Demonstration Of Spatial Dirichlet Process And Its Brief Application To Gyeonggi-do Fine Dust Data

Gyenghun Kang

06 Dec 2021

Contents

1 Introduction

The purpose of this project is to provide a contemporary prediction of air pollution level (pm10) in areas of Gyeonggi-do where the data was not collected. There are currently 136 measurement sites spread over Gyeonggi-do, including Seoul, where hourly observations of various metrics related to air quality are made. Based on these flow of observations, it is of practical importance to estimate and predict air quality of area contiguous to those sites, and extend broader to the entire Gyeonggi-do region. Note that the project does not aim to provide a temporal prediction of air quality level in yet to be observed future. The main objective is to explore and implement a robust statistical model for a spatial prediction (kriging).

After a long dreadful tormentous labor, I had managed to implement the code in the paper, tested the algorithm on a simulated dataset, and the result was as expected. However, when applied to the real dataset, the result was far from ideal. The model structure below does not specify the site-specific mean process, i.e. $\mu_t(s_i)$ for each site s_i. What this means is that unless there are enough covariates to explain the variance of the response, the latent process and the measurement error would have to explain the bulk of the varainces which might cause the corresponding parameters to be unusally large. This was not the problem for a simulated dataset where all the covariates are known. An extension of the model structure below to include the site-specific mean process would be necessary for real data application.

Having said this, I would mainly discuss in this report a background of SDP and its demonstration in a simulated data, along with a brief application to the real data.

2 Spatial Dirichlet Process

2.1 Model Structure

The primary methodology of this project is a Spatial Dirichlet Process, first introduced in [Gelfand et al., 2005] and extended later in [Duan et al., 2007]. One of the most widely used assumptions in modelling a latent spatial process is a Gaussian Process, which is a parametric model with a homogeneous variance and stationary covariance function. Spatial Dirichlet Process propose a nonparametric spatial process that is neither stationary nor homogeneous. In this model, a latent spatial process follows a Dirichlet process with a Gaussian base measure.

Precisely, let $\{(Y_t(s_j), X_t(s_j))\}_{1:T}$ be replications of the observations and the covariates over $t = 1, ..., T$ at locations $s^{(n)} = (s_1, ..., s_n)$. For the moment we assume these locations are identical for all t, but this can be loosened([Duan et al., 2007]). Denote a vector and matrix of data on all sites as $\mathbf{Y}_t, \mathbf{X}_t$. Then the model structure is as follows ([Gelfand et al., 2005])

$$\begin{aligned}
\mathbf{Y}_t \mid \theta_t, \beta, \tau^2 &\sim N_n\big(\mathbf{Y}_t \mid \mathbf{X}_t\beta + \theta_t, \tau^2 I_n\big) \\
\theta_t \mid G^{(n)} &\sim G^{(n)} \\
G^{(n)} \mid \nu, \sigma^2, \phi &\sim DP(\nu G_0^{(n)}) \\
G_0^{(n)}(\cdot \mid \sigma^2, \phi) &= N_n\big(\cdot \mid 0_n, \sigma^2 \mathbf{H}_n(\phi)\big) \\
\nu &\sim \Gamma(\nu \mid a_\nu, b_\nu) \\
\beta, \tau^2 &\sim N_p\big(\beta \mid \beta_0, \Sigma_\beta\big) \times \Gamma^{-1}(\tau^2 \mid a_\tau, b_\tau) \\
\sigma^2, \phi &\sim \Gamma^{-1}(\sigma^2 \mid a_\sigma, b_\sigma) \times unif(0, b_\phi]
\end{aligned}$$

Since the base measure $G_0^{(n)}$ is conjugate to the likelihood, we can obtain posterior samples of the parameters by marginalizing out the random mixing measure $G^{(n)}$. Gibbs sampler algorithm in the paper is essentially a marginal Gibbs Sampler for sampling DP as introduced in BDA, and is available in the appendix of [Gelfand et al., 2005]. The code is available in the file "MCMC_learn.R".

2.2 Bayesian Nonparametric Spatial Prediction

Let $\tilde{s}^{(m)} = (\tilde{s}_1, ..., \tilde{s}_m)$ be a set of locations to be predicted, and denote $\mathbf{Y}_0, \tilde{\mathbf{Y}}_0, \mathbf{X}_0, \tilde{\mathbf{X}}_0$ for the response and covariates corresponding to the observed and the new location. Our purpose is to predict another replication of the Spatial Process, not to interpolate or extrapolate over $t = 1, ..., T$ time period. Hence comes the subscript $_0$. In this sense, the temporal order of t is for our purpose of no importance, and each replication is considered to be conditionally independent given the model parameters.

