Multi-outcome Meshed Gaussian Processes on Projected Inputs

for scalable inference with **EXPOSOME** data

Michele Peruzzi
Postdoc, Dept. Stat. Sci.
Duke University

michele.peruzzi@duke.edu
github.com/mkln

EXPOSOME
data challenges:

Correlated exposures
Multiple outcomes
Outcomes are related
Missing covariate data
Medium-large data size

NEW METHODS
wishlist:

Flexible
Interpretable by construction
Quantify uncertainty
Work with big data
~~Black box~~

A basic model for exposure effects

- One outcome for individual *i*
- Covariates/confounders *z*
- Gaussian iid measurement errors
- Unknown function *h* of several inputs (exposures)
- Gaussian Process (GP) prior model for *h*
- GP with kernel *K*

$$y_i = h(\boldsymbol{x}_i) + z_i^\top \boldsymbol{\gamma} + \varepsilon_i$$

$$\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

$$h \sim GP(0, K(\cdot, \cdot))$$

$$K(\boldsymbol{x}, \boldsymbol{x}') = \exp\left\{ -\sum_j \rho_j (x_j - x_j')^2 \right\}$$

>> Computations via Markov chain Monte Carlo (MCMC)

A basic model for exposure effects
some key issues

- One outcome for individual *i* — multiple & related
- Covariates/confounders *z* — some are missing
- Gaussian iid measurement errors — what about counts, binary, discrete…
- Unknown function *h* of several inputs (exposures) — some are correlated
- Gaussian Process (GP) prior model for *h* — scales poorly to big data
- GP with kernel *K*

$$y_i = h(\boldsymbol{x}_i) + z_i^\top \boldsymbol{\gamma} + \varepsilon_i$$

$$\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

$$h \sim GP(0, K(\cdot, \cdot))$$

$$K(\boldsymbol{x}, \boldsymbol{x}') = \exp\left\{ -\sum_j \rho_j (x_j - x_j')^2 \right\}$$

>> Computations via Markov chain Monte Carlo (MCMC) — inefficient, slow mixing/convergence

Jointly modeling multiple outcomes

- *r=1, …, q* outcomes for individual *i*
- Outcome-specific covariates/confounders *z*
- Gaussian iid measurement errors
- Unknown outcome-specific functions h_r, *r = 1, …, q*
- GPs with outcome-specific kernel

$$y_{i,1} = h_1(\boldsymbol{x}_{i,1}) + z_{i,1}^{\top}\boldsymbol{\gamma}_1 + \varepsilon_{i,1}$$

$$\vdots$$

$$y_{i,q} = h_q(\boldsymbol{x}_{i,q}) + z_{i,q}^{\top}\boldsymbol{\gamma}_q + \varepsilon_{i,q}$$

$$h_r \sim GP(0, K_r(\cdot,\cdot))$$

$$K_r(\boldsymbol{x}, \boldsymbol{x}') = \exp\left\{ -\sum_j \rho_{j,r}(x_{j,r} - x'_{j,r})^2 \right\}$$

independent outcomes?
no borrowing of information

Jointly modeling multiple outcomes

- *r=1, …, q* outcomes for individual *i*
- Outcome-specific covariates/confounders *z*
- Gaussian iid measurement errors
- Unknown outcome-specific functions h_r, *r = 1, …, q*
- GPs with outcome-specific kernel

$$\begin{bmatrix} y_{i,1} \\ \vdots \\ \vdots \\ y_{i,q} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} h_1(\boldsymbol{x}_{i,1}) \\ \vdots \\ h_q(\boldsymbol{x}_{i,q}) \end{bmatrix} + \begin{bmatrix} z_{i,1}^\top & & \\ & \ddots & \\ & & z_{i,q}^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{\gamma}_1 \\ \vdots \\ \boldsymbol{\gamma}_q \end{bmatrix} + \begin{bmatrix} \varepsilon_{i,1} \\ \vdots \\ \varepsilon_{i,q} \end{bmatrix}$$

$$\boldsymbol{y}_i = I_q \boldsymbol{h}(\boldsymbol{x}_i) + \boldsymbol{Z}_i \boldsymbol{\gamma} + \boldsymbol{\varepsilon}_i$$

