Linking GRF and GMRF Through SPDE v1.0_20240403

$\{y(\mathbf{s}_i);\ \mathbf{s}_i \in \mathscr{D} \subseteq \mathbb{R}^d\} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

♦ **Gaussian processes** (GP) and **Gaussian random fields** (GRF) are used interchangeably in two-dimensional space $d = 2$. The mean vector is $\boldsymbol{\mu} := \big(\mu(\mathbf{s}_i)\big)$, and the covariance matrix is $\boldsymbol{\Sigma} := \Big(C(\mathbf{s}_i, \mathbf{s}_j)\Big) \succ 0$.

Weak Stationarity

- The mean function $\mu(\cdot)$ is constant.
- The covariance function $C(\cdot\,,\cdot)$ depends only on the relative position of two locations, not on their absolute positions in the input space. The process is **isotropic** if $C(\cdot\,,\cdot)$ depends only on the magnitude of distance, regardless of direction.

`SpatialDE` fits response GP models to the gene expression levels of each gene across a finite set of N spatial locations. The spatial component of $\boldsymbol{\Sigma} = \sigma_s^2 \cdot \mathbf{K} + \tau^2 \cdot \mathbf{I}$ employs the squared exponential kernel function, which is also the **Matérn kernel** with $\nu \to +\infty$. However, it is challenged by computational complexity $\mathcal{O}(N^3)$.

$$K_{ij} = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}\delta_{ij}}{l} \right)^{\nu} \cdot K_\nu \left(\frac{\sqrt{2\nu}\delta_{ij}}{l} \right) \text{ and denote } \kappa = \frac{\sqrt{2\nu}}{l}\,,\ \nu, l > 0\,,$$

where $K_\nu(\cdot)$ is the modified Bessel function of the second kind and $\delta_{ij} = \|\mathbf{s}_i - \mathbf{s}_j\|$.

`nnSVG` proposes four modifications:

1. It uses the exponential kernel function, which is also the Matérn kernel with $\nu = 1/2$;
2. It uses a gene-specific length scale l in the kernel;
3. It uses a scalable approximation of $\boldsymbol{\Sigma}$ based on **nearest-neighbor Gaussian processes** (NNGP) in $\mathcal{O}(Nm^3)$, where m is the number of NNs, resulting in a sparse precision matrix $\tilde{\boldsymbol{\Omega}} = \tilde{\boldsymbol{\Sigma}}^{-1}$;
4. The degrees of freedom of the asymptotic chi-squared distribution are 2 instead of 1 for significance testing.

♦ **Markov random fields** (MRF) exhibit the Markov property and can implement the MCMC sampling or fast direct numerical methods, typically in $\mathcal{O}(N^{3/2})$.

$$f\Big(y(\mathbf{s}_i) = y_i \mid y(\mathbf{s}_j) = y_j,\ j \neq i\Big) = f\Big(y(\mathbf{s}_i) = y_i \mid y(\mathbf{s}_j) = y_j,\ j \in \partial i\Big) \text{ for } i = 1, \cdots, N,$$

where ∂i are the indicators of a set of neighbors to location $\mathbf{s}_i$.

♦ The **integrated nested Laplace approximation** (INLA) method has gained popularity in spatial data analysis primarily due to its innovative approach to approximating the posterior distributions of latent Gaussian models:

1. The GMRF provides a computationally efficient and flexible framework for modeling spatial dependencies;
2. The **stochastic partial differential equation** (SPDE) approach extends this flexibility to continuous spatial processes and enables INLA to handle non-stationary processes.

