Appendix: Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets

Abhirup Datta, Sudipto Banerjee, Andrew O. Finley and Alan E. Gelfand

A1 Densities on directed acyclic graphs

We will show that if $\mathcal{G} = (\mathcal{S}, N_{\mathcal{S}})$ is acyclic then $\tilde{p}(\mathbf{w}_{\mathcal{S}})$ defined in (3) corresponds to a true density over $\mathcal{S}$. For any directed acyclic graph, there exists a node with zero in-degree i.e. no directed edge pointing towards it. We denote this node by $\mathbf{s}_{\pi(1)}$ This means $\mathbf{s}_{\pi(1)}$ does not belong to the neighbor set of any other location in $\mathcal{S}$. The only term where it appears on the right hand side of (2) is $p(\mathbf{w}(\mathbf{s}_{\pi(1)} \,|\, \mathbf{w}_{N(\mathbf{s}_{\pi(1)})})$ which integrates out to one with respect to $d\mathbf{w}(\mathbf{s}_{\pi(1)})$. We now have a new acyclic directed graph $\mathcal{G}_1$ obtained by removing vertex $\mathbf{s}_{\pi(1)}$ and its directed edges from $\mathcal{G}$. Now we can find a new vertex $\mathbf{s}_{\pi(2)}$ with zero out-degree in $\mathcal{G}_1$ and continue as before to get a permutation $\pi(1), \pi(2), \ldots, \pi(k)$ of $1, 2, \ldots, k$ such that

$$\int \prod_{i=1}^{k} p(\mathbf{w}(\mathbf{s}_i) \,|\, \mathbf{w}_{N(\mathbf{s}_i)}) d\mathbf{w}(\mathbf{s}_{\pi(1)}) d\mathbf{w}(\mathbf{s}_{\pi(2)}) \ldots d\mathbf{w}(\mathbf{s}_{\pi(k)}) = 1$$

An easy application of Fubini's theorem now ensures that this is a proper joint density.

A2 Properties of $\tilde{\mathbf{C}}_{\mathcal{S}}^{-1}$

If $p(\mathbf{w}_{\mathcal{S}}) = N(\mathbf{w}_{\mathcal{S}} \,|\, \mathbf{0}, \mathbf{C}_{\mathcal{S}})$, then $\mathbf{w}(\mathbf{s}_i) \,|\, \mathbf{w}_{N(\mathbf{s}_i)} \sim N(\mathbf{B}_{\mathbf{s}_i}\mathbf{w}_{N(\mathbf{s}_i)}, \mathbf{F}_{\mathbf{s}_i})$, where $\mathbf{B}_{\mathbf{s}_i}$ and $\mathbf{F}_{\mathbf{s}_i}$ are defined in (3). So, the likelihood in (2) is proportional to

$$\frac{1}{\prod_{i=1}^{k}\sqrt{\det(\mathbf{F}_{\mathbf{s}_i})}} \exp\left(-\frac{1}{2}\sum_{i=1}^{k}(\mathbf{w}(\mathbf{s}_i) - \mathbf{B}_{\mathbf{s}_i}\mathbf{w}_{N(\mathbf{s}_i)})'\mathbf{F}_{\mathbf{s}_i}^{-1}(\mathbf{w}(\mathbf{s}_i) - \mathbf{B}_{\mathbf{s}_i}\mathbf{w}_{N(\mathbf{s}_i)})\right)$$

For any matrix $\mathbf{A}$, let $\mathbf{A}[, j : j']$ denote the submatrix formed using columns j to j' where $j < j'$. For $j = 1, 2, \ldots, k$, we define $q \times q$ blocks $\mathbf{B}_{\mathbf{s}_i,j}$ as

$$\mathbf{B}_{\mathbf{s}_i,j} = \begin{cases} \mathbf{I}_q \text{ if } j = i; \\ -\mathbf{B}_{\mathbf{s}_i}[, (l-1)q + 1 : lq] \text{ if } \mathbf{s}_j = N(\mathbf{s}_i)(l) \text{ for some } l; \\ \mathbf{O} \text{ otherwise,} \end{cases}$$

where, for any location $\mathbf{s}$, $N(\mathbf{s})(l)$ is the l-th neighbor of $\mathbf{s}$. So, $\mathbf{w}_{\mathbf{s}_i} - \mathbf{B}_{\mathbf{s}_i}\mathbf{w}_{N(\mathbf{s}_i)} = \mathbf{B}^*_{\mathbf{s}_i}\mathbf{w}_{\mathcal{S}}$, where $\mathbf{B}^*_{\mathbf{s}_i} = [\mathbf{B}_{\mathbf{s}_i,1}, \mathbf{B}_{\mathbf{s}_i,2}, \ldots, \mathbf{B}_{\mathbf{s}_i,k}]$ is $q \times kq$ and sparse with at most $m + 1$ non-zero blocks. Then,

$$\sum_{i=1}^{k}(\mathbf{w}(\mathbf{s}_i) - \mathbf{B}_{\mathbf{s}_i}\mathbf{w}_{N(\mathbf{s}_i)})'\mathbf{F}_{\mathbf{s}_i}^{-1}(\mathbf{w}(\mathbf{s}_i) - \mathbf{B}_{\mathbf{s}_i}\mathbf{w}_{N(\mathbf{s}_i)}) = \sum_{i=1}^{k}\mathbf{w}'_{\mathcal{S}}(\mathbf{B}^*_{\mathbf{s}_i})'\mathbf{F}_{\mathbf{s}_i}^{-1}\mathbf{B}^*_{\mathbf{s}_i}\mathbf{w}_{\mathcal{S}} = \mathbf{w}'_{\mathcal{S}}\mathbf{B}'_{\mathcal{S}}\mathbf{F}_{\mathcal{S}}^{-1}\mathbf{B}_{\mathcal{S}}\mathbf{w}_{\mathcal{S}}\ ,$$

where $\mathbf{F} = diag(\mathbf{F}_{\mathbf{s}_1}, \mathbf{F}_{\mathbf{s}_2}, \ldots, \mathbf{F}_{\mathbf{s}_k})$ and $\mathbf{B}_{\mathcal{S}} = ((\mathbf{B}^*_{\mathbf{s}_1})', (\mathbf{B}^*_{\mathbf{s}_2})', \ldots, (\mathbf{B}^*_{\mathbf{s}_k})')'$. So, we have:

$$(\tilde{\mathbf{C}}_{\mathcal{S}})^{-1} = \mathbf{B}'_{\mathcal{S}}\mathbf{F}_{\mathcal{S}}^{-1}\mathbf{B}_{\mathcal{S}} \tag{A1}$$

From the form of $\mathbf{B}_{\mathbf{s}_i,j}$, it is clear that $\mathbf{B}_{\mathcal{S}}$ is sparse and lower triangular with ones on the diagonals. So, $\det(\mathbf{B}_{\mathcal{S}}) = 1$, $\det((\mathbf{B}'_{\mathcal{S}}\mathbf{F}_{\mathcal{S}}^{-1}\mathbf{B}_{\mathcal{S}})^{-1}) = \prod\det(\mathbf{F}_{\mathbf{s}_i})$ and (2) simplifies to $N(\mathbf{w}_{\mathcal{S}} \,|\, \mathbf{0}, \tilde{\mathbf{C}}_{\mathcal{S}})$.

Let $\tilde{\mathbf{C}}_{\mathcal{S}}^{ij}$ denote the $(i, j)^{th}$ block of $\tilde{\mathbf{C}}_{\mathcal{S}}^{-1}$. Then from equation (A1) we see that for $i < j$,

$\tilde{\mathbf{C}}_{\mathcal{S}}^{ij} = \sum_{l=j}^{k}(\mathbf{B}^*_{\mathbf{s}_l,i})'\mathbf{F}_{\mathbf{s}_l}^{-1}\mathbf{B}^*_{\mathbf{s}_l,j}$. So, $\tilde{\mathbf{C}}_{\mathcal{S}}^{ij}$ is non-zero only if there exists at least one location $\mathbf{s}_l$ such that $\mathbf{s}_i \in N(\mathbf{s}_l)$ and $\mathbf{s}_j$ is either equal to $\mathbf{s}_l$ or is in $N(\mathbf{s}_l)$. Since every neighbor set has at most m elements, there are at most $km(m+1)/2$ such pairs (i,j). This demonstrates the sparsity of $\tilde{\mathbf{C}}_{\mathcal{S}}^{-1}$ for $m \ll k$.

A3 Kolmogorov Consistency for NNGP

Let $\{\mathbf{w}(\mathbf{s}) \,|\, \mathbf{s} \in \mathcal{D}\}$ be a random process over some domain $\mathcal{D}$ with density p and let $\tilde{p}(\mathbf{w}_{\mathcal{S}})$ be a probability density for observations over a fixed finite set $\mathcal{S} \subset \mathcal{D}$. The conditional density $\tilde{p}(\mathbf{w}_{\mathcal{U}} \,|\, \mathbf{w}_{\mathcal{S}})$ for any finite set $\mathcal{U} \subset \mathcal{D}$ outside of $\mathcal{S}$ is defined in (4).

