Bayesian Meta-Learning Approach for Feasible Large Spatial Analysis

Luca Presicce[1] and Sudipto Banerjee[2]

[1] University of Milano-Bicocca, Department of Management, Economics and Statistics, Milano, Italy,
l.presicce@campus.unimib.it,
[2] University of California Los Angeles, Department of Biostatistics, Los Angeles CA, USA

Abstract. Geostatistical modeling is afflicted by onerous computational effort when the number of observed locations is very large. While there exists a burgeoning literature today that attempts to tackle the so-called "big-n" problems, spatial inference remains unfeasible for moderate data sets on modest computing architectures. Our current contribution resides in the domain of "meta-" approaches where a massive data set is split into smaller data sets, each data set is analyzed independently and the inference from these individual data are combined to approximate fully model-based Bayesian inference. Our specific contribution is to introduce Bayesian predictive stacking in spatial meta-analysis in the context of univariate spatial data. Furthermore, we aim to make inference and uncertainty quantification feasible without excessively demanding hardware settings. Introducing new methodologies, and exploiting existing techniques, the analysis of a massive data set with observations in millions is illustrated.

Keywords: Bayesian predictive stacking, meta-learning, geostatistical modeling, conjugate models

1 Introduction

A primary challenge in spatial data science is the analysis of big georeferenced sets of data. This stems from the demanding Gaussian likelihood computations involving matrix factorizations and determinant computations for large spatial covariance matrices. This is referred to as the "Big N" problem in spatial statistics. In the same way, spatial prediction, also called kriging, relies on Gaussian process regression methodologies, that scale very badly with massive sets of data. To manage this problem, burgeoning literature on the analysis of large spatial data sets focuses mostly on a few bins of solutions, at whose foundation there are common ideas. Many of them are in the direction of some sort of dimensionality reduction. One customary idea aims to reduce the dimension of the spatial covariance matrix by imposing a low-rank structure. A second class of approaches considers sparsity as a solution as well. However, they assume

sparsity over the inverse of the spatial covariance matrix, usually achieved by imposing conditional independence, or also with composite likelihood. Another school of thought works reduction by considering smaller problems instead. Partitioning the domain into subregions, and then hierarchically combining models ensures the borrowing of information. Even if there are similarities with the last, the current contribution follows a different idea. The here-introduced accelerated spatial meta-kriging (ASMK) lies in a class of methodology that more recently appeared in the literature, as [4,5]. These contributions, chase a meta-learning-based approach, exploiting Divide-and-Conquer procedures.

The contents of the paper evolves as follows. In Section 2 we briefly present the proposed accelerated spatial meta-kriging approach. Section 3, illustrate an analysis on sea surface temperature with over 1 million observed locations. Conclusive considerations end the work in Section 4.

2 Accelerating spatial meta-kriging

To better understand the contents in Section 2, we remind the reader to the concepts of Bayesian predictive stacking (BPS) of predictive densities. A comprehensive dissertation can be found in [6], and for further details see also [3]. Hereafter, we introduce the accelerated spatial meta-kriging approach, for univariate modeling. The ASMK is a Divide-Conquer strategy, composed as follows. In the divide step, after partitioning the data, we perform the BPS on each subset (within-subsets), obtaining a set of stacked posterior distributions. Then in the conquer step, we serve of BPS procedure, but on stacked posterior distributions (between-subsets), to combine them. By first BPS, we obtain subset inferences, while the second BPS provides full posterior inference. Let us consider a customary regression model for a spatially indexed outcome $y(s)$ at a location s in a bounded region $\mathcal{D} \subset \mathbb{R}^d$,

$$y(\mathbf{s}) = \mathbf{x}(\mathbf{s})^\top \beta + \omega(\mathbf{s}) + \varepsilon(\mathbf{s}), \tag{1}$$

