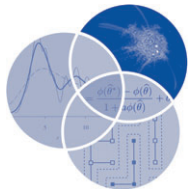
Overview



On nearest-neighbor Gaussian process models for massive spatial data

Abhirup Datta,[1] Sudipto Banerjee,[2*] Andrew O. Finley[3,4] and Alan E. Gelfand[5]

Gaussian Process (GP) models provide a very flexible nonparametric approach to modeling location-and-time indexed datasets. However, the storage and computational requirements for GP models are infeasible for large spatial datasets. Nearest Neighbor Gaussian Processes (Datta A, Banerjee S, Finley AO, Gelfand AE. Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. *J Am Stat Assoc* 2016., JASA) provide a scalable alternative by using local information from few nearest neighbors. Scalability is achieved by using the neighbor sets in a conditional specification of the model. We show how this is equivalent to sparse modeling of Cholesky factors of large covariance matrices. We also discuss a general approach to construct scalable Gaussian Processes using sparse local kriging. We present a multivariate data analysis which demonstrates how the nearest neighbor approach yields inference indistinguishable from the full rank GP despite being several times faster. Finally, we also propose a variant of the NNGP model for automating the selection of the neighbor set size.

How to cite this article:
WIREs Comput Stat 2016, 8:162–171. doi: 10.1002/wics.1383

Keywords: Bayesian methods and theory, computational Bayesian methods, image and spatial data, data structures

INTRODUCTION: GAUSSIAN PROCESSES

The growing capabilities of Geographical Information Systems (GIS) have resulted in a deluge of geo-indexed datasets observed over a very large number of locations. Gaussian Process-based models provide a very flexible non-parametric approach to capture the spatial patterns in such datasets. A Gaussian Process (e.g., Refs 1,2) over a spatial domain $\mathcal{D} \subset \mathbb{R}^d$ is a random surface $w(s)$ over D with a well-defined probability law. It is customarily specified as $w(s) \sim GP(m_\theta(\cdot), C_\theta(\cdot,\cdot))$, where $m_\theta(s)$ is a mean function and $C_\theta(s, t)$ is a covariance function. This can be extended to a multivariate setup, where $w(s) = (w_1(s), w_2(s), \ldots, w_q(s))'$ is a $q \times 1$ vector and $C_\theta(\cdot,\cdot)$ is a valid cross-covariance function (e.g., Ref 3), that is, for any two locations s and t, $C_\theta(s, t)$ is a $q \times q$ cross-covariance matrix whose (i, j)th entry is $\text{cov}\{w_i(s), w_j(t)\}$. The advantage of a Gaussian Process is that its finite dimensional realizations follow a multivariate Gaussian distribution. Specifically, if $w = (w(s_1)', w(s_2)', \ldots, w(s_n)')'$ is the $nq \times 1$ vector of realizations of $w(s)$ over a set of locations in $S = \{s_1, s_2, \ldots, s_n\}$, then

*Correspondence to: baner009@umn.edu

[1]Department of Biostatistics, University of Minnesota, Minneapolis, MN, USA

[2]Department of Biostatistics, University of California, Los Angeles, CA, USA

[3]Department of Forestry, Michigan State University, East Lansing, MI, USA

[4]Department of Geography, Michigan State University, East Lansing, MI, USA

[5]Department of Statistical Science, Duke University, Durham, NC, USA

Conflict of interest: The authors have declared no conflicts of interest for this article.

$$p(w|\theta) \propto \frac{1}{\det(C_\theta)} \exp\left(-\frac{1}{2}(w-m_\theta)' C_\theta^{-1}(w-m_\theta)\right) \quad (1)$$

where $m_\theta = E[w|\theta]$ is the $nq \times 1$ mean vector and $C_\theta = cov(w|\theta)$ is the $nq \times nq$ covariance matrix with (i, j)th block $C_\theta(s_i, s_j)$.

The popularity of Gaussian Processes among non-parametric models is largely indebted to their unparalleled out-of-sample predictive performance and ability to produce a stochastically interpolated surface with uncertainty quantified predictions at new locations. The latter (known as kriging) can be simply achieved using properties of multivariate Gaussian distributions. For example, if $m_\theta(\cdot)$ is fixed and known over the entire domain, then, given the values of $w(s_i)$ for locations in S, we can interpolate the value of the process at any arbitrary point s using

$$w(s) = m_\theta(s) + \sum_{j=1}^{n} A_{\theta j}(s)\left(w(s_j) - m_\theta(s_j)\right) + \eta(s) \quad (2)$$

where $\eta(s) \overset{ind}{\sim} N\left(0, C_\theta(s,s) - C_\theta(s,S)\, C_\theta^{-1}\, C_\theta(S,s)\right)$, $C_\theta(s, S)$ is the $1 \times n$ block-matrix with $C_\theta(s, s_i)$ as the ith block, $C_\theta(S, s)) = C_\theta(s, S)'$ and the $A_{\theta j}(s)$s are $q \times q$ 'kriging weights' obtained by solving the linear system $C_\theta A_\theta(s) = C_\theta(S, s)$ for $A_\theta(s) = [A_{\theta 1}(s) : A_{\theta 2}(s) : \ldots : A_{\theta n}(s)]'$. This results in $m_\theta(s_i) + \sum_{j=1}^{n} A_{\theta j}(s_i)\left(w(s_j) - m_\theta(s_j)\right) = w(s_i)$ for each s_i in S. In fact, the right hand side of Eq. (2) is a *deterministic interpolator* since var $(\eta(s_i)) = 0$ for each s_i in S.

In practice, we often observe a $q \times 1$ variate response $y(s)$ along with a set of explanatory variables $X(s)$ and the GP $w(s)$ is used to capture the spatial patterns beyond the observed covariates. In a Bayesian setting, inference proceeds from the posterior distribution of all unknown parameters and spatial random effects, which is proportional to

$$p(\theta, \tau, \beta) \times N(w|0, C_\theta) \times \prod_{i=1}^{n} N(y(s_i)|X(s_i)\beta + w(s_i), D(\tau)) \quad (3)$$

The noise covariance $D(\tau)$ represents measurement error or microscale variation and depends upon a set of variance parameters τ. The regression component $X(s_i)\beta$ conveniently retains the nice interpretive properties, whereas $w(s)$ represents effects from unknown or unobserved factors. Without loss of generality the mean function of the GP is now taken to be zero as it corresponds to the residual process after adjusting for the regression. Additionally, in a Bayesian setup, priors for β, θ, and τ are included to complete the joint model specification.

