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

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

where β 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 $(\beta, \tau^2, \sigma^2, \theta)$:

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

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

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

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

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 THE COMPUTATIONAL PROBLEM

From an **approximation** point of view, the approximation of the likelihood in either the spatial or spectral domain is fundamental solution to overcome computational obstacles. These solutions focus on model development, the design of efficient and parallel algorithms, 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(bias correction, 2008); Michael L. Stein(Statistical Properties, 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. Dave Higdon(2002); Bruno Sansó(2009); Francky Fouedjio(2016).

Algorithms and platforms:

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

Other scholars:Christopher K. Wikle, Robert B. Gramacy, Bradley P. Carlin, Jonathan R. Stroud, YongTao Guan, Alan E. Gelfand, Peter J. Diggle, Huiyan Sang, Andrw Finley, Matthew J. Heaton, Ying Sun, Furong Sun.

2.1 Likelihood approximations in the spatial domain by low rank methods

Comment: The reduced rank based methods usually fail to accurately capture the local, small scale dependence structure [Sang and Huang, 2012].

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 K ($\ll n$) **multi-resolution bisquare** basis functions:

$$w(s) = \sum_{r=1}^{R} \sum_{k=1}^{K_r} h_{rk}(s) w^{\star}_{rk} \tag{4}$$

where $K = \sum_{r=1}^{R} K_r$ and the coefficients $w^{\star}$ is an K-dimensional Gaussian vector with mean zero and exponential covariance defined by some small basic areal units. which ensures that all estimation and prediction equations only contain inverses of matrices of size $K \times K$.

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 equation in (3). The predictive process projects the original process, $w(s)$ in (1), onto a subspace generated by realizations of the original process at a specified set of locations (or knots), e.g. $s^{\star}_1, \cdots, s^{\star}_K, K \ll n$. 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, i.e.

$$w(s) = \mathbf{c}'_{s_0} \mathbf{\Sigma}^{-1}_{w^{\star}} w^{\star}_{rk} \tag{5}$$

where $\mathbf{c}_{s_0} = \left(\mathbf{C}(s_0, s^{\star}_1), \ldots, \mathbf{C}(s_0, s^{\star}_K)\right)'$, and then $\mathbf{c}'_{s_0} \mathbf{\Sigma}^{-1}_{w^{\star}}$ plays the same role as h in (4), and thus can be regarded as some basis functions.

Comment:

1) 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 $C(\cdot, \cdot)$, and note that the subsequent Multiresolution approximations is in step with GPP in this regard.

2) 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 Likelihood approximations in the spatial domain by 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] and some R packages: spam, 2010; KriSp, 2006; fields, 2017.

Comment: The covariance tapering has shown great computational gains, but it also has its own drawbacks.

1) 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, and the application of tapering techniques to **multivariate random fields** remains to be explored due to the lack of flexible compactly supported cross-covariance functions, see Porcu et al. [2012].

2.2.2 *Spatial partitioning*

By dividing region $\mathbf{D}$ into m disjoint subregions ($d = 1, 2, \cdots, m$), and then the modeling approach based on spatial partitioning is to again assume the model in (6) but take on the assumption of independence between observations across subregions.

$$y_d = \beta X_d + H_d w + \xi_d + \epsilon_d \tag{6}$$

where H_d is a matrix of spatial basis function for subregions d.
Comment:

1) Notice that, in (6) each subregion shares common β and w parameters which allows smoothing across subregions in spite of the independence assumption. Further, the assumption of independence across subregions effectively creates a block-diagonal structure for Σ and allows the likelihood to be computed in parallel (with one node per subregion), thereby facilitating computation.

2) The key to implementing the spatial partitioning approach is the choice of partition and the literature is replete with various options. A priori methods to define the spatial partitioning include partitioning the region into equal areas ([Sang et al., 2011]), partitioning based on centroid clustering ([Kim et al., 2005]), hierarchical clustering based on spatial gradients ([Heaton et al., 2017]). Alternatively, model-based approaches to spatial partitioning include treed regression([Konomi et al., 2014]) and mixture modeling (Neelon et al. [2014]), but these approaches typically require more computation.

2.3 Likelihood approximations in the spatial domain by sparse precision methods

The sparse precision methods focus on basis function and conditional likelihood.

2.3.1 *LatticeKrig*

LatticeKrig (LK, [Nychka et al., 2015]) uses nearly the same setup as is employed by FRK. Here is just a list of the differences:

1) FRK uses families of **bisquare** basis functions that are organized on are organized on irregular basic areal units (BAUs), and LK uses families of **radial** basis functions that are organized on regular grids of increasing resolution.

2) Covariance matrix, $\mathbf{\Sigma}_r$, of the coefficients $w_r^\star$ in (4) is constrained by a exponential function, the exponential function is defined on the centroids of the BAUs, and precision matrix, $\mathbf{Q}_r$, of the coefficients $w_r^\star$ in LK is constrained to a spatial autoregressive (SAR) model, SAR model is defined on the regular lattice.

3) Furthermore, because $\mathbf{Q}_r$ is sparse, LK can set K to be very large (e.g. $K > n$) without much additional computational cost. However, the FRK does not achieve this effect.

