Journal of Advanced Scientific Research

Available online through http://www.sciensage.info/jasr

ISSN
0976-9595
Research Article

CONVERGENCE DIAGNOSTICS IN MCMC FOR ZERO-INFLATED POISSON MODELS FOR AIR POLLUTION DATA

Haydar Koç*, Mehmet Ali Cengiz , Tuba Koç
Ondokuz Mayıs University Faculty of Science Department of Statistics Samsun, Turkey
**Corresponding author: haydarkoc@omu.edu.tr*

ABSTRACT

Convergence diagnostics help to decide whether the Markov chain has reached its stationary and to determine the number of iterations to keep after the Markov chain has reached stationary. There are no conclusive tests that can tell you when the Markov chain has converged to its stationary distribution. In this study, we examine some convergence diagnostics for zero inflated Poisson models for air pollution data.

Keywords: MCMC, Convergence diagnostics, Zero inflated model, air pollution

1. INTRODUCTION

Poisson regression model is the most common model that used for the analysis of the count data. One of the assumptions of the Poisson distribution is that mean and variance are equal. However, if the data contains an excess number of zeros, Poisson regression models will exhibit over dispersion. Using zero-inflated Poisson model that allow for excess zero would be more appropriate if the over dispersion is due to an excess number of zeros. Bayesian analysis is a field of statistics that based on the notion of conditional probability. In the literature several authors have recently proposed Bayesian alternatives to fitting zero inflated models. The Markov Chain Monte Carlo (MCMC) method is the most common method for the Bayesian analysis. The basic idea of MCMC is to generate samples from the posterior distribution and uses these samples to approximate expectations of quantities of interest. For the process there are usually two issues. First, to decide whether the Markov chain has reached its stationary, or the desired posterior distribution and second, to determine the number of iterations to keep after the Markov chain has reached stationarity. Convergence diagnostics help to resolve these issues. Several statistical diagnostic tests like Raftery and Lewis, Geweke test and Heidelberger and Welch test can help to assess Markov chain convergence [1-4]. In this study, we examine some convergence diagnostics for zero inflated models for air pollution data.

2. MATERIAL AND METHODS

2.1. Zero Inflated Poisson Model

Zero-inflated poisson (ZIP) model, well described by Lambert is a simple mixture model for count data with excess zeros [5]. Specifically if *Yi* are independent random variables having a zero-inflated Poisson distribution, the zeros are assumed to arise in two ways corresponding to distinct underlying states. The first state occurs with probability w_i and produces only zeros, while the other state occurs with probability (1-w_i) and leads to a standard Poisson count with mean λ and hence a chance of further zeros. This two-state process gives a simple two-component mixture distribution with p.m.f

$$\Pr(y_i|x_i,\mu_i,w_i) = \begin{cases} w_i + (1-w_i)e^{-\mu_i}, & y_i = 0 \\ (1-w_i)\dfrac{e^{-\mu_i}\,\mu_i^{y_i}}{y_i!}, & y_i = 1,2,3,\ldots \end{cases} \qquad 0 \le w_i \le 1$$

The mean and variance of Y_i are

$$E(y_i) = \mu_i(1-w_i)$$
$$Var(y_i) = (1-w_i)(\mu_i + w_i{\mu_i}^2)$$

indicating that the marginal distribution of *Yi* exhibits over-dispersion, if w_i> 0. It is clear that this reduces to the standard Poisson model when w_i= 0. For a ZIP model the log-likelihoodfunction is given by

$$\mathcal{L}_{ZIP}(\mu,w;y) = \sum_{i=1}^{n} \begin{Bmatrix} I(y_i = 0)\ln[w_i + (1-w_i)e^{-\mu_i}] + \\ I(y_i > 0)[ln(1-w_i) - \mu_i + \text{y}\ln\mu_i - \ln(y!)] \end{Bmatrix}$$

2.2. Bayesian Analysis

Bayesian analysis is a field of statistics that based on the notion of conditional probability. In general, Bayesian statistical methods start with a "prior" distribution for all unknown parameters, updates this prior distribution in the light of the data (i.e., using likelihood) to construct the "posterior" distribution, and then uses the "posterior" distribution for inferential decisions. The posterior density or distribution given by

$$f(\theta|y) \propto f(y|\theta) \times f(\theta)$$

Where, $f(\theta)$ is prior distribution and $f(y|\theta)$ is the log-likelihood function.