$$h_r \sim GP(0, K_r(\cdot, \cdot))$$

independent outcomes?
no borrowing of information

Jointly modeling multiple outcomes

- $r=1, \ldots, q$ outcomes for individual i
- Outcome-specific covariates/confounders z
- Gaussian iid measurement errors
- Unknown outcome-specific linear combinations of functions h_r, $r = 1, \ldots, k < q$
- Identifiability constraints on A like in factor models

$$\begin{bmatrix} y_{i,1} \\ \vdots \\ \vdots \\ y_{i,q} \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & & \vdots \\ \vdots & & \vdots \\ a_{q1} & \cdots & a_{qk} \end{bmatrix} \begin{bmatrix} h_1(\boldsymbol{x}_{i,1}) \\ \vdots \\ h_k(\boldsymbol{x}_{i,k}) \end{bmatrix} + \begin{bmatrix} z_{i,1}^\top & & \\ & \ddots & \\ & & z_{i,q}^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{\gamma}_1 \\ \vdots \\ \boldsymbol{\gamma}_q \end{bmatrix} + \begin{bmatrix} \varepsilon_{i,1} \\ \vdots \\ \varepsilon_{i,q} \end{bmatrix}$$

$$\boldsymbol{y}_i = \boldsymbol{A}\boldsymbol{h}(\boldsymbol{x}_i) + \boldsymbol{Z}_i\boldsymbol{\gamma} + \boldsymbol{\varepsilon}_i$$

$$h_r \sim GP(0, K_r(\cdot,\cdot))$$

Use k < q functions

Missing covariate values & Covariates measured with error

- Imputing missing values is complicated
- Difficult to account for uncertainty if using 2-stage procedures (impute, then fit model)
- We can model outcome and covariate **jointly**

$$\widetilde{y}_i = h_1(\boldsymbol{x}_i) + \widetilde{\varepsilon}_i$$

$$y_i = \underbrace{h_2(\boldsymbol{x}_i)}_{\text{function of the inputs}} + \underbrace{a_{11} h_1(\boldsymbol{x}_i)}_{\text{linear effect of covariate on outcome}} + \varepsilon_i$$

function of the inputs

linear effect of covariate on outcome

Example with 1 outcome y_i

- Suppose j^* is the covariate with missing values for some i
- Label that as outcome $\widetilde{y}_i$ i.e. $\widetilde{y}_i = z_{i,j^*}$
- Then this becomes a model with 2 outcomes

Missing covariate values & Covariates measured with error

- Imputing missing values is complicated
- Difficult to account for uncertainty if using 2-stage procedures (impute, then fit model)
- We can model outcome and covariate **jointly**

<u>Example</u> with 1 outcome y_i

- Suppose *j** is the covariate with missing values for some *i*
- Label that as outcome $\widetilde{y}_i$ i.e. $\widetilde{y}_i = z_{i,j^*}$
- Then this becomes a model with 2 outcomes

$$\widetilde{y}_i = h_1(\boldsymbol{x}_i) + \widetilde{\varepsilon}_i$$

$$y_i = \underbrace{h_2(\boldsymbol{x}_i)}_{\text{function of the inputs}} + \underbrace{a_{11} h_1(\boldsymbol{x}_i)}_{\text{linear effect of covariate on outcome}} + \varepsilon_i$$

$$y_i = h(\boldsymbol{x}_i) + z_i^\top \boldsymbol{\gamma} + \varepsilon_i$$

univariate model as seen before

Missing covariate values & Covariates measured with error

- Imputing missing values is complicated
- Difficult to account for uncertainty if using 2-stage procedures (impute, then fit model)
- We can model outcome and covariate **jointly**

Example with 1 outcome y_i

- Suppose *j** is the covariate with missing values for some *i*
- Label that as outcome $\widetilde{y}_i$ i.e. $\widetilde{y}_i = z_{i,j^*}$
- Then this becomes a model with 2 outcomes

Compactly:

$$\begin{bmatrix} \widetilde{y}_i \\ y_i \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ a_{11} & 1 \end{bmatrix} \begin{bmatrix} h_1(\boldsymbol{x}_i) \\ h_2(\boldsymbol{x}_i) \end{bmatrix} + \begin{bmatrix} \widetilde{\varepsilon}_i \\ \varepsilon_i \end{bmatrix}$$

Multi-outcome model with latent Gaussian Processes

k GPs

p covariates without missing values

q outcomes $\{$ $$\boldsymbol{y}_i = \boldsymbol{A}\boldsymbol{h}(\boldsymbol{x}_i) + \boldsymbol{Z}_i\boldsymbol{\gamma} + \boldsymbol{\varepsilon}_i$$

$q \times k$ matrix
(with constraints)

d inputs

Computing multi-outcome latent GP models is hard!