♦ Hierarchical GP Regression Model

$\mathbf{y}(\mathbf{s}) := \big(y(\mathbf{s}_1), \cdots, y(\mathbf{s}_N)\big)^\top = \boldsymbol{\mu}(\mathbf{s}) + \mathbf{b}(\mathbf{s}) + \boldsymbol{\epsilon}(\mathbf{s})$ for each gene

$\mathbf{b}(\mathbf{s}) \sim \mathcal{N}_N(\mathbf{0},\, \sigma_s^2 \cdot \mathbf{K}) \perp\!\!\!\perp \boldsymbol{\epsilon}(\mathbf{s}) \sim \mathcal{N}_N(\mathbf{0},\, \tau^2 \cdot \mathbf{I})$ A latent spatial process capturing spatial dependence

- Latent GP: $\mathcal{N}_N\big(\mathbf{y}(\mathbf{s}) \mid \boldsymbol{\mu}(\mathbf{s}) + \mathbf{b}(\mathbf{s}),\, \tau^2 \cdot \mathbf{I}\big) \times \mathcal{N}_N\big(\mathbf{b}(\mathbf{s}) \mid \mathbf{0},\, \sigma_s^2 \cdot \mathbf{K}\big) \times \pi(\sigma_s^2,\, l)$
- Response GP: $\mathcal{N}_N\big(\mathbf{y}(\mathbf{s}) \mid \boldsymbol{\mu}(\mathbf{s}),\, \sigma_s^2 \cdot \mathbf{K} + \tau^2 \cdot \mathbf{I}\big) \times \pi(\sigma_s^2,\, l)$

♦ SPDE Representation of a Continuous GRF

A GRF $y(\mathbf{s})$ with the Matérn covariance $\boldsymbol{\Sigma} = \sigma_s^2(\nu, \kappa, d) \cdot \mathbf{K}$ is a solution to the SPDE:

$$\big(\kappa^2(\mathbf{s}) - \Delta\big)^{\frac{\alpha}{2}}\big(\tau(\mathbf{s}) \cdot y(\mathbf{s})\big) = \epsilon(\mathbf{s}), \quad \epsilon(\mathbf{s}) \sim \mathcal{N}(0,\, 1), \quad \alpha = \nu + \frac{d}{2} \geqslant 2, \quad \Delta f(\mathbf{s}) := \nabla^2 f(\mathbf{s}) := \sum_{i=1}^{d} \frac{\partial^2}{\partial s_i^2} f(\mathbf{s})$$

Scaling

◇ <u>Non-stationarity</u> is achieved when $\kappa^2(\mathbf{s})$ and/or $\tau(\mathbf{s})$ are non-constant over $\mathbf{s} \in \mathcal{D} \subseteq \mathbb{R}^d$.

$$\ln \kappa^2(\mathbf{s}) = \sum_i \beta_i^{(\kappa^2)} \cdot B_i^{(\kappa^2)}(\mathbf{s}) \qquad \text{and} \qquad \ln \tau(\mathbf{s}) = \sum_i \beta_i^{(\tau)} \cdot B_i^{(\tau)}(\mathbf{s})$$

♦ As for a GMRF, the result $y(\mathbf{s}) = \sum_{q=1}^{N} w_q \cdot \psi_q(\mathbf{s})$ is a basis function representation

with piecewise linear basis functions $\psi_q(\mathbf{s})$, and Gaussian weights w_q with Markov dependences determined by a general triangulation of the domain $\mathcal{D}$ (finite element methods).

Lindgren, F. (2023, July 18-21). *Embedding stochastic PDEs in Bayesian spatial statistics software* [Keynote address]. Spatial Statistics, Boulder, Colorado, United States. https://www.maths.ed.ac.uk/~flindgre/talks/Lindgren_Boulder2023.pdf

Lindgren, F., Rue, H., & Lindström, J. (2011). 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*, 423-498. DOI: 10.1111/j.1467-9868.2011.00777.x

Svensson, V., Teichmann, S. A., & Stegle, O. (2018). SpatialDE: Identification of spatially variable genes. *Nature Methods*, *15*, 343-346. DOI: 10.1038/nmeth.4636

Weber, L. M., Saha, A., Datta, A., Hansen, K. D., & Hicks, S. C. (2023). nnSVG for the scalable identification of spatially variable genes using nearest-neighbor Gaussian processes. *Nature Communications*, *14*, 1-12. DOI: 10.1038/s41467-023-39748-z

Zhang, L. (2018). *Nearest Neighbor Gaussian Processes (NNGP) Based Models in Stan*. https://mc-stan.org/users/documentation/case-studies/nngp.html