Graphical Gaussian Process Models for Highly Multivariate Spatial Data

Debangan Dey, Abhirup Datta, Sudipto Banerjee

JSM 2021

Johns Hopkins Bloomberg School of Public Health

BIOSTATS

Introduction: Multivariate spatial data

- Most spatial data collection in climatology, forestry, environmental health target multiple variables of interest.
- The goal is to estimate associations over spatial locations for each variable and those among the variables.

Introduction: Multivariate spatial data

- Most spatial data collection in climatology, forestry, environmental health target multiple variables of interest.
- The goal is to estimate associations over spatial locations for each variable and those among the variables.
- Typical marginal spatial regression model:

$$y_i(s) = x_i(s)^{\top}\beta_i + w_i(s) + \epsilon_i(s)\ , \quad i = 1, 2, \ldots, q,\ s \in \mathcal{D} \tag{1}$$

1. q outcomes measured at each location s.
2. $x_i(s)$ is a $p_i \times 1$ vector of predictors.
3. $\epsilon_i(s) \overset{ind}{\sim} N(0, \tau_i^2)$ is the random noise in outcome i.

Introduction: Multivariate Gaussian Processes

- $w(s) = (w_1(s), w_2(s), \ldots, w_q(s))^{\top}$ is modeled as a zero-centred multivariate Gaussian process (GP).
- The cross-covariance is a matrix-valued function $C = (C_{ij}) : \mathcal{D} \times \mathcal{D} \mapsto \mathbb{R}^{q \times q}$ with $C_{ij}(s, s') = \mathrm{Cov}(w_i(s), w_j(s'))$
- C must ensure that for any finite set of locations $\mathcal{S} = \{s_1, \ldots, s_n\}$, the $nq \times nq$ matrix $C(\mathcal{S}, \mathcal{S}) = (C(s_i, s_j))$ is positive definite.

Introduction: Multivariate Gaussian Processes

- $w(s) = (w_1(s), w_2(s), \ldots, w_q(s))^\top$ is modeled as a zero-centred multivariate Gaussian process (GP).
- The cross-covariance is a matrix-valued function $C = (C_{ij}) : \mathcal{D} \times \mathcal{D} \mapsto \mathbb{R}^{q \times q}$ with $C_{ij}(s, s') = \mathrm{Cov}(w_i(s), w_j(s'))$
- C must ensure that for any finite set of locations $\mathcal{S} = \{s_1, \ldots, s_n\}$, the $nq \times nq$ matrix $C(\mathcal{S}, \mathcal{S}) = (C(s_i, s_j))$ is positive definite.

Our contribution:
Highly-multivariate setting with the number of dependent outcomes (q), possibly, in the tens or hundreds at each spatial location.

Motivation

- Most of the research attention is on with massive number of locations (large n) so far while the highly multivariate setting fosters separate computational issues.

Motivation

- Most of the research attention is on with massive number of locations (large n) so far while the highly multivariate setting fosters separate computational issues.
- Likelihoods for popular cross-covariance functions, such as the the multivariate Matérn, involve $O(q^2)$ parameters, and $O(q^3)$ floating point operations (flops).
- Optimizing over or sampling from high-dimensional parameter spaces is inefficient even for modest values of n.

Motivation

- Most of the research attention is on with massive number of locations (large n) so far while the highly multivariate setting fosters separate computational issues.
- Likelihoods for popular cross-covariance functions, such as the the multivariate Matérn, involve $O(q^2)$ parameters, and $O(q^3)$ floating point operations (flops).
- Optimizing over or sampling from high-dimensional parameter spaces is inefficient even for modest values of n.
- Illustrations of multivariate Matérn models have typically been restricted to applications with $q \leq 5$. [GKS10, AGS12]

Our approach

- Can we use the graphical structure between variables in our advantage?

Our approach

- Can we use the graphical structure between variables in our advantage?
- How can we define a graph between component processes of a multivariate Gaussian processes?

Our approach

- Can we use the graphical structure between variables in our advantage?
- How can we define a graph between component processes of a multivariate Gaussian processes?
- Can we preserve the marginal distributions of the component processes in the process?

Our approach

- Can we use the graphical structure between variables in our advantage?
- How can we define a graph between component processes of a multivariate Gaussian processes?
- Can we preserve the marginal distributions of the component processes in the process?
- How does the graph dictate the cross-covariances between pairs of variables?

Our approach

- Can we use the graphical structure between variables in our advantage?
  Yes. In non-spatial settings, Gaussian graphical models are extensively used as a dimension-reduction tool and we can extend it for processes.
- How can we define a graph between component processes of a multivariate Gaussian processes?
- Can we preserve the marginal distributions of the component processes in the process?
- How does the graph dictate the cross-covariances between pairs of variables?

