Bayesian Transfer Learning for Artificially Intelligent Geospatial Systems: A Predictive Stacking Approach

Luca Presicce

Ph.D. student in Statistics

University of Milano-Bicocca, Department of Economics, Management & Statistics

KIT - Karlsruher Institut für Technologie 2025, May 20

This is a joint work with

Sudipto Banerjee

University of California, Los Angeles

- R. Guhaniyogi and S. Banerjee (2018) "Meta-kriging: Scalable bayesian modeling and inference for massive spatial dataset", Technometrics, vol. 60.
- Y. Yao, A. Vehtari, D. Simpson and A. Gelman (2018) "Using stacking to average Bayesian predictive distributions", Bayesian analysis, vol. 13.
- S. Banerjee (2020) "Modeling massive spatial datasets using a conjugate bayesian linear modeling framework", Spatial Statistics, vol. 37.
- L. Zhang, W. Tang and S. Banerjee (2023) "Bayesian Geostatistics Using Predictive Stacking", arXiv:2304.12414
- L. Presicce and S. Banerjee (2024+) "Bayesian Transfer Learning for Artificially Intelligent Geospatial Systems: A Predictive Stacking Approach", arXiv:2410.09504

Introduction

- Geospatial artificial intelligence (GeoAI) is a rapidly evolving discipline at the interface of statistical learning and spatial data science, attempting to harness the analytical capabilities of Artificial Intelligence (AI) to analyze massive amounts of geographic data.

- A fundamental question concerns the role of formal statistical inference in GeoAI, since spatial-temporal random fields enjoy a prominent presence of theoretical developments within classical and Bayesian paradigms [see, e.g., 3, 13, 7, 4, 1].

- GeoAI offers space to accommodate rigorous probabilistic learning as a crucial constituent in artificially intelligent data analysis systems.

Challanges in probabilistic Geospatial systems

- Spatial modeling relies upon Gaussian processes (GPs). Despite great flexibility, their covariance kernels do not yield computationally exploitable structures, and full inference becomes impracticable for massive datasets.

- Full inference typically requires Markov chain Monte Carlo (MCMC) [5], variational approximations [11, 16, 2], or Gaussian Markov random field approximations [12, 10, and references therein] along with integrated nested Laplace approximations (INLA).

- Focusing upon the richness of statistical inference involves a significant amount of human intervention. Building a GeoAI system will require minimizing human intervention to offer a robust spatial data analysis framework.

- 👉 We devise a spatial data analytic framework that holds significant promise for GeoAI. This approach relies upon two basic tenets:
  - ☞ model-based statistical inference for underlying spatial processes in a robust and largely automated manner with minimal human input;
  - ☞ achieving such inference for truly massive amounts of data without resorting to iterative algorithms (such as in MCMC).
- 👉 Retaining the benefits of Bayesian hierarchical models while performing spatial data analysis at unprecedented scales requires some concessions from conventional decision-theoretic paradigms.

Bayesian Transfer Learning for GeoAI

- Transfer learning (TL) broadly refers to propagating knowledge from a task to accomplish a different task. We look at transferring inference from one subset of spatial data to the next in a stream of subsets to assimilate inference for the entire data set.

- This resembles "divide and conquer" methods that divide a computationally unfeasible problem into tractable sub-problems. Wishing to reproduce the inference, as we would be able to analyze the entire dataset with a desired model.

- This is achieved only in special cases. Learning from spatial random fields is complicated because of the complex inherent dependencies in the data.

- We consider the multivariate conjugate matrix-variate Bayesian linear regression model

$$\mathbf{Y} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma} \sim \mathrm{MN}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}, \boldsymbol{\Sigma}), \\ \boldsymbol{\beta} \mid \boldsymbol{\Sigma} \sim \mathrm{MN}(\mathbf{M}_0\mathbf{m}_0, \mathbf{M}_0, \boldsymbol{\Sigma}), \quad \boldsymbol{\Sigma} \sim \mathrm{IW}(\boldsymbol{\Psi}_0, \nu_0) \tag{1}$$

