Submitted to the Annals of Applied Statistics
arXiv: arXiv:0000.0000

SUPPLEMENT: NON-SEPARABLE DYNAMIC NEAREST-NEIGHBOR GAUSSIAN PROCESS MODELS FOR LARGE SPATIO-TEMPORAL DATA WITH AN APPLICATION TO PARTICULATE MATTER ANALYSIS

By Abhirup Datta, Sudipto Banerjee, Andrew O. Finley, Nicholas A.S. Hamm and Martijn Schaap

S1. Eligible Sets. We provide formal definitions of the eligible sets. Recall from Section 4.1 that, for any location s, $A(\mathbf{s}, V, m)$ is the set of m-nearest neighbors of $\mathbf{s}$ in V. So $\mathbf{s} \in V$ implies that $\mathbf{s} \in A(\mathbf{s}, V, m)$ for all $m \geq 1$. For each $(\mathbf{s}_i, t_j)$ in $\mathcal{R}$, we define the *eligible set*

$$E(\mathbf{s}_i, t_j) = \bigcup_{k=1}^{m} \{(\mathbf{s}, t_{j-k}) \,|\, \mathbf{s} \in A(\mathbf{s}_i, S, [m/k])\} \bigcup \{(\mathbf{s}, t_j) \,|\, \mathbf{s} \in A(\mathbf{s}_i, S_i, m)\} \tag{S1}$$

where for any positive number x, $[x]$ denotes the greatest integer not exceeding x. So the eligible set for a space-time point consists of m-nearest neighbors from the time levels j and $j-1$, $[m/2]$ nearest neighbors from time level $j-2$ and so on upto $[m/m] = 1$ nearest neighbor from time level $j-m$. This is also illustrated in Figure 3(c). So the size of $E(\mathbf{s}_i, t_j)$ does not exceed $m + \sum_{k=1}^{m} [m/k]$. As m is typically chosen to be around 20, this sum is approximately $4m$.

For any point t outside T, let $t[k]$ denote the k^{th} nearest time point of t in T. Then, we define the eligible set for any $(\mathbf{s}, t)$ outside $\mathcal{R}$ as

$$E(\mathbf{s}, t) = \bigcup_{k=1}^{m} \{(\mathbf{s}, t[k]) \,|\, \mathbf{s} \in A(\mathbf{s}, S, [m/k])\} \tag{S2}$$

The eligible sets do not depend on the covariance parameters $\boldsymbol{\theta}$. We now show that for any point $(\mathbf{s}, t)$ in $\mathcal{L}$, the eligible set $E(\mathbf{s}, t)$ defined by Equations S1 and S2 contains m-nearest neighbors of $(\mathbf{s}, t)$ for all values of $\boldsymbol{\theta}$ as long as the underlying covariance function $C(h, u \,|\, \boldsymbol{\theta})$ possess natural monotonicity.

Proposition S1. *If $C(h, u \,|\, \boldsymbol{\theta})$ satisfies natural monotonicity defined in Section 4.1 for every value of $\boldsymbol{\theta}$, then, for every $(\boldsymbol{s}, t)$, the eligible set $E(\boldsymbol{s}, t)$ defined in Equations S1 and S2 contains $N_\theta(\boldsymbol{s}, t)$ for all $\boldsymbol{\theta}$*

Proof. We only prove for $(\mathbf{s}, t) = (\mathbf{s}_i, t_j) \in \mathcal{R}$. The proof for $(\mathbf{s}, t) \notin \mathcal{R}$ is similar. We assume that $(\mathbf{s}_u, t_{j-k}) \in N_\theta(\mathbf{s}_i, t_j)$ for some $\boldsymbol{\theta}$, $u \leq N$ and $k \geq 1$. Also let $\mathbf{s}_i[l]$ denote the l^{th} nearest neighbor of $\mathbf{s}_i$ among $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N\}$. So, $\mathbf{s}_u = \mathbf{s}_i[l]$ for some $l \geq 1$. Therefore, by natural monotonicity of C, we have $C((\mathbf{s}_i, t_j), (\mathbf{s}_i[a], t_{j-k}) \,|\, \boldsymbol{\theta}) \geq C((\mathbf{s}_i, t_j), (\mathbf{s}_i[l], t_{j-k}) \,|\, \boldsymbol{\theta})$ for all $1 \leq a \leq l$. One more application of natural monotonicity implies that $C((\mathbf{s}_i, t_j), (\mathbf{s}_i[a], t_{j-b}) \,|\, \boldsymbol{\theta}) > C((\mathbf{s}_i, t_j), (\mathbf{s}_i[a], t_{j-k}) \,|\, \boldsymbol{\theta})$ for all $1 \leq b \leq k$. As $(\mathbf{s}_u, t_{j-k}) \in N_\theta(\mathbf{s}_i, t_j)$, then so does $(\mathbf{s}_i[a], t_{j-b})$ for all $a \leq l$ and $b \leq k$. Therefore, $lk \leq m$ i.e. $l \leq [m/k]$. □