If the number of locations n is large, traditional GP models run into multiple computational issues. Evaluation of the Gaussian density for w in Eqs (1) or (3) involves storing C_θ and computing its inverse and determinant. Memory requirements for storing the $nq \times nq$ matrix C_θ is $O(n^2q^2)$ and may exhaust storage resources for large n. Even if C_θ can be stored, computing the inverse and determinant are both very expensive operations. This is typically best achieved using the Cholesky decomposition $C_\theta = L\Lambda L'$, where L is lower triangular and Λ is diagonal. However, the Cholesky decomposition takes $O(n^3q^3)$ floating point operations (flops), so even for an univariate response ($q = 1$) and modestly large n ($\approx 50{,}000$) the computational demands cannot be met by a modern computer. This is often termed the 'big-N' problem in spatial statistics. This has necessitated computationally efficient alternatives to the traditional GP models.

In the next section, we provide a brief overview of the existing, and still growing, literature on modeling spatial and spatio-temporal datasets. Subsequently, we focus upon a class of highly scalable sparsity-inducing Gaussian process models for massive space-time data. This approach has recently been explored in Refs 4,5, called Nearest-Neighbor Gaussian Processes (NNGP) using sparse Gaussian distributions on directed acyclic graphs. In this article, we offer an alternate development of how to construct a well-defined NNGP using essentially the kriging equation. This construction ensures a legitimate spatial process and one does not need to prove Kolmogorov's existence or consistency conditions as described in Ref 4.

The remainder of this article evolves as follows. In Section *Modeling Large Spatial Data* we present a brief survey of approaches for modeling large spatial data, Section *Nearest-Neighbor Gaussian Processes* presents the construction of the NNGP, Section *Multivariate Data Analysis* presents a synthetic application of the NNGP to bivariate spatial outcomes and Section *Choosing m* extends the NNGP to learn about the number of neighbors (as opposed to fixing them). Finally, Section *Conclusion* concludes the article with some remarks.

MODELING LARGE SPATIAL DATA

There is a burgeoning literature on statistical modeling of large spatial and spatio-temporal datasets. Some approaches try to approximate the likelihood in Eq. (1) to obtain a computationally tractable

pseudo-likelihood. Subsequently the covariance parameters θ are estimated using maximum composite likelihood approaches. The likelihood simplifications include pairwise differences composite likelihoods,[6,7] block composite likelihood,[8] small conditioning sets[9,10] among others. Subsequently, kriging at a new location s_0 proceeds by imputing these estimates in Eq. (2). If n is large this can present a computational roadblock as one now needs to store the estimated covariance matrix $C_{\hat{\theta}}$ to calculate the kriging weights $C_{\hat{\theta}}(s_0,S)C_{\hat{\theta}}^{-1}$. For kriging at N new locations, the total flop count post estimation is still $O(n^3 + Nn^2)$. So for large n and N these requirements remain prohibitive.

Compactly supported covariances (CSC)[11–15] are also popularly used to reduce the computational burden. CSC yields sparse correlation and precision structures to expedite calculations. Despite the sparsity, computing $\det(C_\theta)$ may still remain problematic. Also, empirical studies revealed that tapered covariance-based approaches may sometimes perform worse than simple alternatives like block diagonal covariance matrices[16] or nearest-neighbor type local approximations.[17] Yet another approach embeds the irregular locations in a larger regular lattice to construct computationally tractable covariance matrices utilizing spectral properties[18] or Gaussian Markov Random Fields.[19] Inferences from these methods are limited to the resolution of the embedding lattice and cannot interpolate at finer resolutions.

In machine learning, Gaussian Process regression is used to facilitate uncertainty-quantified interpolation and smoothing of an outcome observed at a large number of inputs (locations). When the input-dimension d is large, Ref 20 exploits dimension-wise additive or multiplicative structure in the covariance functions to reduce computation using a multitude of fast algorithms and optimizations. The computational complexity of their algorithms decrease with increasing dimensionality and becomes asymptotically scalable ($O(n)$) at $d = \infty$. However, for large spatial data ($d = 2$), there is little computational gain accrued. Very recently, the study of Ref 21 demonstrated that exact kriging can be performed in a scalable fashion using a suitable sparse grid design for the locations. While it is feasible to control the input locations for computer experiments, most spatial datasets commonly found in forestry or ecology are observed at a very large number of irregular locations and any computationally scalable exact kriging method remains elusive.

For irregularly located large datasets, there is a considerable amount of literature on low-rank approaches for large spatial data.[2,22–33] Low-rank models approximate the covariance matrix C_θ as the sum of low rank and sparse (diagonal) matrices, that is, $C_\theta \approx BKB' + \Delta$ where B is $n \times r$, K is $r \times r$ and Δ is diagonal. Likelihood computations now only involve inversion of the $r \times r$ matrix K and, as $r \ll n$, the flop count is limited to $O(nr^2 + r^3)$. Some low rank approaches[23,26] can be formulated as well-defined Gaussian Processes over the domain. This is very convenient as it provides an unified platform for parameter estimation and kriging at arbitrary resolutions. Furthermore, a full rank GP prior for the spatial random effects can be simply replaced with a low-rank GP prior in any hierarchical setup like Eq. (3). However, as demonstrated in,[4,34] low rank processes struggle to emulate the inference from the expensive full-rank GPs. The gain in computational efficiency is often offset by oversmoothed estimates of the spatial surface.

Localized GP regression based on few nearest neighbors has also been used to obtain fast kriging estimates. Ref 35 provides fast updates of kriging equations after adding a new location to the input set. Iterative application of their algorithm yields a localized kriging estimate based on a small set of locations (including few nearest neighbors). The local estimate often provides an excellent approximation to the global kriging estimate which uses data observed at all the locations to predict at a new location. However, this assumes that the parameters associated with the mean and covariance of the GP are known or already estimated. Local Approximation GPs (LAGP)[17] extend this further to estimate the parameters at each new location. LAGP thus essentially provides a non-stationary local approximation to a Gaussian Process at every predictive location and can be used to interpolate or smooth the observed data.