- Let $\mathscr{D} = \{\mathscr{D}_1, \ldots, \mathscr{D}_K\}$, where each $\mathscr{D}_k = \{Y_k, X_k\}$. By distribution theory, starting with $\mathrm{MNIW}(\beta, \Sigma \mid M_k m_k, M_k, \Psi_k, \nu_k)$ at $k = 0$ (the prior), the Bayesian updating

$$p(\beta, \Sigma \mid \mathscr{D}_{1:k+1}) \propto p(\beta, \Sigma \mid \mathscr{D}_{1:k}) \times p(Y_{k+1} \mid X_{k+1}\beta, V_{k+1}, \Sigma) \tag{2}$$

- leads to $\beta, \Sigma \mid \mathscr{D}_{1:k+1} \sim \mathrm{MNIW}(M_{k+1}m_{k+1}, M_{k+1}, \Psi_{k+1}, \nu_{k+1})$ Until $k = K$, with

$$M_{k+1}^{-1} = M_k^{-1} + X_{k+1}^\top V_{k+1}^{-1} X_{k+1}, \quad m_{k+1} = m_k + X_{k+1}^\top V_{k+1}^{-1} Y_{k+1}, \\ \nu_{k+1} = \nu_k + n_{k+1}, \quad \Psi_{k+1} = \Psi_k + Y_{k+1}^\top V_{k+1}^{-1} Y_{k+1} + m_k^\top M_k m_k - m_{k+1}^\top M_{k+1} m_{k+1}. \tag{3}$$

- However, We exactly recover the posterior distribution $p(\beta, \Sigma \mid \mathscr{D})$ only if Y_k's are assumed to be independent. Spatial and spatiotemporal random field models (more generally correlated data) immediately present a challenge.

- We devise a method for assimilating the learning from each block, exploiting the fact that V_k is indexed by a few parameters. Fixing these parameters for each k, yielding closed-form posterior inference on β and Σ.

- Stacking combines these analytically accessible distributions, reconstructing the posterior (and predictive) distributions for the spatial random field without imposing block independence.

- Bayesian predictive stacking (BPS) assimilates models using a weighted distribution in the convex hull, $C = \left\{\sum_{j=1}^{J} w_j p(\cdot \mid \mathscr{D}, \mathscr{M}_j) : \sum_j w_j = 1, w_j \geq 0\right\}$, of individual posterior distributions by maximizing the score [8, 17] to fetch

$$(w_1, \ldots, w_J)^\top = \arg\max_{w \in S_1^J} \frac{1}{n} \sum_{i=1}^{n} \log \sum_{j=1}^{J} w_j p\left(Y_i \mid \mathscr{D}_{-i}, \mathscr{M}_j\right) \ , \tag{4}$$

- where $\mathscr{D}_{-i}$ is the dataset excluding the i-th observation, and $\mathscr{M} = (\mathscr{M}_1, \ldots, \mathscr{M}_J)$ are J different models. Each element of $\mathscr{M}$ corresponds to fixed spatial correlation kernel parameters in V.

- Solving (4) minimizes the Kullback-Leibler divergence from the true predictive distribution, easily executable using convex optimization [9, 6]. Since the true predictive distribution is unknown, we use a leave-one-out (LOO) estimate of the expected value of the score [17].

- Let $\mathcal{S} = \{s_1, \ldots, s_n\}$ be a set of n locations yielding observations on q possibly correlated outcomes. We collect them into matrix $Y\,(n \times q)$. Let $X\,(n \times p)$ consisting of $p < n$ explanatory variables (rank p).

- We cast this into (1) introducing a multivariate latent spatial process $\mathbf{\Omega}\,(n \times q)$ as

$$\begin{aligned} \mathbf{Y} \mid \boldsymbol{\beta}, \mathbf{\Omega}, \mathbf{\Sigma} &\sim \mathrm{MN}(\mathbf{X}\boldsymbol{\beta} + \mathbf{\Omega}, (\alpha^{-1} - 1)\mathbb{I}_n, \mathbf{\Sigma}) \\ \boldsymbol{\beta} \mid \mathbf{\Sigma} &\sim \mathrm{MN}(\mathbf{M}_0\mathbf{m}_0, \mathbf{M}_0, \mathbf{\Sigma}) \\ \mathbf{\Omega} \mid \mathbf{\Sigma} &\sim \mathrm{MN}(\mathbf{0}, \mathbf{V}, \mathbf{\Sigma}) \\ \mathbf{\Sigma} &\sim \mathrm{IW}(\mathbf{\Psi}_0, \nu_0). \end{aligned} \tag{5}$$

- $V\,(n \times n)$ is a spatial correlation matrix with (i, j)-th element is the value of a positive definite spatial correlation function $\rho(s_i, s_j; \phi)$ indexed by ϕ.