- Suppose $i = 1, \ldots, n$ subjects
- q outcomes
- Effective data dimension is $\boldsymbol{nq}$
- Posterior computation scales as $\boldsymbol{O(n^3 q^3)}$

$$\begin{bmatrix} y_{i,1} \\ \vdots \\ \vdots \\ y_{i,q} \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & & \vdots \\ \vdots & & \vdots \\ a_{q1} & \cdots & a_{qk} \end{bmatrix} \begin{bmatrix} h_1(\boldsymbol{x}_{i,1}) \\ \vdots \\ h_k(\boldsymbol{x}_{i,k}) \end{bmatrix} + \begin{bmatrix} z_{i,1}^\top & & \\ & \ddots & \\ & & z_{i,q}^\top \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_q \end{bmatrix} + \begin{bmatrix} \varepsilon_{i,1} \\ \vdots \\ \varepsilon_{i,q} \end{bmatrix}$$

$$\boldsymbol{y}_i = \boldsymbol{A}\boldsymbol{h}(\boldsymbol{x}_i) + \boldsymbol{Z}_i\boldsymbol{\gamma} + \boldsymbol{\varepsilon}_i$$

Common strategy (e.g. BKMR):

>> use low-rank *aka* “predictive” GPs

(e.g. Banerjee et al 2008 JRSSB)

→

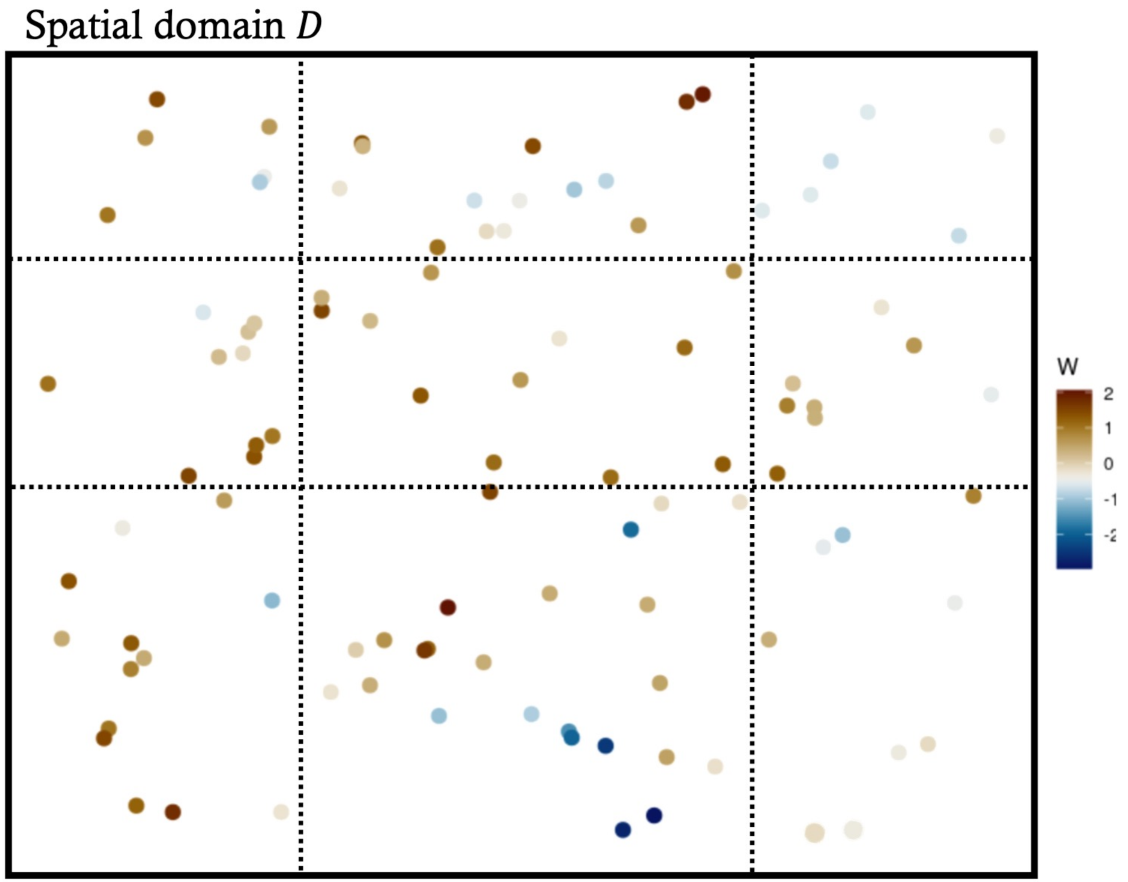
- Low-rank methods depend on defining *knots*
- Number of knots $n^* \ll n$
- Issues in approximating GP when d is large (input dimension)
- Oversmooth surfaces when n very large