2.3.2 *Multi-resolution approximations*

In contrast to FRK or LatticeKrig, the multi-resolution approximation (MRA) basis functions and the prior distribution of the corresponding weights $w_r^\star$ are chosen using the predictive-process approach to automatically adapt to any given covariance function $\mathbf{C}_r$, and so the MRA can adjust flexibly to a desired spatial smoothness and dependence structure. The MRA allows the number of basis functions to be approximately the same as the data by the two ways to do. The one is by increasing sparsity of the covariance matrices of the corresponding weights (achieved by tapering method, see [Katzfuss and Hammerling, 2017]), the other is by recursively partitioning the spatial domain (see [Katzfuss, 2017]).

2.3.3 *Gaussian Markov random field approximations by SPDE*

Rue and Held [2005] rigorously suggest to approximate Gaussian random fields through Markov fields with a huge increase in speed for the simulations, and then [Lindgren et al., 2011] show the construction of the corresponding GMRFs can be used to represent the Matérn field on a triangulated lattice. The drawback of [Lindgren et al., 2011] is 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. Subsequently, Bolin and Kirchner [2020] extend the model to fractional order by the rational approximation.

Comment:

1) 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.

2) 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], 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.

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. Comment:

1) Total operations is $O((n+k)m^3)$ where $m(\approx 20)$ is the size of conditioning set or neighbor set, and n and k is sample size and reference set size.

2) NNGP can be extended to large spatio-temporal data (Datta et al. [2016b]) and non-stationary process(Konomi et al. [2019]).

2.4 Likelihood approximations in the spectral domain: circulant embedding and periodic embeddings

[Yang and Bradley, 2020]

Comment: 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.

D. Bolin and K. Kirchner. The rational spde approach for gaussian random fields with general smoothness. *Journal of Computational and Graphical Statistics*, 29(2):274–285, 2020. URL https://chenyw68.github.io/Literature/[2020]TheRationalSPDEApproachforGaussianRandomFieldsWithGeneralSmoothness.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, W. F. Christensen, and M. A. Terres. Nonstationary gaussian process models using spatial hierarchical clustering from finite differences. *Technometrics*, 59(1):93–101, 2017. URL https://chenyw68.github.io/Literature/[2017]NonstationaryGaussianProcessModelsUsingSpatialHierarchicalClusteringfromFiniteDifferences.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. URL https://chenyw68.github.io/Literature/[2017]AMulti-ResolutionApproximationforMassiveSpatialDatasets.pdf.

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.

H.-M. Kim, B. K. Mallick, and C. Holmes. Analyzing nonstationary spatial data using piecewise gaussian processes. *Journal of the American Statistical Association*, 100(470):653–668, 2005. URL https://chenyw68.github.io/Literature/[2005]Analyzingnonstationaryspatialdatausingpiecewise Gaussianprocesses.pdf.

B. Konomi, A. Hanandeh, P. Ma, and E. Kang. Computationally efficient nonstationary nearest-neighbor gaussian process models using data-driven techniques. *Environmetrics*, 30(8):e2571, 2019. URL https://chenyw68.github.io/Literature/[2020]Computationallyefficientnonstationarynearest-neighborGaussianprocessmodelsusingdata-driventechpdf.

B. A. Konomi, H. Sang, and B. K. Mallick. Adaptive bayesian nonstationary modeling for large spatial datasets using covariance approximations. *Journal of Computational and Graphical Statistics*, 23 (3):802–829, 2014. URL https://chenyw68.github.io/Literature/[2014]AdaptiveBayesianNonstationaryModelingforLargeSpatialDatasetsUsingCovarianceApproximations.pdf.

F. Lindgren, H. Rue, and J. Lindström. An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011. URL https://chenyw68.github.io/Literature/[2011]AnexplicitlinkbetweenGFandGMRFstheSPDEapproach.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.

B. Neelon, A. E. Gelfand, and M. L. Miranda. A multivariate spatial mixture model for areal data: examining regional differences in standardized test scores. *Journal of the Royal Statistical Society. Series C, Applied statistics*, 63(5):737, 2014. URL https://chenyw68.github.io/Literature/[2014]Amultivariatespatialmixturemodelforarealdata.pdf.

D. Nychka, S. Bandyopadhyay, D. Hammerling, F. Lindgren, and S. Sain. A multiresolution gaussian process model for the analysis of large spatial datasets. *Journal of Computational and Graphical Statistics*, 24(2):579–599, 2015. URL https://chenyw68.github.io/Literature/[2015]AMultiresolutionGaussianProcessModelfortheAnalysisofLargeSpatialDatasets.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. Rue and L. Held. *Gaussian Markov random fields: theory and applications*. CRC press, 2005. URL https://chenyw68.github.io/Literature/[2005]Gaussianmarkovrandomfieldstheoryandapplications.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.

H. Sang, M. Jun, and J. Z. Huang. Covariance approximation for large multivariate spatial data sets with an application to multiple climate model errors. *The Annals of Applied Statistics*, pages 2519–2548, 2011. URL https://chenyw68.github.io/Literature/[2011]Covarianceapproximationforlargemultivariatespatialdatasets.pdf.

H.-C. Yang and J. R. Bradley. Bayesian inference for big spatial data using non-stationary spectral simulation. *arXiv preprint arXiv:2001.06477*, 2020. URL https://chenyw68.github.io/Literature/[2020]BayesianInferenceforBigSpatialDataUsingNon-stationarySpectralSimulation.pdf.