- $\alpha \in [0, 1]$ is the proportion of variability due to the spatial process and introduces discontinuity in the spatial correlation.

- Let assume

$$\gamma, \Sigma \sim \mathrm{MNIW}\left(\mu_\gamma, V_\gamma, \Psi_0, \nu_0\right), \tag{6}$$

with $\gamma^\top = \left[\beta^\top, \Omega^\top\right]$, $\mu_\gamma^\top = \left[m_0^\top M_0, 0_{q\times n}\right]$ and $V_\gamma =$ blockdiag $\left\{M_0, \rho_\phi(\mathcal{S}, \mathcal{S})\right\}$.

- The MNIW prior is conjugate with respect to the matrix-normal likelihood. Thus, for any fixed $\{\alpha, \phi\}$ and hyperparameters in the prior density, we obtain the MNIW posterior density

$$\gamma, \Sigma \mid \mathscr{D} \sim \mathrm{MNIW}\left(\gamma, \Sigma \mid \mu_\gamma^\star, V_\gamma^\star, \Psi^\star, \nu^\star\right), \tag{7}$$

where $V_\gamma^\star = \begin{bmatrix} \frac{\alpha}{1-\alpha} X^\top X + M_0^{-1} & \frac{\alpha}{1-\alpha} X^\top \\ \frac{\alpha}{1-\alpha} X & \rho_\phi^{-1}(\mathcal{S}, \mathcal{S}) + \frac{\alpha}{1-\alpha}\mathbb{I}_n \end{bmatrix}^{-1}$, $\quad \mu_\gamma^\star = V_\gamma^\star \begin{bmatrix} \frac{\alpha}{1-\alpha} X^\top Y + m_0 \\ \frac{\alpha}{1-\alpha} Y \end{bmatrix}$

and $\Psi^\star = \Psi_0 + \frac{\alpha}{1-\alpha} Y^\top Y + m_0^\top M_0 m_0 - \mu_\gamma^{*\top} V^{*-1} \mu_\gamma^\star, \quad \nu^\star = \nu_0 + n.$

- Let $\mathcal{U} = \{u_1, \ldots, u_{n'}\}$ be a set of n' unobserved locations where we seek to predict the value of Y based upon $X_{\mathcal{U}}$ ($n' \times p$). The joint posterior predictive for $Y_{\mathcal{U}}$ and the unobserved latent process $\Omega_{\mathcal{U}}$ is

$$Y_{\mathcal{U}}, \Omega_{\mathcal{U}} \mid \mathscr{D} \sim \mathrm{T}_{2n',q}(\nu^\star, \mu^\star, V^\star, \Psi^\star). \tag{8}$$

- Which is a matrix-variate Student's t with degrees of freedom $\nu^\star$, location matrix $\mu^\star = M\mu_\gamma^\star$, row-scale matrix $V^\star$, and column-scale matrix $\Psi^\star$.

where $M_{\mathcal{U}} = \rho_\phi(\mathcal{U},\mathcal{S})\rho_\phi^{-1}(\mathcal{S},\mathcal{S})$ and $V_{\Omega_{\mathcal{U}}} = \rho_\phi(\mathcal{U},\mathcal{U}) - \rho_\phi(\mathcal{U},\mathcal{S})\rho_\phi^{-1}(\mathcal{S},\mathcal{S})\rho_\phi(\mathcal{S},\mathcal{U})$

$$M = \begin{bmatrix} 0 & M_{\mathcal{U}} \\ X_{\mathcal{U}} & M_{\mathcal{U}} \end{bmatrix} \text{ and } V^\star = MV_\gamma^\star M^\top + V_E, \quad V_E = \begin{bmatrix} V_{\Omega_{\mathcal{U}}} & V_{\Omega_{\mathcal{U}}} \\ V_{\Omega_{\mathcal{U}}} & V_{\Omega_{\mathcal{U}}} + (\alpha^{-1} - 1)\mathbb{I}_{n'} \end{bmatrix}.$$

- This tractability is only possible if $\{\alpha, \phi\}$ are fixed, which are inconsistently estimable [18] resulting in poorer convergence. We pursue exact inference using (7) and (8), stacking the inference over the different fixed values of $\{\alpha, \phi\}$.