Scalability via spatial meshing

Meshed GPs

(MP et al 2020 JASA)

- Take a set of knots with $n^* \approx n$ or even larger
- **Partition** into disjoint blocks
- Link partitions via "nice" **directed acyclic graph** (DAG)



Scalability via spatial meshing

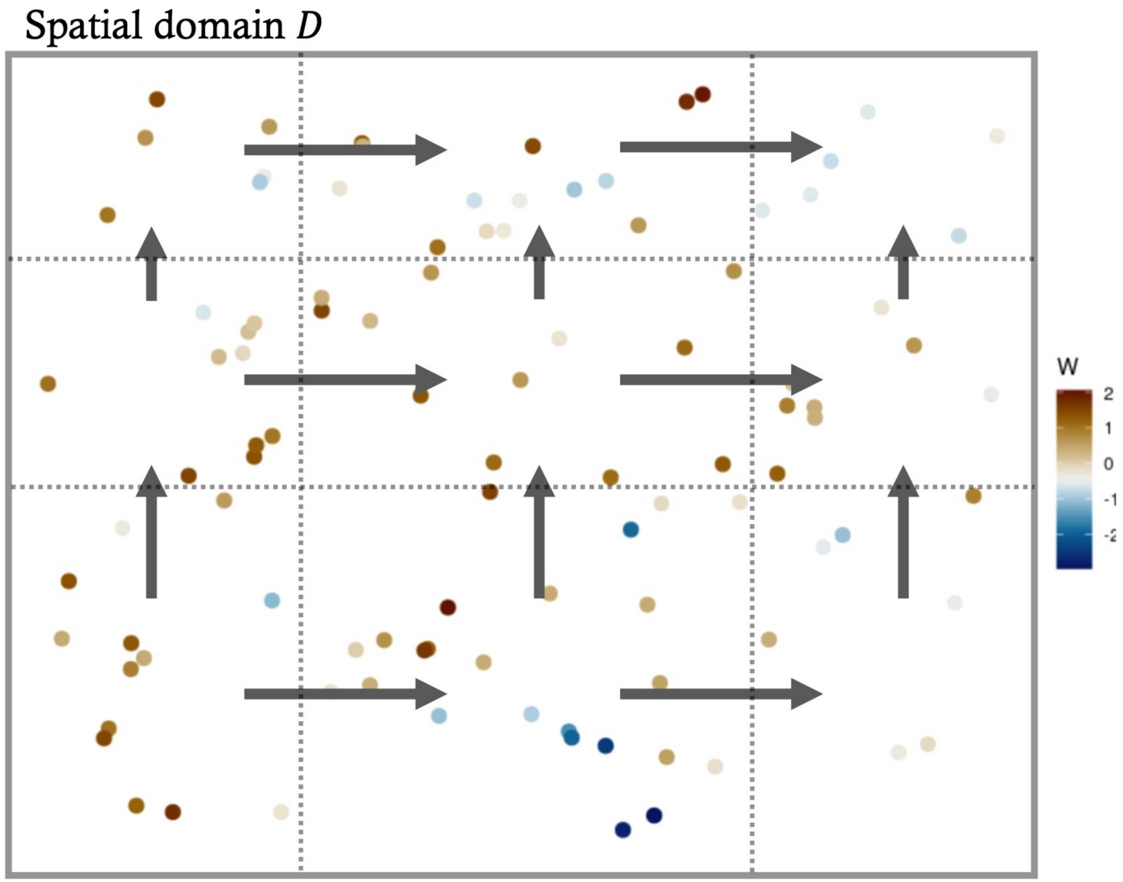
Meshed GPs

(MP et al 2020 JASA)

