UNDERSTANDING OF SPATIAL PROBLEM FOR LARGE DATASETS

YEWEN CHEN

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 GAUSSIAN RANDOM FIELD AND ITS CHALLENGE

Gaussian spatial processes has been popular for decades in spatial data contexts like geostatistics where they are known as kriging, and in computer experiments where they are deployed as surrogate models or emulators. More recently, they have become a popular prediction engine in the machine learning literature. The reasons are many, but the most important are probably that: the Gaussian structure affords a large degree of analytic capability not enjoyed by other general-purpose approaches to nonparametric nonlinear modeling; and because they perform well in out-of-sample tests.

Assume there is a response or dependent variable $Y(s)$ at a generic location $s \in \mathbf{D} \subset \mathcal{R}^2$ along with a $p \times 1$ vector of spatially referenced predictors $\mathbf{X}(s)$. A spatial regression model has the form

$$y(s) = \boldsymbol{\beta}\mathbf{X}(s) + w(s) + \epsilon(s) \tag{1}$$

where $\boldsymbol{\beta}$ is the vector of regression coefficients. The residual from the regression is decomposed into two independent parts: a spatial process, $w(s)$, modelling spatial association, and an independent process, $\epsilon(s)$, also known as the nugget effect, modelling measurement error.

The nugget effect $\epsilon(s)$ is often assumed to follow a normal distribution with variance τ^2 for every location s. The spatial process $w(s)$ in (1) is often referred to as spatial random effects, capturing the effect of unmeasured or unobserved covariates with spatial pattern.

The most common specification for $w(s)$ is $w(s) \sim GP(0, \mathbf{C}(\cdot,\cdot))$, a zero-mean Gaussian process with a valid covariance function $\mathbf{C}(s, s')$. It is often reasonable to assume a constant process variance and thus we specify $\mathbf{C}(s, s') = \sigma^2\rho(s, s'; \theta)$, where $\rho(s, s'; \theta)$ is a correlation function and θ is a vector of correlation parameters which needs to be estimated from a finite number of observations, $\mathbf{Y} = (y(s_1).\cdots, y(s_n))'$.

Therefore, $y(s)$ follows a spatial Gaussian process, and thus we have the log-likelihood function for $(\boldsymbol{\beta}, \tau^2, \sigma^2, \theta)$:

$$l(\boldsymbol{\beta}, \tau^2, \sigma^2, \theta) \propto -\frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{Y} - \boldsymbol{\beta}\mathbf{X})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \boldsymbol{\beta}\mathbf{X}) \tag{2}$$

where $\boldsymbol{\Sigma} = \mathbf{C} + \tau^2\mathbf{I}$. Then best linear unbiased prediction (BLUP) at an unobserved location s_0 can be obtained by the kriging equation:

$$\hat{y}(s_0) = \boldsymbol{\beta}\mathbf{X}(s_0) + \mathbf{c}'_{s_0}\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \boldsymbol{\beta}\mathbf{X}) \tag{3}$$

where $\mathbf{c}_{s_0} = (\mathbf{C}(s_0, s_1), \ldots, \mathbf{C}(s_0, s_n))'$.

However, with large or massive data, direct implementation of these statistical process, including parameter etimation by (2) and interpolation by (3), becomes computationally prohibitive, since evaluating the log-likelihood in (2) and solving the kriging equation (3) involve the Cholesky factorization of an $n \times n$ covariance matrix for data of size n, which requires $O(n^3)$ operations and $O(n^2)$ memory in general(see Porcu et al. [2012], Sang and Huang [2012], Heaton et al. [2019], Ma and Kang [2020]).

2 SEVERAL APPROACHES TO OVERCOME THIS LARGE MATRIX PROBLEM

Solutions to this computational intractability focus on model development, the design of efficient and parallel algorithms (see Katzfuss and Hammerling [2017] and Katzfuss [2017]), and the improvement and efficient use of modern computing platforms (e.g. Using TensorFlow and GPU).

Model:

1) Conditional distributions: Michael L. Stein (The restricted likelihood, 2004);
2) Sparse covariance by tapering method: Reinhard Furrer, 2006; Douglas Nychka, 2008; Michael L. Stein(2013);
3) Low-Rank methods: Noel Cressie (FRK, 2008), Sudipto Banerjee (Predictive process, 2008);
4) Sparse precision: Douglas Nychka(lattice kriging, 2015) by basis-function; Matthias Katfuss(multiresolution approximation, 2017) by basis-function; Finn Lindgren (GRMF Approximations, 2011) by SPDE; Abhi Datta (NNGP, 2016a, 2016b, 2016c, 2019) by conditional distributions;
5) Spectral method: Montserrat Fuentes and Joe Guinness(Circulant embedding, 2017 and Periodic Embeddings, 2007, 2019).
6) Discrete process convolutions.