In this study we use the uniform distribution as prior distribution for the parameter of the ZIP model. In the literature several authors have recently proposed Bayesian alternatives to fitting zero inflated models. The Markov Chain Monte Carlo (MCMC) method is the most common method for the Bayesian analysis. The basic idea of MCMC is to generate samples from the posterior distribution and uses these samples to approximate expectations of quantities of interest. For the process there are usually two issues. First, to decide whether the Markov chain has reached its stationary, or the desired posterior distribution and second, to determine the number of iterations to keep after the Markov chain has reached stationarity. Convergence diagnostics help to resolve these issues.

2.3. Markov Chain Convergence Diagnostics

2.3.1. Geweke Diagnostic

The Geweke test compares the sample mean in the early segment of the Markov chain to the mean in the latter segment of the chain in order to detect failure of convergence [2]. This is a two-sided test, and large absolute *z*-scores indicate convergence problems. The statistic upon which this diagnostic is based has the general form

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\hat{S}_1(0)/n_1 + \hat{S}_2(0)/n_2}}$$

where the variance estimate $\hat{S}(0)$ is calculated as the spectral density at frequency zero to account for serial correlation in the sampler output.

2.3.2. Heidelberger and Welch Diagnostic

The stationarity test is one-sided; rejection occurs when the *p*-value is greater than 1 - alpha. To perform the half-width test, we need to select an alpha level and a predetermined accuracy value. If the calculated relative half width of the confidence interval is greater than the accuracy value, we conclude that there are not enough data to accurately estimate the mean with 1-alpha confidence under that specific accuracy value [3,4]. Given an MCMC chain the null hypothesis of convergence is based on Brownian bridge theory and uses the Cramer-von-Mises test statistic

$$\int_0^1 B_n(t)^2 dt$$

Where

$$B_n(t) = \frac{T_{\lfloor nt \rfloor} - \lfloor nt \rfloor \bar{x}}{\sqrt{nS(0)}} \qquad T_k = \begin{cases} 0 & k = 0 \\ \sum_{j=1}^{k} x_j & k \neq 0 \end{cases}$$

,

and $S(0)$ is the spectral density evaluated at frequency zero.

2.3.3. Raftery and Lewis Diagnostic

The methods of Raftery and Lewis are designed to estimate the number of MCMC samples needed when quantiles are the posterior summaries of interest [1]. Their diagnostic is applicable for the univariate analysis of a single parameter and chain.

i. **N**min is the minimum number of iterations required to estimate the quantile of interest with the prespecified accuracy under the assumption of independence (i.e., with zero autocorrelation).
ii. **N** is the total number of iterations that the chain must run.
iii. **M** is the number of burnin iterations.
iv. **I** is the dependence factor given by **I** = **N**/**N**min, which indicates the relative increase of the total sample due to autocorrelations. If I is equal to one, then the generated values are independent. On the other hand, values greater than 5 often indicate a problematic behavior; for details, see [6].