- Take a set of knots with $n^* \approx n$ or even larger
- **Partition** into disjoint blocks
- Link partitions via "nice" **directed acyclic graph** (DAG)

- In geostatistics (i.e. $d = 2$ or $d = 3$):

  Meshed GP scales to **data in the millions:**
  - ✓ **multivariate** outcomes
  - ✓ **multi-type**/non-Gaussian (counts, binary, discrete…)
  - ✓ **misaligned** outcomes (different outcomes measured at different inputs)
  - ✓ **parallel** computing of expensive steps



Scalability via spatial meshing

Meshed GPs

(MP et al 2020 JASA)

- Take a set of knots with $n^* \approx n$ or even larger
- **Partition** into disjoint blocks
- Link partitions via “nice” **directed acyclic graph** (DAG)

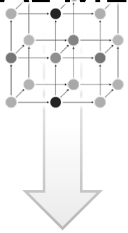
- In geostatistics (i.e. $d = 2$ or $d = 3$):
  Meshed GP scales to **data in the millions:**
  - ✓ **multivariate** outcomes
  - ✓ **multi-type**/non-Gaussian (counts, binary, discrete…)
  - ✓ **misaligned** outcomes (different outcomes measured at different inputs)
  - ✓ **parallel** computing of expensive steps

collect all $\boldsymbol{h}(\boldsymbol{x}_i)$ into vector $\boldsymbol{h}$, then

$$p(\boldsymbol{h} \mid \boldsymbol{\rho}) = \text{MVN}(\boldsymbol{h}; \boldsymbol{0}, \boldsymbol{K}_{\boldsymbol{\rho}})$$

BIG

SPATIAL MESHING



$$\widetilde{p}(\boldsymbol{h} \mid \boldsymbol{\rho}) = \prod_{g \in \text{Graph}} \text{MVN}(\boldsymbol{h}_g; \underset{\text{mean}}{\text{conditional}}, \underset{\text{variance}}{\text{conditional}})$$

small

$$= \prod_{g \in \text{Graph}} p(\boldsymbol{h}_g \mid \boldsymbol{h}_{\text{parents of } g}, \boldsymbol{\rho})$$

Scalability via spatial meshing *on projected inputs (pi)*

Meshed GPs

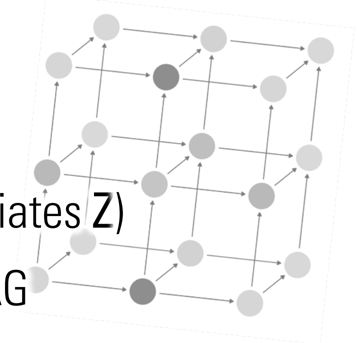
(MP et al 2020 JASA)

- Partitioning & building DAG is difficult when $d \gg 2$
- Exposome data lacks natural spatial domain

  > No geolocation info {Longitude, Latitude} for partitioning!
- Number of exposures is $d \gg 2$
- Exposures are correlated

πMeshed GPs

(MP et al, 2021+)



- Construct **low-dimensional** space D^*

  (build via **PCA projection** of inputs **X** or covariates **Z**)
- Partition D^* and map partitions to "nice" DAG
- Sparse DAG used <u>only for scalability to big data!*</u>

$$\widetilde{p}(\boldsymbol{h} \mid \boldsymbol{\rho}) = \prod_{g \in \text{Graph}} \text{MVN}(\boldsymbol{h}_g; \begin{matrix}\text{conditional} \\ \text{mean}\end{matrix}, \begin{matrix}\text{conditional} \\ \text{variance}\end{matrix})$$

<u>*the GP kernels remain defined on the *full* input domain</u>

$$K_r(\boldsymbol{x}, \boldsymbol{x}') = \exp\left\{ -\sum_j \rho_{j,r}(x_{j,r} - x'_{j,r})^2 \right\}$$