where $\mathbf{x}(\mathbf{s})$ is a $p \times 1$ vector of spatially referenced predictors, β is a $p \times 1$ vector of slopes measuring the trend, $\omega(\mathbf{s}) \sim \mathrm{GP}\left(0, \sigma^2 \boldsymbol{\rho}_\psi(\cdot,\cdot)\right)$ is a zero-centred spatial Gaussian process on $\mathbb{R}^d$ with spatial correlation function $\boldsymbol{\rho}_\psi(\cdot,\cdot)$ depending on spatial range parameter ψ, and σ^2 is a scale (spatial variance) parameter. The white noise process $\varepsilon(\mathbf{s}) \sim \mathrm{N}\left(0, \tau^2\right)$ with variance τ^2 captures measurement error. Let $\mathcal{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\} \in \mathcal{D}$ be a set of n spatial locations, yielding to $\mathbf{y} = (y(\mathbf{s}_1), \ldots, y(\mathbf{s}_n))$ with predictors at these locations collected in the $n \times p$ matrix $\mathbf{X} = [\mathbf{x}(\mathbf{s}_1) : \cdots : \mathbf{x}(\mathbf{s}_n)]^\top$. Defined the finite-dimensional realization of the spatial process as $\boldsymbol{\omega} = (\omega(\mathbf{s}_1), \ldots, \omega(\mathbf{s}_n))^\top$, and let $\boldsymbol{\rho}_\psi(\mathcal{S}, \mathcal{S})$ be the $n \times n$ spatial correlation matrix constructed from $\boldsymbol{\rho}_\psi(\cdot,\cdot)$. A Bayesian hierarchical model, including the specification of prior distributions, is constructed as

$$\mathbf{y} \mid \boldsymbol{\omega}, \beta, \sigma^2 \sim \mathrm{N}\left(\mathbf{X}\beta + \boldsymbol{\omega}, \delta^2\sigma^2 \mathbb{1}_n\right), \tag{2}$$

$$\boldsymbol{\omega} \mid \sigma^2 \sim \mathrm{N}\left(0, \sigma^2 \boldsymbol{\rho}_\psi(\mathcal{S}, \mathcal{S})\right) \tag{3}$$

$$\beta \mid \sigma^2 \sim \mathrm{N}\left(\mu_\beta, \sigma^2 V_\beta\right), \quad \sigma^2 \sim \mathrm{IG}\left(a_\sigma, b_\sigma\right). \tag{4}$$

Where the spatial correlation parameter ψ and the noise-to-spatial variance ratio $\delta^2 := \tau^2/\sigma^2$ are considered fixed. While $\mu_\beta, V_\beta, a_\sigma$, and b_σ are fixed hyperparameters specifying the prior distributions for β and σ^2. Considering to fix $\{\psi, \delta^2\}$, ensures closed-form conjugate posterior inference for this hierarchical specification, for further details on the results presented below see [2,1].

Let $\boldsymbol{\gamma} = [\beta^\top, \boldsymbol{\omega}^\top]^\top$, then a conjugate Bayesian model is obtained by considering the joint prior on $\{\boldsymbol{\gamma}, \sigma^2\}$, denoted as $p\left(\boldsymbol{\gamma}, \sigma^2 \mid \mu_\beta, V_\beta, a_\sigma, b_\sigma\right)$. For any fixed $\{\psi, \delta^2\}$, we have the closed-form posterior density

$$p\left(\boldsymbol{\gamma}, \sigma^2 \mid \mathbf{y}\right) = p\left(\sigma^2 \mid \mathbf{y}\right) p\left(\boldsymbol{\gamma} \mid \sigma^2, \mathbf{y}\right) \tag{5}$$

$$= \mathrm{IG}\left(\sigma^2 \mid a_\sigma^\star, b_\sigma^\star\right) \mathrm{N}\left(\boldsymbol{\gamma} \mid \hat{\boldsymbol{\gamma}}, \sigma^2 M_\star\right), \tag{6}$$

$$\text{where} \quad a_\sigma^\star = a_\sigma + n/2, \quad b_\sigma^\star = b_\sigma + 1/2\left(\mathbf{y}_\star - \mathbf{X}_\star\hat{\boldsymbol{\gamma}}\right)^\top V_\star^{-1}\left(\mathbf{y}_\star - \mathbf{X}_\star\hat{\boldsymbol{\gamma}}\right), \tag{7}$$

$$M_\star^{-1} = \mathbf{X}_\star^\top V_\star^{-1}\mathbf{X}_\star, \quad \hat{\boldsymbol{\gamma}} = M_\star\mathbf{X}_\star^\top V_\star^{-1}\mathbf{y}_\star, \tag{8}$$

