Discussion on Competition for Spatial Statistics for Large Datasets

Yasumasa MATSUDA

The team of Tohoku University attended sub-competition 2b in the competition on spatial statistics for large datasets, where prediction on 100,000 testing points were to be constructed conditional on 900,000 training points. We chose a covariance tapering approach in a simplified way to manage one million spatial data points. Dividing $[0, 1]^2$ into 30×30 sub-regions with equal area, we construct predictors separately in each sub-region conditional on training data over the extended sub-region with length enlarged by $\sqrt{2}$ by fitting Matérn class covariances.

Key Words: Covariance tapering; Cross validation; Matern class.

1. METHOD

We would like to express our deepest respect and gratitude to the great trial of the competition for huge spatial dataset. We are sure the interesting methods tried in the competition will work as standard benchmarks helpful for future works in large spatial data analysis.

The team of Tohoku University is composed of the three spatial statisticians working mainly on econometrics applications in social science fields. Our choice of approach is a modified covariance tapering to manage one million spatial data points in a feasible way. Covariance tapering was originally proposed by Kaufman et al. (2008) and was used in a modified way by Zhang et al. (2013). We followed the way of Zhang et al. (2013) in a simplified way to improve prediction performances around edges. While Zhang et al. (2013) decomposed spatial variations into high and low frequency components, we focused on high frequency components only for better prediction performances.

Dividing the whole region $[0, 1]^2$ into $30 \times 30 = 900$ disjoint subregions with equal area, which we call lattice A, we extend each of the subregions by enlarging the length by $\sqrt{2}$, which we denote lattice B. Our basic idea to manage large data points is an use of training points in a lattice B for kriging of testing points in a lattice A by Matérn class covariance.

Y. Matsuda (✉) Graduate School of Economics and Management, Tohoku University, Sendai, Japan (E-mail: *yasumasa.matsuda.a4@tohoku.ac.jp*).

© 2021 International Biometric Society
Journal of Agricultural, Biological, and Environmental Statistics, Volume 26, Number 4, Pages 612–613
https://doi.org/10.1007/s13253-021-00463-1

More precisely, let x be training points in a lattice $a \in A$ and y be testing points in the corresponding lattice $b \in B$ including a. We shall conduct a kriging of unobserved $Z(y)$ by observed $Z(x)$. We evaluate $K = cov(Z(x), Z(x))$ and $L = cov(Z(y), Z(x))$ by the Matérn class in the equation (1) with $\tau = 0$ in the paper, where the parameters of smoothness and range (scale) are selected by CV as

$$\text{Case 1:}\nu = 1.5, h = 0.0133,$$
$$\text{Case 2:}\nu = 1.0, h = 0.0133.$$

Then the predictors are constructed by

$$\bar{Z} + LK^{-1}(Z(x) - \bar{Z}),$$

for $\bar{Z}$ given by the sample mean of $Z(x)$.

We do hope it will help as a benchmark for future developments of large spatial data analysis.

[Received June 2021. Revised June 2021. Accepted July 2021. Published Online July 2021.]

REFERENCES

Kaufman C, Schervish M, Nychka D (2008) Covariance tapering for likelihood-based estimation in large spatial data sets. J. Am. Statist. Assoc. 103:1545–1555

Zhang B, Sang H, Huang JZ (2013) Full-scale approximations of spatio-temporal covariance models for large datasets. Statist. Sin. 25:99–114

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.