- For GeoAI we seek to minimize human intervention using a set of J candidate values $\{\alpha_j, \phi_j\}$ specifying model $\mathscr{M}_j$ for $j = 1, \ldots, J$.

- We now obtain analytical closed forms for $p(\beta, \Sigma \mid \mathscr{D}, \mathscr{M}_j)$ for each j, and use BPS of predictive densities to evaluate the stacked posterior distribution.

- For each subset of the data, we compute the stacking weights $\{z_{k,j}\}$ as

$$\max_{z_k \in S_1^J} \frac{1}{n_k} \sum_{i=1}^{n_k} \log \sum_{j=1}^{J} z_{k,j} p\left(Y_{k,i} \mid \mathscr{D}_{k,[-l]}, \mathscr{M}_j\right), \tag{9}$$

where $Y_{k,i}$ is the i-th row of $Y_{k,[l]} \in \mathscr{D}_{k,[l]}$ (the l-th fold within the k-th dataset), and $l = 1, \ldots, L$, with L number of folds for K-fold cross-validation.

- Since $p\left(Y_{k,i} \mid \mathscr{D}_{k,[-l]}, \mathscr{M}_j\right) = \mathrm{T}_{1,q}(Y_{k,i} \mid \nu^\star_{[-l]}, \mu^\star_i, V^\star_i, \Psi^\star_{[-l]})$ available in closed-form, computations are very efficient.

- For each of $k = 1, \ldots, K$ dataset, $\mathscr{D}_k$, we compute:
  - stacked posterior: $\hat{p}(\cdot \mid \mathscr{D}_k) = \sum_{j=1}^{J} \hat{z}_{k,j}\, p(\cdot \mid \mathscr{D}_k, \mathscr{M}_j)$ for $j = 1, \ldots, J$
  - stacking weights: $\hat{z}_k = \{\hat{z}_{k,j}\}_{j=1,\ldots,J}$.

- For inference on the full spatial dataset, we apply BPS a second time, which is equivalent to solving the following convex optimization problem:

$$\max_{w \in S_1^K} \frac{1}{n} \sum_{i=1}^{n} \log \sum_{k=1}^{K} w_k \hat{p}\left(Y_i \mid \mathscr{D}_{k,[-l]}\right) = \max_{w \in S_1^K} \frac{1}{n} \sum_{i=1}^{n} \log \sum_{k=1}^{K} w_k \sum_{j=1}^{J} \hat{z}_{k,j} p\left(Y_{k,i} \mid \mathscr{D}_{k,[-l]}, \mathscr{M}_j\right). \tag{10}$$

- Once the stacking weights $\hat{w} = \{\hat{w}_k\}_{k=1,\ldots,K}$ are obtained, estimation of any posterior or posterior predictive distribution of interest is achieved as

$$\hat{p}(\cdot \mid \mathscr{D}) = \sum_{k=1}^{K} \hat{w}_k \sum_{j=1}^{J} \hat{z}_{k,j}\, p\left(\cdot \mid \mathscr{D}_k, \mathscr{M}_j\right) . \tag{11}$$

- Given the two sets of weights derived from double stacking, the estimated full posterior distribution is a mixture of finite mixtures.