where $\mathbf{y}_\star, \mathbf{X}_\star, V_\star$ composed the augmented linear system cast from the spatial model in Equation (2). For more specifications about these results see [1,9]. Let $\mathcal{U} = \{\mathbf{u}_1, \ldots, \mathbf{u}_{n'}\} \in \mathcal{D}$ be a set of n' of unknown points, with $\boldsymbol{\omega}_\mathcal{U}$ and $\mathbf{y}_\mathcal{U}$ be the $n' \times 1$ vectors with elements $\omega\left(\mathbf{u}_i\right)$ and $y\left(\mathbf{u}_i\right)$ for $i = 1, 2, \ldots, n'$. Let $\mathbf{X}_\mathcal{U} = \left(\mathbf{x}\left(\mathbf{u}_1\right) : \ldots : \mathbf{x}\left(\mathbf{u}_{n'}\right)\right)^\top$ be the $n' \times p$ matrix of predictors and let $\boldsymbol{\rho}_\psi(\mathcal{U},\mathcal{U})$ the spatial correlation matrix at $\mathcal{U}$. Then, Bayesian inference proceeds from exact posterior samples obtained from Equation (5). We first draw values of $\sigma^2 \sim \mathrm{IG}\left(a_\sigma^\star, b_\sigma^\star\right)$ followed by a single draw of $\boldsymbol{\gamma} \sim \mathrm{N}\left(\hat{\boldsymbol{\gamma}}, \sigma^2 M_\star\right)$ for each drawn value of σ^2. This yields samples $\left\{\boldsymbol{\gamma}, \sigma^2\right\}$ from Equation (5). While predictive inference for the latent process $\boldsymbol{\omega}_\mathcal{U}$ and the outcome $\mathbf{y}_\mathcal{U}$ is obtained by drawing a value of $\boldsymbol{\omega}_\mathcal{U} \sim N\left(\mu_\omega(\boldsymbol{\gamma}), \sigma^2 V_\omega\right)$ with $\mu_\omega(\boldsymbol{\gamma}) := \boldsymbol{\rho}_\psi(\mathcal{U},\mathcal{S})\boldsymbol{\rho}_\psi(\mathcal{S},\mathcal{S})^{-1}\boldsymbol{\omega}$ and $V_\omega := \boldsymbol{\rho}_\psi(\mathcal{U},\mathcal{U}) - \boldsymbol{\rho}_\psi(\mathcal{U},\mathcal{S})\boldsymbol{\rho}_\psi(\mathcal{S},\mathcal{S})^{-1}\boldsymbol{\rho}_\psi^\top(\mathcal{U},\mathcal{S})$ for each value of $\left\{\boldsymbol{\gamma}, \sigma^2\right\}$ drawn above (see Section 3.4 in [1]), then drawing a value of $\mathbf{y}_\mathcal{U} \sim N\left(\mathbf{X}_\mathcal{U}\beta + \boldsymbol{\omega}_\mathcal{U}, \sigma^2\delta^2\mathbb{1}_{n'}\right)$ for each drawn value of β (extracted from $\boldsymbol{\gamma}$), σ^2 and $\boldsymbol{\omega}_\mathcal{U}$. As anticipated, this tractability is only possible if the range decay ψ and the noise-to-spatial variance ratio δ^2 are fixed. So far, alternative approaches using K-fold cross-validation have been explored with limited success [2]. The aforementioned conjugate modeling is then used in each subset of the data. Given this framework, in the Divide-step we use BPS to "accelerate" the subset modeling, and then the entire inference. Indeed, BPS is a particular case of Bayesian model averaging approaches, which allows us to "integrate out" the hyperparameters $\{\delta^2, \psi\}$. This avoids the need for simulation-based approaches, such as MCMC, exploiting exact conjugate modeling, and resulting in an acceleration of the inferences. However, as presented in [6], to perform BPS of predictive densities one has to solve the optimization problem:

$$\max_{\mathbf{z}_k \in \mathbf{S}_1^J} \frac{1}{n_k} \sum_{i=1}^{n_k} \log \sum_{j=1}^{J} z_{k,j} p\left(y_{k,i} \mid \mathbf{y}_{k,-i}, \mathbf{M}_j\right), \tag{9}$$

where $\mathbf{S}_1^J$ is the J-dimensional simplex, and J represents the number of model configurations $\mathbf{M}_j$, i.e. the cardinality of the set composed by a grid of hyperpa-

rameter $\{\delta^2, \psi\}$ values. In Equation (9), the quantities $\mathbf{y}_k$ represent the subvector of $\mathbf{y}$ that corresponds to the k-th subset, n_k is its length, and $y_{k,i}, \mathbf{y}_{k,-i}$ are the i-th element of $\mathbf{y}_k$ and the vector itself without the i-th element, respectively. Once the stacking weights $\hat{\mathbf{z}}_k = \{\hat{z}_{k,j}\}_{j=1,\dots,J}$, for $k = 1, \dots, K$ are obtained, the stacked estimate of the predictive density, within each partition, can be recovered as

$$\hat{p}(\tilde{\mathbf{y}} \mid \mathbf{y}_k) = \sum_{j=1}^{J} \hat{z}_{k,j}\; p\left(\tilde{\mathbf{y}} \mid \mathbf{y}_k, \mathbf{M}_j\right), \quad k = 1, \dots, K, \quad j = 1, \dots, J. \tag{10}$$