Our approach

- Can we use the graphical structure between variables in our advantage?
  Yes. In non-spatial settings, Gaussian graphical models are extensively used as a dimension-reduction tool and we can extend it for processes.
- How can we define a graph between component processes of a multivariate Gaussian processes?
  We define *process-level conditional independence* for a multivariate GP $w(\cdot) = (w_1(\cdot), \dots, w_q(\cdot))^\top$ over $\mathcal{D}$ adapting the analogous definition for multivariate discrete time-series in [Dah00].
- Can we preserve the marginal distributions of the component processes in the process?
- How does the graph dictate the cross-covariances between pairs of variables?

Our approach

- Can we use the graphical structure between variables in our advantage?
  Yes. In non-spatial settings, Gaussian graphical models are extensively used as a dimension-reduction tool and we can extend it for processes.
- How can we define a graph between component processes of a multivariate Gaussian processes?
  We define *process-level conditional independence* for a multivariate GP $w(\cdot) = (w_1(\cdot), \ldots, w_q(\cdot))^{\mathsf{T}}$ over $\mathcal{D}$ adapting the analogous definition for multivariate discrete time-series in [Dah00].
- Can we preserve the marginal distributions of the component processes in the process?
  Stitching
- How does the graph dictate the cross-covariances between pairs of variables?
  Stitching

Stitching Prereqs: Process-level conditional independence

Let $\mathcal{V} = \{1, \dots, q\}$, $B \subset \mathcal{V}$ and $w_B(\mathcal{D}) = \{w_k(s) : k \in B,\ s \in \mathcal{D}\}$.

Two processes $w_i(\cdot)$ and $w_j(\cdot)$ are conditionally independent given the processes $\{w_k(\cdot) \,|\, k \in V \setminus \{i,j\}\}$ if -

$$\mathrm{Cov}(z_{iB}(s), z_{jB}(s')) = 0 \text{ for all } s, s' \in \mathcal{D} \text{ and } B = \mathcal{V} \setminus \{i,j\},$$
$$\text{where } z_{kB}(s) = w_k(s) - \mathrm{E}[w_k(s) \,|\, \sigma(\{w_j(s'') : j \in B,\ s'' \in \mathcal{D}\})].$$

Stitching Prereqs: Graphical Gaussian Processes

A $q \times 1$ GP $w(\cdot)$ is a Graphical Gaussian Process (GGP) with respect to a graph $\mathcal{G}_{\mathcal{V}} = (\mathcal{V}, E_{\mathcal{V}})$ when the univariate GPs $w_i(\cdot)$ and $w_j(\cdot)$ are conditionally independent for every $(i, j) \notin E_{\mathcal{V}}$. We denote such processes as GGP($\mathcal{G}_{\mathcal{V}}$).

Stitching Prereqs: Graphical Gaussian Processes

A $q \times 1$ GP $w(\cdot)$ is a Graphical Gaussian Process (GGP) with respect to a graph $\mathcal{G}_\mathcal{V} = (\mathcal{V}, E_\mathcal{V})$ when the univariate GPs $w_i(\cdot)$ and $w_j(\cdot)$ are conditionally independent for every $(i,j) \notin E_\mathcal{V}$. We denote such processes as GGP$(\mathcal{G}_\mathcal{V})$.

Theorem

(a) *There exists a unique $q \times 1$ GGP$(\mathcal{G}_\mathcal{V})$ $w(\cdot)$ with cross-covariance function $M(h) = (M_{ij}(h))$ such that $M_{ij}(h) = C_{ij}(h)$ for $i = j$ and for all $(i,j) \in E_\mathcal{V}$;*

(b) *If $\tilde{F}(\omega)$ denotes the SDM of $w(\cdot)$ and $\mathcal{F}$ is the set of SDMs of all possible GGP$(\mathcal{G}_\mathcal{V})$, then*

$$\tilde{F}(\cdot) = \arg\min_{K(\cdot) \in \mathcal{F}} \int_\omega d_{KL}(F(\omega) \| K(\omega)) d\omega \ ,$$

where $d_{KL}(F\|K) = tr(K^{-1}F) + \log\det(K)$ denotes the Kullback-Leibler divergence between two positive definite matrices F and K.

Stitching: Can we get the optimal GGP?