Recently, Ref 4 proposed a NNGP for modeling large spatial data. NNGP is a well-defined Gaussian Process over a domain $D \subset \mathbb{R}^d$ and yields finite dimensional Gaussian densities with sparse precision matrices. This has been also extended to a dynamic NNGP for massive spatio-temporal data.[5] Like the Gaussian Predictive Process, NNGP enjoys all the benefits of being a proper GP. It delivers massive scalability both in terms of parameter estimation and kriging. Most importantly it does not oversmooth like low-rank processes and accurately emulates the inference from the full-rank GPs. In the next section, we provide a general method to construct scalable Gaussian processes and demonstrate how particular choices lead to the NNGP and dynamic NNGP models.

NEAREST-NEIGHBOR GAUSSIAN PROCESSES

The computational bottleneck in evaluating Eq. (1) lies in computing the Cholesky decomposition of C_θ. We show here how the NNGP models proposed by Refs 4,5 can be constructed within an alternative framework by appropriately choosing the kriging weights in Eq. (2). Assume, without loss of generality, that $m_\theta = 0$. We will construct the process in two steps. First, we specify a multivariate Gaussian distribution for the realizations of the process over a fixed finite set $S = \{s_1, s_2, \ldots, s_n\}$. Subsequently we extend it to a process over the domain through kriging equations based on the locations in S. These two steps are described below:

$$\begin{aligned} w_S &= \left(w(s_1)', w(s_2)', \ldots, w(s_n)'\right)' \sim N(0, K_\theta) \\ w(s) &= \sum_{j=1}^{n} A_j(s) w(s_j) + \eta(s) \\ \eta(s) &\overset{\text{ind}}{\sim} N(0, \Gamma_\theta(s)) \text{ for any } s \notin S \end{aligned} \tag{4}$$

where K_θ is a $nq \times nq$ positive definite matrix, each $A_{\theta j}(s)$ is a $q \times q$ matrix and each $\Gamma_\theta(s)$ is a $q \times q$ positive definite matrix.

It is straightforward to see that the process defined by Eq. (4) is a well-defined Gaussian Process on the domain D. In fact, as demonstrated in,[5] Eq. (4) subsumes a large class of Gaussian Processes. For example, if we choose $K_\theta = C_\theta$ (the covariance of w_S assuming they are the realizations of the full GP), $A_\theta(s)$ to be the kriging weights $C_\theta(s,S)C_\theta^{-1}$ (as defined in Eq. (2)) and $\Gamma_\theta(s)$ to be the corresponding kriging variance $C_\theta(s,s) - C_\theta(s,S)C_\theta^{-1}C_\theta(S,s)$, then $w(s)$ has the same distribution as the full GP with covariance function $C_\theta(\cdot,\cdot)$. If S is chosen to be a small set of 'knots,' $w(s)$ becomes a low-rank Gaussian Process. Instead of opting for low-rank specifications, we will use Eq. (4) to construct a class of sparsity-inducing NNGP. This construction of the NNGP is different from and perhaps more intuitive than Ref 4 and clearly reveals the NNGP as a fully model-based extension of alternative nearest neighbor approaches such as Refs 9,10.

The construction in Eq. (4) yields a legitimate process for any finite collection of spatial locations S. One convenient choice is to let S be the set of observed locations. Then the choice of K_θ determines the scalability of parameter estimation whereas choices for the matrices $A_\theta(s)$ and $\Gamma(s)$ determine the scalability for kriging. Let $U = \{u_1, u_2, \ldots, u_N\}$ be any finite set of locations outside S. Then, from Eq. (4) we have

$$P(w_{S\cup U}) = N(w_S|0, K_\theta) \times N(w_U|A_U w_S, \Gamma_U) \tag{5}$$

where $A_U = (A_\theta(u_1)', A_\theta(u_2)', \ldots, A_\theta(u_N)')'$ and $\Gamma_U = \text{diag}(\Gamma_\theta(u_1), \Gamma_\theta(u_2), \ldots, \Gamma_\theta(u_N))$ In a full GP, the conditional variance is $C(w_U|w_S) = C_\theta(u,u) - C_\theta(u,S)C_\theta^{-1}C_\theta(S,u)$. So, if two predictive locations u_i and u_j are nearby, their correlation is preserved in the conditional distribution. However, in practice, if the number of predictive locations is large, such joint kriging is rarely used as $C(w_U|w_S)$ is large and generating predictive samples becomes computationally prohibitive. Instead, for large datasets it is customary to carry out predictions independently at each new location. Then the predictive variance becomes $C_\theta(u,u) - C_\theta(u,S)C_\theta^{-1}C_\theta(S,u)$. Independent kriging circumvents large matrix calculations and can also leverage embarrassingly parallel computing if available. Since, our scalable Gaussian Process is designed for large datasets, we build the predictive distribution by incorporating this conditional independence in the second equation of Eq. (4). If two predictive locations are sufficiently close, their neighbor sets will be same and, hence, although conditionally independent, marginally they will exhibit strong correlation. Also, a simulation study comparing the performance of full GP independent kriging, full GP joint kriging and NNGP kriging provided in (Appendix G)[4] reveals that joint kriging hardly yields any noticeable benefits.

The Gaussian Process defined in Eq. (4) will be scalable if, for any finite U the storage and computation requirements for evaluating the joint likelihood $P(w_{S\cup U})$ are both $O(n + N)$, that is, linear in the total number of locations. A prerequisite for this is to construct K_θ^{-1} with $O(n)$ non-zero entries. How can we achieve this?

One approach would be to avoid computing the Cholesky factorization and directly *model* $K_\theta = L\Lambda L'$, where L is unit lower-triangular (i.e., has I_q's or $q \times q$ identity matrices along the diagonal) and Λ is block-diagonal with $q \times q$ positive-definite submatrices along the diagonal. In other words, we model the elements of L and Λ to construct K_θ. Note that $\det(K_\theta)$ is the product of the diagonal entries of Λ. Since $w_S \sim N(0, L\Lambda L')$, we have $L^{-1}w_S \sim N(0, \Lambda)$. Letting $L^{-1} = I - A$ we have $w_S = Aw_S + N(0, \Lambda)$. Since, L is lower triangular, so is A with $q \times q$ blocks $A_{i,j}$. If Λ_i denotes the ith diagonal block of Λ, we have

$$w(s_i) \sim N\left(\sum_{j=1}^{i-1} A_{i,j} w(s_j), \Lambda_i\right) \tag{6}$$

Let $S_i = \{s_1, s_2, \ldots, s_{i-1}\}$. If $\{A_{i,j}|j = 1, 2, \ldots, i-1\}$ are chosen to be the kriging weights of $w(s_i)$ based

on w_{S_i} (analogous to Eq. (2)), and Λ_i the corresponding kriging variance, we obtain $K_\theta = C_\theta$. However, this involves inverting $(i-1)q \times (i-1)q$ matrices $C_\theta(S_i, S_i)$ which will be computationally cumbersome for larger i.