Algorithms and platforms:

1) INLA: Haavard Rue.
2) Parallel algorithm: Matthias Katfuss.
3) TensorFlow: Andrew Zammit-Mangion.

2.1 Low rank methods

Comments: The reduced rank based methods usually fail to accurately capture the local, small scale dependence structure. The subsequent multiresolution (MR) methods (e.g. LatticeKrig, Multiresolution Approximations) compensates for this problem to some extent.

2.1.1 *Fixed Rank Kriging*

FRK (Cressie and Johannesson [2006], Cressie and Johannesson [2008]) aims to approximate the spatial process $w(s)$ in (1) by a linear combination of r ($<< n$) basis functions which ensures that all estimation and prediction equations only contain inverses of matrices of size $r \times r$.

2.1.2 *Gaussian Predictive Processes (GPP)*

With regard to the challenge of computational cost on covariance matrices, Banerjee et al. [2008] proposed a class of models based on the idea of a spatial predictive process which is motivated by kriging ideas or kriging equation. The predictive process projects the original process onto a subspace generated by realizations of the original process at a specified set of locations (or knots). The approach is in the same spirit as process modeling approaches using basis functions and kernel convolutions, that is, specifications which attempt to facilitate computations through lower dimensional process representations.

Comments: One advertised advantage of using the GPP approach as opposed to FRK or LatticeKrig is that the GPP basis functions are completely determined by the choice of covariance function $C(\cdot,\cdot)$.

At the same time, however, when $C(\cdot,\cdot)$ is governed by unknown parameters (which is nearly always the case) the GPP basis functions need to be calculated iteratively rather than once as in FRK or LatticeKrig which will subsequently increase computation time.

2.2 Sparse covariance methods

2.2.1 *Tapering*

Including tapering for estimation (e.g. approximation log-likelihood function by a tapered covariance. See Kaufman et al. [2008]) and tapering for Kriging (or for interpolation or for prediction, e.g. approximation kriging equation by replacing the original covariance by a tapered version. See Furrer et al. [2006]).

Comments:

1) The covariance tapering has shown great computational gains, but it also has its own drawbacks. The covariance tapering may not be effective in accounting for spatial dependence with long range.

2) The accuracy of the tapering approximation for nonstationary problems remains an open question.

3) The application of tapering techniques to multivariate random fields remains to be explored due to the lack of flexible compactly supported cross-covariance functions.

2.2.2 *Spatial Partitioning*

The approximation of the likelihood in either the spatial or spectral domain is another solution to overcome computational obstacles.

2.3 Sparse precision methods

Likelihood Approximations by Conditional method in the Spatial Domain.

2.3.1 *LatticeKrig*

2.3.2 *Multiresolution Approximations*

2.3.3 *SPDE/INLA*

The numerical factorization of the precision matrix using sparse matrix algorithms can be done at a typical cost of $O(n^{3/2})$ for two-dimensional GMRFs.

Comments: The drawback of this approachis that we can only find the explicit form of GMRFs for those Gaussian random fields that have a Matérn covariance structure at certain integer smoothnesses. they can be extended to model Matérn covariances on the sphere, nonstationary locally isotropic Gaussian random fields, Gaussian random fields with oscillating correlation functions, and non-isotropic fields.

2.3.4 *NNGP*

The nearest neighbor Gaussian process (Datta et al. [2016a], Datta et al. [2016b], Finley et al. [2019]) is defined from the conditional specification of the joint distribution of spatial process $w(s)$ in (1), one forms of composite likelihoods which is motivated by Vecchia (1988) ideas. The total flop counts is of the order $(n+k)m^3$, where $m(\approx 20)$, NNGP is much faster than the full Gaussian model which requires $O(n^3)$ flops.

Composite likelihoods which points out the difficulty in choosing conditioning sets and evaluation of the approximation accuracy, but with the advent of the NNGP approach, the problem has been solved.