At each $b = 1, ..., B$ iteration of MCMC sampling, having obtained posterior samples of the parameters in (2.1), we can draw random samples of the latent variable $(\theta_0, \tilde{\theta}_0)$ and hence $(\mathbf{Y}_0, \tilde{\mathbf{Y}}_0)$. We can average this prediction across B iterations to obtain a posterior mean estimate of the posterior predictive distribution.

In summary, given the posterior samples of $(\theta_t^{(b)}, W^{(b)}, \nu^{(b)}, \sigma^{(b)}, \tau^{(b)})$ for $b = 1, ..., B$ iterations where W is a vector of cluster assignment,

1. Get $\theta_j^{*(b)}$ for $j = 1, ..., T^*$ where θ_j^* is a parameter associated with each of $T^*(\leq T)$ number of clusters.

2. Sample $\tilde{\theta}_j^{*(b)} \mid \theta_j^{*(b)}$ for a new set of locations $\tilde{s}^{(m)} = (\tilde{s}_1, ..., \tilde{s}_m)$ via conditional normal distribution.

3. Given $(\theta_j^{*(b)}, \tilde{\theta}_j^{*(b)})$ for observed and new locations each, sample $(\theta_0^{(b)}, \tilde{\theta}_0^{(b)})$ by polya-urn sampling, i.e. sample $(\theta_j^{*(b)}, \tilde{\theta}_j^{*(b)})$ w.p. $\frac{T_j}{\nu^{(B)}+T}$ or sample from new cluster $G_0^{(n+m)}((\theta^{*(b)}_{T^{*(b)}+1}, \tilde{\theta}^{*(b)}_{T^{*(b)}+1}) \mid \sigma^{(b)}, \phi^{(b)})$ w.p. $\frac{\nu^{(B)}}{\nu^{(B)}+T}$.

4. Given $(\theta_0^{(b)}, \tilde{\theta}_0^{(b)})$, sample $(\mathbf{Y}_0, \tilde{\mathbf{Y}}_0)$.

Hence we have a total of B number of predicted datasets, and we would use a posterior mean over B for krigging. Note that the model clusters within T, NOT $s^{(n)} = (s_1, ..., s_n)$! Moreover, this sampling scheme enables **automatic imputation for missing values**. Imputation for specific sites in specific times are done in an exactly the same manner as above. Below we demonstrate this with a simulated dataset.

3 Simulated Dataset

As in the paper, I have generated a nonstationary Gaussian process with a covariance function

$$\mathrm{cov}(Y(s_i), Y(s_j)) = \tilde{\sigma}(s_i)\tilde{\sigma}(s_j)\exp(-\phi\|s_i - s_j\|)$$

which is not a function of $\|s_i - s_j\|$ only. Additionally, I set $\beta = [1, 0.1, -0.1]^T, \sigma = 0.05, \phi = 1, \tau = 0.5$ for covariate of intercept, longitude(x-axis) and latitude(y-axis). Below is the traceplot along with predicted $\tilde{\mathbf{Y}}_0$) drawn against the data from the true distribution. Data was generated from measurement sites in Gyeonggi-do and Seoul, with 60 number of replications.

