Spatio-Temporal Analysis of Particulate Matter based on the Quantile Factor Model

Minji Kim Joonpyo Kim Hee-Seok Oh

Department of Statistics, Seoul National University, Seoul, Korea

SCAN ME

Introduction

- The ultimate goal of this study is to understand the behavior of large-scale fine particulate matter ($PM_{2.5}$) data. We aim to find the hidden factor structure of large spatio-temporal data based on a quantile factor model (QFM) that admits cross-sectional and serial dependence and heteroscedasticity ([2]).
- A novel quantile factor model is carried out for spatio-temporal analysis of $PM_{2.5}$ values to reduce the dimension of the data while summarizing the relations among multivariate time series data. Also, we further expand it to some extremal levels that capture the tail variables of the data and estimate the entire conditional distribution of $PM_{2.5}$ values

Quantile Factor Model

- We consider a model,

$$X = F_\tau \Lambda_\tau^T + \nu_\tau \text{ for } \tau \in (0,1),$$

  where $F_\tau = [f_1(\tau), \ldots, f_T(\tau)]^T$ is a $T \times r$ matrix of factors and $\Lambda_\tau = [\lambda_1(\tau), \ldots, \lambda_N(\tau)]^T$ is an $N \times r$ matrix of factor loadings for r factors.
- The idiosyncratic error $\nu_{it}(\tau) = (\nu_\tau)_{it}$ is assumed to satisfy $P[\nu_{it}(\tau) \le 0 | f_t(\tau)] = \tau$, implying the conditional quantile function of $X_{it} = (X)_{it}$ is

$$Q_{X_{it}}[\tau | f_t(\tau)] = \lambda_i(\tau)^T f_t(\tau).$$

- The parameters in the model are estimated by simultaneously minimizing the objective function,

$$S_{NT}(F, \Lambda) = \frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} \rho_\tau (X_{it} - \lambda_i^T f_t)$$

  under the condition $\frac{1}{T} F_\tau^T F_\tau = I_{\hat{r}}$ and $\Lambda_\tau^T \Lambda_\tau$ is a diagonal matrix with non-increasing elements.

Extremal Quantile

- Extreme modeling is of high interest in analyzing climate data.
- Let F_{it} be the distribution of X_{it} which satisfies that for random samples $\{Z_m\}_{m=1}^n$ from F_{it}, there exists sequences $\{a_n\}$ and $\{b_n\}$ such that

$$\mathrm{P}(a_n^{-1}(\max_{1 \le m \le n} Z_m - b_n) \le z) \to \exp(-(1+\gamma z)^{-1/\gamma})$$

  as n grows to infinity for some γ and z such that $1 + \gamma z \ge 0$.
- Then, the extremal level quantile τ_{ex} of a random variable Y can be estimated as

$$\hat{Q}_{\tau_{ex}}(Y) = \left(\frac{1-\tau_0}{1-\tau_{ex}} \right)^\gamma \hat{Q}_{\tau_0}(Y)$$

  using intermediate quantile estimates $\hat{Q}_{\tau_0}(Y)$ from the QFM result ([5]).
- For estimation of the tail index γ_i at the ith station, we use Hill estimator with the ordered statistics of $\{X_{it}\}_{t=1}^T$, $\{X_{i(j)}\}_{j=1}^T$.
- Thus, γ_i is estimated as

$$\hat{\gamma}_i = \frac{1}{k} \sum_{j=1}^{k} \log X_{i(j)} - \log X_{i(k)}$$

  for a suitable k. Motivated from [5], we consider $k = [cT^{1/3}]$ with various values of c, and choose the best one using five-fold cross-validation.
- The estimated number of common factors set to 2 for $\tau \in \mathcal{T}_l$, and the validation procedure choose $c = 2$ for estimating the tail index by the Hill estimator.

The Analysis Procedure

1. Impute missing values
2. For each $\tau \in \mathcal{T} = \{0.05l; 1 \le l \le 19\}$, initialize $\hat{F}_\tau^{(0)}$ using the principal component estimates subject to the normalization $\frac{1}{T} F_\tau^T F_\tau = I_r$.
3. Iterate (a) and (b) until $S_{NT}(F_\tau, \Lambda_\tau)$ converges for each $\tau \in \mathcal{T}$.
   (a) Given $\hat{F}_\tau^{(m)}$, perform quantile regression of $\{X_{it}\}_{t=1}^T$ on $\hat{F}_\tau^{(m)}$ to estimate $\hat{\lambda}_i^{(m+1)}$ for $i = 1, 2, \ldots, N$.
   (b) Given $\hat{\Lambda}_\tau^{(m+1)}$, use quantile regression of $\{X_{it}\}_{i=1}^N$ on $\hat{\Lambda}_\tau^{(m+1)}$ to estimate $\hat{f}_t^{(m+1)}$ for $t = 1, 2, \ldots, T$.
4. Normalize the estimators.
5. Sort the quantiles for $\tau \in \mathcal{T}$ in an increasing order.
6. Estimate upper extreme quantiles $\hat{Q}_{it}(\tau)$ for $\tau \in \mathcal{T}_{ex}$ based on $\tau_0 = 0.95$.
7. Linearly interpolate $(\hat{Q}_{it}^*(\tau))_{\tau \in \mathcal{T} \cup \mathcal{T}_{ex}}$ to obtain the quantile process $\tau \mapsto \hat{Q}_{it}^*(\tau)$
8. Estimate the distribution function as $\hat{\mathrm{F}}_{it} = (\hat{Q}_{it}^*)^{-1}$