3 LIKELIHOOD APPROXIMATIONS

Likelihood Approximations in the Spectral Domain

Comments: The spectral methods are computationally efficient by avoiding the calculation of determinants and can be easily adapted to model nonstationary processes as a mixture of independent stationary processes. [15] presented a version of Whittle's approximation to the Gaussian negative log-likelihood by introducing a lattice process which can be used to deal with irregularly spaced data. Additional computational savings were obtained by truncating the spectral representation of the lattice process. If n is the total number of observations of the process Y, m is lattice size, the calculation requires $O(m \log_2 m + n)$ operations rather than $O(n^3)$ for the exact likelihood of Y.

However, they do not overcome the difficulty in prediction with massive data.

REFERENCES

S. Banerjee, A. E. Gelfand, A. O. Finley, and H. Sang. Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):825–848, 2008. URL https://chenyw68.github.io/Literature/[2008]Gaussianpredictiveprocessmodelsforlargespatialdatasets.pdf.

N. Cressie and G. Johannesson. Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):209–226, 2008. URL https://chenyw68.github.io/Literature/[2008]Fixedrankkrigingforverylargespatialdatasets.pdf.

N. A. Cressie and G. Johannesson. Spatial prediction for massive datasets. *Proceedings of the Australian Academy of Science Elizabeth and Frederick White Conference*, 2006. URL https://chenyw68.github.io/Literature/[2006]Spatialpredictionformassivedatasets.pdf.

A. Datta, S. Banerjee, A. O. Finley, and A. E. Gelfand. Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111 (514):800–812, 2016a. URL https://chenyw68.github.io/Literature/[2016]Hierarchicalnearest-neighborGaussianprocessmodelsforlargegeostatisticaldatasets.pdf.

A. Datta, S. Banerjee, A. O. Finley, N. A. Hamm, and M. Schaap. Nonseparable dynamic nearest neighbor gaussian process models for large spatio-temporal data with an application to particulate matter analysis. *The annals of applied statistics*, 10(3):1286, 2016b. URL https://chenyw68.github.io/Literature/[2016]NonseparabledynamicNNGPmodelsforlargespatio-temporaldata.pdf.

A. O. Finley, A. Datta, B. D. Cook, D. C. Morton, H. E. Andersen, and S. Banerjee. Efficient algorithms for bayesian nearest neighbor gaussian processes. *Journal of Computational and Graphical Statistics*, 28(2):401–414, 2019. URL https://chenyw68.github.io/Literature/[2019]Efficientalgorithmsforbayesiannearestneighborgaussianprocesses.pdf.

R. Furrer, M. G. Genton, and D. Nychka. Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, 15(3):502–523, 2006. URL https://chenyw68.github.io/Literature/[2006]Covariancetaperingforinterpolationoflargespatialdatasets.pdf.

M. J. Heaton, A. Datta, A. O. Finley, R. Furrer, J. Guinness, R. Guhaniyogi, F. Gerber, R. B. Gramacy, D. Hammerling, M. Katzfuss, et al. A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological and Environmental Statistics*, 24 (3):398–425, 2019. URL https://chenyw68.github.io/Literature/[2018]Acasestudycompetitionamongmethodsforanalyzinglargespatialdata.pdf.

M. Katzfuss. A multi-resolution approximation for massive spatial datasets. *Journal of the American Statistical Association*, 112(517):201–214, 2017.

M. Katzfuss and D. Hammerling. Parallel inference for massive distributed spatial data using low-rank models. *Statistics and Computing*, 27 (2):363–375, 2017. URL https://chenyw68.github.io/Literature/[2017-formal]Parallelinferenceformassivedistributedspatialdatausinglow-rankmodels.pdf.

C. G. Kaufman, M. J. Schervish, and D. W. Nychka. Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association*, 103(484):

1545–1555, 2008. URL https://chenyw68.github.io/Literature/[2008]CovarianceTaperingforLikelihoodBasedEstimationinLargeSpatialDataSets.pdf.

P. Ma and E. L. Kang. A fused gaussian process model for very large spatial data. *Journal of Computational and Graphical Statistics*, pages 1–11, 2020. URL https://chenyw68.github.io/Literature/[2020]AFusedGaussianProcessModelforVeryLargeSpatialData.pdf.

E. Porcu, J.-M. Montero, and M. Schlather. *Advances and challenges in space-time modelling of natural events*, volume 207. Springer Science & Business Media, 2012. URL https://chenyw68.github.io/Literature/[2012]AdvancesandChallengesinSpace-timeModellingofNaturalEvents.pdf.

H. Sang and J. Z. Huang. A full scale approximation of covariance functions for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1):111–132, 2012. URL https://chenyw68.github.io/Literature/[2012]Afullscaleapproximationofcovariancefunctionsforlargespatialdatasets.pdf.