Given any $\mathcal{G}_{\mathcal{V}}$ and a cross-covariance function C, we seek a multivariate GP $w(\cdot)$ that -

(i) exactly preserves the marginal distributions specified by C, i.e., $w_i(\cdot) \sim GP(0, C_{ii})\, \forall i$,

(ii) is a GGP, i.e., satisfies process-level conditional independence as specified by the $\mathcal{G}_{\mathcal{V}}$,

(iii) exactly or approximately retains the cross-covariances specified by C for pairs of variables included in the graph $\mathcal{G}_{\mathcal{V}}$, i.e., for $(i,j) \in E_{\mathcal{V}}$, $\mathrm{Cov}(w_i(s), w_j(s')) \approx C_{ij}(s,s')$.

Stitching: Visulization

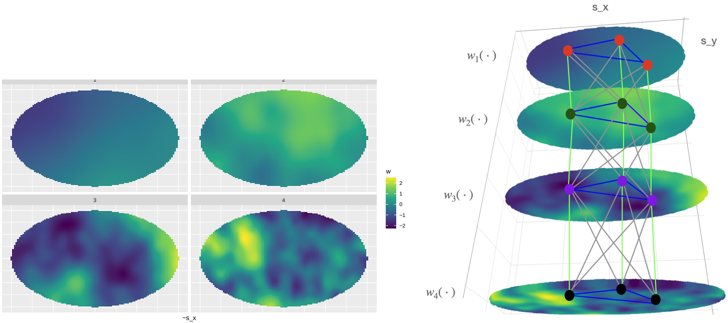


Figure: Stitching Gaussian Processes.
Left: Realizations of 4 univariate GPs.
Right: Realization of a multivariate (4-dimensional) GGP created by stitching together the 4 univariate GPs from the left figure using the strong product graph over the 4 variables and 3 locations.

Stitching: Sparse graph between variables

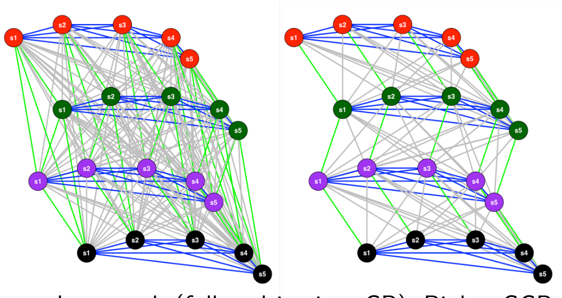


Figure: Left: GGP with complete graph (full multivariate GP), Right: GGP with path graph between variables.

4 node colors represent 4 variables. We use blue lines for edges between locations of same variable, green for edges between different variables at same site, gray for edges between different variables at different locations.

Stitching: The story

- We begin our construction on $\mathcal{L}$, a finite, but otherwise arbitrary, set of locations in $\mathcal{D}$ (the set of 3 locations in previous slide).
- We stitch together the variables at the 3 locations in $\mathcal{L}$ such that there is a *thread* (edge) between two variable-location pairs if and only if there is an edge between the two variables in the graph
- We then stitch each of the remaining surfaces independently so that they have the same distribution as the univariate surfaces from the left panel and preserves the graphical model at the process-level.

Stitching: The math

To fulfill our three requirements, we model $w(\mathcal{L}) \sim N(0, M(\mathcal{L}, \mathcal{L}))$ seeking a p.d. matrix $M(\mathcal{L}, \mathcal{L})$ such that

(a) $M_{ii}(\mathcal{L}, \mathcal{L}) = C_{ii}(\mathcal{L}, \mathcal{L})$ for all $i = 1, \ldots, q$, to satisfy (i),

(b) $(M(\mathcal{L}, \mathcal{L})^{-1})_{ij} = 0$ for all $(i, j) \notin E_{\mathcal{V}}$ to satisfy (ii),

(c) $M_{ij}(\mathcal{L}, \mathcal{L}) = C_{ij}(\mathcal{L}, \mathcal{L})$ for all $(i, j) \in E_{\mathcal{V}}$, to satisfy (iii).

Existence of such a matrix $M(\mathcal{L}, \mathcal{L})$ is guaranteed by Dempster's seminal result [Dem72] in covariance selection problems.

Stitching: The math

To fulfill our three requirements, we model $w(\mathcal{L}) \sim N(0, M(\mathcal{L}, \mathcal{L}))$ seeking a p.d. matrix $M(\mathcal{L}, \mathcal{L})$ such that

(a) $M_{ii}(\mathcal{L}, \mathcal{L}) = C_{ii}(\mathcal{L}, \mathcal{L})$ for all $i = 1, \ldots, q$, to satisfy (i),

(b) $(M(\mathcal{L}, \mathcal{L})^{-1})_{ij} = 0$ for all $(i, j) \notin E_{\mathcal{V}}$ to satisfy (ii),

(c) $M_{ij}(\mathcal{L}, \mathcal{L}) = C_{ij}(\mathcal{L}, \mathcal{L})$ for all $(i, j) \in E_{\mathcal{V}}$, to satisfy (iii).

