Air Pollution Distribution Analysis for Beijing Haze Gibbs Sampling Algorithm Application

Xingchen Ling
Department of Economics
Duke University
Durham, NC 27701
xl122@duke.edu

Bo Wang
Department of Economics
Duke University
Durham, NC 27701
bw133@duke.edu

December 6, 2015

Abstract

The purpose of this project is to build a robust Lognormal model to estimate air pollution levels in Beijing, the capital city of China. Using the real air pollution data from U.S. Department of State Air Quality Monitoring Program, we applied Gibbs sampling to estimate the parameters of the Lognormal model. Considering the significant difference of air condition between weekdays and weekends, we also fitted two lognormal distributions for weekdays and weekends separately.

Introduction

The serious air pollution issue in Beijing has gained increasing attention during the last few years. Our project is focused on building a robust Lognormal model to estimate air pollution levels in Beijing. Lognormal distribution is often used to model positive right-skewed random variables. One example of such a variable is air pollution measurements, which are taken in micrograms per cubic meter. Pollution levels tend to be right skewed with a heavy right tail. We used weak prior for the two parameters: μ and σ^2, and used Gibbs sampling by putting in the real air pollution data from U.S. Department of State Air Quality Monitoring Program. By noticing there is significant difference between air conditions in weekdays and those in weekends (which makes sense since there are more people driving, factories working, etc. on weekdays), we fitted two separate lognormal distributions: one for weekdays, and one for weekends. From the trace plots of the two parameters, we obviously see that our sampler has converged and explored the space reasonably well, and we can also get a lot of sensible statistics from the model, such as the posterior probability that $\mu_1 > \mu_2$, and $\sigma_1 > \sigma_2$.

Methodology

We construct a model using a lognormal likelihood for data, with weak priors of your μ and σ^2 and the algorithm is as follows.

weak priors: $\mu \sim \mathcal{N}(0, 100)$, $\sigma^2 \sim$ InvGamma$(0.001, 0.001)$
Full conditional distributions:

(1) $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$ (in this case, $\mu_0 = 0, \sigma_0^2 = 100$)

x$|\mu, \sigma^2 \sim \ln\mathcal{N}(\mu, \sigma^2)$

$p(x) = \frac{1}{x\sigma\sqrt{2\pi}} exp[-\frac{(\ln x-\mu)^2}{2\sigma^2}]$

$$\begin{aligned} p(\mu|x_{1:n}, \sigma^2, \mu_0, \sigma_0^2) &\propto p(x_{1:n}|\mu, \sigma^2)p(\mu|\mu_0, \sigma_0^2) \\ &= \prod_{i=1}^n \frac{1}{x_i\sigma\sqrt{2\pi}} exp[-\frac{(\ln x_i-\mu)^2}{2\sigma^2}] \frac{1}{\sigma\sqrt{2\pi}} exp[-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}] \\ &\propto exp[-\frac{\sum_{i=1}^n (\ln x_i-\mu)^2}{2\sigma^2}] exp[-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}] \\ &= exp[-\frac{\sigma_0^2 \sum_{i=1}^n (\ln x_i-\mu)^2+\sigma^2(\mu-\mu_0)^2}{2\sigma^2\sigma_0^2}] \\ &\propto \mathcal{N}(\frac{\sigma_0^2 \sum lnx_i+\sigma^2\mu_0}{n\sigma^2+n\sigma_0^2}, \ \frac{2\sigma^2\sigma_0^2}{n\sigma^2+n\sigma_0^2}) \end{aligned}$$

(2) $\sigma^2 \sim$ InvGamma(α_0, β_0)

$p(\sigma^2) = \frac{\beta^2}{\Gamma(\alpha)} (\sigma^2)^{-\alpha-1} exp(-\frac{\beta}{\sigma^2})$

$$\begin{aligned} p(\sigma^2|x_{1:n}, \mu, \alpha_0, \beta_0) &\propto p(x_{1:n}|\mu, \sigma^2)p(\sigma^2|\alpha_0, \beta_0) \\ &= \frac{1}{\sigma^n} exp[-\frac{\sum \ln x_i^2-2\mu 2\ln x_i+n\mu^2}{2\sigma^2}] \sigma^{-2\alpha_0-2} exp(-\frac{\beta_0}{\sigma^2}) \\ &= (\sigma^2)^{-\alpha_0-\frac{n}{2}-1} exp[-\frac{\frac{1}{2}\sum_i (\ln x_i-\mu)^2+\beta_0}{\sigma^2}] \\ &\propto \text{InvGamma}(\alpha_0 + \tfrac{n}{2}, \ \tfrac{1}{2}\textstyle\sum_i (\ln x_i - \mu)^2 + \beta_0) \end{aligned}$$

in this case, $\alpha_0 = 0.001$, $\beta_0 = 0.001$

By noticing there is significant difference between air conditions in weekdays and those in weekends, we fitted two separate lognormal distributions: one for weekdays, and one for weekends and the algorithm is as follows.