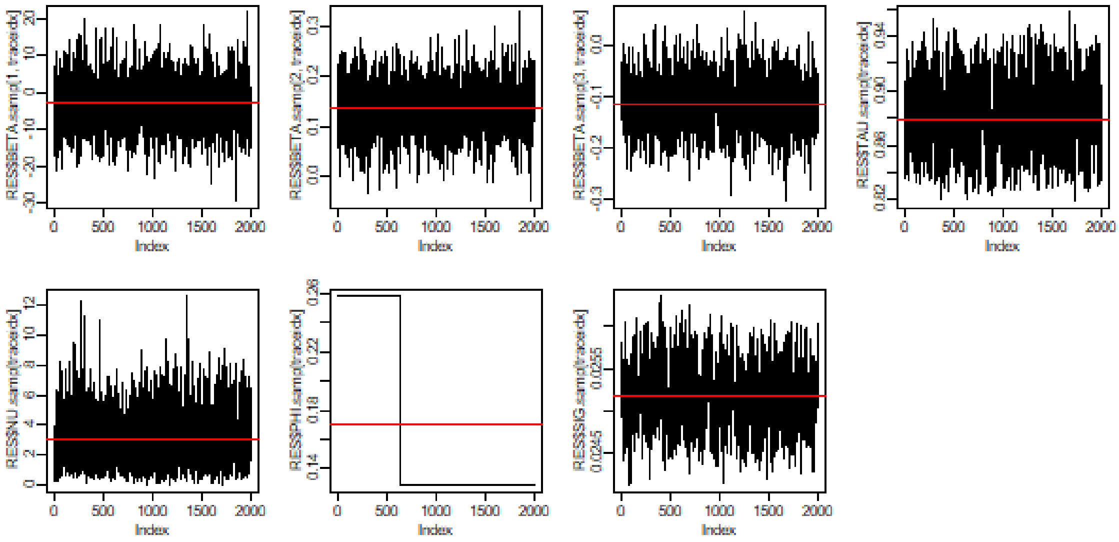

Figure 3.1: Traceplots

Mixing has to be pathetic for ϕ as it is sampled from a discrete grid of phi with probability proportional to

$$p(\phi)|H_n(\phi)|^{-1}\exp\Big(-\sum_{j=1}^{T^*}\theta_j^{*\prime}H_n^{-1}(\phi)\theta_j^*/(2\sigma^2)\Big)$$

which requires costly computation of matrix inversion and determinant calculation for, say, L number of grid values of ϕ. For the sake of time, I had to limit my choice of b_phi, a range of the grid and the size of grid length. A finer and broader grid would result in better mixing. β appears to have converged around the true value except for the intercept, which is possibly due to a difference in covariance structure. For other latent process parameters, we cannot assess its convergence because the covariane structure of the data generating process is not SDP. σ, τ, ϕ in MCMC samples bear no relation at all to the true values in the data generating process.

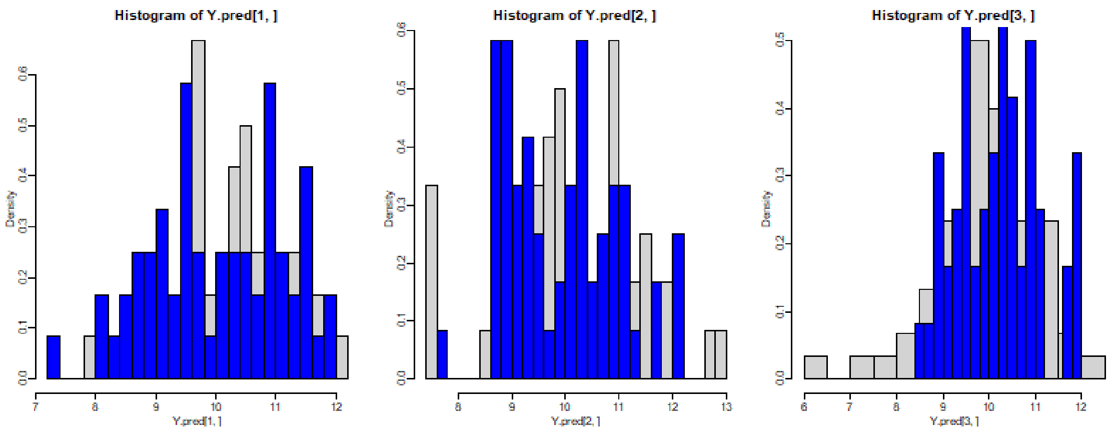

Figure 3.2: Predictions

For unobserved sites, predictions averaged over B iterations are marked in blue and compared to the unobserved data in gray from the true data generating process, for a total of $t = 1, ..., T$ replications. Since all the fixed effects were taken account of, the prediction is quite accurate. This is not the case for the real data unfortunately.

4 Real Data: Winter Gyeonggi-do Fine Dust Measurements

I have prepared a daily average of pm10 measurements over 134 measurement sites spread across Seoul and Gyeoggi-do, with a total of 90 replications collected over 2020 Nov, Dec and Jan. Temporal order was not considered here. These months are chosen because these are in the same season. Again, with covariates fed in as intercept, longitude(x-axis) and latitude(y-axis), the results are follows.