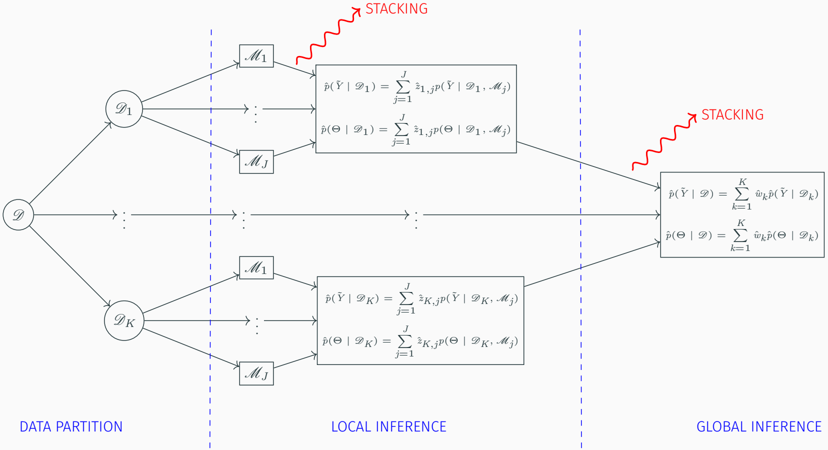
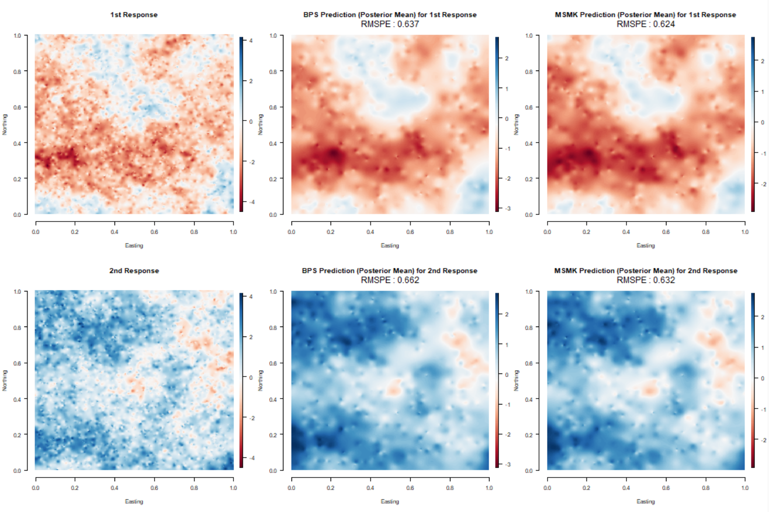


Figure 1: Double Bayesian predictive stacking approach representation

Findings

- Implementing the methodology, we also developed the spBPS R package, available on CRAN, and provide functions to automately perform double BPS for hierarchical model in Equation (5).
- Reproducible programs can be found on GitHub, in the repository lucapresicce/Bayesian-Transfer-Learning-for-GeoAI
- Simulations, and data analyses were executed on a laptop running an Intel Core I7-8750H CPU with 5 available cores for parallel computation, and 16 Gb of RAM.
- We investigate theoretical computational complexity, memory management issues, efficient parameter sampling schemes, and sensitivity on data shards K.



👉 We conduct multiple simulation experiments on synthetic data generated from (5) to evaluate inferential performance while underscoring comparisons with existing approaches.

Figure 2: from left to right: comparison between the true generated response surfaces, the surfaces predicted from BPS and SMK (posterior mean), with RMSPE. For $n = 5000$, $K = 10$.

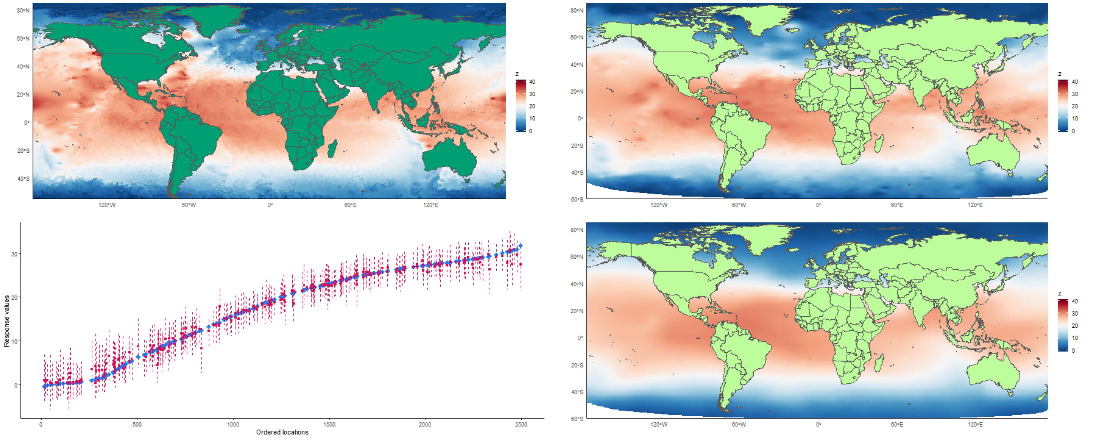


Figure 3: from left to right: comparison between training (top left panel), test (top right panel), and predicted surface (bottom right panel). In addition, the empirical coverage for the response (bottom left panel). Results for $K = 2,000$.