$y_1|\mu_1, \sigma_1^2 \sim \ln\mathcal{N}(\mu_1, \sigma_1^2)$ $\quad y_2|\mu_2, \sigma_2^2 \sim \ln\mathcal{N}(\mu_2, \sigma_2^2)$

$\mu_1 \sim \mathcal{N}(m_1, V_1)$ $\quad \sigma_1^2 \sim$ InvGamma (a_1, b_1)

$\mu_2|\mu_1 \sim \mathcal{N}(m_2, V_2)\mathbb{1}(\mu_1 > \mu_2)$ $\quad \sigma_2^2 \sim$InvGamma(a_2, b_2)

Setting priors:

$a_1 = b_1 = a_2 = b_2 = 0.01$

$V_1 = V_2 = 100$

both m_1 and m_2 are sample means.

Calculating the full conditionals:

To Sample μ_1:

$$\begin{aligned} \text{p}(\mu_1|y_1,\sigma_1^2,\mu_2) &\propto p(y_1|\mu_1,\sigma_1^2)p(\mu_1)p(\mu_2|\mu_1) \\ &= exp[-\tfrac{1}{2V_1}(\mu_1-m_1)^2]exp[-\tfrac{1}{2\sigma_1^2}\textstyle\sum_i(\ln y_{1i}-\mu_1)^2]\mathbb{1}(\mu_1>\mu_2) \\ &\propto exp[-\tfrac{1}{2}(n_1/\sigma_1^2+1/V_1)\mu_1^2-2(\textstyle\sum_{i=1}^{n_1}\ln y_{1i}-\mu_1)^2]\mathbb{1}(\mu_1>\mu_2) \\ &\propto \mathcal{N}(\hat{m_1},\hat{V_1})\mathbb{1}(\mu_1>\mu_2) \end{aligned}$$

where $\hat{m_1}=\frac{\sum_i \ln y_{1i}/\sigma_1^2+m_1/V_1}{n_1/\sigma_1^2+1/V_1}$ $\quad \hat{V_1}=(n_1/\sigma_1^2+1/V_1)^{-1}$

To Sample μ_2:

$$\begin{aligned} \text{p}(\mu_2|y_2,\sigma_2^2,\mu_1) &\propto p(y_2|\mu_2,\sigma_2^2)p(\mu_2)p(\mu_2|\mu_1) \\ &\propto \mathcal{N}(\hat{m_2},\hat{V_2})\mathbb{1}(\mu_1>\mu_2) \end{aligned}$$

where $\hat{m_2}=\frac{\sum_i \ln y_{2i}/\sigma_2^2+m_2/V_2}{n_2/\sigma_2^2+1/V_2}$

For σ_1^2:

$$\begin{aligned} \text{p}(\sigma_1^{-2}|y_1,\mu_1) &\propto p(y_1|\mu_1,\sigma_1^2)p(\sigma_1^{-2}) \\ &= exp[-\tfrac{1}{2}\sigma_1^{-2}\textstyle\sum_i(\ln y_{1i}-\mu_1)^2](\sigma_1^{-2})^{a_1-1}exp(-b_1\sigma_1^{-2}) \\ &\propto (\sigma_1^{-2})^{a_1+\frac{n_1}{2}-1}exp\{-\sigma_1^{-2}[b_1+\tfrac{\sum_{i=1}^{n_1}(\ln y_{1i}-\mu_1)^2}{2}]\} \\ &\propto Gamma(\hat{a_1},\hat{b_1}) \end{aligned}$$

where $\hat{a_1}=a_1+\frac{n_1}{2}$ $\quad \hat{b_1}=b_1+\frac{\sum_i(\ln y_{1i}-\mu_1)^2}{2}$

Thus, $\sigma_1^2 \sim$ InvGamma $(\hat{a_1},\hat{b_1})$

Similarly, $\sigma_1^2 \sim$ InvGamma $(\hat{a_1},\hat{b_1})$ $\quad$ where $\hat{a_2}=a_2+\frac{n_2}{2}$ $\quad \hat{b_2}=b_2+\frac{\sum_i(\ln y_{2i}-\mu_2)^2}{2}$

Results

Figure 1 shows the Traceplots and Running Average for μ and σ^2. As is shown both in the Traceplots and Running Average graphs, our samplers have converged and explored the space reasonably well.

