Bayesian Model Checking in GLSM for Count Data

Liang Jing and Victor De Oliveira

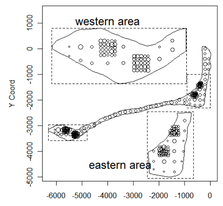
liang.jing@utsa.edu, victor.deoliveira@utsa.edu

Abstract

Model checking and selection in hierarchical models remain difficult problems due to the unobservable latent process. Here Bayesian model checking methods are introduced and compared with a new model checking method based on transformed residuals. Our simulation study reveals that Bayesian model checking methods fails on GLSM due to difficulty of choosing appropriate diagnostic statistics and conservatism of posterior predictive p-value, while the new model checking method provides more promising results.

Geostatistical Data

Each observation consists of **two attributes**:
- sampling location $\boldsymbol{s}_i$, where the response is observed;
- observed value Y_i, count of certain event.



Ronglap data set were collected from Rongelap Island, 2500 miles south-west of Hawaii. U.S. nuclear weapons testing programme generated heavy fallout over the island in the 1950s and Rongelap island has been uninhabited since 1985. There are 157 observations:
- y_i is the photon emission count
- $\boldsymbol{s}_i$ identifies spatial location
- l_i is the time (in seconds) over which y_i was accumulated

Hierarchical Model

Generalized linear spatial models were first proposed by Diggle et al. (1998) and widely used since then,

$$Y_i|\theta_i \overset{\text{id}}{\sim} Pois(\cdot|e^{\theta_i}),\ \ i=1,...,n;$$
$$\boldsymbol{\theta}|\boldsymbol{\eta} \sim MVN(D\boldsymbol{\beta}, \boldsymbol{\Sigma}),\ \ \Sigma_{ij} = \sigma^2\rho(u_{ij};\phi,\kappa)$$
$$\boldsymbol{\eta} = (\boldsymbol{\beta}, \sigma^2, \phi, \kappa)$$

- $\theta(\boldsymbol{s})$ is stationary Gaussian process;
- $D\boldsymbol{\beta}$ is the mean structure where D is covariate matrix (usually related to locations) and $\boldsymbol{\beta}$ is coefficient vector;
- $\rho(u_{ij};\phi,\kappa)$ is a correlation function, determined by Euclidean distance $u_{ij} = \|s_i - s_j\|$ between the i^{th} and j^{th} locations s_i, s_j.

For example, matern family:

$$\rho(u) = [2^{\kappa-1}\Gamma(\kappa)]^{-1}(u/\phi)^{\kappa}K_{\kappa}(u/\phi)$$

where $K_\kappa(\cdot)$ denotes a modified Bessel function of order κ.

R Package Development

▶ **Features** (under development)
- performs posterior sampling for parameter estimation, prediction, and model checking in hierarchical models with correlated latent variables;
- C++ programs are seamlessly embedded to handle heavy computational tasks of Markov chain generation and large matrices computation;
- parallel computing techniques are implemented to further speed up estimation and prediction;
- results are displayed by a combination of numerical and graphical summaries.

Robust MCMC Algorithm

Common Hastings-within-Gibbs algorithms fail to generate well-behaviored posterior samples because of the large number of latent variables that are highly correlated and influenced by response variables. We implemented robust MCMC algorithm proposed by Christensen et al. (2006)

▶ **Data-based parameterization**
after parameterization, the components of $\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\eta}}$ are approximately uncorrelated, and have zero means and unit variances

$$\boldsymbol{\theta} \to \tilde{\boldsymbol{\theta}}(\boldsymbol{\theta};\boldsymbol{\beta},\boldsymbol{\eta},Y)$$
$$\boldsymbol{\beta} \to \tilde{\boldsymbol{\beta}}(\boldsymbol{\beta};\boldsymbol{\eta},Y)$$
$$\boldsymbol{\eta} \to \tilde{\boldsymbol{\eta}}(\boldsymbol{\eta};Y)$$

▶ **Langevin-Hastings algorithm**
the gradient information of the target density is used into the proposal density to improve convergence.