We will first show that for every finite set $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ in $\mathcal{D}$, $n \in \{1, 2, \ldots\}$ and for every permutation $\pi(1), \pi(2), \ldots, \pi(n)$ of $1, 2, \ldots, n$ we have,

$$\tilde{p}\left(\mathbf{w}(\mathbf{v}_1), \mathbf{w}(\mathbf{v}_2), \ldots, \mathbf{w}(\mathbf{v}_n)\right) = \tilde{p}\left(\mathbf{w}(\mathbf{v}_{\pi(1)}), \mathbf{w}(\mathbf{v}_{\pi(2)}), \ldots, \mathbf{w}(\mathbf{v}_{\pi(n)})\right) \ .$$

. We begin by showing that for any finite set $\mathcal{V}$, the expression given in (5) is a proper density. Let $\mathcal{U} = \mathcal{V} \setminus \mathcal{S}$. Since $\mathcal{V} \cup (\mathcal{S} \setminus \mathcal{V}) = \mathcal{S} \cup \mathcal{U}$, we obtain

$$\begin{aligned}&\int \tilde{p}(\mathbf{w}_{\mathcal{V}}) \prod_{\mathbf{v}_i \in \mathcal{V}} d(\mathbf{w}(\mathbf{v}_i)) = \int \tilde{p}(\mathbf{w}_{\mathcal{U}} \,|\, \mathbf{w}_{\mathcal{S}}) \tilde{p}(\mathbf{w}_{\mathcal{S}}) \prod_{\mathbf{v}_i \in \mathcal{U}} d(\mathbf{w}(\mathbf{v}_i)) \prod_{\mathbf{s}_i \in \mathcal{S}} d(\mathbf{w}(\mathbf{s}_i)) \\ &= \int \tilde{p}(\mathbf{w}_{\mathcal{S}}) \left(\int \tilde{p}(\mathbf{w}_{\mathcal{U}} \,|\, \mathbf{w}_{\mathcal{S}}) \prod_{\mathbf{v}_i \in \mathcal{U}} d(\mathbf{w}(\mathbf{v}_i)) \right) \prod_{\mathbf{s}_i \in \mathcal{S}} d(\mathbf{w}(\mathbf{s}_i)) = \int \tilde{p}(\mathbf{w}_{\mathcal{S}}) \prod_{\mathbf{s}_i \in \mathcal{S}} d(\mathbf{w}(\mathbf{s}_i)) = 1\end{aligned}$$

Note that $\mathcal{S}$ is fixed. Therefore, the expression for the joint density of $\mathbf{w}_{\mathcal{V}}$ depends only on the the neighbor sets $N(\mathbf{v}_i)$ for $\mathbf{v}_i \in \mathcal{U}$. So the NNGP density for $\mathcal{V}$ is invariant under any permutation of locations inside $\mathcal{V}$.

We now prove that for every location $\mathbf{v}_0 \in \mathcal{D}$, we have, $\tilde{p}(\mathbf{w}_{\mathcal{V}}) = \int \tilde{p}(\mathbf{w}_{\mathcal{V} \cup \{\mathbf{v}_0\}}) d(\mathbf{w}(\mathbf{v}_0))$. let $\mathcal{V}_1 = \mathcal{V} \cup \{\mathbf{v}_0\}$. We split the proof into two cases. If $\mathbf{v}_0 \in \mathcal{S}$, then using the fact

$\mathcal{V}_1 \setminus \mathcal{S} = \mathcal{V} \setminus \mathcal{S} = \mathcal{U}$, we obtain

$$\begin{aligned}\int \tilde{p}(\mathbf{w}_{\mathcal{V}_1}) d(\mathbf{w}(\mathbf{v}_0)) &= \int \tilde{p}(\mathbf{w}_{\mathcal{S}}) \tilde{p}(\mathbf{w}_{\mathcal{V}_1 \setminus \mathcal{S}} \,|\, \mathbf{w}_{\mathcal{S}}) \prod_{\mathbf{s}_i \in \mathcal{S} \setminus \mathcal{V}_1} d(\mathbf{w}(\mathbf{s}_i)) d(\mathbf{w}(\mathbf{v}_0) \\ &= \int \tilde{p}(\mathbf{w}_{\mathcal{S}}) \tilde{p}(\mathbf{w}_{\mathcal{V} \setminus \mathcal{S}} \,|\, \mathbf{w}_{\mathcal{S}}) \prod_{\mathbf{s}_i \in \mathcal{S} \setminus \mathcal{V}} d(\mathbf{w}(\mathbf{s}_i)) = \tilde{p}(\mathbf{w}_{\mathcal{U}}) \; .\end{aligned}$$

If $\mathbf{v}_0 \notin \mathcal{S}$, then $\mathbf{w}(\mathbf{v}_0)$ does not appear in the neighborhood set of any other term. So, $p(\mathbf{w}(\mathbf{v}_0) \,|\, \mathbf{w}_{\mathcal{S}})$ integrates to one with respect to $d(\mathbf{w}(\mathbf{v}_0))$. The result now follows from $\int p(\mathbf{w}_{\mathcal{V}_1} \,|\, \mathbf{w}_{\mathcal{S}}) d(\mathbf{w}(\mathbf{v}_0)) = p(\mathbf{w}_{\mathcal{V}} \,|\, \mathbf{w}_{\mathcal{S}})$.

A4 Properties of NNGP

Standard Gaussian conditional distribution facts reveal that the conditional distribution $\mathbf{w}(\mathbf{u}_i) \,|\, \mathbf{w}_{\mathcal{S}} \sim N(\mathbf{B}_{\mathbf{u}_i} \mathbf{w}_{N(\mathbf{u}_i)}, \mathbf{F}_{\mathbf{u}_i})$ where $\mathbf{B}_{\mathbf{u}_i}$ and $\mathbf{F}_{\mathbf{u}_i}$ be defined analogous to (3) based on the neighbor sets $N(\mathbf{u}_i)$. From (4), we see that

$$\tilde{p}(\mathbf{w}_{\mathcal{U}} \,|\, \mathbf{w}_{\mathcal{S}}) = \frac{1}{\prod_{i=1}^{r} \sqrt{\det(\mathbf{F}_{\mathbf{u}_i})}} \exp\left(-\frac{1}{2} \sum_{i=1}^{r} (\mathbf{w}(\mathbf{u}_i) - \mathbf{B}_{\mathbf{u}_i} \mathbf{w}_{N(\mathbf{u}_i)})' \mathbf{F}_{\mathbf{u}_i}^{-1} (\mathbf{w}(\mathbf{u}_i) - \mathbf{B}_{\mathbf{u}_i} \mathbf{w}_{N(\mathbf{u}_i)}) \right)$$

It then follows that $\tilde{p}(\mathbf{w}_{\mathcal{U}} \,|\, \mathbf{w}_{\mathcal{S}}) \sim N(\mathbf{B}_{\mathcal{U}} \mathbf{w}_{\mathcal{S}}, \mathbf{F}_{\mathcal{U}})$ where $\mathbf{B}_{\mathcal{U}} = (\mathbf{B}'_{\mathbf{u}_1}, \mathbf{B}'_{\mathbf{u}_2}, \ldots, \mathbf{B}'_{\mathbf{u}_r})'$ and $\mathbf{F}_{\mathcal{U}} = diag(\mathbf{F}_{\mathbf{u}_1}, \mathbf{F}_{\mathbf{u}_2}, \ldots, \mathbf{F}_{\mathbf{u}_r})$. Since each row of $\mathbf{B}_{\mathcal{U}}$ has at most m non-zero entries, $\mathbf{B}_{\mathcal{U}}$ is sparse for $m \ll k$.