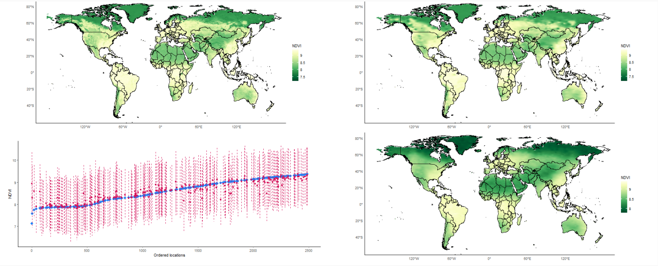


Figure 4: from left to right: comparison between training (top left panel), test (top right panel), and predicted surface (bottom right panel), for NDVI response. In addition, the empirical coverage for outcomes (bottom left panel). Results for $K = 2,000$.

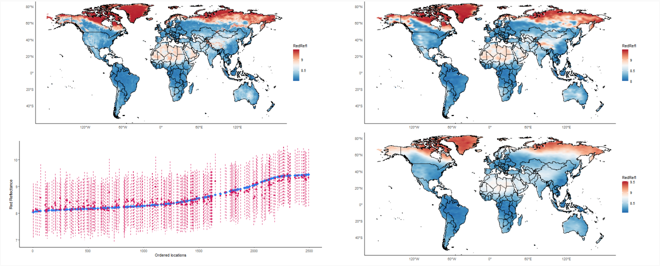


Figure 5: from left to right: comparison between training (top left panel), test (top right panel), and predicted surface (bottom right panel), for red reflectance response. In addition, the empirical coverage for outcomes (bottom left panel). Results for $K = 2,000$.

Conclusion

- Existing approaches rely on iterative algorithms for estimating weakly identifiable parameters [18, 14]. We, instead, propose a transfer learning double-stacking approach:
  - we first obtain stacked inference for each of the parameters within each subset;
  - then we assimilate inference across subsets by a second stacking algorithm.
- We harness analytical closed-form distribution theory to deliver inference. This is possible by “fixing” spatial correlation kernel parameters.
- Our approach carries out fast and exact inference based on the matrix-variate distributions and assimilates the inference using Bayesian predictive stacking [15, 17].

Ongoing works

joint work with Sudipto Banerjee (UCLA).

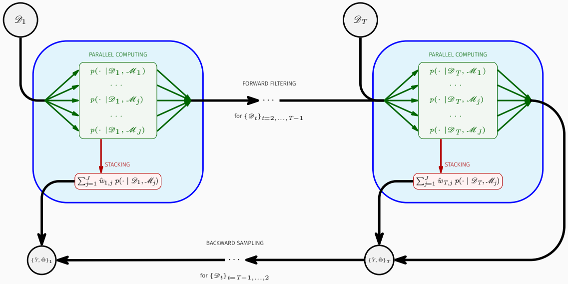


Figure 6: Data shards dynamics dependences representation

☛ Scalable Dynamic linear model strategy to manage massive online streaming of multivariate spatiotemporal datasets, ensures an adaptive framework for modeling Markovian dependence through time.

☛ Conjugate inference through predictive stacking, combining sequential and parallel processing of temporal slices: each unit passes assimilated information forward, then back-smoothed.

joint work with Federico Castelletti (UCSC)

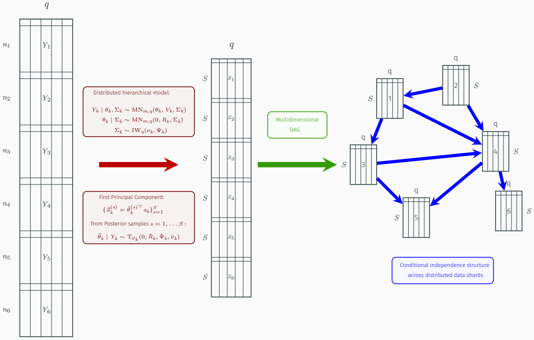


Figure 7: From gathering distributed settings to model the conditional dependencies across datasets.