Computational efficiency can be achieved by limiting the number of non-zero $A_{i,j}$ blocks in each row of A to at most m where $m \ll n$. Let $N(s_i)$ denote $\{s_j | A_{i,j} \neq O\}$. The non-zero $A_{i,j}$s can then be obtained as the kriging weights of $w(s_i)$ based on $w_{N(s_i)}$. The kriging variance Λ_i becomes $\Lambda_i = C_\theta(s_i, s_i) - C_\theta(s_i, N(S_i))\ C_\theta(N(S_i), N(S_i))^{-1}\ C_\theta(N(S_i), s_i)$. So, calculating $A_{i,j}$s, Λ_is now only require inverting $mq \times mq$ matrices. The determinant of K_θ can now be easily calculated as $\det(K_\theta) = \prod_{i=1}^{n} \det(\Lambda_i)$. Hence, the total flop count for evaluating the density $N(w_S | 0, K_\theta)$ can be shown to be $O(nm^3q^3)$.

The size of the neighbor sets control the sparsity of K_θ^{-1} whereas the choice of the neighbor sets plays a vital role in determining how accurately K_θ can approximate C_θ. Refs 4,10 used m nearest neighbors of s_i among S_i to construct $N(s_i)$. As nearest neighbors corresponds to points with highest spatial correlation, this choice produced excellent approximations. In a spatio-temporal domain without any definition of distance, Ref 5 selected adaptive neighbor sets $N_\theta(s_i)$ based on the strength of the correlation function $C_\theta(\cdot,\cdot)$.

NNGP uses nearest neighbors to create the small conditioning sets. Nearest neighbors have been shown to be sub-optimal for predicting at a new location—theoretically, for some special designs, in Ref 9 and empirically in Refs 17,35 However, the alternatives proposed do not guarantee better performance. For example, Ref 9 proposes using few nearest neighbors and few distant ones. This often performs worse than nearest neighbors (Ref 4 Appendix I). The study of Ref 17 chooses the neighbor set sequentially by maximizing the reduction in predictive variance at each step of expanding the neighbor set. Ref 35 demonstrated that the optimal neighbor sets are not necessarily nested as their size increases. Hence, a greedy forward selection procedure like the one proposed in Ref 17 is not guaranteed to yield the optimal subset of size m. The construction of scalable Gaussian Processes does not require the neighbor sets to be nearest neighbors. Any subset of size m—including the choices used by Ref 9 can be used to construct valid sparse Gaussian Processes. However, through extensive simulations over a very wide range of scenarios documented in Ref 4, the simple choice of nearest neighbors is vindicated by the consistently accurate approximations of the full GP by NNGP.

Finally, we turn to define $A_\theta(s) = [A_{\theta 1}(s) : A_{\theta 2}(s) : \ldots : A_{\theta n}(s)]$ and $\Lambda_\theta(s)$ for every $s \notin S$. Once again sparsity is achieved by defining a small neighbor set $N(s) \subset S$ and choosing the $q \times q$ blocks $A_{\theta j}(s) = O$ if $s_j \notin N(s)$. The non-zero $A_{\theta j}(s)$s are simply obtained as the kriging weights for $w(s)$ on $w_{N(s)}$. Choosing $\Lambda_\theta(s)$ as the kriging variance $\mathrm{var}(w(s) | w_{N(s)})$ completes the process specification. Evaluating the likelihood $N(w(s) | A_\theta(s) w_S, \Lambda_S)$ only requires inverting an $mq \times mq$ matrix. So for inference over S and N locations outside S, the total storage and computational requirements are $O((N+n)m^2q^2)$ and $O((n+N)m^3q^3)$, respectively. As these requirements are linear in $N+n$ and m is typically very small (≈ 20) this ensures that the resulting GP is scalable to massive datasets.

As NNGP is a proper Gaussian process, we can use it as a prior for the spatial random effects in any hierarchical model formulation. For example, the Bayesian model in Eq. (3) now becomes:

$$p(\theta,\tau,\beta) \times N(w | 0, K_\theta) \times \prod_{i=1}^{n} N(y(s_i) | X(s_i)\beta + w(s_i), D(\tau))$$

An efficient Markov chain Monte Carlo (MCMC) sampler using Gibbs steps and random-walk Metropolis steps described in Ref 4 can now be used for updating $w(s)$, β and the covariance parameters θ. Predictions at arbitrary locations in $U = \{u_1, u_2, \ldots, u_N\}$ are performed by sampling from the posterior predictive distribution

$$\int N(y_U | X_U\beta + w_U, D_U(\tau)) \times N(w_U | A_\theta(U) w_S, \Gamma_\theta(U)) \times p(w_S, \beta, \tau, \theta | y_S)$$

where $y_U = (y(u_1)', y(u_2)', \ldots, y(u_N)')'$, X_U is the corresponding covariance matrix, $D_U(\tau) = \mathrm{diag}(D(\tau), D(\tau), \ldots, D(\tau))$, $A_\theta(U) = (A_\theta(u_1)', A_\theta(u_2)', \ldots, A_\theta(u_N)')'$ is the sparse kriging weights matrix and $\Gamma_\theta(U) = \mathrm{diag}(\Gamma_\theta(u_1), \Gamma_\theta(u_2), \ldots, \Gamma_\theta(u_N))'$. We refer the reader to Refs 4,5 for details on these sampling methods.