Existence of such a matrix $M(\mathcal{L}, \mathcal{L})$ is guaranteed by Dempster's seminal result [Dem72] in covariance selection problems.

Extend it to infinite-dimensional GP [BGFS08, FSBG09]-
(Predictive process + Independent residual process)

$$w_i(s) = w_i^*(s) + z_i(s) = C_{ii}(s, \mathcal{L}) C_{ii}(\mathcal{L}, \mathcal{L})^{-1} w_i(\mathcal{L}) + z_i(s) \quad \text{for all} \quad s \in \mathcal{D} \setminus \mathcal{L}\,, \qquad (2)$$

Stitching: What we achieve

Theorem

Given a cross-covariance function C and an inter-variable graph $\mathcal{G}_{\mathcal{V}}$, stitching creates a valid multivariate GGP $w(\cdot)$ with a valid (p.d.) cross-covariance function M such that:

(a) $w_i(\cdot) \sim GP(0, C_{ii})$, *i.e.,* $M_{ii}(s, s') = C_{ii}(s, s')$ *for all* $s, s' \in \mathcal{D}$ *and for each* $i = 1, \ldots, q$,

(b) $w(\cdot)$ *is a* $GGP(\mathcal{G}_{\mathcal{V}})$ *on* $\mathcal{D}$,

(c) *if* $(i, j) \in E_{\mathcal{V}}$, *then* $M_{ij}(s, s') = C_{ij}(s, s')$ *for all* $s, s' \in \mathcal{L}$.

Stitching produces a multivariate GP $w(\cdot)$ that exactly satisfies the first two conditions sought in optimal GGP. Condition (iii) is satisfied exactly on $\mathcal{L}$ and approximately on $\mathcal{D} \setminus \mathcal{L}$ for the stitched GP.

Application: Multivariate Matérn

The isotropic multivariate Matérn cross-covariance function on a d-dimensional domain is $C_{ij}(s,s') = \sigma_{ij} H_{ij}(\|s - s'\|)$, where $H_{ij}(\cdot) = H(\cdot \mid \nu_{ij}, \phi_{ij})$, H being the Matérn correlation function [AGS12].

To ensure a valid multivariate Matérn cross-covariance function, it is sufficient to constrain the intra-site covariance matrix $\Sigma = (\sigma_{ij})$ to be of the form (Theorem 1 of [AGS12]).

$$\sigma_{ij} = \quad b_{ij} \frac{\Gamma(\frac{1}{2}(\nu_{ii}+\nu_{jj}+d))\Gamma(\nu_{ij})}{\phi_{ij}^{2\Delta_A+\nu_{ii}+\nu_{jj}}\Gamma(\nu_{ij}+\frac{d}{2})} \text{ where } \Delta_A \geq 0, \text{ and } B = (b_{ij}) > 0, \text{ i.e., is p.d.} \tag{3}$$

This is equivalent to Σ being constrained as $\Sigma = (B \odot (\gamma_{ij}))$

Multivariate Matérn: Computational consideration for stitching

- Stitching needs to constrain $B = (b_{ij})$ to be p.d. on an $O(q^2)$-dimensional parameter space.
- Searching in such a high-dimensional space is difficult for large q and verifying positive definiteness of B incurs an additional cost of $O(q^3)$ flops.
- Evaluating $w(\mathcal{L}) \sim N(0, M(\mathcal{L}, \mathcal{L}))$ involves matrix operations for the $nq \times nq$ matrix $M(\mathcal{L}, \mathcal{L})$. While the precision matrix, $M(\mathcal{L}, \mathcal{L})^{-1}$, is sparse because of $\mathcal{G}_V$, its determinant is usually not available in closed form and the calculation can become prohibitive even for small n.