As the nearest neighbor densities of $\mathbf{w}_{\mathcal{S}}$ and $\mathbf{w}_{\mathcal{U}} \,|\, \mathbf{w}_{\mathcal{S}}$ for every finite $\mathcal{U}$ outside $\mathcal{S}$ are Gaussian, all finite dimensional realizations of an NNGP process will be Gaussian. Let $\mathbf{v}_1$ and $\mathbf{v}_2$ be any two locations in $\mathcal{D}$ and let $\widetilde{E}$ and $\widetilde{Cov}$ denote, respectively, the expectation and covariance operator for a NNGP. Then, if $\mathbf{v}_1 = \mathbf{s}_i$ and $\mathbf{v}_2 = \mathbf{s}_j$ are both in $\mathcal{S}$ then we obviously have $\widetilde{Cov}(\mathbf{w}(\mathbf{v}_1), \mathbf{w}(\mathbf{v}_2)\,|\,\boldsymbol{\theta}) = \tilde{\mathbf{C}}_{\mathbf{s}_i,\mathbf{s}_j}$. If $\mathbf{v}_1$ is outside $\mathcal{S}$ and $\mathbf{v}_2 = \mathbf{s}_j$, then

$$\begin{aligned} \widetilde{Cov}(\mathbf{w}(\mathbf{v}_1), \mathbf{w}(\mathbf{v}_2)\,|\,\boldsymbol{\theta}) &= \widetilde{E}(\widetilde{Cov}(\mathbf{w}(\mathbf{v}_1), \mathbf{w}(\mathbf{v}_2)\,|\,\mathbf{w}_{\mathcal{S}}, \boldsymbol{\theta})) + \widetilde{Cov}(\widetilde{E}(\mathbf{w}(\mathbf{v}_1)), \widetilde{E}(\mathbf{w}(\mathbf{v}_2))\,|\,\mathbf{w}_{\mathcal{S}}, \boldsymbol{\theta})) \\ \therefore \tilde{\mathbf{C}}(\mathbf{v}_1, \mathbf{v}_2\,|\,\boldsymbol{\theta}) &= 0 + \widetilde{Cov}(\mathbf{B}_{\mathbf{v}_1}\mathbf{w}_{N(\mathbf{v}_1)}, \mathbf{w}(\mathbf{s}_j)\,|\,\boldsymbol{\theta}) = \mathbf{B}_{\mathbf{v}_1}\tilde{\mathbf{C}}_{N(\mathbf{v}_1),\mathbf{s}_j} \end{aligned}$$

If both $\mathbf{v}_1$ and $\mathbf{v}_2$ are outside $\mathcal{S}$, then $\tilde{\mathbf{C}}(\mathbf{v}_1, \mathbf{v}_2\,|\,\boldsymbol{\theta}) = \delta(\mathbf{v}_1 = \mathbf{v}_2)\mathbf{F}_{\mathbf{v}_1} + \mathbf{B}_{\mathbf{v}_1}\tilde{\mathbf{C}}_{N(\mathbf{v}_1),N(\mathbf{v}_2)}\mathbf{B}'_{\mathbf{v}_2}$, which yields (7).

For any two set of locations A and B, let $||A, B||$ denote the pairwise Euclidean distance matrix. Let $\mathcal{Z}_1$ denote set of all points $\mathbf{v}$ such that $\mathbf{v}$ is equidistant from any two points in $\mathcal{S}$. Since $\mathcal{S}$ is finite, the set $\mathcal{Z}_2 = (\mathcal{Z}_1 \times \mathcal{Z}_1) \cup \{(\mathbf{v}, \mathbf{v})\,|\,\mathbf{v} \in \mathcal{D}\}$ has Lebesgue measure zero in the Euclidean domain $\Re^d \times \Re^d$. We will show that $\tilde{\mathbf{C}}(\mathbf{v}_1, \mathbf{v}_2\,|\,\boldsymbol{\theta})$ is continuous for any pair $(\mathbf{v}_1, \mathbf{v}_2)$ in $\mathcal{D} \times \mathcal{D} \setminus \mathcal{Z}_2$. Observe that for any pair of points $(\mathbf{v}_1, \mathbf{v}_2)$ in $\mathcal{D} \times \mathcal{D} \setminus \mathcal{Z}_2$, it is easy to verify that $\lim\limits_{\mathbf{h}_i \to 0} ||(\mathbf{v}_i + \mathbf{h}_i, N(\mathbf{v}_i + \mathbf{h}_i)|| \to ||\mathbf{v}_i, N(\mathbf{v}_i)||$, for $i = 1, 2$, and $\lim\limits_{\mathbf{h}_1 \to 0, \mathbf{h}_2 \to 0} ||N(\mathbf{v}_1 + \mathbf{h}_1), N(\mathbf{v}_2 + \mathbf{h}_2)|| \to ||N(\mathbf{v}_1), N(\mathbf{v}_2)||$. We prove the continuity of $\tilde{\mathbf{C}}(\mathbf{v}_1, \mathbf{v}_2\,|\,\boldsymbol{\theta})$ for the case when $\mathbf{v}_1$ is outside $\mathcal{S}$ and $\mathbf{v}_2 = \mathbf{s}_j$. The other cases are proved similarly. We assume that the covariance function for the original GP is isotropic and continuous. The two distance results yield $\mathbf{B}_{\mathbf{v}_1+\mathbf{h}_1} = \mathbf{C}_{\mathbf{v}_1+\mathbf{h}_1, N(\mathbf{v}_1+\mathbf{h}_1)}\mathbf{C}^{-1}_{N(\mathbf{v}_1+\mathbf{h}_1)} \to \mathbf{C}_{\mathbf{v}_1, N(\mathbf{v}_1)}\mathbf{C}^{-1}_{N(\mathbf{v}_1)} = \mathbf{B}_{\mathbf{v}_1}$. Also, as $\mathbf{v}_2 + \mathbf{h}_2 \to \mathbf{v}_2 = \mathbf{s}_j$, then $\mathbf{s}_j \in N(\mathbf{v}_2 + \mathbf{h}_2)$ for small enough $\mathbf{h}_2$. Let $\mathbf{s}_j = N(\mathbf{v}_2 + \mathbf{h}_2)(1)$ and, hence, $\mathbf{C}_{\mathbf{v}_2+\mathbf{h}_2, N(\mathbf{v}_2+\mathbf{h}_2)}\mathbf{C}^{-1}_{N(\mathbf{v}_2+\mathbf{h}_2)} \to \mathbf{e}_1$ where $\mathbf{e}_1 = (1, 0, \ldots, 0)_{m \times 1}$. Therefore,

$$\begin{aligned} \lim_{\mathbf{h}_1 \to 0, \mathbf{h}_2 \to 0} \tilde{\mathbf{C}}(\mathbf{v}_1 + \mathbf{h}_1, \mathbf{v}_2 + \mathbf{h}_2\,|\,\boldsymbol{\theta}) &= \mathbf{B}_{\mathbf{v}_1}\ lim_{\mathbf{h}_1 \to \mathbf{0}, \mathbf{h}_2 \to \mathbf{0}}\ \widetilde{Cov}(\mathbf{w}_{N(\mathbf{v}_1+\mathbf{h}_1)}, \mathbf{w}_{N(\mathbf{v}_2+\mathbf{h}_2)}\,|\,\boldsymbol{\theta})\mathbf{e}_1 \\ &= \mathbf{B}_{\mathbf{v}_1}\ lim_{\mathbf{h}_1 \to \mathbf{0}}\ \widetilde{Cov}(\mathbf{w}_{N(\mathbf{v}_1+\mathbf{h}_1)}, \mathbf{w}(\mathbf{s}_j)\,|\,\boldsymbol{\theta}) = \mathbf{B}_{\mathbf{v}_1}\tilde{\mathbf{C}}_{N(\mathbf{v}_1),\mathbf{s}_j} = \tilde{\mathbf{C}}(\mathbf{v}_1, \mathbf{v}_2\,|\,\boldsymbol{\theta})\ . \end{aligned}$$