MULTIVARIATE DATA ANALYSIS

We present the results of a multivariate simulation study to demonstrate the accuracy of scalable Gaussian Processes. The synthetic data comprises of $q = 2$ responses at each of $n = 1000$ locations within a unit square domain. The bivariate responses were generated as:

TABLE 1 | Multivariate Synthetic Data Analysis Parameter Estimates and Computing Time in Minutes for Candidate Models. Parameter Posterior Summary 50 (2.5, 97.5) Percentiles

		NNGP		Full Gaussian Process
	True	$m = 5$	$m = 10$	
$\beta_{1,0}$	1	0.81 (0.22, 1.43)	0.64 (–0.05, 1.45)	0.82 (–0.11, 1.71)
$\beta_{1,1}$	–5	–4.94 (–5.02, –4.85)	–4.95 (–5.04, –4.86)	–4.95 (–5.04, –4.86)
$\beta_{2,0}$	1	1.03 (0.15, 2.02)	1.31 (0.26, 2.37)	0.95 (–0.27, 2.18)
$\beta_{2,1}$	5	5.03 (4.91, 5.15)	5.02 (4.89, 5.14)	5.01 (4.89, 5.13)
$AA'_{1,1}$	4	3.88 (3.14, 5.13)	4.06 (3.20, 5.80)	4.20 (3.21, 5.47)
$AA'_{2,1}$	–4	–3.58 (–4.80, –2.87)	–3.66 (–5.50, –2.86)	–3.79 (–4.96, –2.87)
$AA'_{2,2}$	8	7.31 (6.02, 9.15)	7.43 (6.07, 9.70)	7.18 (5.74, 9.15)
τ_1^2	0.1	0.07 (0.02, 0.25)	0.06 (0.02, 0.25)	0.07 (0.02, 0.24)
τ_2^2	0.1	0.10 (0.02, 0.87)	0.08 (0.02, 0.70)	0.09 (0.03, 1.33)
ϕ_1	6	6.98 (4.05, 15.33)	7.09 (3.21, 14.61)	6.95 (3.79, 12.19)
ϕ_2	6	4.14 (3.20, 8.07)	4.80 (3.14, 9.87)	5.47 (3.41, 12.07)
ν_1	0.25	0.25 (0.19, 0.38)	0.27 (0.20, 0.37)	0.27 (0.21, 0.37)
ν_2	0.25	0.22 (0.16, 0.42)	0.22 (0.15, 0.37)	0.23 (0.16, 0.68)
DIC	–	845.47	747.82	934.73
GPD	–	30,666.13	30,782.05	36,182.24
RMSPE	–	1.68	1.67	1.67
% CI coverage	–	94.7	94.7	94.1
Mean 95% CI width	–	6.31	6.24	6.14
Time (in minutes)	–	18.82	75.62	369.10

$$y(s) = \begin{pmatrix} y_1(s) \\ y_2(s) \end{pmatrix} = N\left(\begin{pmatrix} \beta_{10} + x_1(s)\beta_{11} + w_1(s) \\ \beta_{20} + x_1(s)\beta_{21} + w_2(s) \end{pmatrix}, \begin{pmatrix} \tau_1^2 & 0 \\ 0 & \tau_2^2 \end{pmatrix} \right) \tag{7}$$

The spatial random effects $w(s) = (w_1(s), w_2(s))'$ are generated from a bivariate Gaussian process with a isotropic cross-covariance specification $C(s_i, s_j|\theta) = A\Delta A'$, where A is 2×2 lower-triangular with positive diagonal elements, Δ is 2×2 diagonal with $\rho(s_i, s_j; \phi_b, \nu_b)$ as the bth diagonal entry, $b = 1, 2$ (e.g., Ref 36). Here $\rho(s_i, s_j; \phi_b, \nu_b)$ denote the Matérn correlation function, that is,

$$\rho(s_i, s_j; \phi, \nu) = \frac{1}{2^{\nu-1}\Gamma(\nu)} \left(||s_i - s_j||\phi\right)^{\nu} K_{\nu}\left(||s_i - s_j||\phi\right); \quad \phi > 0, \nu > 0 \tag{8}$$

where $||s_i - s_j||$ is the Euclidean distance between locations s_i and s_j with ϕ controlling the decay in spatial correlation and ν controlling the process smoothness.

The *true* parameter values used to generate the data are shown in the first column of Table 1. The non-intercept predictor columns x_1 and x_2 were drawn from $N(0, 1)$. In Table 1, the first subscript on the regression coefficients index the response variable, for example, $\beta_{1,0}$ and $\beta_{1,1}$ are the intercept and slope associated with the first response variable. Here too, response-specific subscripts index the elements of the cross-covariance matrix AA', nuggets τ_1^2 and τ_2^2, and correlation parameters $\{(\phi_b, \nu_b)|b = 1, 2\}$. Candidate models included the full Gaussian Process (*Full GP*) and NNGP with $m = 5$ and $m = 10$. The models were assessed using model fit and out-of-sample prediction. Metrics used were Deviance Information Criterion (DIC),[37] GPD score[38] or Root Mean Square Predictive Error (RMSPE).[39]

For all models, the intercept and slope regression parameters were given *flat* prior distributions. The variance components τ_1^2 and τ_2^2 were assigned inverse-Gamma $IG(2, 1)$ priors, AA' followed an inverse-Wishart $IW(3, 0.1)$, and the Matérn spatial decay and smoothness parameters received uniform prior supports $U(3, 30)$ and $U(0.1, 1)$, respectively. These prior distributions on ϕ and ν correspond to support between approximately 0.05 and 1.3 domain distance units.