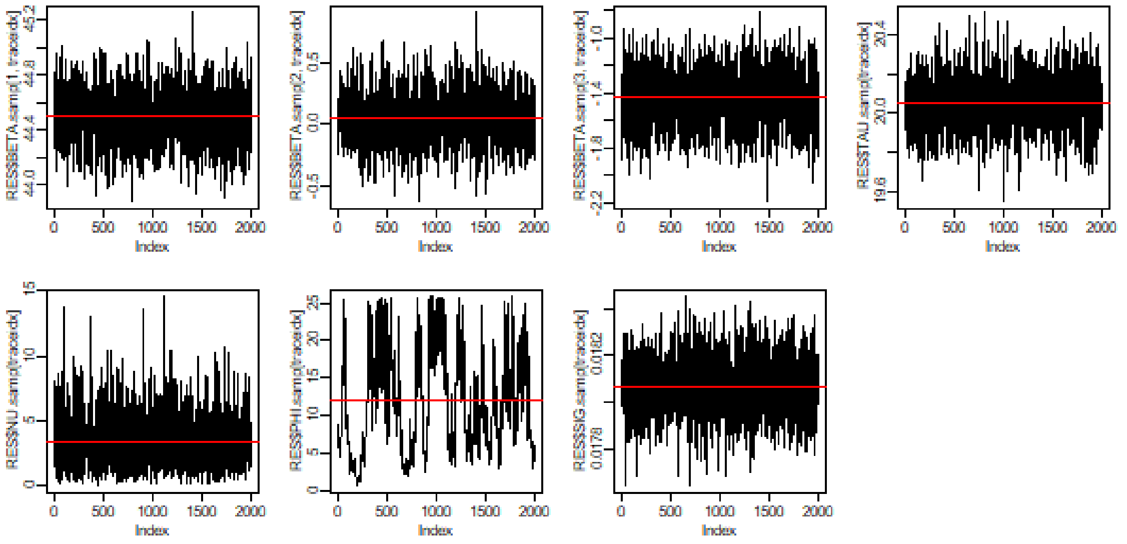

Figure 4.1: Traceplots

With 3000 iterations and 1000 burn-in, mixing appears to be acceptable except for *phi*. Since *phi* is sampled from a grid of values, there is necessarily an upper bound of sampled *phi*. It is interesting to note that a fixed effect for latitude is negative, meaning a lower pollution level further up north, but this is to be taken with a grain of salt given the dearth of covariates and site-specific process, as explained in the Introduction.

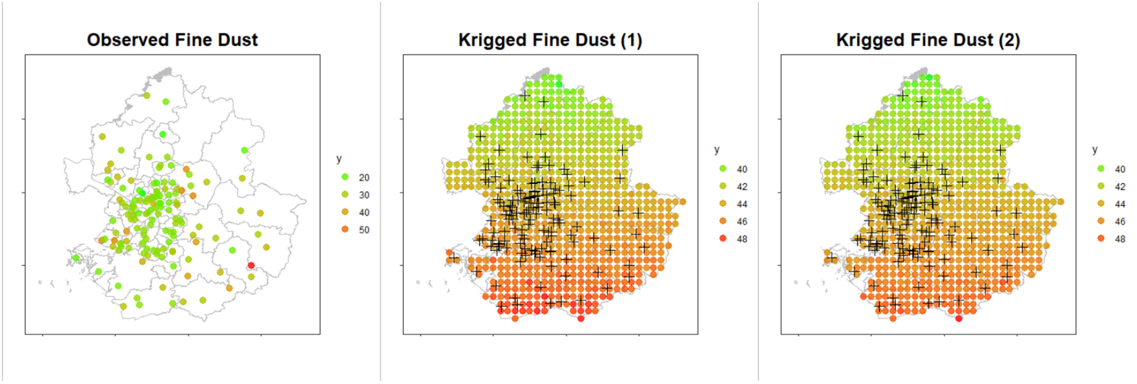

Figure 4.2: Krigged

Spatial predictions at grid locations were averaged over B number of iterations, and randomly selected t of $(\tilde{Y}_t(s))$ were shown above, among a total of $t = 1, ..., 90$ replications, with missing values imputed in MCMC process. For the lack of site-specific mean process, an absolute level of predictions are quite a bit off from the data.

5 Rooms For Improvement

- Expansion of SDP model with site-specific mean process
- Collecting more covariates for observed and to be predicted locations
- Denser and wider support for ϕ for better mixing, or perhaps simply substituting Empirical Bayes Estimator

The codes and data for this project can be accessed at `https://github.com/hun-learning94/SDP_Gelfand2005`.

References

[Duan et al., 2007] Duan, J. A., Guindani, M., and Gelfand, A. E. (2007). Generalized spatial dirichlet process models. *Biometrika*, 94(4):809–825.

[Gelfand et al., 2005] Gelfand, A. E., Kottas, A., and MacEachern, S. N. (2005). Bayesian nonparametric spatial modeling with dirichlet process mixing. *Journal of the American Statistical Association*, 100(471):1021–1035.