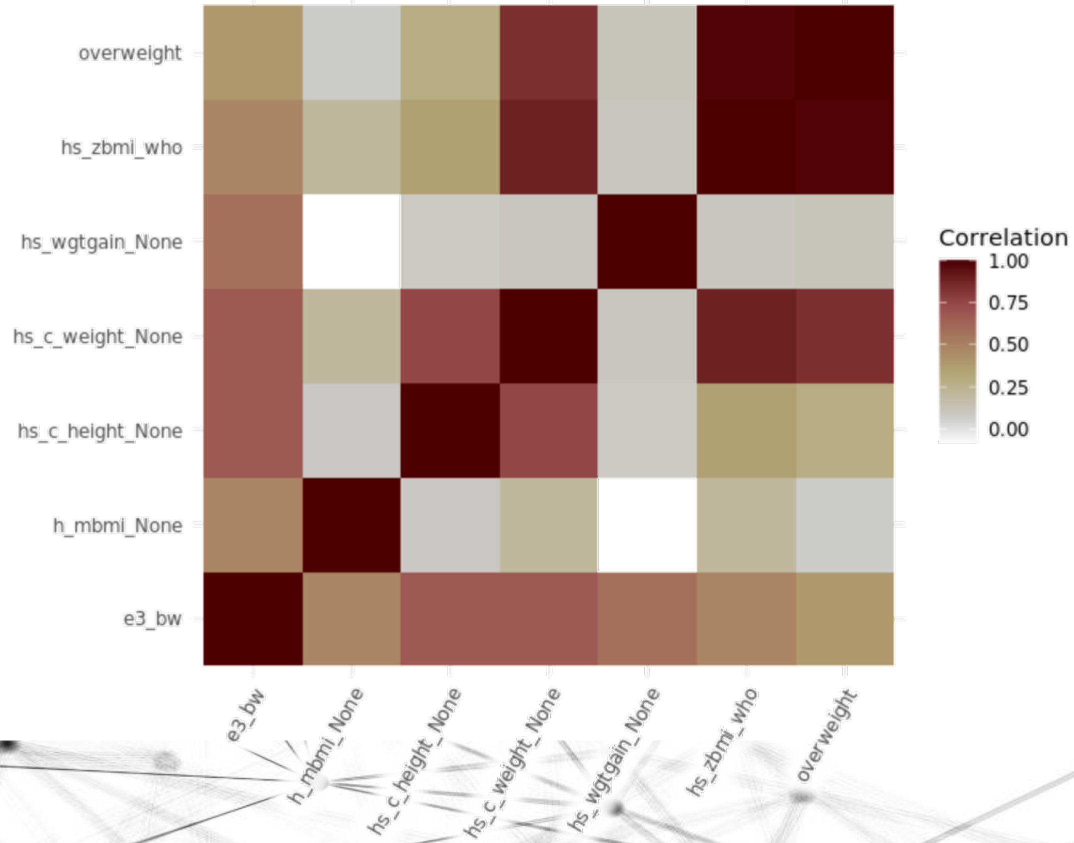
soon available as R package at
github.com/mkln/meshed

πMeshed GP on EXPOSOME data

Benchmarks (on 2000 MCMC iterations)

meshed::pimeshed	bkmr::kmbayes
1 outcome: 5.8 seconds	1 outcome: 188 seconds (32x slower)
7 outcomes: 32 seconds	7 outcomes: N/A