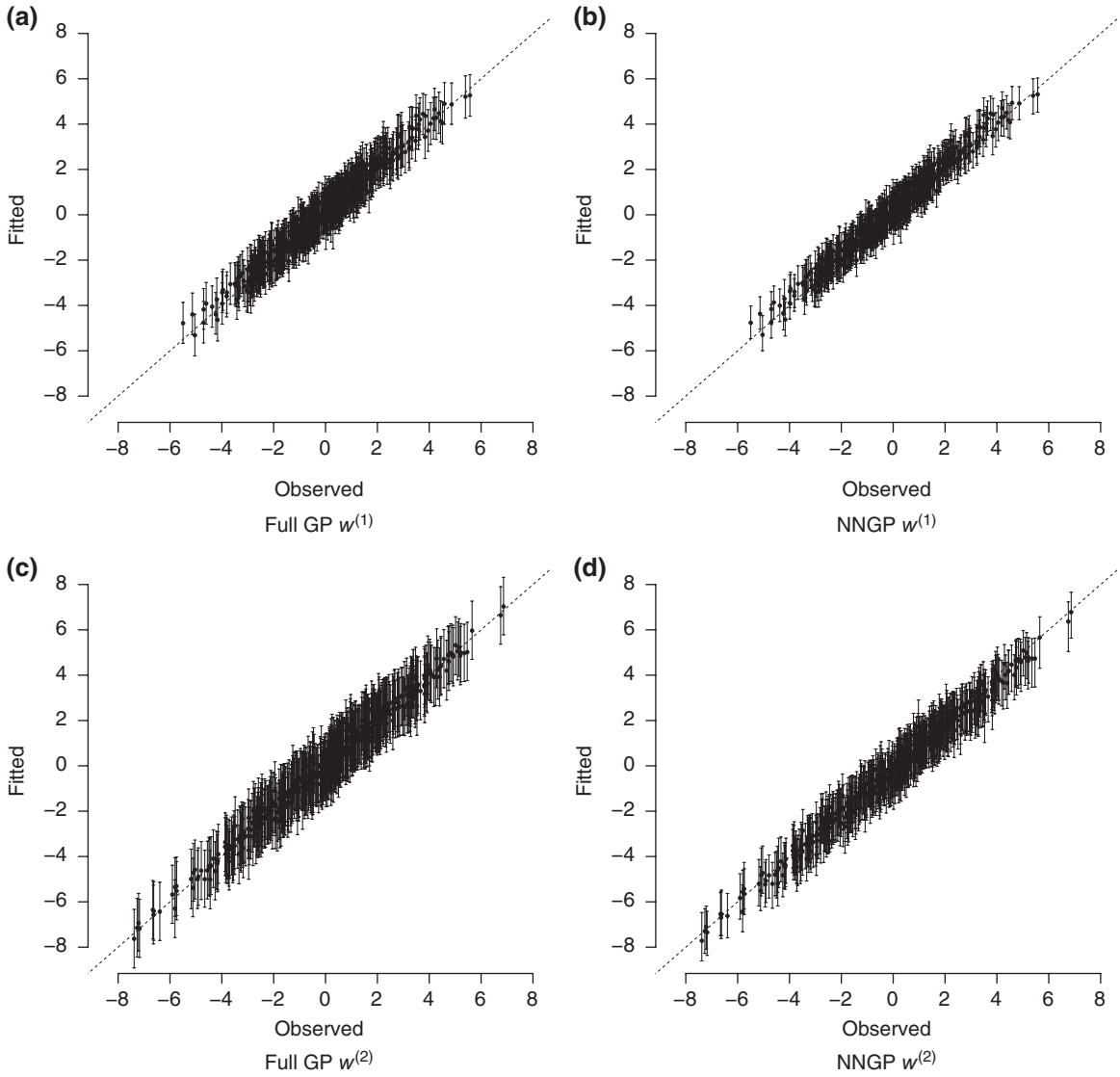


FIGURE 1 | Multivariate synthetic data analysis true versus fitted spatial random effects w for the full Gaussian Process (Full GP) and nearest neighbor Gaussian process (NNGP) for $m = 5$. Superscripts on $w^{(1)}$ and $w^{(2)}$ correspond to the random effects associated with the first and second response variables. (a) Full GP $w^{(1)}$, NNGP $w^{(1)}$, (c) Full GP $w^{(2)}$ and (d) NNGP $w^{(2)}$

Candidate model parameter estimates and performance metrics based on 25,000 iterations are provided in Table 1. We discarded the first 5000 iterations as burn-in samples.

All model specifications produce similar posterior median estimates and 95% credible intervals that contain the *true* parameter values. DIC and GPD scores suggest the NNGP models have improved fit over the Full GP model. Figure 1 provides scatter plots of the response specific *true* random effects versus posterior estimates from the Full GP and NNGP models. These plots show that both models closely approximate the true multivariate process. Further, both models produce ∼99% coverage rates for the random effects associated with the first and second response variables. However, the NNGP model provides marginally narrower credible interval widths, that is, NNGP mean widths of 1.66 and 1.77 for the first and second response, respectively, versus 1.8 and 2.6 from the Full GP model. The narrowing of the credible intervals is particularly apparent when comparing Figures 1a and 1b.

Turning to the out-of-sample prediction results, the NNGP and Full GP models produced comparable RMSPE and mean 95% credible interval widths as shown in Table 1. All models showed appropriate 95% credible interval coverage rates.

The last row in Table 1 provides model computing times for one chain of 25,000 iterations. Clearly the reduced dimensionality and computational complexity of the NNGP models provide enormous reductions in computing time over the conventional GP sampler.

CHOOSING *m*

The size of the neighbor sets m has direct impact on the storage and computation of a NNGP model. Simulation experiments detailed in Ref 4 demonstrates how one can run NNGP models for different choices of m (possibly in parallel) and choose m which minimizes some model evaluation metrics like RMSPE. Nevertheless, for large datasets or in absence of parallel resources, running the MCMC for multiple values of m may remain a computational challenge. We propose an alternative method that bypasses the need for multiple runs. We consider m as a parameter and update it dynamically in the Gibbs' sampler. The neighbor sets and consequently the NNGP covariance K_θ now change with m. We denote them by $N_m(s)$ and $K_\theta(m)$ respectively. If $\pi(m)$ denotes a prior for the discrete parameter m, we can generate samples using the full conditional

$$p(m|\cdot) \propto \pi(m)N(w|0, K_\theta(m)) \tag{9}$$

The support of $\pi(m)$ is chosen to be $[1, m_0]$ where m_0 can be determined from the available computational resources, that is, choose m_0 for which $O(nm_0^2)$ storage is available and $O(nm_0^3)$ time is reasonable. Since, Eq. (9) does not correspond to any standard distribution, m can be updated using a Metropolis random walk step within the Gibbs' sampler.

We illustrate the performance of this variable-m NNGP in a small simulation experiment. We generated data at 2000 locations within the unit square from the univariate model $y(s) = \beta_1 x_1(s) + \beta_2 x_2(s) + w(s) + \varepsilon(s)$, where x_1 and x_2 were generated from $N(0, 1)$, $\varepsilon(s) \overset{iid}{\sim} N(0, \tau^2)$ and $w(s) \sim$ a GP with exponential covariance structure (Matern covariance with $\nu = 0.5$) with decay ϕ. For model fitting, we used a NNGP prior for $w(s)$, improper prior ($\propto 1$) for β_1 and β_2, uniform prior for ϕ and inverse-gamma prior for σ^2 and τ^2.