2.3.4. MCMC error

For each of the parameters the following inequality must be hold. Monte Carlo Standard Errors<5% of standard deviations

3. RESULTS AND DISCUSSION

A part of the data was used by Cengiz and Cengiz and Terzi [7, 8]. The data set obtained from Afyon Respiratory Disease Hospital and Afyon Environmental Department Air Pollution Unit between 1 October 2006 - 30 September 2010. The data examines the relations between the numbers of admissions with respiratory disease who applied to Afyon Respiratory Disease Hospital and the measures of air pollution (SO_2 (Sulfur dioxide) - PM10 (Particulate matter) values) at the city centre. After performing Bayesian ZIP analysis using uniform prior for all parameter in the model, the results for convergence diagnostics for all parameters with different number of iterations are obtained as per the following table:

Table1*. *Results for convergence diagnostics for each parameter for different iteration number

		Number of iteration							
		500	1000	2000	5000	10000	20000	30000	40000
Geweke Diagnostics	Intercept	+	-	+	+	+	+	+	+
	SO_2	-	+	+	+	+	+	+	+
	PM10	-	-	+	+	+	+	+	+
	Intercept (inflation parameter)	+	+	-	+	+	+	+	+
Raftery-Lewis Diagnostics	Intercept	-	-	-	-	-	+	+	+
	SO_2	-	-	-	-	-	-	-	+
	PM10	-	-	-	-	-	-	-	+
	Intercept (inflation parameter)	-	-	-	-	-	-	-	+
Heidelberger-Welch Diagnostics	Intercept	-	-	-	-	+	+	+	+
	SO_2	-	-	-	+	+	+	+	+
	PM10	+	+	+	+	+	+	+	+
	Intercept (inflation parameter)	-	-	-	-	+	+	+	+
MCSE/SD	Intercept	-	-	-	-	+	+	+	+
	SO_2	-	-	-	-	+	+	+	+
	PM10	-	-	-	-	+	+	+	+
	Intercept (inflation parameter)	-	-	-	-	-	+	+	+

"+ " shows that the Markov Chain reached convergence, "-" shows that the Markov Chain did not reached convergence. After 40000 iterations, the Markov chain for all parameters reached convergence for all diagnostics. The results of diagnostics for convergence are given in Table 2.

Table2. Results for convergence diagnostics for each parameter for 40000 iterations

	Geweke Diagnostics		Raftery-Lewis Diagnostics	Heidelberger-Welch Diagnostics		MCSE/SD
	z	Pr >|z|	Dependence Factor	Cramer-von-Mises Stat	p	
Intercept	-0,9863	0,324	1,0758	0,0506	0,8724	0,0201
SO_2	-0,3933	0,6941	1,1209	0,0432	0,916	0,0199
PM10	1,0696	0,2848	1,0980	0,0722	0,7381	0,0196
Intercept (inflation parameter)	-1,0416	0,2976	1,1209	0,1135	0,5226	0,0236

As shown in table 2 for 40000 iterations the diagnostic statistics all show that the Markov chain reached convergence. The Geweke statistics not significant, the Raftery-Lewis statistics show an adequate sample size, and Heidelberger-Welch diagnostics all passed The ratio of the Monte Carlo standard errors and the standard deviations is much smaller than 0,05.

We produce a number of graphs which also aid convergence diagnostic checks. As an example Figure 1 shows diagnostic plots for PM10. From the trace plots we can say that the mean of the Markov chain has stabilized and appears constant over the graphs. The plots show that the chains appear to have reached convergence. The posterior autocorrelations are quite small and the posterior density appears bell-shaped