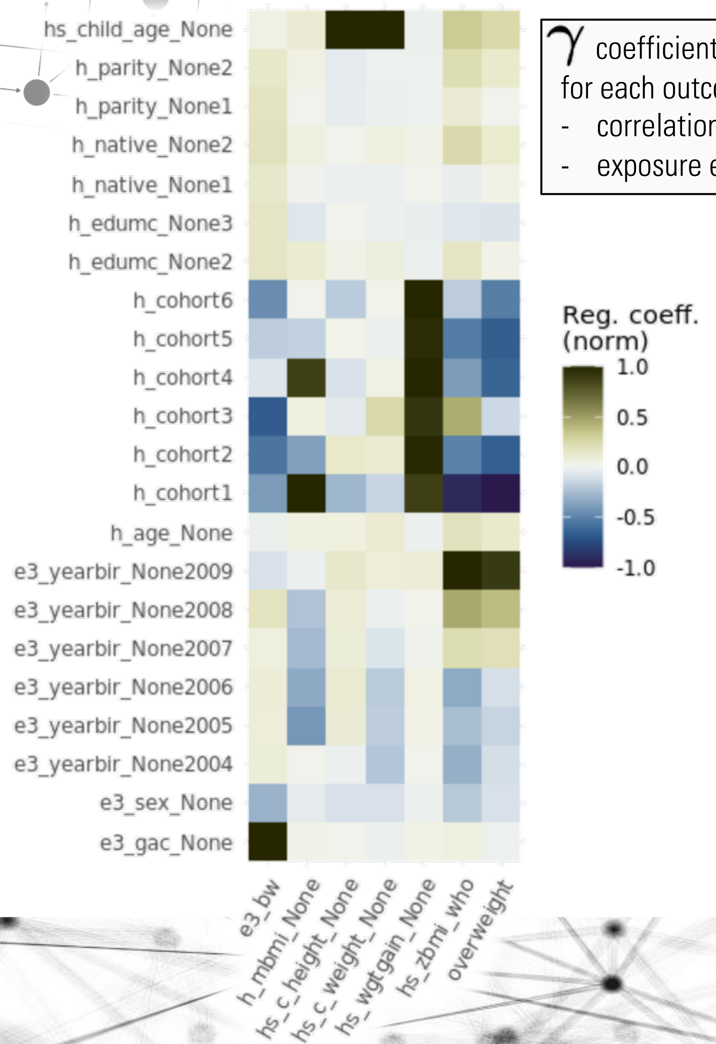
Correlations across outcomes
after accounting for covariates and exposures



$$\underbrace{q \text{ outcomes}}\ \{\ \boldsymbol{y}_i = \underbrace{\boldsymbol{A}}_{q\times k \text{ matrix}}\overbrace{\boldsymbol{h}}^{k \text{ GPs}}(\underbrace{\boldsymbol{x}_i}_{d \text{ inputs}}) + \overbrace{\boldsymbol{Z}_i}^{p \text{ covariates without missing values}}\gamma + \boldsymbol{\varepsilon}_i$$

- $d = 17$ exposures X
- $p = 9$ covariates/confounders with non-missing data Z
- $n = 1227$ subjects of which 1076 with fully observed outcomes
- $q = 7$ outcomes, of which 4 from covariates with missing data
- Effective data size $nq = 8589$
- $k = 4$

πMeshed GP on EXPOSOME data



γ coefficients on the covariates/confounders Z for each outcome & accounting for
- correlations across outcomes
- exposure effects

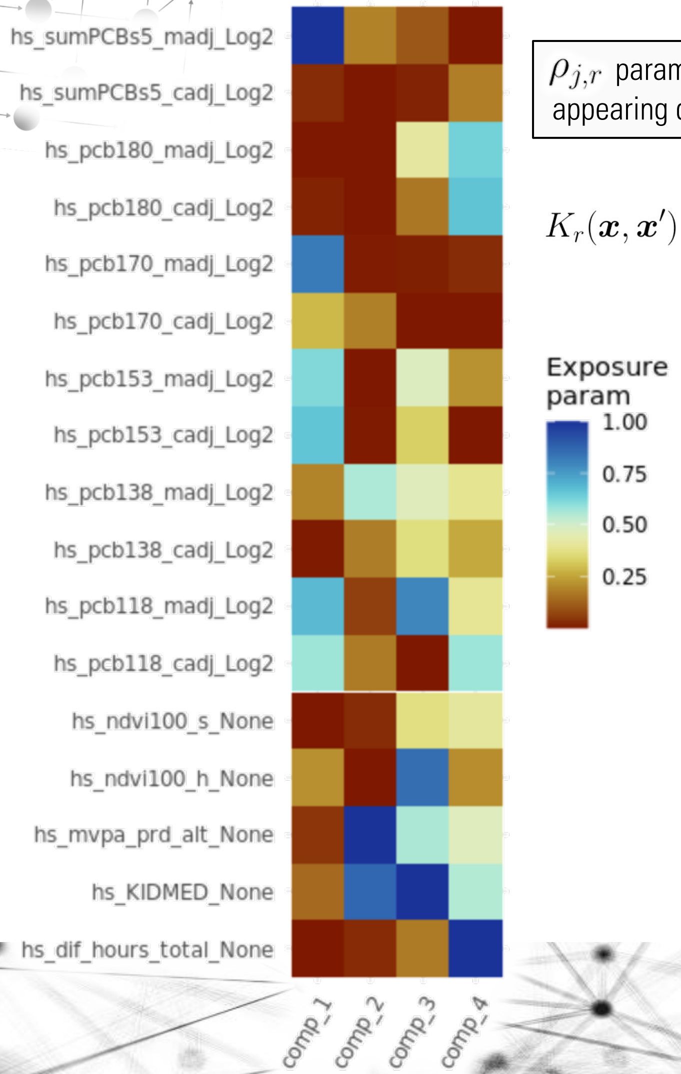
Benchmarks (on 2000 MCMC iterations)

