Local Agencies," .