Mutivariate Graphical Matérn: Decomposable graphs

- Considering only on decomposable graphs factorizes the likelihood and reduce computational complexity.
- If the decomposable graph $\mathcal{G}_{\mathcal{V}}$ has a perfect clique sequence $\{K_1, K_2, \cdots, K_p\}$ with separators $\{S_2, \ldots, S_m\}$, then the GGP likelihood on $\mathcal{L}$ can be decomposed as

$$f_M(w(\mathcal{L})) = \frac{\Pi_{m=1}^{p} f_C(w_{K_m}(\mathcal{L}))}{\Pi_{m=2}^{p} f_C(w_{S_m}(\mathcal{L}))}\ , \tag{4}$$

Mutivariate Graphical Matérn: Decomposable graphs

- Considering only on decomposable graphs factorizes the likelihood and reduce computational complexity.
- If the decomposable graph $\mathcal{G}_{\mathcal{V}}$ has a perfect clique sequence $\{K_1, K_2, \cdots, K_p\}$ with separators $\{S_2, \ldots, S_m\}$, then the GGP likelihood on $\mathcal{L}$ can be decomposed as

$$f_M(w(\mathcal{L})) = \frac{\Pi_{m=1}^{p} f_C(w_{K_m}(\mathcal{L}))}{\Pi_{m=2}^{p} f_C(w_{S_m}(\mathcal{L}))}\,, \tag{4}$$

- the precision matrix of $w(\mathcal{L})$ satisfies [Lau96]

$$M(\mathcal{L},\mathcal{L})^{-1} = \sum_{m=1}^{p} [C^{-1}_{[K_m \boxtimes \mathcal{G}_{\mathcal{L}}]}]^{\mathcal{V}\times\mathcal{L}} - \sum_{m=2}^{p} [C^{-1}_{[S_m \boxtimes \mathcal{G}_{\mathcal{L}}]}]^{\mathcal{V}\times\mathcal{L}}\,, \tag{5}$$

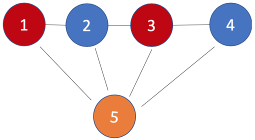
Mutivariate Graphical Matérn: Decomposable graphs

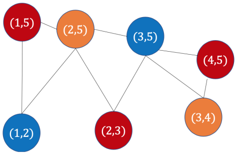
Table: Properties of any q-dimensional multivariate Matérn GP of [GKS10] or [AGS12] and a multivariate graphical Matérn GP stitched using a decomposable graph $\mathcal{G}_V$ with largest clique size q^* (typically $\ll q$), length of perfect ordering p, and maximal number of cliques p^* sharing a common vertex.

Model attributes	Multivariate Matérn	Multivariate Graphical Matérn
Number of parameters	$O(q^2)$	$O(\lvert E_{\mathcal{V}}\rvert + q)$
Parameter constraints	$O(q^3)$	$O(p^*(q^{*3}))$ (worst case)
Storage	$O(n^2 q^2)$	$O(pn^2 q^{*2})$ (worst case)
Time complexity	$O(n^3 q^3)$	$pn^3 q^{*3}$ (worst case)
Conditionally independent processes	No	Yes
Univariate components are Matérn GPs	Yes	Yes

Implementations

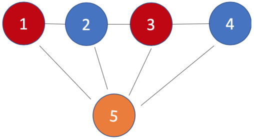
- **Known graph:** We implement a chromatic Gibbs sampler which uses the graph coloring to facilitate parallel simulation of both the latent spatial processes and correlation parameters (b_{ij}) respectively.

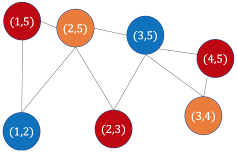




Implementations

- **Known graph:** We implement a chromatic Gibbs sampler which uses the graph coloring to facilitate parallel simulation of both the latent spatial processes and correlation parameters (b_{ij}) respectively.





- **Unknown graph:** We augment the sampler above with a reversible jump MCMC sampler[BL13] which moves between junction trees[GT13] in the graph spaces to infer about the graphs.

Simulations: Competing models

(a) PM: Parsimonious Multivariate Matérn of [GKS10];

(b) MM: Multivariate Matérn of [AGS12].

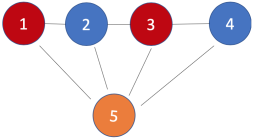
(c) GM: Graphical Matérn (GGP on the latent process, stitched using multivariate Matérn model (b)).

All models consider $\nu_{ij} = \nu_{ii} = \nu_{jj} = \frac{1}{2}$.

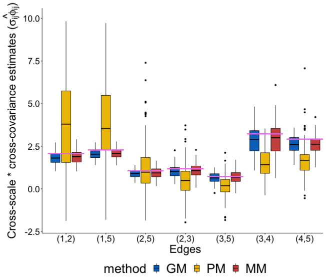
Simulation: Scenario

Table: Different simulation scenarios considered for the comparison between methods.