A5 Simulation Experiment: Robustness of NNGP to ordering of locations

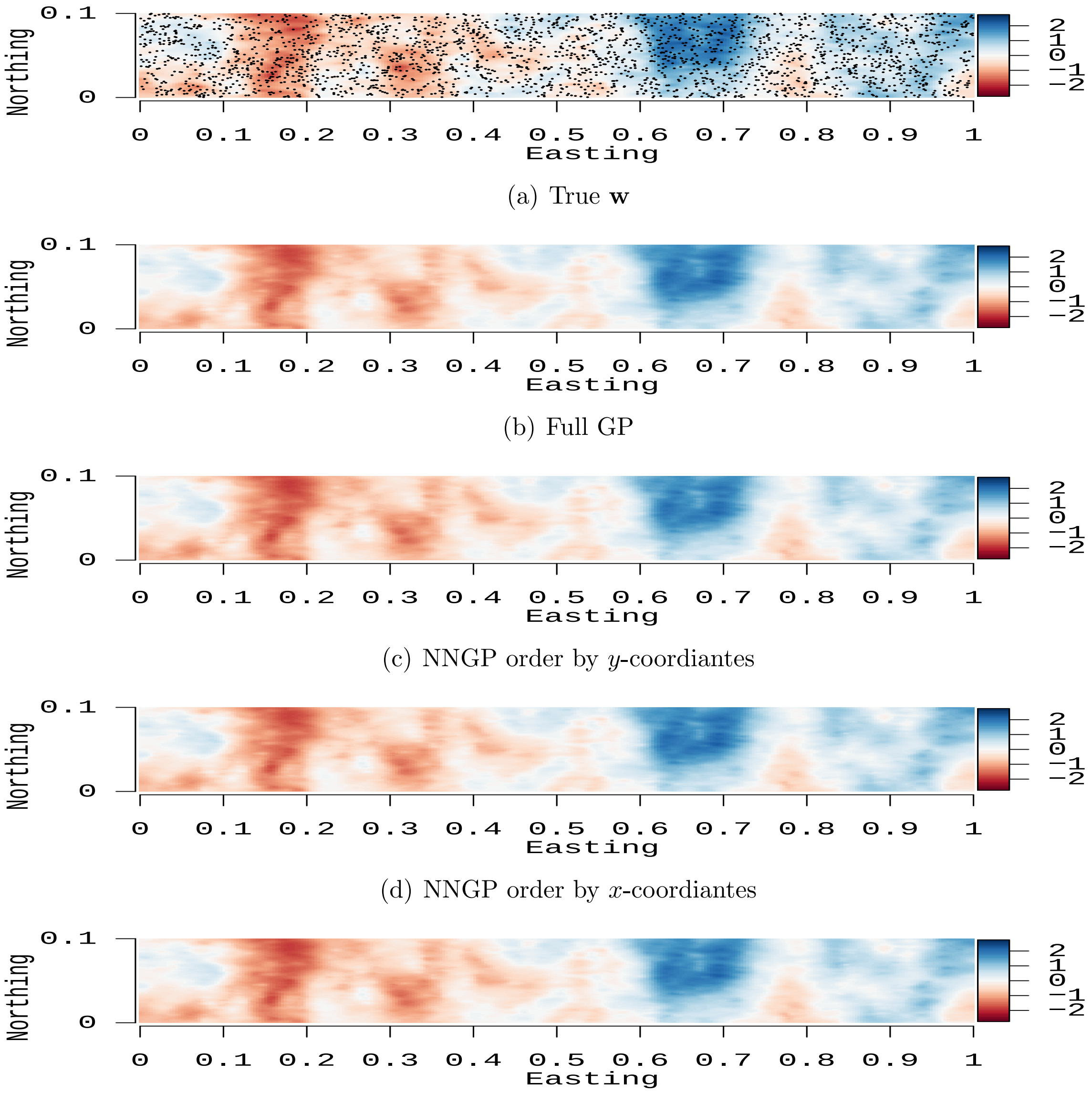
We conduct a simulation experiment demonstrating the robustness of NNGP to the ordering of the locations. We generate the data for $n = 2500$ locations using the model in Section 5.1. However instead of a square domain we choose a long skinny domain (see Figure 1(a)) which can bring out possible sensitivity to ordering due to scale disparity between the x and y axes. We use three different orderings for the locations: ordering by x-coordinates, by y-coordinates and by the function $f(x, y) = x + y$.

Table A1 demonstrates that the point estimates and the 95% credible intervals for the process parameters from all three NNGP models are extremely consistent with the estimates from the full Gaussian process model.

Posterior estimates of the spatial residual surface from the different models are shown in Figure A1. Again, the impact of the different ordering is negligible. We also plotted the difference between the posterior estimates of the random effects of the true GP and NNGP for all 3 orderings in Figure A2. It was seen that this difference was negligible compared to the difference between the true spatial random effects and full GP estimates. This shows the inference obtained from the NNGP (using any ordering) closely emulates the corresponding full GP inference.

Table A1: Univariate synthetic data analysis parameter estimates and computing time in minutes for NNGP m=10 and full GP models. Parameter posterior summary 50 (2.5, 97.5) percentiles.

				NNGP ($\mathcal{S} = \mathcal{T}$)	
		Full	Order by	Order by	Order by
	True	Gaussian Process	y-coordinates	x-coordinates	$x + y$-coordinates
σ^2	1	0.640 (0.414, 1.297)	0.712 (0.449, 1.530)	0.757 (0.479, 1.501)	0.718 (0.464, 1.436)
τ^2	0.1	0.107 (0.098, 0.117)	0.106 (0.097, 0.114)	0.107 (0.099, 0.117)	0.107 (0.098, 0.115)
ϕ	6	8.257 (4.056, 13.408)	8.294 (3.564, 12.884)	7.130 (3.405, 11.273)	7.497 (3.600, 11.911)



(a) True $\mathbf{w}$

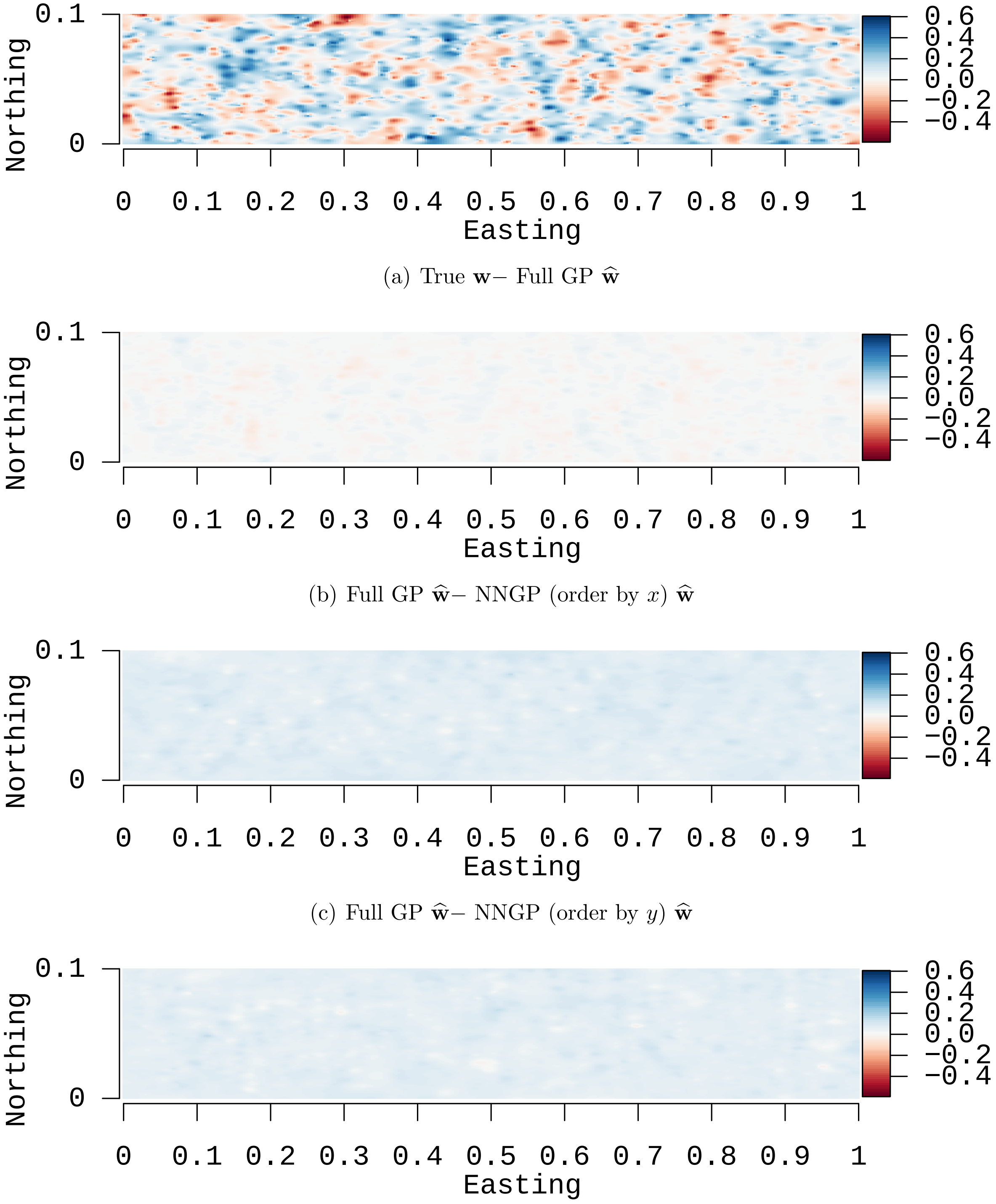
(b) Full GP

(c) NNGP order by y-coordiantes

(d) NNGP order by x-coordiantes

(e) NNGP order by $x + y$-coordiantes

Figure A1: Robustness of NNGP to ordering: Figures (a) and (b) show interpolated surfaces of the true spatial random effects and posterior median estimates for full geostatistical model respectively. Figures (c), (d), and (e) show interpolated surfaces of the posterior median estimates for NNGP model with $\mathcal{S} = \mathcal{T}$, $m = 10$, and alternative coordinate ordering. Corresponding true and estimated process parameters are given in Table A1.



(a) True $\mathbf{w}$− Full GP $\widehat{\mathbf{w}}$

(b) Full GP $\widehat{\mathbf{w}}$− NNGP (order by x) $\widehat{\mathbf{w}}$

(c) Full GP $\widehat{\mathbf{w}}$− NNGP (order by y) $\widehat{\mathbf{w}}$

(d) Full GP $\widehat{\mathbf{w}}$− NNGP (order by $x + y$) $\widehat{\mathbf{w}}$

Figure A2: Difference between Full GP and NNGP estimates of spatial effects: Figure (a) shows the difference between the true spatial random effects and the full GP posterior median estimates. Figures (b), (c) and (d) plots the difference between posterior median estimates of full GP and NNGP ordered by x, y and $x + y$ co-ordinates respectively. All the figures are in the same color scale.