Thus, for each of the K subsets, the BPS procedure provides an estimate of the posterior predictive $\hat{p}(\tilde{\mathbf{y}} \mid \mathbf{y}_k)$, and a set of stacking weights $\hat{\mathbf{z}}_k$. To recover the global inference for the entire data set, we apply the BPS procedure again. Hence, a double stacking will be performed. Then, the double BPS of predictive densities boils down to the solution of the following convex optimization problem:

$$\max_{\mathbf{w} \in \mathbf{S}_1^K} \frac{1}{n} \sum_{i=1}^{n} \log \sum_{k=1}^{K} w_k \sum_{j=1}^{J} \hat{z}_{k,j} p\left(y_{k,i} \mid \mathbf{y}_{k,-i}, \mathbf{M}_j\right), \tag{11}$$

with $\mathbf{S}_1^K$ as the K-dimensional simplex. Again, once the stacking weights $\mathbf{w} = \{w_k\}_{k=1,\dots,K}$, that regulate the combination of the individual subset posterior predictive models are obtained by Equation (11), the double BPS estimation of the full posterior predictive distribution is then recovered:

$$\hat{p}(\tilde{\mathbf{y}} \mid \mathbf{y}) = \sum_{k=1}^{K} \hat{w}_k\; \hat{p}(\tilde{\mathbf{y}} \mid \mathbf{y}_k), \qquad\qquad k = 1, \dots, K. \tag{12}$$

Let $\mathcal{U}$ be a set of n' of unknown points as in Section 2. Thus, the prediction for the new response $\mathbf{y}_{\mathcal{U}}$ can be straightforwardly attained from Equation (12).

$$\hat{p}(\mathbf{y}_{\mathcal{U}} \mid \mathbf{y}) = \sum_{k=1}^{K} \hat{w}_k \sum_{j=1}^{J} \hat{z}_{k,j}\; p\left(\mathbf{y}_{\mathcal{U}} \mid \mathbf{y}_k, \mathbf{M}_j\right). \tag{13}$$

A similar strategy can be adopted for the posterior predictive surface for the spatial process $\boldsymbol{\omega}$. Indeed, the prediction for $\boldsymbol{\omega}_{\mathcal{U}}$ is derived from

$$\hat{p}(\boldsymbol{\omega}_{\mathcal{U}} \mid \mathbf{y}) = \sum_{k=1}^{K} \hat{w}_k\; \hat{p}(\boldsymbol{\omega}_{\mathcal{U}} \mid \mathbf{y}_k), \tag{14}$$

where $\hat{p}(\boldsymbol{\omega}_{\mathcal{U}} \mid \mathbf{y}_k) = \sum_{j=1}^{J} \hat{z}_{k,j}\; p\left(\omega_{\mathcal{U}} \mid \mathbf{y}_k, \mathbf{M}_j\right)$, and $\{\hat{z}_{k,j}\}$ as the same stacking weights obtained before, for $k = 1, \dots, K$ and $j = 1, \dots, J$.

3 Real data illustration

One of the most discussed topics of the last years is global warming and its consequences. Then, institutions have started to monitor key aspects of its evolution,

providing data that allows the research to develop crucial predictive global models. One of the main aspects is the surface sea temperature (SST), which NASA maintains a satellite-based database that extends all over the world. Further details on this can be found on NASA related online portal. In this section, we present an application of ASMK to the SST data. The current offer has the scope to enable researchers to perform geostatistical analysis on massive spatial data sets, even with modest computational and memory resources. For this reason, the data analysis was carried out on a standard laptop. In detail, we used a Windows machine equipped with 16 Gb of RAM, Intel i7-8750H processor (enabling parallel computations over 10 logical cores). As an illustration, we analyze a data set composed of $1,001,000$ georeferenced observations of SST, collected within June 2017. To this end, we take advantage of the data set used in [7], but differently from them, we consider SST data all over the world, instead of a selected subregion. Within this massive data set, we use $n = 1,000,000$ observations for model fitting while the rest for predictive assessment. As explanatory variables, we consider the coordinates after a sinusoidal projection. For how to concern the Divide-Conquer strategy, we take into account $K = 2,000$ subsets of random sampled observations, that corresponds to a subset size of 500 units. Prior choices follow standard quantities, as considered in [7], and [9]. The set of hyperparameter values was defined by considering explanatory spatial data analysis, evaluating the variograms to find insight about ψ and δ^2. In Figure 1, the interpolated maps of the results are reported.