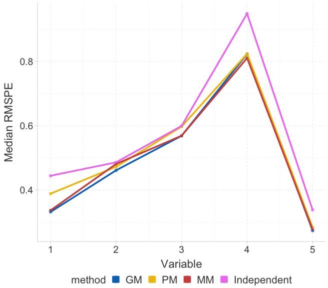
Set	q	Graph $\mathcal{G}_{\mathcal{V}}$	B	Nugget	Locations	Data model	Fitted models
1A	5	Gem (Figure below)	Random	No	Same location for all variables	GM	GM, MM, PM
1B	5	Gem (Figure below)	Random	No	Same location for all variables	MM	GM, MM, PM
2A	15	Path	$b_{i-1,i} = \rho_i$	Yes	Partial overlap in locations for variables	GM	GM, PM
2B	15	Path	$b_{i-1,i} = \rho_i$	Yes	Partial overlap in locations for variables	MM	GM, PM
3A	100	Path	$b_{i-1,i} = \rho_i$	Yes	Partial overlap in locations for variables	GM	GM
3B	100	Path	$b_{i-1,i} = \rho_i$	Yes	Partial overlap in locations for variables	MM	GM



Simulation: Results



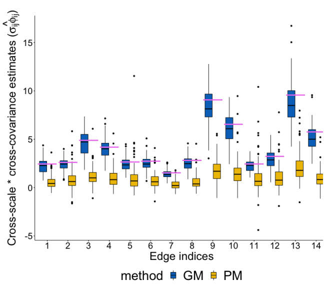
(a) Set 1B



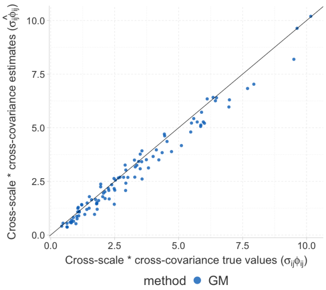
(b) Set 1B

Figure: Performance of graphical Matérn under mis-specification - (a): Estimates of the edge-specific cross-covariance parameters for the set 1B. The pink lines indicate true parameter values. (b): Median RMSPE for GM, MM, PM and Independent GP model for Set 1B.

Simulation: Results



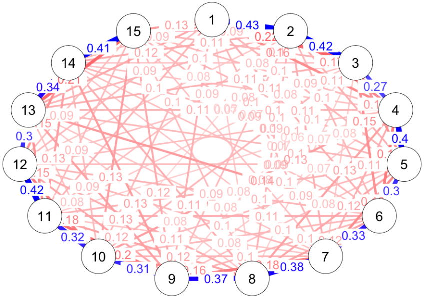
(a) Set 2B



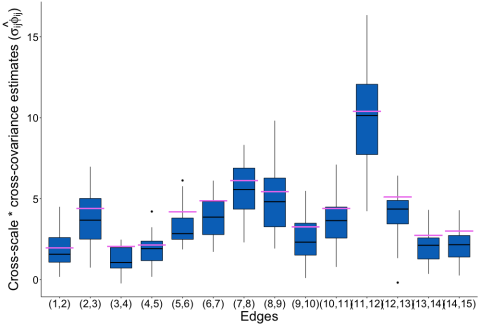
(b) Set 3B

Figure: Performance of graphical Matérn under mis-specification: (a) and (b): Estimates of the cross-covariance parameters $\sigma_{ij}\phi_{ij} = \Gamma(1/2)b_{ij}$, $(i,j) \in E_{\mathcal{V}}$ for the sets 2B and 3B respectively. The pink lines indicate true parameter values.

Simulation: Unknown graph



(a) Posterior edge selection probabilities for Set 2A.



(b) Cross-covariance parameter estimates for Set 2A while estimating the unknown graph

Figure: Performance of GGP with unknown graph for Set 2A. Blue edges denote the true edges and red denotes the non-existent edges. Edges are weighted proportional to the estimated posterior selection probabilities. Horizontal pink lines indicate the true values.

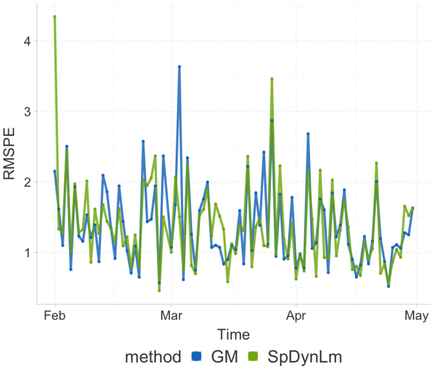
Data analysis: Spatio-temporal modelling of $PM_{2.5}$

- We model daily levels of $PM_{2.5}$ measured at monitoring stations across 11 states of the north-eastern US and Washington DC.
- Duration is a three month period from February, 01, 2020, until April, 30th, 2020.