A6 Simulation experiment: NNGP credible intervals as function of m

From a classical viewpoint, NNGP can be regarded as a computationally convenient approximation to the full GP model. The accuracy of the approximation is expected to improve with increase in m as NNGP model becomes identical to the full model when m equals the sample size. However, we construct the NNGP as an independent model and found that inference from this model closely emulates that from the full GP model. Figure 1 demonstrates how root mean square predictive error and parameter CI width vary with choice of m. We conduct another simulation experiment to investigate how the parameter estimation of the hierarchical NNGP model depends on m.

We generated a dataset of size 1000 using the model described in Section 5.1 for 4 combination of values of ϕ and σ^2. Other parameters and prior choices were similar to those in section A8. Figure A3 gives true values of σ^2 and effective range $(3/\phi)$ alongwith the posterior medians and credible intervals for the full GP, NNGP with $m = 10$ and $m = 100$. We see that the CI's for NNGP $m = 10$ and $m = 100$ are almost identical and are very close to the CI for full GP. This suggests that even for small values of m NNGP, parameter CI's closely resemble full GP parameter CI's.

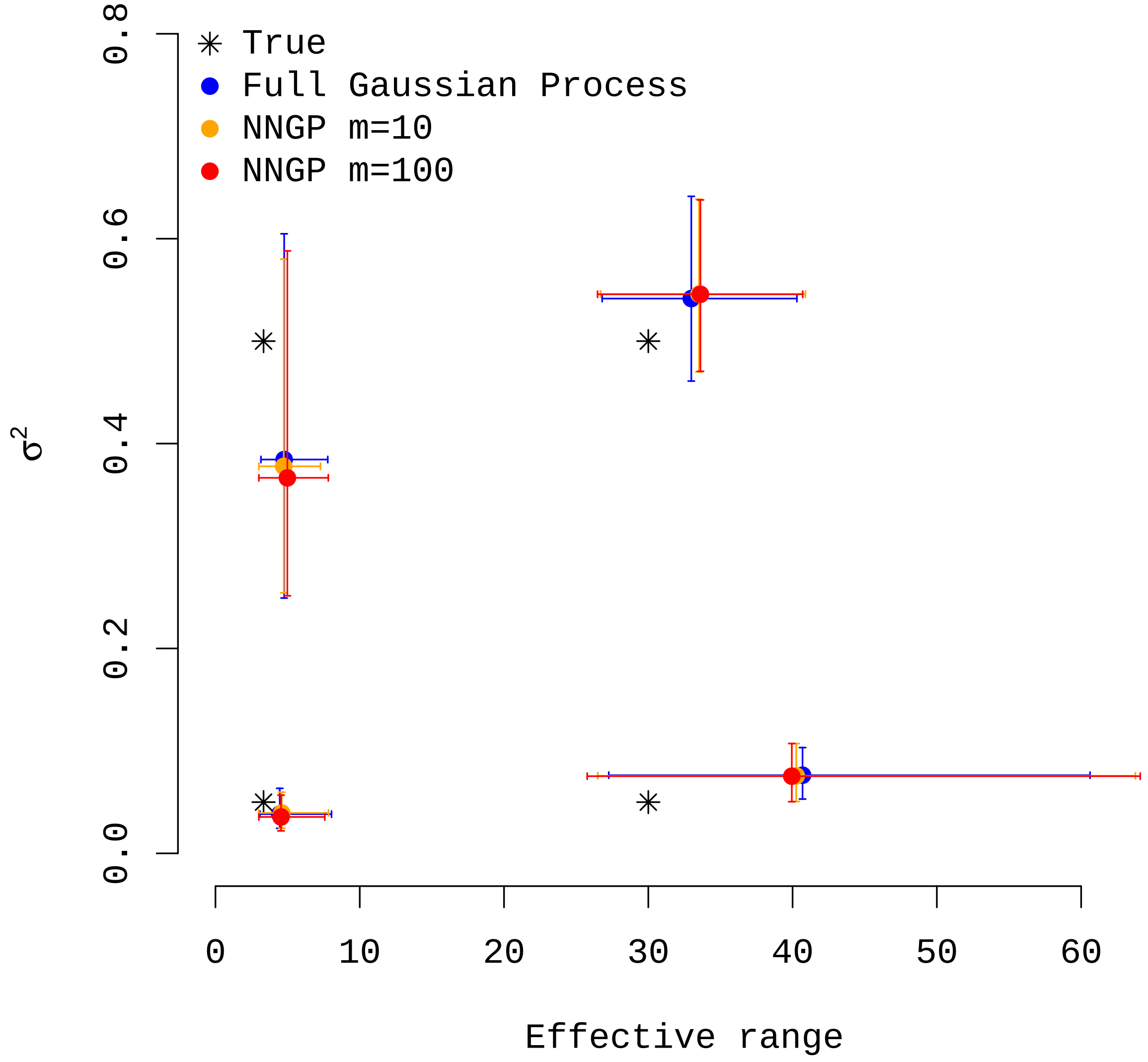


Figure A3: NNGP credible intervals for small and large values of m

A7 Simulation experiment: Data with gaps

One possible area of concern for NNGP is that if the data have large gaps and the NNGP is constructed using the data locations as the reference set $\mathcal{S}$, then NNGP covariance function may be a poor approximation of the full GP covariance function. This arises from the fact that if the reference set has large gaps then two very close locations outside $\mathcal{S}$ can have very different neighbor sets. Since, in a NNGP, locations outside $\mathcal{S}$ are correlated through their neighbors sets this may lead to little correlation among very close points in certain regions of the domain.

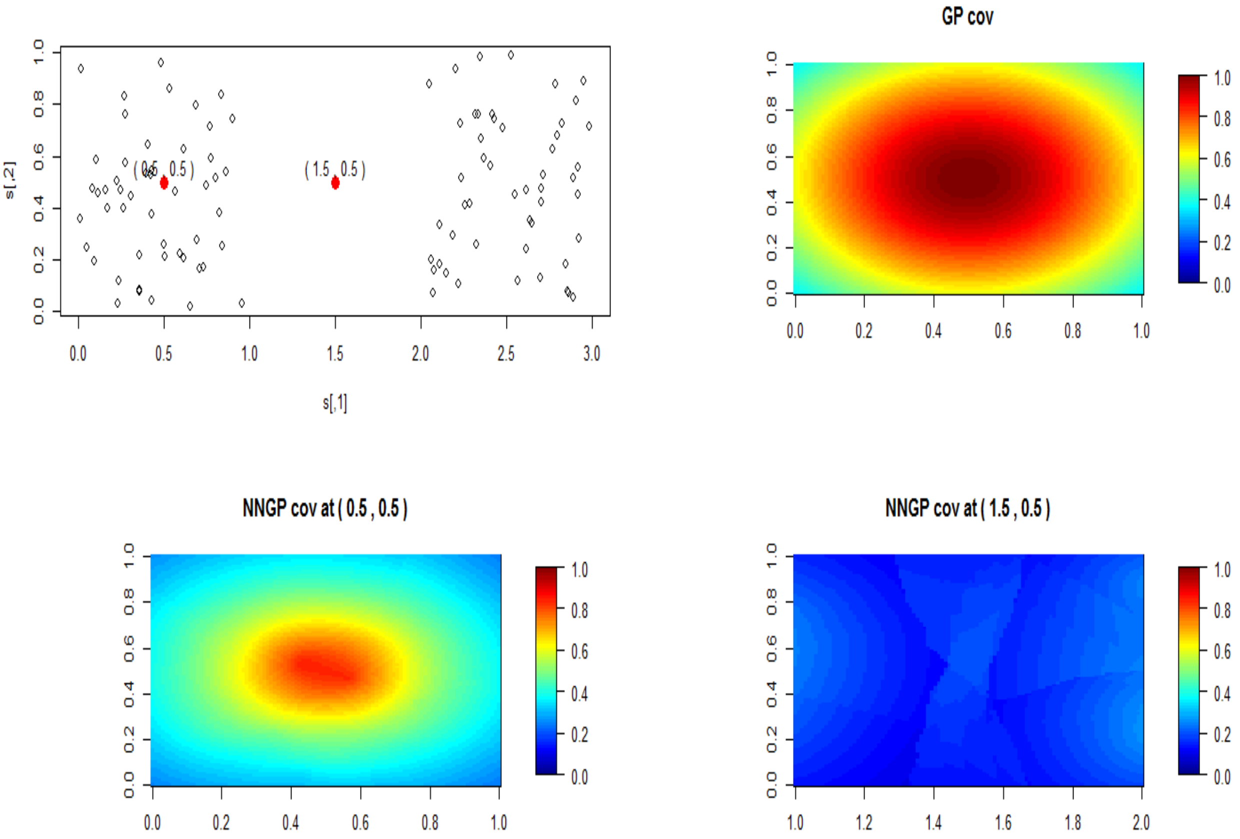


Figure A4: Full GP and NNGP ($m = 10$) covariance function for data with gaps

Figure A4 demonstrates this issue. We generate a set $\mathcal{T}$ of 100 locations (topleft) on the domain $[0,3] \times [0,1]$ with half the locations in $[0,1] \times [0,1]$ and the remaining half in $[2,3] \times [0,1]$. This creates a large gap in the middle where there are no datapoints. The topright panel shows the heatmap of the full GP covariance function with σ^2=1 and $\phi = 2$ (so that the effective range is 1.5). The NNGP is a non-stationary process and the covariance function depends on the locations. We evaluate this covariance at two points (red dots in the topleft figure) — $(0.5, 0.5)$ (which is surrounded by many points in $\mathcal{S}$) and $(1.5, 0.5)$ (which is at the middle of the gap and equidistant from the two sets of locations in $\mathcal{S}$).

