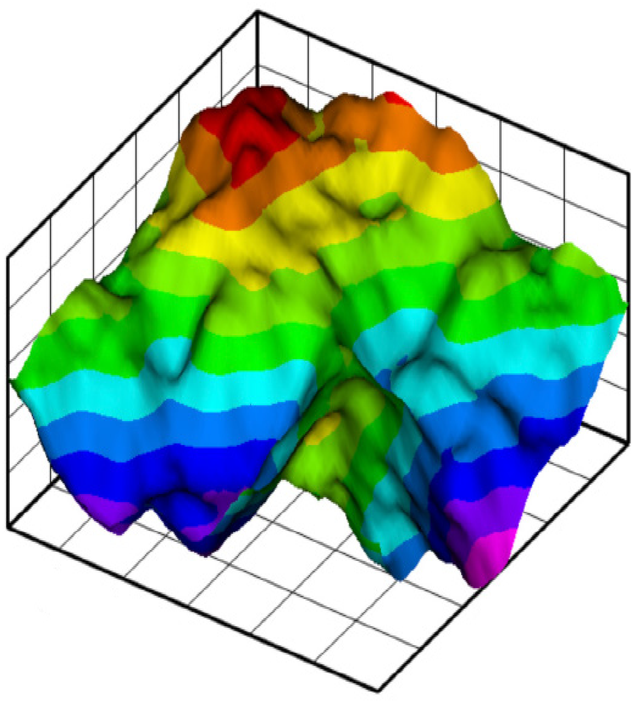


GRAMAR: Graph Machine Regression for assessing the effects of chemical exposures on health outcomes

Visit the GRAMAR poster

5 ratings | 3 answered questions

About this method

- Flexible regression modeling of large scale data in which several correlated inputs may have nonlinear effects on outputs
- Targets estimation of a smooth response surface, similar to standard GP regression
- Eliminates Gaussian process bottlenecks by inducing conditional independence
- Sparse directed acyclic graph (DAG) outlines conditional independence assumptions
- Based on successful family of methods for geostatistical data
- Based on projecting the input space to lower dimensions
- Open source R packages available for a variety of settings (incl. univariate, multivariate, multi-type, GP factor models)

FREE
& Open source

FREE delivery **Wednesday, June 29.**

In Stock.

Qty: 1

Add to Cart

Submit now

Anonymous peer review

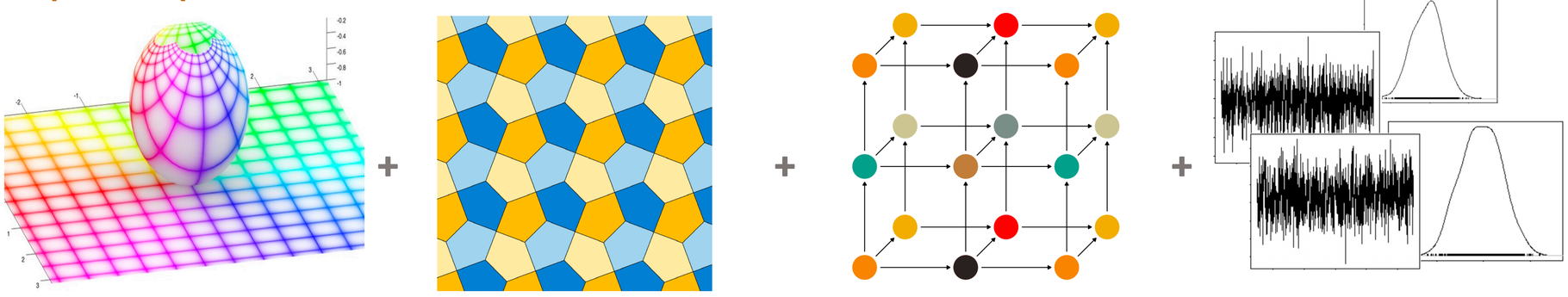
Authored by M.P. & D.B.D. and Funded by NIH and ERC

Rejection policy: Not eligible for rejection

Add to Wish to Submit

Have one to publish?
Submit to BA

Steps to reproduce this method



Project input space into a lower dimensional space:
- PCA
- Laplacian Eigenmaps
- other feature-preserving projections

Partition the newly-created input space
- Voronoi tessellation
- axis-parallel
- naive
- others

Link each region to a node in a DAG. Assume DAG is sparse.
- Fast density evaluation in original input space
- Graph coloring

Posterior sampling via MCMC:
- collapsed: sparse Cholesky
- non-collapsed: parallel Gibbs/Langevin/HMC