meshed::pimeshed	bkmr::kmbayes
1 outcome: 5.8 seconds	1 outcome: 188 seconds (32x slower)
7 outcomes: 32 seconds	7 outcomes: N/A

$$q \text{ outcomes } \left\{ \; \boldsymbol{y}_i = \underbrace{\boldsymbol{A}}_{q\times k \text{ matrix}} \overbrace{\boldsymbol{h}}^{k \text{ GPs}}(\underbrace{\boldsymbol{x}_i}_{d \text{ inputs}}) + \overbrace{\boldsymbol{Z}_i}^{p \text{ covariates without missing values}} \gamma + \boldsymbol{\varepsilon}_i \right.$$

- $d = 17$ exposures X
- $p = 9$ covariates/confounders with non-missing data Z
- $n = 1227$ subjects of which 1076 with fully observed outcomes
- $q = 7$ outcomes, of which 4 from covariates with missing data
- Effective data size $nq = 8589$
- $k = 4$

πMeshed GP on EXPOSOME data



$\rho_{j,r}$ parameters for exposures appearing on the GP kernels

$$K_r(\boldsymbol{x}, \boldsymbol{x}') = \exp\left\{ -\sum_j \rho_{j,r}(x_{j,r} - x'_{j,r})^2 \right\}$$

Benchmarks (on 2000 MCMC iterations)

meshed::pimeshed	bkmr::kmbayes
1 outcome: <u>5.8 seconds</u>	1 outcome: 188 seconds (32x slower)
7 outcomes: 32 seconds	7 outcomes: N/A

$$q \text{ outcomes } \{ \ \boldsymbol{y}_i = \underbrace{\boldsymbol{A}}_{q\times k \text{ matrix}} \overbrace{\boldsymbol{h}}^{k \text{ GPs}}(\underbrace{\boldsymbol{x}_i}_{d \text{ inputs}}) + \overbrace{\boldsymbol{Z}_i}^{p \text{ covariates without missing values}} \gamma + \boldsymbol{\varepsilon}_i$$

- $d = 17$ exposures X
- $p = 9$ covariates/confounders with non-missing data Z
- $n = 1227$ subjects of which 1076 with fully observed outcomes
- $q = 7$ outcomes, of which 4 from covariates with missing data
- Effective data size $nq = 8589$
- $k = 4$ latent GP factors

Thank you!

- **Meshed GPs**: **PM**, Banerjee S, Finley AO (2020).
  Highly Scalable Bayesian Geostatistical Modeling via Meshed Gaussian Processes on Partitioned Domains.
  JASA in press [doi.org/10.1080/01621459.2020.1833889]

- **GriPS for Meshed GPs**: **PM**, Banerjee S, Dunson DB, Finley AO (2021).
  Grid-Parametrize-Split (GriPS) for Improved Scalable Inference in Spatial Big Data Analysis
  [arxiv.org/abs/2101.03579]

- **SPAMTREES**: **PM** & Dunson DB (2020).
  Spatial Multivariate Trees for Big Data Bayesian Regression
  [arxiv.org/abs/2012.00943]

- **Melange** *(meshed Riemannian-manifold Langevin algorithms)*: **PM** & Dunson DB (2021+).
  Spatial Meshing for General Bayesian Multivariate Models
  [soon]

Supported by:
ERC grant 856506
NIH grant R01ES028804