The NNGP field at $(0.5, 0.5)$ (bottomleft) closely resembles the GP field. This is because the neighbors of $(0.5, 0.5)$ are close to the point and provides strong information about the true GP at that point. The NNGP field at $(1.5, 0.5)$ (bottomright) is almost non-existent with near zero correlations even at very small distances. This is an expected consequence of the way NNGP is constructed. Any two points outside $\mathcal{S}$ are correlated via their neigbhor sets only. The neighbors for $(1.5, 0.5)$ are far away from the point it provides weak information about the point as it is in the middle of the gap.

This suggests that a NNGP constructed using a reference set with large gaps is a poor approximation to the full GP as a process in certain regions of the domain. If the data locations do have large gaps, perhaps a NNGP with $\mathcal{S}$ as a grid over the domain provides a much better approximation to the full GP. To observe this we used a 14×7 grid over the domain $[0,3] \times [0,1]$ as $\mathcal{S}$. So the size of this new $\mathcal{S}$ was similar to the original sample size of 100. Figure A5 demonstrates the NNGP covariance function at the two points using this new $\mathcal{S}$. We see that using the grid $\mathcal{S}$, the NNGP covariance function at the two points are very similar and closely resemble the true GP covariance function. This suggests that in order for the NNGP to resemble full GP, the reference set needs to have points uniformly distributed over the domain.

However, from a kriging perspective, if the data have large gaps, inference from a NNGP

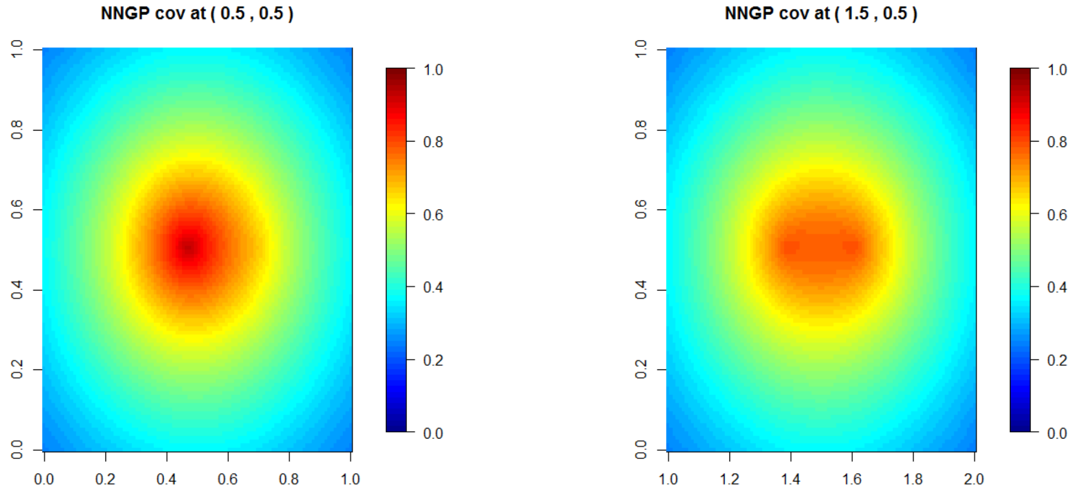


Figure A5: NNGP covariance function using a grid $\mathcal{S}$

with $\mathcal{S} = \mathcal{T}$ may not differ a lot from the full GP inference. Even when one uses the full GP, kriging is usually done one point at a time and thereby ignores the covariances between points outside the data locations and assumes conditional independence. Figure A6 plots

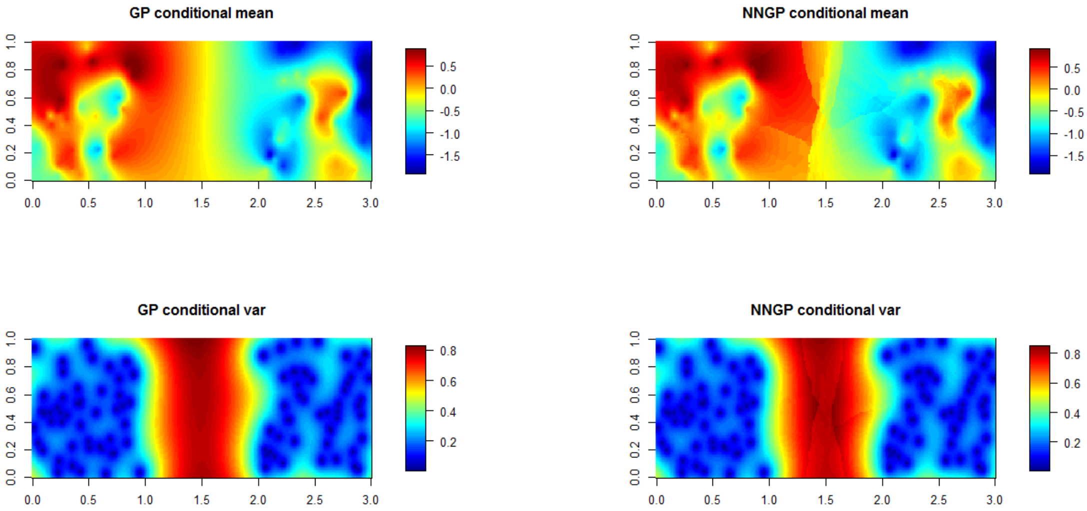


Figure A6: Kriging means and variances for full GP and NNGP ($\mathcal{S}$ = data locations)

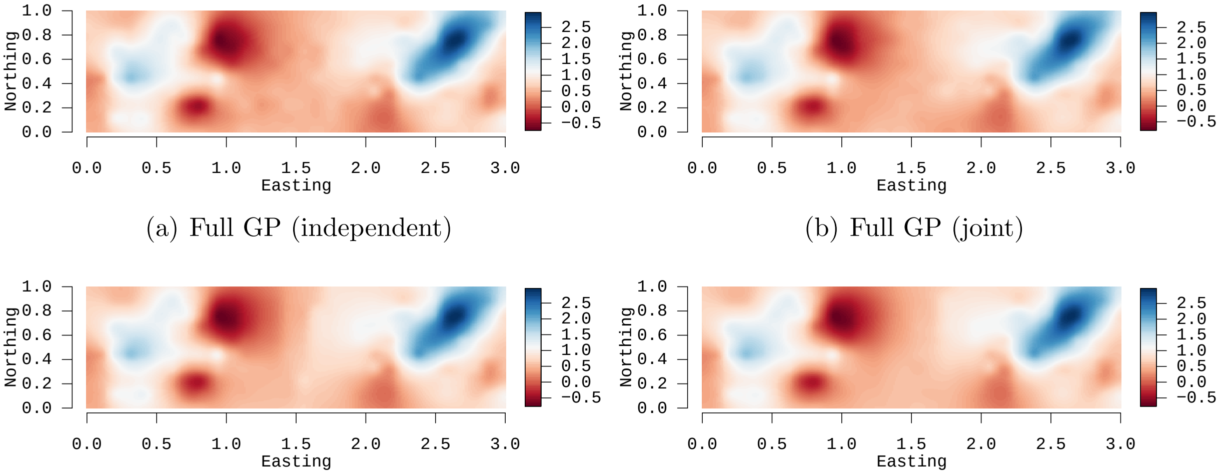
the kriging mean and variances over the entire domain for the full GP and the NNGP. They

	True	Full GP	NNGP m=10	NNGP m=20
β	1	0.72 (0.00, 1.32)	0.65 (-0.14, 1.30)	0.69 (0.02, 1.16)
τ	0.01	0.03 (0.01, 0.05)	0.03 (0.01, 0.06)	0.03 (0.01, 0.06)
σ^2	1	0.63 (0.38, 1.31)	0.65 (0.39, 1.29)	0.62 (0.38, 1.27)
ϕ	2	2.94 (1.27, 5.19)	2.76 (1.27, 5.25)	2.91 (1.34, 5.20)
RMSPE	–	0.58 (ind)	0.57	0.57
	–	0.58 (joint)	–	–
95% CI cover	–	94.00 (ind)	95.66	95.33
	–	95.33 (joint)	–	–
Mean 95% CI width	–	2.12 (ind)	2.12	2.13
	–	2.11 (joint)	–	–

Table A2: Data analysis for locations with gaps

are very close. This suggests even for data with gaps the kriging performance of NNGP and GP are similar.