▶ **Group Updating with Gibbs Sampler**

for *N iterations* **do**
1. Update $\tilde{\boldsymbol{\theta}} \sim p(\tilde{\boldsymbol{\theta}}|\tilde{\boldsymbol{\beta}},\boldsymbol{\eta},Y)$
2. Update $\tilde{\boldsymbol{\beta}} \sim p(\tilde{\boldsymbol{\beta}}|\tilde{\boldsymbol{\theta}},\boldsymbol{\eta},Y)$
3. Update $\tilde{\boldsymbol{\eta}} \sim p(\tilde{\boldsymbol{\eta}}|\tilde{\boldsymbol{\theta}},\tilde{\boldsymbol{\beta}},Y)$

▶ **Prior**
Christensen (2002) suggested flat prior and provided conditions under which the posterior is proper.

Bayesian Model Checking

Bayarri and Castellanos (2007) stated that **existing Bayesian model checking methods** can be seen to correspond to particular choices of the following three components:
- **a diagnostic statistic** $T(\boldsymbol{y})$ that summarizes a feature of the data
- **a specified reference distribution** $h(t)$ that represents an "predictive" distribution for T
- **a way to measure conflict** between $t^{obs} = T(\boldsymbol{y}^{obs})$ and the reference distribution

▶ **Two ways to measure conflict**

$p - value = P^{h(\cdot)}(T(\boldsymbol{Y}) \geq T(\boldsymbol{y}^{obs})), \quad RPS = \frac{h(t^{obs})}{\sup_t\{h(t)\}}$

small values indicate incompatibility.

▶ **The choices of reference distribution**

$$h(t) \leftarrow h(\boldsymbol{y}) = \int p(\boldsymbol{y}|\boldsymbol{\theta})\pi^*(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}$$

thus, choosing $h(t)$ amounts to choosing $\pi^*(\boldsymbol{\theta})$.
- Empirical Bayes distribution: $\pi^{EB}_{prior}(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}|\boldsymbol{\eta} = \hat{\boldsymbol{\eta}})$
- Posterior predictive distribution: $\pi(\boldsymbol{\theta},\boldsymbol{\eta}|\boldsymbol{y}^{obs}) \propto p(\boldsymbol{y}^{obs}|\boldsymbol{\theta})\pi(\boldsymbol{\theta},\boldsymbol{\eta})$
- Partial posterior predictive distribution

$\pi_{ppp}(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}|\boldsymbol{y}^{obs}\backslash t^{obs}) \propto \frac{p(\boldsymbol{y}^{obs}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{p(t^{obs}|\boldsymbol{\theta})}$

▶ **Results**

Simulated Data vs. Assumed Model

	T_1	T_2	T_3	T_4
"explt" vs. "explt"	0.49/0.97	0.53/0.99	0.47/0.99	0.12/0.58
"explt" vs. "expnt"	0.48/0.99	0.52/1.00	0.45/0.97	0.12/0.63
"expQt" vs. "expnt"	0.52/1.00	0.54/1.00	0.54/0.99	0.24/0.70
"expSin" vs. "expnt"	0.51/1.00	0.59/0.95	0.49/0.99	0.73/0.77

Transformed Residual Checking

▶ **Transformed residuals** are defined as

$$e_i = \Phi^{-1}(F_i(Y_i))$$

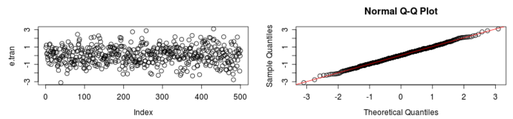
where the CDF $F_i(\cdot)$ for GLSM comes from

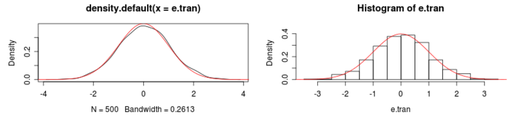
$$f(y_i|\boldsymbol{\eta}) \leftarrow f(\boldsymbol{y}|\boldsymbol{\eta}) = \int \prod_i p(y_i|\theta_i)\pi(\boldsymbol{\theta}|\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\theta}.$$

▶ **Property**

When
(1) the model is correctly specified and
(2) $\boldsymbol{\eta}$ is the true parameter,
transformed residuals will follow standard normal distribution.
Thus, by examining the normality of transformed residuals the information of goodness of fit can be obtained.