Note that the initial estimate $\hat{F}^{(0)}$ in Step 1. is $\sqrt{T}$ times the r largest eigenvalues of $X^T X$ ([1]). Also, the normalization in Step 4. can be done with the rotation matrix $\sqrt{T}(F_\tau^T F_\tau)^{-\frac{1}{2}} P_\tau$, where P_τ is the matrix of eigenvectors of $(F_\tau^T F_\tau)^{\frac{1}{2}} (\Lambda_\tau^T \Lambda_\tau)(F_\tau^T F_\tau)^{\frac{1}{2}}$ ([3]).

Data Pre-processing

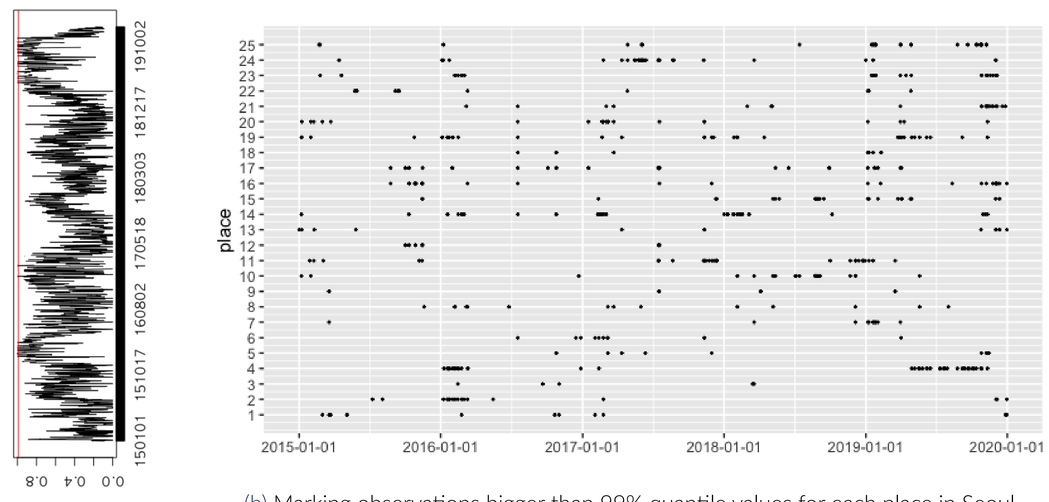
- In this study, we analyze the daily average of $PM_{2.5}$ data observed at $N = 103$ stations in Korea for $T = 1825$ times from 2015.01.01 to 2019.12.31.
- First decompose $Y(s,t)$, the observed $PM_{2.5}$ value at location $s \in \{1, 2, \ldots, N\}$ and time $t \in \{1, 2, \ldots, T\}$, into a mean effect $\mu(s,t)$ and the anomaly $X(s,t)$, i.e.

$$Y(s,t) = \mu(s,t) + X(s,t).$$

- Estimate $\mu(s,t)$ by computing the temperature average for each specific place and each time of the year, and then smooth the estimated mean by computing, for each grid cell separately, a moving average over windows of size 6. We focus on analyzing the anomalies $X(s,t)$ obtained as