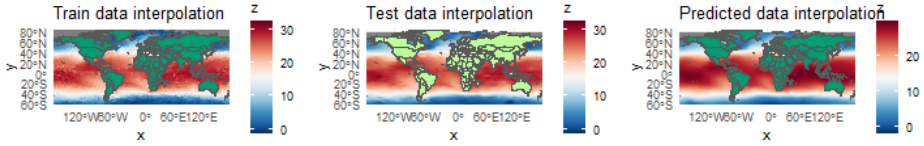

Fig. 1. From left to right: comparison between training, test, and predicted surface.

Figure 1 shows an empirically striking capability to reproduce the test data surface of SST. Moreover, the full model-based Bayesian inference requires only 280 minutes, just more than 4 hours. We assess predictive performance using RMSPE on 1000 withheld observations for the ASMK model, comparing it with the Bayesian linear model (BLM), which does not account for spatial variability. While the BLM performs a RMSPE of 8.981, the accelerated spatial meta-kriging approach achieves a much lower RMSPE of 1.346. This remarks the effectiveness of the proposed procedure, as well as the chance to have Bayesian geostatistical modeling feasible with millions of observations. The entire analysis and graphics were implemented in native `R` and `c++` programming languages within the `Rcpp` package, indeed called `ASMK`, freely available on the corresponding author GitHub repository lucapresicce/ASMK.

4 On going features

Here, we briefly illustrate the principles behind the ASMK for univariate geostatistical modeling. This is currently an ongoing work, and as such, some features need more investigation and elaboration. First of all, a set of simulations are under development now. In particular, we empirically want to show that ASMK is capable of recovering spatial surfaces, without penalizing the posterior learning of the parameters, mainly checking for the inference about the regression coefficients β and the variance σ^2. Moreover, we aim to show that ASMK behaves as the SMK of [4], but that it is faster. Indeed, the reduced computational complexity and the better timing are two of the main strengths of ASMK. Furthermore, ASMK pretends to make feasible the analysis of massive spatial data sets even on modest computational resources. To this matter, the parallelization framework is self-constrained to a standard laptop, as explained in Section 3. The motivation can be traced back to the fact that most of the solutions in the literature use huge computational infrastructures. However, this does not allow computational advantages reproducible in general. As evident from the illustration in Section 3, we could be on the right path. Indeed, we provide a feasible approach to the analysis of millions of points, in a reasonable amount of time without compromising the inferences, and without using extreme computational resources.

References

1. Banerjee, S.: Modeling massive spatial datasets using conjugate Bayesian linear modeling framework. Spat. Stat. 37, (2020). doi.org/10.1016/j.spasta.2020.100417
2. Finley, A. O., Datta, A., Cook, B. D., Morton, D. C., Andersen, H. E., Banerjee, S.: Efficient algorithms for Bayesian nearest neighbor Gaussian Processes. J. Comput. Graph. Stat. 28, (2019). doi.org/10.1080/10618600.2018.1537924
3. Gneiting, T., Raftery, A. E.: Strictly proper scoring rules, prediction, and estimation. J. Amer. Statist. Assoc. 102, (2007). doi.org/10.1198/016214506000001437
4. Guhaniyogi, R., Banerjee, S.: Meta-kriging: scalable Bayesian modeling and inference for massive spatial datasets. Technometrics 60, (2018). doi.org/10.1080/00401706.2018.1437474
5. Guhaniyogi, R., Li, C., Savitsky, T., Srivastava, S.: A divide-and-conquer Bayesian approach to large-scale kriging. Submitted to Stat. Sci., (2019). doi.org/10.48550/arXiv.1712.09767
6. Yao, Y., Vehtari, A., Simpson, D., Gelman, A.: Using stacking to average Bayesian predictive distributions (with discussion). Bayesian Anal. 13, (2018) doi.org/10.1214/17-BA1091
7. Zhang, L., Datta, A., Banerjee, S.: Practical Bayesian modeling and inference for massive spatial data sets on modest computing environments. Stat. Anal. Data Min. 12, (2019). doi.org/10.1002/sam.11413
8. Zhang, L., Banerjee, S.: Spatial factor modeling: A Bayesian matrix-normal approach for misaligned data. Biometrics 78, (2021). doi.org/10.1111/biom.13452
9. Lu Zhang and Wenpin Tang and Sudipto Banerjee: Exact bayesian geostatistics using predictive stacking. arXiv preprint. (2023) arxiv.org/abs/2304.12414