S2. Comparisons with Local Approximation GP. Gramacy and Apley (2015) proposed a Local Approximation Gaussian Process (LAGP) to predict a function $y(\mathbf{s}_0)$ at a location s_0 given the observations $\{y(\mathbf{s}_1), y(\mathbf{s}_2), \ldots, y(\mathbf{s}_n)\}$ for n locations. For predicting at each new location s_0, a small neighbor set of size m is selected from $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\}$. However, unlike Nearest Neighbor Gaussian Processes (Datta et al., 2016), the neighbor sets doesn't merely consist of m nearest neighbors of s_0. Instead, locations are augmented to an initial set of $m_0 < m$ nearest neighbors based on minimization of the predictive Mean Square Error for $y(\mathbf{s}_0)$. Simulation studies detailed in Gramacy and Apley (2015) suggests that LAGP produces improved prediction over the fully nearest neighbor based approach while the computational costs for the two methods are comparable.

Although promising in terms of predictive performance and computational scalability, LAGP is not a proper Gaussian Process. It essentially provides a non-stationary local approximation to a Gaussian Process at every predictive location. LAGP thereby can only be used for predicting the response and lacks the versatility of a full GP based approach.

For the European PM_{10} data, we used a semiparametric regression model with the raw CTM output as a covariate and a spatio-temporally varying random intercept (Equation 5.3) which was modeled using a DNNGP prior. LAGP, in its current form, cannot be used for hierarchical models like spatially varying intercept or spatially varying coefficient models where a GP is used to model latent random effects instead of the response.

Moreover, LAGP uses an isotropic squared exponential covariance function for specifying the GP. In a spatio-temporal setup, this amounts to the assumption that the variation along the spatial and temporal directions are on the same scale. These assumptions are violated when we are modeling an anisotropic or non-separable space-time function. LAGP, unlike Dynamic NNGP, cannot accommodate such space-time non-separable functions, and as seen in simulation experiments detailed in Table S1, performs poorly when

Table S1
Average RMSPE numbers for LAGP and Dynamic NNGP

	LAGP	Simple DNNGP	Adaptive DNNGP
Case 1	0.94	0.92	0.90
Case 2	0.61	0.55	0.56
Case 3	0.89	0.83	0.83

the space and time variations differ significantly.

Finally, one of our main objectives for the spatio-temporal analysis of European PM_{10} dataset was to understand the underlying spatial and temporal dependence of PM_{10}. Indeed, the covariance parameter estimates in Table 2 reveals that the effective spatial range is around 45 to 60 km and the temporal range is around 30 to 33 days (Figure 5). Understanding the spatio-temporal structure after adjusting for the CTM output may help identify the physical processes missing in the CTM and subsequently improve the CTM itself. LAGP is not a proper Gaussian Process and does not have any global space and time decay parameters and is hence unsuitable for such spatio-temporal analysis.

Nevertheless, as recommended by the reviewers, we compared the predictive performance of Dynamic NNGP and LAGP via simulation experiments. We generated data from the model

$$y(\mathbf{s},t) = w(\mathbf{s},t) + \epsilon(\mathbf{s},t) \tag{S1}$$

where $\epsilon \overset{\text{iid}}{\sim} N(0,\tau^2)$ and $w(\mathbf{s},t)$ are the realizations of a GP with a non-separable covariance structure specified in (6.1). We used the same three sets of parameters as detailed in Section 6 which corresponds to 1) short spatial range and long temporal range, 2) long spatial and temporal range, and; 3) long spatial range and short temporal range. For each case we generated the data on a $15 \times 15 \times 15$ space-time grid and computed the RMSPE based on 500 holdout locations. The process was repeated 50 times for each of the three scenarios and the average RMSPE numbers are reported in Table S1. We observe that for Case 3 where the temporal range dominates, we see LAGP tends to perform worse. This is expected as LAGP assumes an isotropic covariance function ignoring the disparity in space and time variation scales. Overall, Dynamic NNGP performs better for all the three scenarios.