$$X(s,t) = Y(s,t) - \hat{\mu}(s,t)$$

Results

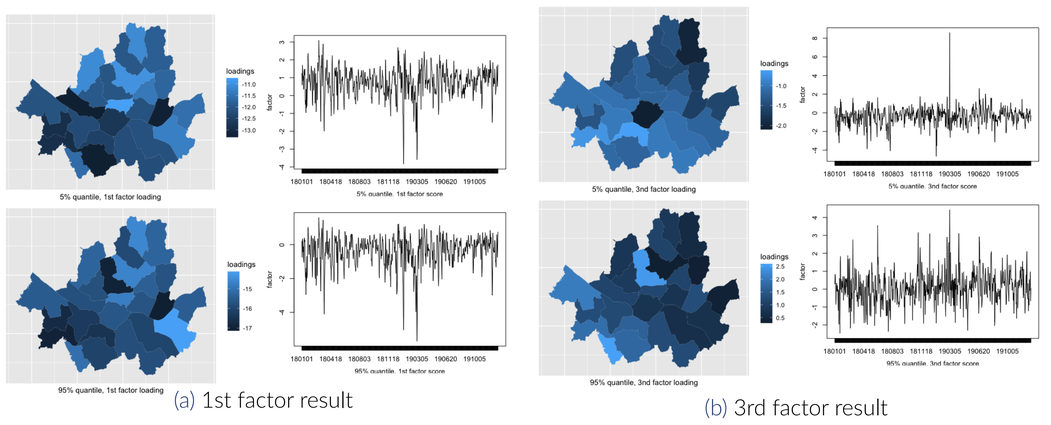


(a) Empirical CDF values at the station 4

(b) Marking observations bigger than 99% quantile values for each place in Seoul

Figure: Information about empirical CDF values bigger than 99%

- Figure 1 (a) displays the empirical CDF value of each observation at Gwangjin-gu district, i.e. probabilities of X_{it} being less than or equal to the observed anomalies based on the estimated distribution function $\hat{\mathrm{F}}_{it} = (\hat{Q}_{it})^{-1}$. The red line denotes a probability of 0.99. Figure 1 (b) plots such observations bigger than estimated 99% quantile values for each station in Seoul, where 25 stations are located as in Figure 3. Each row represents each station. As we can see in both (a) and (b), extreme $PM_{2.5}$ concentration are frequently observed at Gwangjin-gu, located at the right side of Seoul, in Mid-2019. As such, we can identify and compare periods showing excessive particulate matter anomalies differently distributed by place, providing information on the trends beyond the means.



(a) 1st factor result

(b) 3rd factor result

Figure: Factors and loadings distributed in Seoul: $\tau = 5\%$(up) and $\tau = 95\%$(down)

- Figure 2 shows the spatial distribution of two factor loadings at $\tau = 0.05, 0.95$ around Seoul. Each station represents each district. Factor loadings are coefficient values representing each region's degree of response as the factor score values change over time. Plots on the right side of the figure are temporal plots of factor scores at $\tau = 0.05, 0.95$. When factors show similar trends in time over several quantiles, the different distribution of loading values over those quantiles represents the difference in regional sensitivities, that is, each region's degree of response as the quantiles change. Moreover, we observe that the third factor shows different factor trends between the quantiles $\tau = 0.05$ and $\tau = 0.95$, implying there exist different time factors responsible for the extremal events.

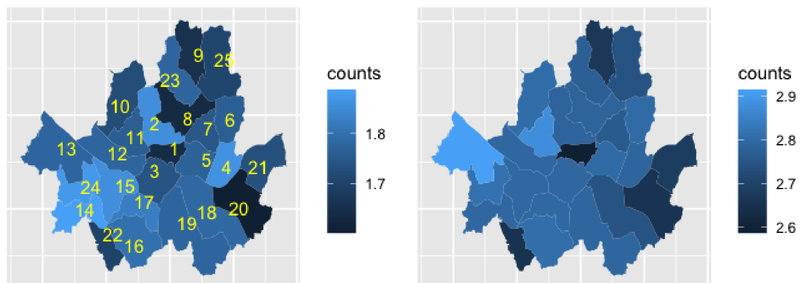


Figure: Log counts of extremes at $\tau = 0.95$ (left) and $\tau = 0.9995$ (right) in Seoul

- Upper extreme quantiles providing information needed for estimating the whole distribution, we further suggest new statistics to analyze the extremal behavior. For example, to evaluate whether the degree of occurrence of extreme conditions varies by region, Figure 3 shows the number of extreme events occurring in Seoul. Let $N_{ex}(i; u, \tau)$ be the count of the extremes,

$$N_{ex}(i; u, \tau) = \sum_t I(\hat{Q}_{it}(\tau) > u) = |T_{i;\tau,u}|$$

  , where $T_{i;\tau,u} = \{t : \hat{Q}_{it}(\tau) > u\}$ at station i. Figure 3 shows the values of $\log_{10}(1 + N_{ex})$ with $u = 40$ and $\tau = 0.95, 0.9995$. We observe that the spatial distributions of the values between the two extremal quantiles are different. For the station 8, for instance, observe $|T_{8;0.95,40}|$ is relatively small, while $|T_{8;0.9995,40}|$ relatively large. $T_{8;0.95,40} \subset T_{8;0.9995,40}$ implies there are relatively many elements in the set $T_{8;0.9995,40} - T_{8;0.95,40}$. As such, we can explore local characteristics of tail distributions by applying several u and τ values.

Overall, the proposed QFM-based analysis is useful in analyzing spatio-temporal data with dependency structures, offering complete estimates and valuable insights beyond the mean. The analysis could further be expanded to predict the missing outcomes or forecast future trends.

References

[1] J. Bai and S. Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70, 01 2001.
[2] G. Chamberlain and M. Rothschild. Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica*, 51(5):1281–1304, 1983.
[3] Liang Chen, Juan Dolado, and Jesus Gonzalo. Quantile factor models. *Working Paper*, 11 2019.
[4] Ivan Fernandez-Val, Victor Chernozhukov, Iv-n Fern-ndez-Val, and Alfred Galichon. Quantile and probability curves without crossing. *Econometrica*, 78:1093–1125, 01 2010.
[5] Huixia Wang, Deyuan Li, and Xuming He. Estimation of high conditional quantiles for heavy-tailed distributions. *Journal of The American Statistical Association*, 107, 12 2012.