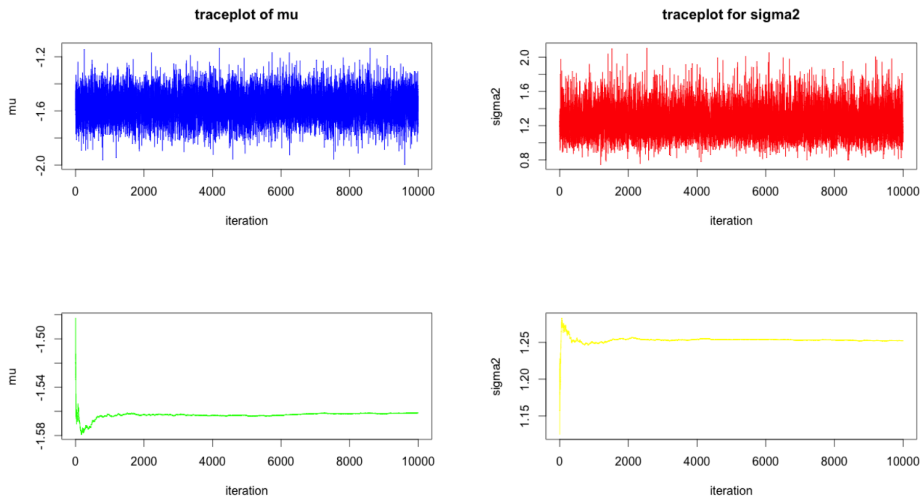

Figure 1: Traceplot and Running Average

By noticing there is significant difference between air conditions in weekdays and those in weekends, we fitted two separate lognormal distributions: one for weekdays, and one for weekends. Figure 2 and Figure 3 shows the Traceplots and Running Average for μ and σ^2 for weekdays and weekends separately.