We also generated a dataset over $\mathcal{T}$ and fitted the full GP and NNGP ($\mathcal{S} = \mathcal{T}$) model to compare parameter estimation and kriging performance. In addition to the conventional independent kriging, we also used the computationally expensive joint kriging for the full GP to see if it improves kriging quality at locations in the gap. Table A2 provide the parameter estimates and model fitting metrics. Figures A7 and A8 gives the posterior median and the variance surface over the domain. We see that the the NNGP and full GP produce very similar parameter estimates and kriging. Hence, for data with large gaps both the full GP and NNGP ($\mathcal{S} = \mathcal{T}$) doesn't provide enough information for locations inside the gaps. So even if NNGP ($\mathcal{S} = \mathcal{T}$) poorly approximates the full GP as a process, in terms of model fitting, their performances are very similar.



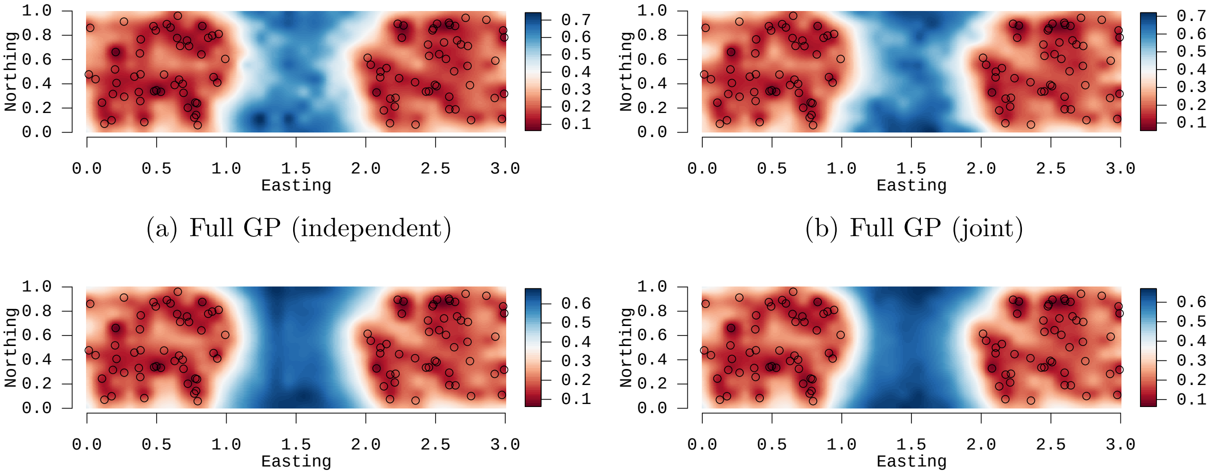
(a) Full GP (independent)

(b) Full GP (joint)

(c) NNGP $m = 10$

(d) NNGP $m = 20$

Figure A7: Posterior median surface for data with gaps



(a) Full GP (independent)

(b) Full GP (joint)

(c) NNGP $m = 10$

(d) NNGP $m = 20$

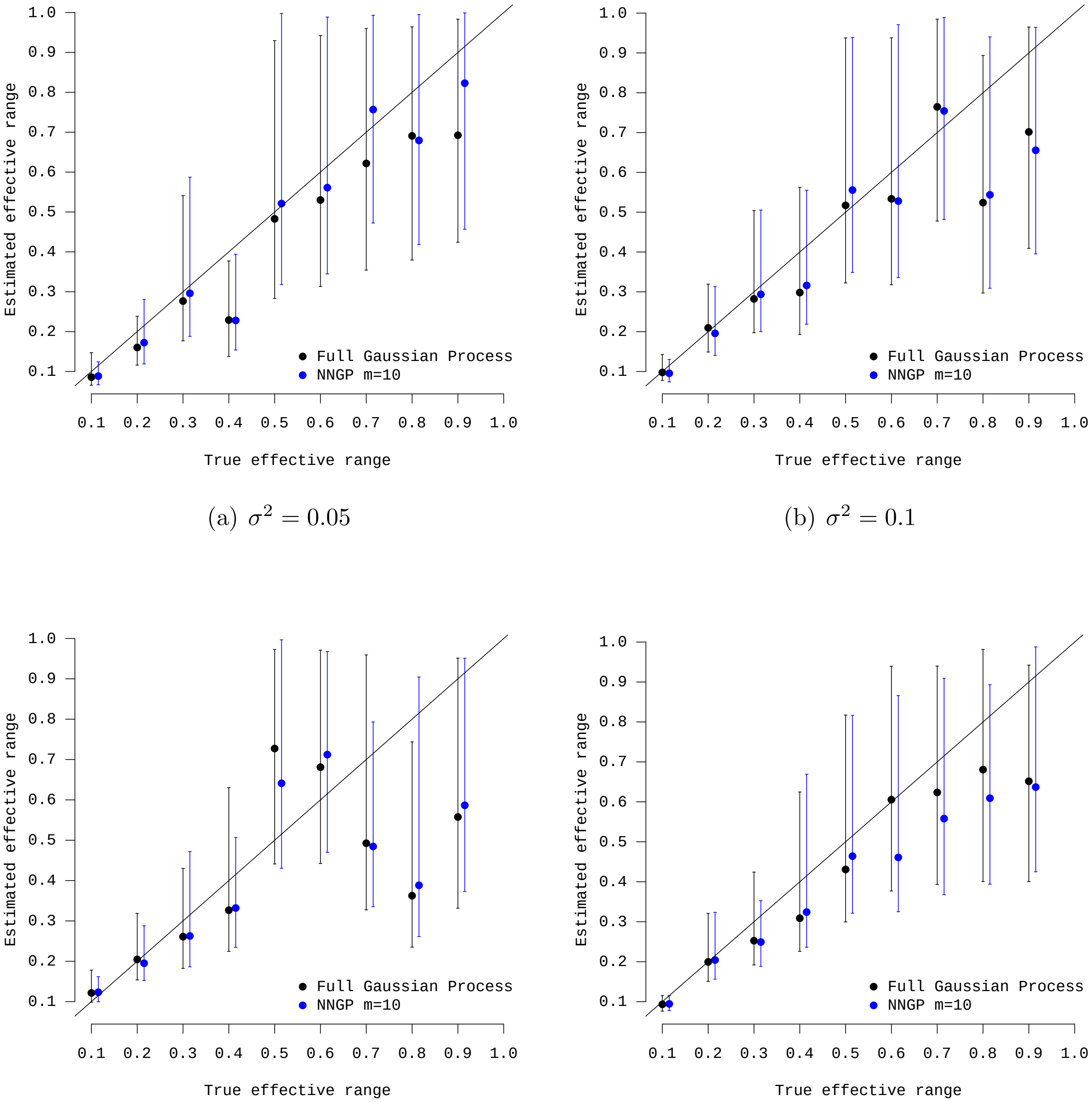
Figure A8: Posterior variance surface for data with gaps

A8 Simulation experiment: Slow decaying covariance functions

We note in Section 2.1 that several valid choices of neighbor sets can be used to construct a NNGP. However, our choice of using m-nearest neighbors to construct neighbor sets performed extremely well for all the data analysis in Section 5. Since, our design of NNGP just includes m-nearest neighbors it is natural to be skeptical of the performance of NNGP when the data arises from a Gaussian process with very flat tailed covariance function. Such a covariance function implies that even distant observations are significantly correlated with the given observation and m-nearest neighbors may fail to capture all the information about the covariance parameters.

We generate datasets of size 2500 in a unit domain using the model described in Section 5.1 for a wide range of values for the parameters σ^2 and ϕ. The marginal variance σ^2 was varied over $(0.05, 0.1, 0.2, 0.5)$ and the 'true effective range' $3/\phi$ phi was varied over $(0.1, 0.2, \ldots, 1)$. Larger values of the 'true effective range' indicate higher correlation between points at large distances. The nugget variance τ^2 was held constant at 0.1. The prior on ϕ was U(3,300) or 0.01 to 1 distance units. Also both τ^2 and σ^2 were given Inverse Gamma$(2, 0.1)$ priors in all cases.

Figure A9 gives the results for NNGP and full GP CIs. We see that for all choices of parameters, the posterior samples from the NNGP and full GP look identical. This strongly suggests that the NNGP model deliver inference similar to that of a full GP even for slow decaying covariance functions and justifies the choice of the neighbor sets.



(a) $\sigma^2 = 0.05$

(b) $\sigma^2 = 0.1$

(c) $\sigma^2 = 0.2$

(d) $\sigma^2 = 0.5$

Figure A9: Univariate synthetic data analysis: true versus posterior 50% (2.5%, 97.5%) percentiles for the effective spatial range simulated for various values of σ^2 and $\tau^2 = 0.1$. NNGP model fit with $\mathcal{S} = \mathcal{T}$ and $m = 10$.