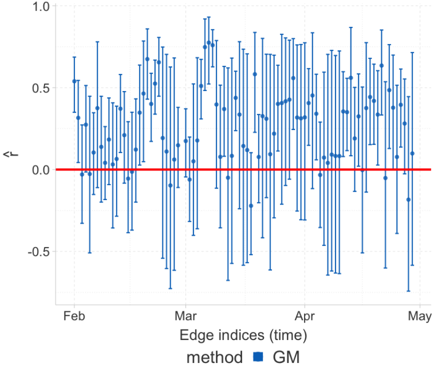
Data analysis: Spatio-temporal modelling of $PM_{2.5}$

- We model daily levels of $PM_{2.5}$ measured at monitoring stations across 11 states of the north-eastern US and Washington DC.
- Duration is a three month period from February, 01, 2020, until April, 30th, 2020.
- We selected $n = 86$ stations with at least two months of measured data for both 2020 and 2019.
- The daily 2019 $PM_{2.5}$ data treated as a baseline covariate for the 2020 $PM_{2.5}$ levels.
- Meteorological variables such as temperature, barometric pressure, wind-speed and relative humidity are adjusted as covariates.

Data Analysis: results



(a) Prediction performance for full analysis



(b) Estimates of time-specific cross-correlations

First two weeks of February

Last two weeks of April

Data analysis: Spatio-temporal modelling of $PM_{2.5}$

- We model daily levels of $PM_{2.5}$ measured at monitoring stations across 11 states of the north-eastern US and Washington DC.
- Duration is a three month period from February, 01, 2020, until April, 30th, 2020.

Data analysis: Spatio-temporal modelling of $PM_{2.5}$

- We model daily levels of $PM_{2.5}$ measured at monitoring stations across 11 states of the north-eastern US and Washington DC.
- Duration is a three month period from February, 01, 2020, until April, 30th, 2020.
- We selected $n = 86$ stations with at least two months of measured data for both 2020 and 2019.
- The daily 2019 $PM_{2.5}$ data treated as a baseline covariate for the 2020 $PM_{2.5}$ levels.
- Meteorological variables such as temperature, barometric pressure, wind-speed and relative humidity are adjusted as covariates.

Data analysis: Spatio-temporal modelling of $PM_{2.5}$

- We model daily levels of $PM_{2.5}$ measured at monitoring stations across 11 states of the north-eastern US and Washington DC.
- Duration is a three month period from February, 01, 2020, until April, 30th, 2020. $(q = 89)$

Data analysis: Spatio-temporal modelling of $PM_{2.5}$

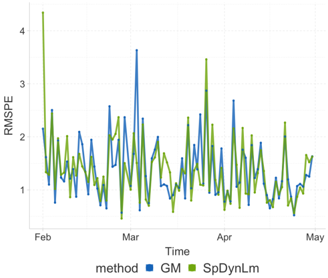
- We model daily levels of $PM_{2.5}$ measured at monitoring stations across 11 states of the north-eastern US and Washington DC.
- Duration is a three month period from February, 01, 2020, until April, 30th, 2020. $(q = 89)$
- We selected $n = 86$ stations with at least two months of measured data for both 2020 and 2019.
- The daily 2019 $PM_{2.5}$ data treated as a baseline covariate for the 2020 $PM_{2.5}$ levels.
- Meteorological variables such as temperature, barometric pressure, wind-speed and relative humidity are adjusted as covariates.

Data analysis: Spatio-temporal modelling of $PM_{2.5}$

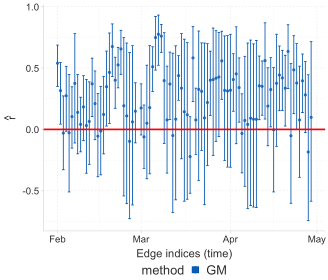
- We model daily levels of $PM_{2.5}$ measured at monitoring stations across 11 states of the north-eastern US and Washington DC.
- Duration is a three month period from February, 01, 2020, until April, 30th, 2020. ($q = 89$)
- We selected $n = 86$ stations with at least two months of measured data for both 2020 and 2019.
- The daily 2019 $PM_{2.5}$ data treated as a baseline covariate for the 2020 $PM_{2.5}$ levels.
- Meteorological variables such as temperature, barometric pressure, wind-speed and relative humidity are adjusted as covariates.
- Neither Parsimonious Matern or Multivariate Matern couldn't be fit for full data as they involve $89^2/2 \approx 4000$ cross-covariance parameters and 8000×8000 matrix computations.

Data Analysis: results

We compare GGP with the spatial dynamic linear model (SpDynLm) in [FBG12], (doesn't have autocorrelation parameters and assumes increasing variance with time).