We used 10,000 MCMC iterations and discarded the first 5000 as burn-ins. For the variable-m NNGP we used a discrete uniform prior for m with support $\{1, 2, \ldots, 20\}$. We also fit regular NNGP models with $m = 10$ and $m = 20$ as benchmarks. The true and fitted values of the parameters are shown in Table 2. We observe that the coefficient estimates for β_1, β_2, and τ^2 were similar for all three models, whereas the estimate of σ^2 was much better for the variable-m NNGP. Also the posterior 95% credible intervals for σ^2 and ϕ were significantly narrower than those for the regular NNGPs. This observation is slightly surprising as the possible range of values of m includes both 10 and 20—the size of the neighbor sets for the regular NNGPs. Nevertheless, we observed that varying the size of the neighbor sets can yield improved confidence intervals in addition to saving additional computations for running multiple NNGPs.

Computationally, the variable-m NNGP adds little to the storage and memory requirements of the standard NNGP. If m_0 denotes the upper bound for the support of m, then the storage requirements are $O(nm_0^2)$ and the computational complexity for evaluating the NNGP likelihood at the lth MCMC iteration will be $O(nm(l)^3)$ where $m(l)$ denotes the updated value of m at the l^{th} iteration.

Figure 2 shows the posterior frequency distribution of m in the variable m NNGP. While the distribution of m is not interesting for any inferential purposes, the figure shows that although the prior is discrete uniform $\{1, 2, \ldots, 20\}$, the posterior shows heavy left skewness with higher mass for larger values of m. This is expected as the data are generated from a full-rank GP which essentially corresponds to $m = n$. However, if computational resources demand that more mass towards smaller values of m is desirable, one can add a suitable prior penalizing large m.

TABLE 2 | Parameter Estimates (50% [2.5%, 97.5%] percentiles) for Fixed and Variable m NNGP Models

	True	NNGP $m = 10$	NNGP $m = 20$	NNGP Variable-m
β_1	1	0.98 (0.95 1.02)	0.99 (0.95 1.02)	0.99 (0.95 1.02)
β_2	5	4.98 (4.95 5.02)	4.98 (4.95 5.02)	4.98 (4.95 5.02)
σ^2	1	1.64 (0.78 6.39)	1.36 (0.65 5.72)	1.17 (0.71 2.17)
τ^2	0.5	0.47 (0.43 0.5)	0.47 (0.44 0.51)	0.47 (0.44 0.51)
ϕ	1	1.01 (0.26 2.55)	1.13 (0.26 2.87)	1.36 (0.66 2.21)

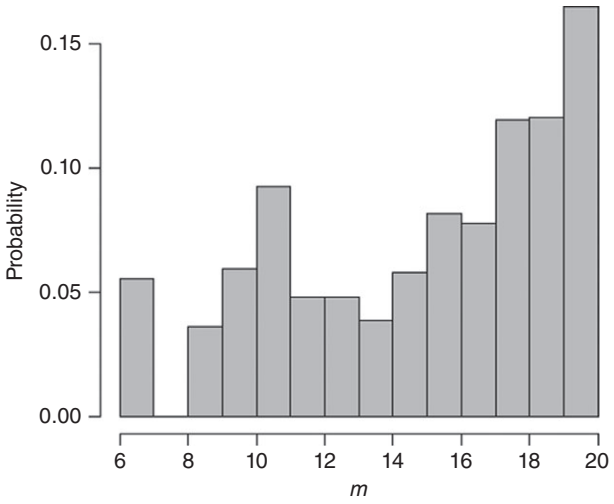


FIGURE 2 | Posterior distribution of m for variable-m NNGP.

CONCLUSIONS

Fully model-based Bayesian inference for large spatial and spatiotemporal datasets is challenging because of expensive computations involving matrices without apparent exploitable structure. Recently Ref 4 proposed a NNGP which is a GP whose realizations over any finite collection of locations will have a sparse precision matrix. In this expository article, we have provided some further insight into the NNGP models and how they can be easily constructed as linear transformations of the process realizations over a fixed set of locations. We not only illustrate their use in Bayesian hierarchical modeling but also add some additional insight by modeling (rather than fixing) the number of neighbors.

The construction of NNGP requires a preordering of the spatial locations. While the choice of ordering has been empirically shown to have little impact on the performance of the NNGP[4] or other nearest neighbor based approaches,[9,10] it remains an annoyance from a purely theoretical perspective as spatial locations do not have any natural ordering.

Also, NNGP, in its current form is only constructed using stationary (or isotropic) covariance functions. Extending NNGP to create sparse Gaussian Processes for modeling non-stationary spatial surfaces also remains challenging. Kriging based on few Euclidean nearest neighbors may no longer remain accurate if the stationarity assumption is not valid. LAGP[17] possess advantage in this aspect as the covariance parameters are estimated at each predictive location thereby incorporating non-stationarity. The implementation of LAGP in the R-package *laGP* is also extremely efficient.

The storage and computational requirements of NNGP are linear in size of the dataset and the dimension of the locations. Hence, it can easily scale to datasets with hundreds of thousands or possibly millions of high-dimensional locations. One potential area of concern is that, the spatial random effects are updated sequentially in the MCMC algorithm for NNGP described in.[4] While empirical observations reveal that convergence is achieved very fast, a block update of the spatial random effects may speed up the MCMC significantly. Two possible algorithms were described in Ref 4 but their performances were not evaluated. More research is being undertaken to fully explore these possibilities. We also plan to migrate our lower level C++ code into an R package for wider accessibility.

ACKNOWLEDGMENTS

The authors thank the Associate Editor and anonymous reviewers for their suggestions which considerably improved the manuscript. Sudipto Banerjee was supported by NSF DMS-1513654. Andrew Finley was supported by National Science Foundation (NSF) DMS-1513481, EF-1137309, EF-1241874, and EF-1253225, as well as NASA Carbon Monitoring System grants.