▶ **Comparing Distributions**
- Graphical methods: QQ-plot, histogram, relative density plot





- Numerical methods: we propose to use Hellinger distance.
  Hellinger distance is defined as

$$d_H(P,Q) = \frac{1}{2}\int |\sqrt{dP} - \sqrt{dQ}|^2$$

where dP and dQ are two PDFs to be compared.

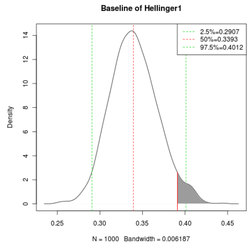
▶ **The Procedure of Calculation**
- calculate the Hellinger distance between empirical distribution of transformed residuals and standard normal.

$$(\boldsymbol{e}^{obs}, N(0,1)) \to d^{obs}$$

- obtain the *"baseline distribution"*, which is the distribution of Hellinger distances computed between theoretical standard normal distribution and empirical distributions of standard normal samples.

$$x_i^{(k)} \sim N(0,1) \quad i = 1,...,n$$
$$(\boldsymbol{x}^{(k)}, N(0,1)) \to d^{(k)} \quad k = 1,...,N$$

- calculate "one-side p-value", $p = P(d \geq d^{obs})$



Small p-value rejects the normality of transformed residuals which means the observed data is NOT compatible with the assumed model.

▶ **Results**

Mis-fit

	p-value	r.rate
"explt" vs. "expnt"	0.02	0.92
"expQt" vs. "expnt"	0.00	1.00
"expSin" vs. "expnt"	0.00	1.00

Correct-fit

	p-value	r.rate
"explt" vs. "explt"	0.26	0.01
"expQt" vs. "expQt"	0.28	0.01
"expSin" vs. "expSin"	0.16	0.08

Rongelap data

	"expnt"	"explt"
$\hat{\boldsymbol{\beta}}$	1.83	(1.87, -0.49, 0.31)
$\hat{\sigma}$	0.55	0.60
$\hat{\phi}$	0.02	0.03
p-value	0.027	0.000
r.rate	0.90	1.00

Weed data

	"expnt"	"explt"
$\hat{\boldsymbol{\beta}}$	4.07	(4.35, 0.10, -1.51)
$\hat{\sigma}$	1.09	1.22
$\hat{\phi}$	0.17	0.19
p-value	0.315	0.004
r.rate	0.005	1.00

▶ **Conclusions**
- Bayesian model checking can detect model failure only when the diagnostic statistic is well chosen which is usually very difficult without deep understanding of data.
- Transformed residual checking has much better performance. In the case the mean structure is misspecified, model failures are successfully detected.
- Transformed residual checking suggests GLSM without any trend in mean structure for "Weed" data and rejects GLSM with either linear trend or no trend for "Rongelap" data.

▶ **Improvement & Future Work**

Problem: however, when the normality of $\boldsymbol{e}$ is rejected, we cannot determine the rejection is due to condition (1) or (2), or both.

Goal: find a way to determine
- if condition (1) is satisfied?
- if yes, if condition (2) is satisfied?

References

[1] Diggle, P. J., Tawn, J. A. and Moyeed, R. A. (1998). Model-based geostatistics. *J. Roy. Statist. Soc. Ser. C*, 47, 299-350.

[2] Christensen, O. F., Roberts, G. O. and Sköld, M. (2006). Robust Markov chain Monte Carlo methods for spatial generalized linear mixed models. *J. of Computational and Graphical Statistics*, Vol. 15, No. 1, 1-17.

[3] Bayarri, M. J. and Castellanos, M. E. (2007). Bayesian checking of the second levels of hierarchical models. *Statistical Science*, Vol.22 No. 3. 322-343.