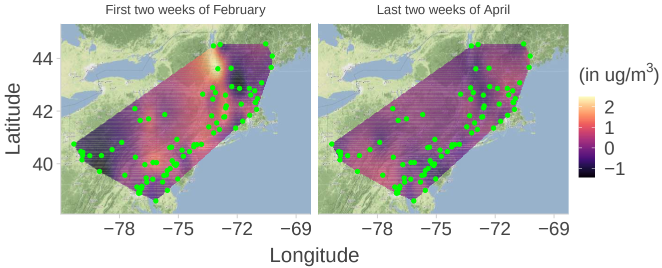
(a) Prediction performance for full analysis



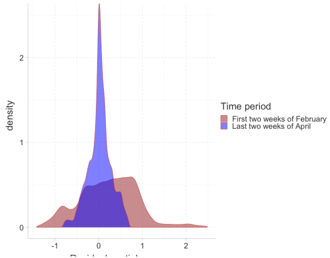
(b) Estimates of time-specific cross-correlations

Figure: $PM_{2.5}$ analysis: (a) Daily RMSPE for the full analyses, (b) Estimates and credible intervals of the cross-correlation parameters $r_{t,t-1}$ (corresponding to the cross-covariances $b_{t,t-1}$)

Data Analysis: Results



(a) Average residual surfaces



(b) Density of residual spatial processes

Figure: $PM_{2.5}$ analysis: (a) Estimates of the residual spatial processes from GM (after adjusting for covariates), (b) Density of residual spatial process values (across locations) for two different time periods

Summary

- Existence, uniqueness, and optimality of graphical Gaussian Processes (GGP) with process-level conditional independence.

Summary

- Existence, uniqueness, and optimality of graphical Gaussian Processes (GGP) with process-level conditional independence.
- Stitching prescribes a practical construction of this Optimal GGP.

Summary

- Existence, uniqueness, and optimality of graphical Gaussian Processes (GGP) with process-level conditional independence.
- Stitching prescribes a practical construction of this Optimal GGP.
- Decomposable graph assumption reduces parameter space, computational complexity and work for tens or hundreds of variables.
- A recipe for unknown graph estimation for moderately large number of variables ($q = 15$). Future work would aim higher q.

Summary

- Existence, uniqueness, and optimality of graphical Gaussian Processes (GGP) with process-level conditional independence.
- Stitching prescribes a practical construction of this Optimal GGP.
- Decomposable graph assumption reduces parameter space, computational complexity and work for tens or hundreds of variables.
- A recipe for unknown graph estimation for moderately large number of variables ($q = 15$). Future work would aim higher q.
- We can extend it to spatial factor models, spatial time-series, asymmetric or non-stationary processes.

Questions?

Thank you!

Preprint: `https://arxiv.org/pdf/2009.04837.pdf`

Email: ddey1@jhu.edu

Twitter: debangan07

References I

- Tatiyana V Apanasovich, Marc G Genton, and Ying Sun, *A valid matérn class of cross-covariance functions for multivariate random fields with any number of components*, Journal of the American Statistical Association **107** (2012), no. 497, 180–193.
- Sudipto Banerjee, Alan E Gelfand, Andrew O Finley, and Huiyan Sang, *Gaussian predictive process models for large spatial data sets*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **70** (2008), no. 4, 825–848.
- Richard J Barker and William A Link, *Bayesian multimodel inference by rjmcmc: A gibbs sampling approach*, The American Statistician **67** (2013), no. 3, 150–156.
- Rainer Dahlhaus, *Graphical interaction models for multivariate time series*, Metrika **51** (2000), no. 2, 157–172.
- Arthur P Dempster, *Covariance selection*, Biometrics (1972), 157–175.

References II

- Andrew O Finley, Sudipto Banerjee, and Alan E Gelfand, *Bayesian dynamic modeling for large space-time datasets using gaussian predictive processes*, Journal of geographical systems **14** (2012), no. 1, 29–47.
- Andrew O Finley, Huiyan Sang, Sudipto Banerjee, and Alan E Gelfand, *Improving the performance of predictive process modeling for large datasets*, Computational statistics & data analysis **53** (2009), no. 8, 2873–2884.
- Tilmann Gneiting, William Kleiber, and Martin Schlather, *Matérn cross-covariance functions for multivariate random fields*, Journal of the American Statistical Association **105** (2010), no. 491, 1167–1177.
- Peter J Green and Alun Thomas, *Sampling decomposable graphs using a markov chain on junction trees*, Biometrika **100** (2013), no. 1, 91–110.
- Steffen L Lauritzen, *Graphical models*, vol. 17, Clarendon Press, 1996.