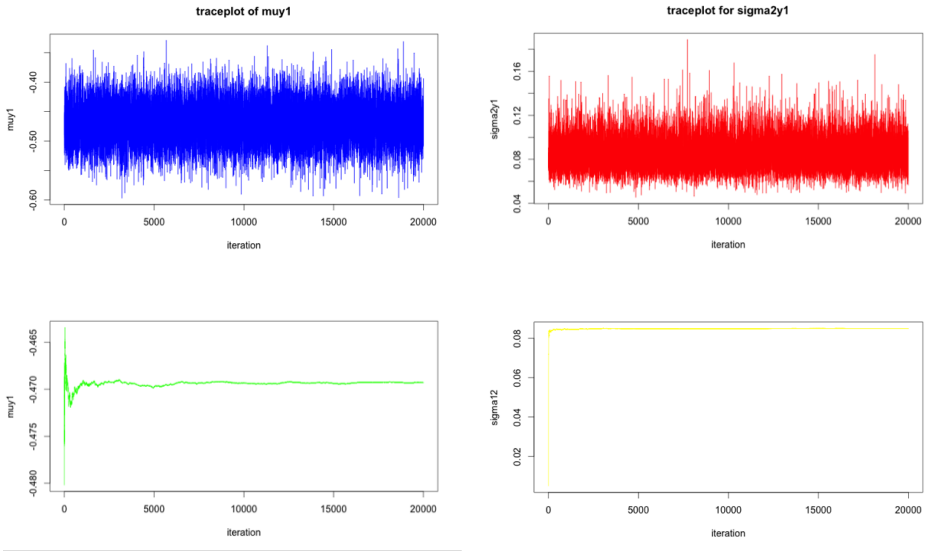

Figure 2: Traceplot and Running Average for Weekdays

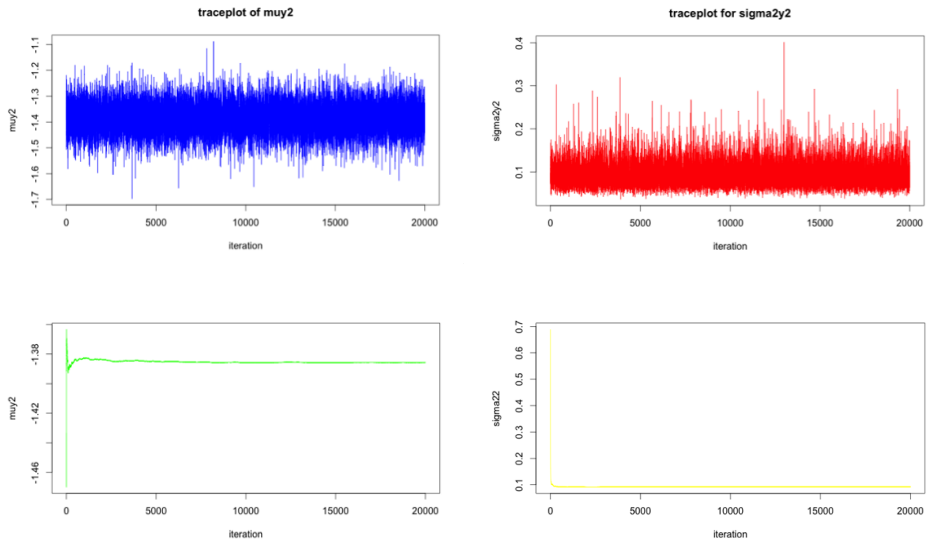

Figure 3: Traceplot and Running Average for Weekends

As is shown in the traceplots as well as in the running average plots, the sampler has converged and explored the space reasonably well.

The posterior point estimates and 95% confidence intervals for each of the four parameters is as follows:

$$\mu_1 : -0.469, [-0.536, -0.402]$$
$$\mu_2 : -1.385, [-1.499, -1.272]$$
$$\sigma_1 : 0.291, [0.247, 0.344]$$
$$\sigma_2 : 0.301, [0.231, 0.398]$$

The posterior probability that $\mu_1 > \mu_2$, and the posterior probability that $\sigma_1 > \sigma_2$ are 1, 0.439, which says there is a significant difference between the mean of air pollution in week days compare to that of weekends, but there is no much difference in the variance. (Mind here that μ and σ is not the mean or variance of distribution of air pollution).
The posterior probability that the pollution level on a randomly chosen future Tuesday is higher than the pollution level on a randomly chosen future Saturday 0.9824.

Conclusion

The key idea we developed here is that by using Bayesian inference statistics and Gibbs sampling, we are able to build a robust model for estimating the air pollution distribution in Beijing. As is shown in the trace plots and the running average plots, the

estimators converge well, and since we had weak priors for the parameters, we are confident that the posteriors we got can explain the real data on air pollutions that we have connected very well. One of the improvements is that instead of just fitting two models for week days and weekends, we can introduce time-series analysis here and make the estimation more accurate.

References

[1] Mukherjee, Sayan. *Probabilistic Machine Learning*. Department of Statistical Science, Duke University, 19 Nov. 2015. Web. 6 Dec. 2015. <https://stat.duke.edu/ sayan/561/2015/stat_ml.pdf>

[2] Hoff, Peter D. *A first course in Bayesian statistical methods*. Springer Science Business Media, 2009.

[3] Limpert, Eckhard, Werner A. Stahel, and Markus Abbt. "Log-normal Distributions across the Sciences: Keys and Clues On the charms of statistics, and how mechanical models resembling gambling machines offer a link to a handy way to characterize log-normal distributions, which can provide deeper insight into variability and probability?normal or log-normal: That is the question." *BioScience* 51.5 (2001): 341-352.

[4] Congdon, Peter. *Bayesian statistical modelling*. Vol. 704. John Wiley Sons, 2007.

[5] Richard O. Gilbert. *Statistical methods for environmental pollution monitoring*. John Wiley Sons, 1987.

[6] Bouguila, Nizar, Djemel Ziou, and Ernest Monga. "Practical Bayesian estimation of a finite beta mixture through Gibbs sampling and its applications." *Statistics and Computing* 16.2 (2006): 215-225.

[7] Kan, H. D., and Bing-Heng Chen. "Statistical distributions of ambient air pollutants in Shanghai, China." *Biomedical and Environmental Sciences*" 17.3 (2004): 366-372.

[8] Upadhyay, S. K., and M. Peshwani. "Posterior analysis of lognormal regression models using the Gibbs sampler." *Statistical Papers* 49.1 (2008): 59-85.

[9] Martn, J., and C. J. Prez. "Bayesian analysis of a generalized lognormal distribution." *Computational Statistics Data Analysis* 53.4 (2009): 1377-1387.

[10] Blackwood, Larry G. "The lognormal distribution, environmental data, and radiological monitoring." *Environmental monitoring and assessment* 21.3 (1992): 193-210.

[11] Sedek, Jannatul Naemah Mohamed, Nor Azam Ramli, and Ahmad Shukri Yahaya. "Air quality predictions using log normal distribution functions of particulate matter in Kuala Lumpur." *Malaysian Journal of Environmental Management* 7 (2006): 33-41.

[12] Hammitt, James K., and Ying Zhou. "The economic value of air-pollution-related health risks in China: a contingent valuation study." *Environmental and Resource Economics* 33.3 (2006): 399-423.