Technical specifications

Input space dimension	Suppose $z_i \in \mathbb{R}^d$ is a vector of inputs for subject i. We store all subjects' inputs in matrix $\boldsymbol{Z}$.
Target model	$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + f(z_i) + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, assume $h(\cdot) \sim \Pi_{\boldsymbol{\theta}}$. At $\boldsymbol{Z}$, in vector form $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{f}_{\boldsymbol{Z}} + \boldsymbol{\varepsilon}$.
Projection step	Project $\boldsymbol{Z}$ to $\boldsymbol{Z}^*$ so that $z_i^* \in \mathbb{R}^2$. On the new space, we can build a sparse DAG process (MGP, NNGP, MRA, SpamTrees, Vecchia-like).
Sparse GP choice	If MGP, then $\boldsymbol{Z}^*$ is partitioned into m blocks. If NNGP, then for every z_i^* we have a list of neighbors.
DAG factorization	Partition & sparse DAG factorization: $\pi_{\boldsymbol{\theta}}(\boldsymbol{f}_{\boldsymbol{Z}}) = N(\boldsymbol{f}_1; \boldsymbol{0}, \boldsymbol{C}_1) N(\boldsymbol{f}_2; \boldsymbol{H}_2 \boldsymbol{f}_{\mathrm{Pa}(2)}, \boldsymbol{R}_2) \cdots N(\boldsymbol{f}_m; \boldsymbol{H}_m \boldsymbol{f}_{\mathrm{Pa}(m)}, \boldsymbol{R}_m)$. Note that $\boldsymbol{Z}$ (and thus, $\boldsymbol{f}_{\boldsymbol{Z}}$) is also partitioned into m blocks $\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m\}$, each of which is 1:1 with the partition in $\boldsymbol{Z}^*$ space.
Notation for conditional Gaussians	Product of conditional Gaussians: $\boldsymbol{H}_i = \boldsymbol{C}_{i,\mathrm{Pa}(i)} \boldsymbol{C}_{\mathrm{Pa}(i)}^{-1}, \quad \boldsymbol{R}_i = \boldsymbol{C}_i - \boldsymbol{H}_i \boldsymbol{C}_{\mathrm{Pa}(i),i}$ using a pre-specified kernel operating on $\boldsymbol{Z}$ (thus, unrelated to chosen projection)
MCMC computations	Graph coloring leads to parallel Gibbs. Sparse DAG leads to sparse precision matrix and fast collapsed samplers.

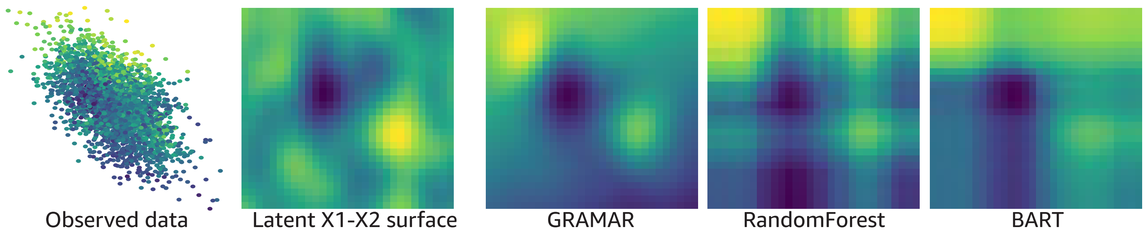
User reviews

Yurin Competent

★★★★★ **Almost gave up on GPs, until I found GRAMAR!**

Size: n=3000, d=15 | **Verified analysis**

I tried this on some simulated data with 15 correlated inputs. Lo and behold this thing crunches numbers F A S T, 50x faster than BKMR!



Observed data | Latent X1-X2 surface | GRAMAR | RandomForest | BART

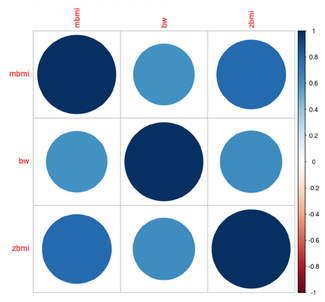
Recovered surfaces

Imma Jenyius

★★★★★ **Works well on HELIX exposure data**

Size: n=1096, d=51 | **Verified analysis**

I used GRAMAR on a multivariate outcome to estimate latent correlations. Fantastic! It only took 33 seconds when BKMR doesn't even work on multivariate outcomes and takes minutes for univariate data!



Method	RMSE	MAE	Covg95
GRAMAR-LE	0.4516	0.3410	0.9635
GRAMAR-PCA	0.4748	0.3522	0.9600
GRAMAR-PCA (a/p)	0.4533	0.3386	0.9598
BART	0.4555	0.3417	0.9406
BKMR	0.4599	0.3468	0.9479
RandomForest	0.4685	0.3535	NA
XGBOOST	0.4755	0.3552	NA

Back to top

Get to Know Us

Michele Peruzzi
Postdoctoral Associate
Department of Statistical Science, Duke University

David Dunson
Arts and Sciences Distinguished Professor of Statistical Science
Department of Statistical Science, Duke University

Analyze Data with Us

The **gramar** R package
github.com/mkln/gramar

The **meshed** R package
CRAN.R-project.org/package=meshed

Let Us Help You

michele.peruzzi@duke.edu
https://mkln.it

References

Peruzzi M, Banerjee S, Finley AO (2022) Meshed GPs. *JASA*. doi: 10.1080/01621459.2020.1833889
Peruzzi M, Dunson DB (2022) SpamTrees. *JMLR*. https://jmlr.org/papers/v23/20-1361.html
Peruzzi M, Dunson DB (2022) Spatial meshing in general Bayesian models. arXiv: 2201.10080
Datta A, Banerjee S, Finley AO, Gelfand AE (2016). NNGPs. *JASA*. doi: 10.1080/01621459.2015.1044091
Katzfuss M, Guinness J (2021). General Vecchia *StatSci*. doi: 10.1214/19-STS755
Vecchia AV (1988). Estim. cont. sp. proc. *JRSS:B*. doi: 10.1111/j.2517-6161.1988.tb01729.x