S3. Non-stationary covariances. The Gneiting covariance function (6.1) used to construct the Dynamic NNGP is a stationary covariance function. However, we can also use Dynamic NNGP to approximate certain class of non-stationary covariance functions. Gelfand et al. (2004) proposed a Spatially Varying Linear Model of Coregionalization (SVLMC) model to create non-stationary spatial processes as spatially weighted sum of stationary processes. For univariate responses, this reduces to $w(\mathbf{s},t) = \sigma^2(\mathbf{s},t)v(\mathbf{s}, t)$ where $v(\mathbf{s},t) \sim GP$ with mean 0 and covariance function defined in (6.1) with $\sigma^2 = 1$. Subsequently, there are several ways of modeling $\sigma^2(\mathbf{s},t)$ If a covariate $x(\mathbf{s},t)$ is observed, $\sigma^2(\mathbf{s},t)$ is often modeled as $x(\mathbf{s},t)^\phi\sigma^2$ for some unknown $\phi \geq 0$. This introduces covariate dependent non-stationarity. In such case, due to non-stationarity, physical nearest neighbors of a location may not correspond to locations with highest correlations for the process $w(\mathbf{s},t)$. Hence, the simple or the adaptive neighbor selection methods described in Section 4 which relies on the above mentioned principle, may lead to conditioning sets with very few points with high correlation with a given point. Hence, we suggest modeling the stationary process $v(\mathbf{s},t)$ as a DNNGP instead of $w(\mathbf{s},t)$. Letting $\mathbf{v}$ the vector formed by stacking up $v(\mathbf{s},t)$ over the set of observed locations, the joint likelihood will be specified as

$$\begin{aligned}
&\prod_{(\mathbf{s},t)} N(y(\mathbf{s},t)\,|\,x(\mathbf{s},t)\beta + \sigma x(\mathbf{s},t)^{\phi/2}v(\mathbf{s},t),\tau^2) \times N(\mathbf{v}\,|\,\mathbf{0},\tilde{\mathbf{C}}(\boldsymbol{\theta})) \\
&\times p(\boldsymbol{\theta}) \times p(\sigma) \times p(\phi) \times IG(\tau^2\,|\,a_\tau,b_\tau) \times N(\boldsymbol{\beta}\,|\,\boldsymbol{\mu}_\beta,\mathbf{V}_\beta)\ , \qquad \text{(S1)}
\end{aligned}$$

where $\tilde{\mathbf{C}}(\boldsymbol{\theta})$ is the DNNGP covariance matrix created from the original non-separable covariance matrix for the process $v(\mathbf{s},t)$. The Gibbs' sampler steps are similar to those detailed in Sections 5.1 and 5.2. The spatio-temporal random effects $v(\mathbf{s},t)$ are updated sequentially and have conjugate normal distributions similar to those defined in (5.4). The covariance parameters $\boldsymbol{\theta}$ are updated using the random walk Metropolis step (5.5). The additional parameters σ and ϕ are also easily updated using the Metropolis random walk step.

Alternatively, the non-stationary variance $\sigma^2(\mathbf{s},t)$ can be modeled as a smooth spatial process for e.g. $\log\sigma^2(\mathbf{s},t) \sim GP$. In such a case, we can use DNNGPs to approximate the stationary covariances of both $v(\mathbf{s},t)$ and $u(\mathbf{s},t) = \log\sigma^2(\mathbf{s},t)$. However, unlike $v(\mathbf{s},t)$ the full conditional for $u(\mathbf{s},t)$ is not Gaussian. Instead, $p(u(\mathbf{s},t)\,|\,\cdot)$ is proportional to

$$N(y(\mathbf{s},t)\,|\,x(\mathbf{s},t)\beta + \exp(u(\mathbf{s},t)/2)\ v(\mathbf{s},t),\tau^2) \times N(\mathbf{u}\,|\,\mathbf{0},\tilde{\mathbf{C}}_u(\boldsymbol{\theta}_u))\ ,$$

where $\tilde{\mathbf{C}}_u(\boldsymbol{\theta}_u)$ is the DNNGP covariance matrix for $\mathbf{u}$ – the vector of random effects $u(\mathbf{s},t)$, and $\boldsymbol{\theta}_u$ are the associated covariance parameters. Evaluating this conditional likelihood for all $(\mathbf{s},t)$ for Metropolis updates will still require $O(n)$ flops and scalability will be retained.

Although, theoretically DNNGP can be easily extended to model some non-stationary covariances in a scalable manner, extensive simulation studies needs to be conducted to actually determine the accuracy of parameter estimation and kriging in a non-stationary setup. MCMC convergence behavior for the spatio-temporal random effects in absence of conjugate Gibbs' updates also needs to be investigated. Furthermore, more general non-stationary structures as specified in Paciorek and Schervish (2006) also needs to be considered. These explorations, however, do not fall within the scope of the current paper and we identify them as areas of future research.

References.

Datta, A., Banerjee, S., Finley, A. O. and Gelfand, A. E. (2016). Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets. *Journal of the American Statistical Association* In Press.

Gelfand, A., Schmidt, A., Banerjee, S. and C., S. (2004). Nonstationary multivariate process modeling through spatially varying coregionalization. *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research* **13** 263-312.

Gramacy, R. B. and Apley, D. W. (2015). Local Gaussian Process Approximation for Large Computer Experiments. *Journal of Computational and Graphical Statistics* **24** 561-578.

Paciorek, C. J. and Schervish, M. J. (2006). Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics* 483–506.