- Graph-based methodology to broadcast propagation within Bayesian distributed learning paradigms.

- Conditional dependencies structure learning for multivariate analysis of correlated data, e.g., tensor of spatiotemporal data.

References

[1] Banerjee, S., B. P. Carlin, and A. E. Gelfand (2015). *Hierarchical modeling and analysis for Spatial Data*. Chapman & Hall/CRC.

[2] Cao, J., M. Kang, F. Jimenez, H. Sang, F. T. Schaefer, and M. Katzfuss (2023, 23–29 Jul). Variational sparse inverse cholesky approximation for latent Gaussian processes via double kullback-leibler minimization. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett (Eds.), *Proceedings of the 40th International Conference on Machine Learning*, Volume 202 of *Proceedings of Machine Learning Research*, pp. 3559–3576. PMLR.

[3] Cressie, N. (1993). *Statistics For Spatial Data, Revised Edition*. John Wiley & Sons.

[4] Cressie, N. and C. K. Wikle (2011). *Statistics for spatio-temporal data*. Wiley.

[5] Finley, A. O., A. Datta, B. D. Cook, D. C. Morton, H.-E. Andersen, and S. Banerjee (2019). Efficient algorithms for Bayesian nearest neighbor Gaussian processes. *Journal of Computational and Graphical Statistics 28*(2), 401–414.

[6] Fu, A., B. Narasimhan, and S. Boyd (2020). CVXR: An R package for disciplined convex optimization. *Journal of Statistical Software 94*(14), 1–34.

[7] Gelfand, A. E., P. Diggle, P. Guttorp, and M. Fuentes (2010). *Handbook of Spatial Statistics*. Taylor & Francis.

[8] Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association 102*(477), 359–378.

[9] Grant, M. and S. Boyd (2008). Graph implementations for nonsmooth convex programs. In V. Blondel, S. Boyd, and H. Kimura (Eds.), *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pp. 95–110. Springer-Verlag Limited.

[10] Lindgren, F., H. Rue, and J. Lindström (2011, 08). 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.

[11] Ren, Q., S. Banerjee, A. O. Finley, and J. S. Hodges (2011). Variational bayesian methods for spatial data analysis. *Computational Statistics & Data Analysis 55*(12), 3197–3217.

[12] Rue, H., S. Martino, and N. Chopin (2009, 04). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society Series B: Statistical Methodology 71*(2), 319–392.

[13] Stein, M. L. (1999). *Interpolation of spatial data*. Springer Series in Statistics. New York: Springer-Verlag. Some theory for Kriging.

[14] Tang, W., L. Zhang, and S. Banerjee (2021). On identifiability and consistency of the nugget in Gaussian spatial process models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 83*(5), 1044–1070.

[15] Wolpert, D. H. (1992). Stacked generalization. *Neural Networks 5*(2), 241–259.

[16] Wu, L., G. Pleiss, and J. P. Cunningham (2022, 17–23 Jul). Variational nearest neighbor Gaussian process. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato (Eds.), *Proceedings of the 39th International Conference on Machine Learning*, Volume 162 of *Proceedings of Machine Learning Research*, pp. 24114–24130. PMLR.

[17] Yao, Y., A. Vehtari, D. Simpson, and A. Gelman (2018). Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Analysis 13*(3), 917–1007.

[18] Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association 99*(465), 250–261.

- Let $Y_{n\times q}$ be an $n \times q$ random matrix that is endowed with a probability law from the matrix-normal distribution, $\mathrm{MN}(M, V, U)$, with probability density function

$$p(Y \mid M, V, U) = \frac{\exp\left[-\frac{1}{2}\,\mathrm{tr}\left\{U^{-1}(Y-M)^{\top}V^{-1}(Y-M)\right\}\right]}{(2\pi)^{\frac{np}{2}}\mid U\mid^{\frac{n}{2}}\mid V\mid^{\frac{q}{2}}}, \tag{12}$$

- where $\mathrm{tr}(\cdot)$ is the trace operator on a square matrix.

- The matrix-variate Gaussian distribution is often parametrized as follows: M is the mean matrix, and V and U are the $n \times n$ row-covariance and $q \times q$ column covariance matrices, respectively.