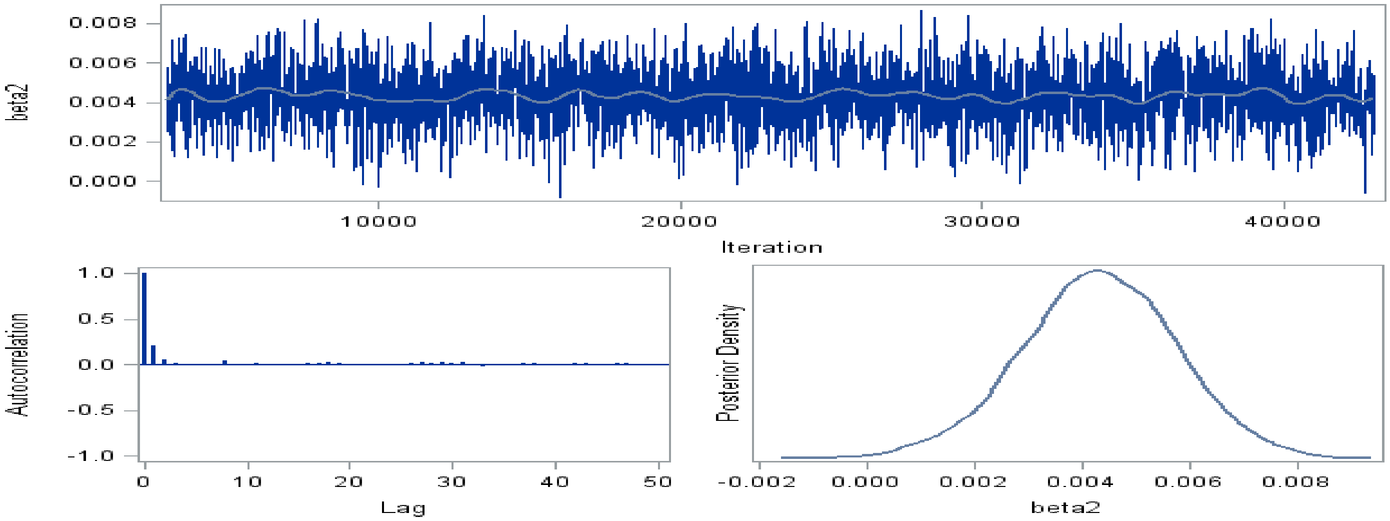

Fig.1. Diagnostic Plots for convergence check

4. CONCLUSION

Convergence diagnostics help to decide whether the Markov chain has reached its stationary and to determine the number of iterations to keep after the Markov chain has reached stationary. There are no conclusive tests that can tell you when the Markov chain has converged to its stationary distribution. In this study, we examine some convergence diagnostics for zero inflated model for air pollution data. We show that after 40000 iterations, the Markov chain for all parameters reached convergence for all diagnostics.

We suggest that you should proceed with caution. Meanwhile, note that you should check the convergence of all parameters. With some models, certain parameters can appear to have very good convergence behaviour, but that could be misleading due to the slow convergence of other parameters. If some of the parameters have bad mixing, you cannot get accurate posterior inference for parameters that appear to have good mixing.

5. REFERENCES

1. Raftery A. E. & Lewis S. M. The Number of Iterations, Convergence Diagnostics and Generic Metropolis Algorithms, in W. R. Gilks, D. J. Spiegelhalter, and S. Richardson, eds., Markov Chain Monte Carlo in Practice, London, UK: Chapman & Hall; 1996.
2. Geweke J. Evaluating the Accuracy of Sampling-Based Approaches to Calculating Posterior Moments,” in J. M. Bernardo, J. O. Berger, A. P. Dawiv, and A. F. M. Smith, eds., Bayesian Statistics, volume 4, Oxford, UK: Clarendon Press; 1992.
3. Heidelberger P. & Welch P. D. A Spectral Method for Confidence Interval Generation and Run Length Control in Simulations, Communication of the ACM, 1981; 24:233-245.
4. Heidelberger P, Welch PD. *Operations Research*, 1983; 31: 1109-1144.
5. Lambert D. *Technometrics*, 1992; 34:1.
6. Best N, Cowles M, and Vines K. CODA: Convergence Diagnostics and Output Analysis Software for Gibbs Sampling Output, Version 0.30, MRC Biostatistics Unit, Institute of Public Health, Cambridge, UK; 1996.
7. Cengiz MA. *Pol.Jou.Environ.Stud*, 2012; 21, Suppl3: 565-568.
8. Cengiz MA, Terzi Y. *Cent. Eur. Jou. Public Health*, 2012; 20, Suppl 4: 167-173.