A9 Simulation experiment: Wave covariance function

We have restricted most of our simulation experiments to Matérn (or in particular exponential) covariance functions. Matérn covariance functions like many other covariance functions decrease monotonically with distance and hence nearest neighbors of a location have highest correlation with that location. We wanted to investigate the performance of NNGP for covariance functions which do not monotonically decrease with distance. We use the two-dimensional damped cosine covariance function given by:

$$C(d) = \exp(-d/a)\cos(\phi d)\ ,\ a \leq 1/\phi \tag{A2}$$

First, we generated the Kullback-Leibler (KL) divergence numbers for the NNGP model with respect to the full GP model using damped cosine covariance. In addition to the default neighbor selection scheme, we also used an alternate scheme described by Stein et al. (2004). This scheme includes $m' = \lceil 0.75m \rceil$ nearest neighbors and $m - m'$ neighbors whose ranked distances from the i^{th} location equal $m + \lfloor l(i-m-1)/(m-m') \rfloor$ for $l = 1, 2, \ldots, m - m'$. Stein et al. (2004) suggested that this scheme choice often improves parameter estimation. The two schemes are referred to as NNGP and NNGP (alt) respectively. We used $\phi = 10$, $a = .099$, sample sizes of $100, 200$ and 500 and varied m from 5 to 50 in increments of 5.

Figure A10 plots the KL divergence numbers (in log-scale) for varying m, n and neighbor selection schemes. We see that larger sample size implies higher KL divergence numbers which is expected as with increasing sample size the size of the neighbor set m becomes smaller in proportion. Also, we see that KL numbers for the alternate neighbor selection scheme are always higher indicating that nearest neighbors perform better even for such wave covariance functions. In general we observed that the KL numbers are quite small for $m \geq 25$ for all n and neighbor selection schemes indicating that the NNGP models closely approximate the true damped cosine GP.

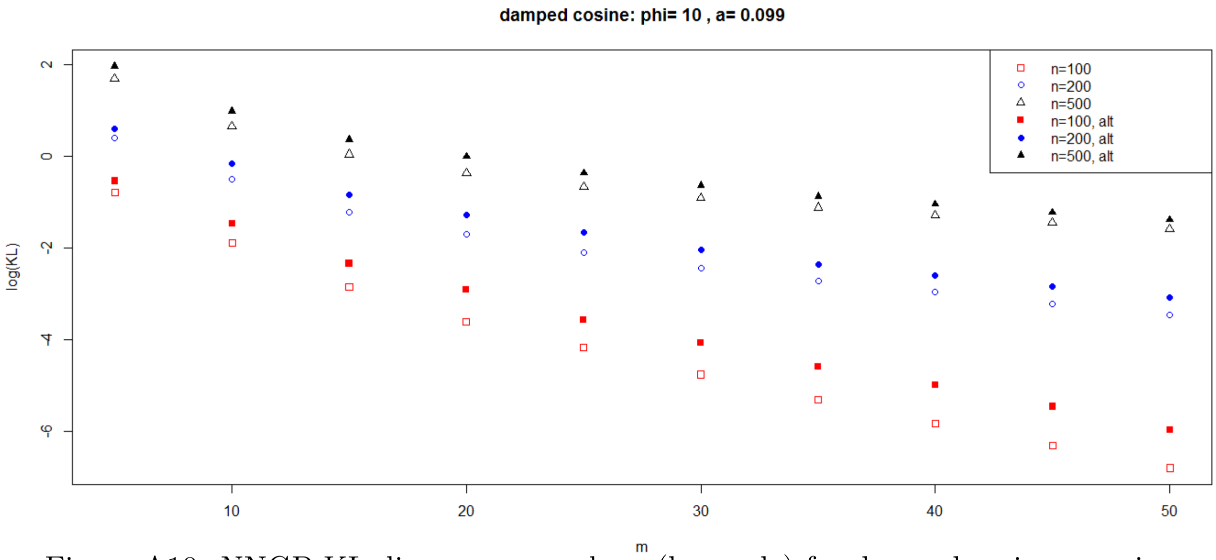


Figure A10: NNGP KL divergence numbers (log scale) for damped cosine covariance

Next, we conducted a data analysis using the wave covariance function. We choose $n = 500$, $m = 10, 20$. The two values of m yielded around 3.4% and 18.7% nearest neighbors which were negatively correlated with the corresponding locations. Table A3 gives the parameter estimates for the NNGP model. Figure A11 demonstrates how the NNGP approximates the wave covariance function while figure A12 plots the true and fitted random effect surface. We observe that NNGP provides an excellent approximation of the the true wave GP in terms of model parameter estimation and kriging.

We could not fit the full GP model due to computation instability of the large wave covariance matrix. NNGP does not involve inverting large matrices and hence we could use it for model fitting.

	True	m=10	m=20
β_0	1	1.03 (0.65, 1.34)	1.06 (0.70, 1.32)
β_1	5	5.00 (4.95, 5.06)	5.00 (4.95, 5.06)
τ^2	0.1	0.06 (0.02, 0.12)	0.05 (0.03, 0.11)
σ^2	1	1.13 (0.90, 1.57)	1.14 (0.90, 1.57)
ϕ	10	7.41 (1.63, 11.59)	6.31 (1.61, 10.50)
a	0.099	0.093 (0.067, 0.135)	0.09 (0.07, 0.14)

Table A3: Damped cosine GP data analysis using NNGP

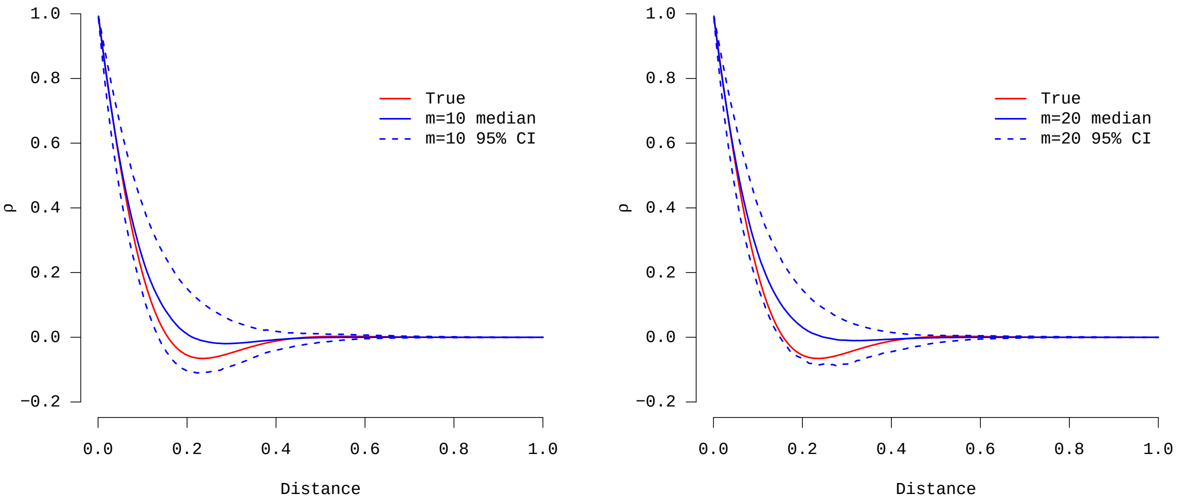
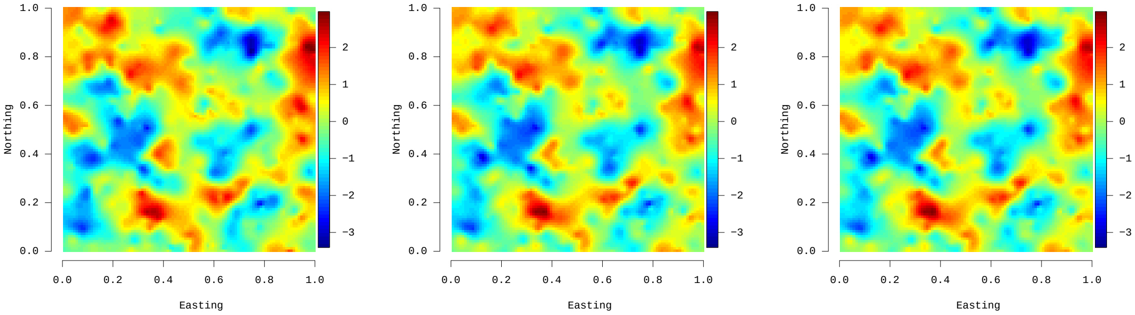


Figure A11: Wave covariance function estimates using NNGP



(a) True

(b) NNGP $m = 10$

(c) NNGP $m = 20$

Figure A12: True and estimated (posterior median) random effect surface of the damped cosine GP

References

Stein, M. L., Chi, Z., and Welty, L. J. (2004), "Approximating Likelihoods for Large Spatial Data Sets," *Journal of the Royal Statistical Society, Series B*, 66, 275–296.