REFERENCES

1. Cressie NAC, Wikle CK. *Statistics for Spatio-temporal Data*. Wiley Series in Probability and Statistics. Hoboken, NJ: Wiley; 2011.
2. Rasmussen CE, Williams CKI. *Gaussian Processes for Machine Learning*. 1st ed. Cambridge, MA: The MIT Press; 2005.
3. Banerjee S, Carlin BP, Gelfand AE. *Hierarchical Modeling and Analysis for Spatial Data*. 2nd ed. Boca Raton, FL: Chapman & Hall/CRC; 2014.
4. Datta A, Banerjee S, Finley AO, Gelfand AE. Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. *J Am Stat Assoc* 2016. doi:10.1080/01621459.2015.1044091.
5. Datta A, Banerjee S, Finley AO, Hamm NAS, Schaap M. Non-separable dynamic nearest-neighbor gaussian process models for large spatio-temporal data with an application to particulate matter analysis. *Ann Appl Stat* 2016. Available at: http://imstat.org/aoas/next_issue.html.

6. Bai Y, Song PXK, Raghunathan TE. Bayesian dynamic modeling for large space-time datasets using gaussian predictive processes. *J R Stat Soc B* 2012, 74:799–824.
7. Bevilacqua M, Gaetan C, Mateu J, Porcu E. Estimating space and space-time covariance functions for large data sets: a weighted composite likelihood approach. *J Am Stat Assoc* 2012, 107:268–280.
8. Eidsvik J, Shaby BA, Reich BJ, Wheeler M, Niemi J. Estimation and prediction in spatial models with block composite likelihoods. *J Comput Graph Stat* 2014, 23:295–315.
9. Stein ML, Chi Z, Welty LJ. Approximating likelihoods for large spatial data sets. *J R Stat Soc B* 2004, 66:275–296.
10. Vecchia AV. Estimation and model identification for continuous spatial processes. *J R Stat Soc B* 1988, 50:297–312.
11. Du J, Zhang H, Mandrekar VS. Fixed-domain asymptotic properties of tapered maximum likelihood estimators. *Ann Stat* 2009, 37:3330–3361.
12. Furrer R, Genton MG, Nychka D. Covariance tapering for interpolation of large spatial datasets. *J Comput Graph Stat* 2006, 15:503–523.
13. Kaufman CG, Scheverish MJ, Nychka DW. Covariance tapering for likelihood-based estimation in large spatial data sets. *J Am Stat Assoc* 2008, 103:1545–1555.
14. Kaufman CG, Bingham D, Habib S, Heitmann K, Frieman JA. Efficient emulators of computer experiments using compactly supported correlation functions, with an application to cosmology. *Ann Appl Stat* 2011, 5:2470–2492.
15. Shaby BA, Ruppert D. Tapered covariance: Bayesian estimation and asymptotics. *J Comput Graph Stat* 2012, 21:433–452.
16. Stein ML. Statistical properties of covariance tapers. *J Comput Graph Stat* 2013, 22:866–885.
17. Gramacy RB, Apley DW. Local gaussian process approximation for large computer experiments. *J Comput Graph Stat* 2015, 24:561–578.
18. Stroud JR, Stein ML, Lysen S. Bayesian and maximum likelihood estimation for gaussian processes on an incomplete lattice. 2014. Available at: http://arxiv.org/abs/1402.4281.
19. Rue H, Held L. *Gaussian Markov Random Fields: Theory and Applications*. Monographs on Statistics and Applied Probability. Boca Raton, FL: Chapman & Hall/CRC; 2005.
20. Gilboa E, Saatci Y, Cunningham JP. Scaling multidimensional inference for structured Gaussian processes. *IEEE Trans Pattern Anal Mach Intell* 2015, 37:424–436.
21. Plumlee M. Fast prediction of deterministic functions using sparse grid experimental designs. *J Am Stat Assoc* 2014, 109:1581–1591.
22. Banerjee S, Finley AO, Waldmann P, Ericcson T. Hierarchical spatial process models for multiple traits in large genetic trials. *J Am Stat Assoc* 2010, 105:506–521.
23. Banerjee S, Gelfand AE, Finley AO, Sang H. Gaussian predictive process models for large spatial datasets. *J R Stat Soc B* 2008, 70:825–848.
24. Crainiceanu CM, Diggle PJ, Rowlingson B. Bivariate binomial spatial modeling of loa loa prevalence in tropical Africa. *J Am Stat Assoc* 2008, 103:21–37.
25. Cressie N, Johannesson G. Fixed rank kriging for very large data sets. *J R Stat Soc B* 2008, 70:209–226.
26. Finley AO, Banerjee S, McRoberts RE. Hierarchical spatial models for predicting tree species assemblages across large domains. *Ann Appl Stat* 2009, 3:1–32.
27. Higdon D. Space and space time modeling using process convolutions. Technical Report, Institute of Statistics and Decision Sciences, Duke University, Durham, NC, 2001.
28. Kammann EE, Wand MP. Geoadditive models. *Appl Stat* 2003, 52:1–18.
29. Katzfuss M. A multi-resolution approximation for massive spatial datasets. *J Am Stat Assoc* 2016. doi: 10.1080/01621459.2015.1123632.
30. Sang H, Huang JZ. A full scale approximation of covariance functions for large spatial data sets. *J R Stat Soc B Stat Methodol* 2012, 74:111–132.
31. Snelson E, Ghahramani Z. Sparse gaussian processes using pseudo-inputs. In: *Advances in Neural Information Processing Systems*, vol. 18. Cambridge, MA: MIT Press; 2006, 1257–1264.
32. Stein ML. Spatial variation of total column ozone on a global scale. *Ann Appl Stat* 2007, 1:191–210.
33. Stein ML. A modeling approach for large spatial datasets. *J Korean Stat Soc* 2008, 37:3–10.
34. Stein ML. Limitations on low rank approximations for covariance matrices of spatial data. *Spat Stat* 2014, 8:1–19.
35. Emory X. The kriging update equations and their application to the selection of neighboring data. *Comput Geosci* 2009, 13:269–280.
36. Gelfand AE, Banerjee S. Multivariate spatial process models. In: Gelfand AE, Diggle PJ, Fuentes M, Guttorp P, eds. *Handbook of Spatial Statistics*. Boca Raton, FL: CRC Press; 2010, 495–516.
37. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit. *J R Stat Soc B* 2002, 64:583–639.
38. Gelfand AE, Ghosh SK. Model choice: a minimum posterior predictive loss approach. *Biometrika* 1998, 85:1–11.
39. Yeniay O, Goktas A. A comparison of partial least squares regression with other prediction methods. *Hacettepe J Math Stat* 2002, 31:99–111.

