Botzmann-Gibbs distributions beyond statistical physics: Gaussian processes, spatiotemporal statistics and machine learning

Dionissios Hristopulos*

*School of Electrical and Computer Engineering
Technical University of Crete, Chania, Greece

MECO 2023, Stará Lesná, Slovakia

About the Technical University of Crete

TUC Campus, Akrotiri, Chania (Crete)

MSc in Machine Learning and Data Science TUC MLDS

Collaborators

- Andreas Pavlides
- Ivi Tsantili
- Vasiliki Agou
- Emmanouil Varouchakis
- Manolis Petrakis
- Milan Žukovič
- Samuel Elogne
- Denis Allard
- Thomas Opitz
- Emilio Porcu

Table of Contents

1. Introduction
2. Boltzmann-Gibbs Gaussian Processes
3. Local Interactions and GMRFs
4. Boltzmann-Gibbs Processes and Smoothed Particle Hydrodynamics
5. Stochastic Local Interaction Model
6. SLI Interpolation & Applications
7. Connections with Deep Learning?
8. Concluding Remarks

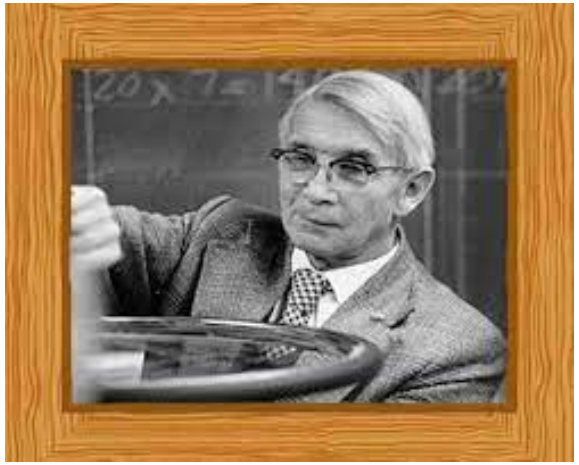
Motivation

Ludwig Boltzmann

Boltzmann-Gibbs distribution

$$p_i \propto \exp(-\epsilon_i / k_B T).$$

Josiah Willard Gibbs



Ernst Ising

Ising Model

$$H = \sum_{<i,j>} J_{i,j} S_i S_j + h \sum_i S_i, \; S_i = \pm 1.$$

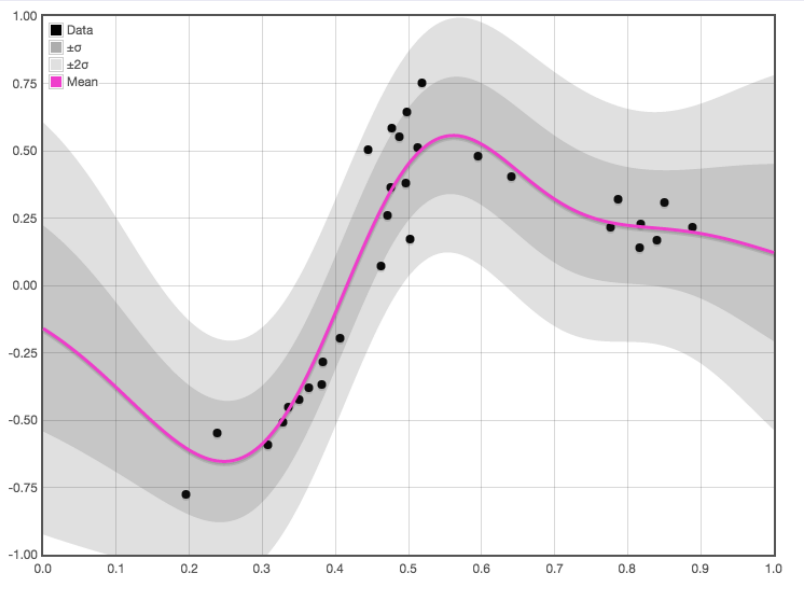
Gaussian field theory

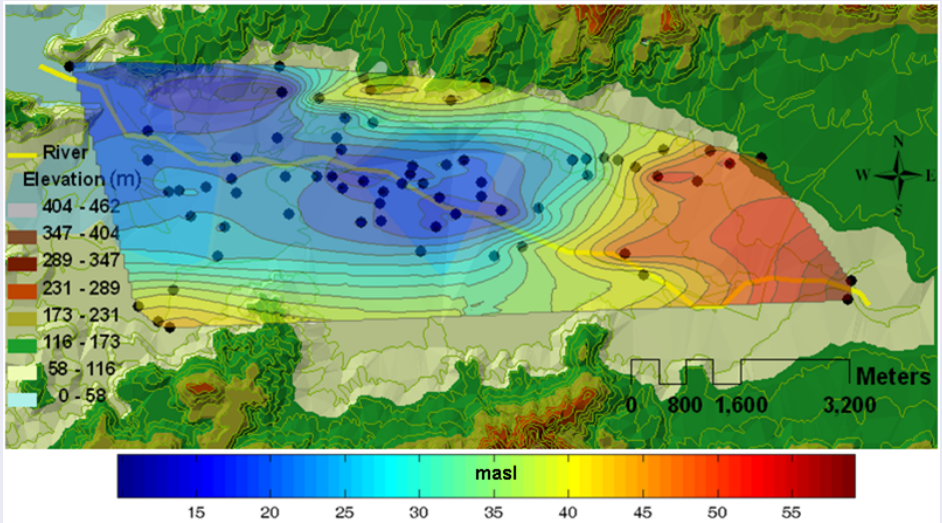
$$p \sim \mathrm{e}^{-\int \mathrm{d}\mathbf{x} \left[\alpha_0 \phi^2(\mathbf{x}) + \alpha_1 (\nabla \phi(\mathbf{x}))^2 \right]}$$

Motivation

Gaussian process regression

A Gaussian process $z(\mathbf{x}) \sim \mathcal{GP}\left(m(\mathbf{x}), C(\mathbf{x}, \mathbf{x}')\right)$ is a general approximation for functions based on sparse data.





- $\mathbf{z} \mid \mathbf{X}, \boldsymbol{\theta} \sim \mathcal{N}(\mathbf{m}, \mathbf{K}), \;\; m_i = m(\mathbf{x}_i), \; K_{i,j} = C(\mathbf{x}_i, \mathbf{x}_j), \; i, j = 1, \ldots, N$
- Prediction requires the inverse covariance kernel $\mathbf{K}^{-1}$

Motivation

- Accurate prediction of space-time processes based on sparse data remains a challenge (geostatistics, machine learning).
- Covariance-based approaches (kriging, Gaussian Process Regression) are computationally expensive due to calculations with large covariance matrices.
- Constructing physically interpretable models for covariance kernels is not straightforward.

- Can we leverage ideas from statistical physics to improve modeling and prediction of space-time processes ?

Motivation

- Accurate prediction of space-time processes based on sparse data remains a challenge (geostatistics, machine learning).
- Covariance-based approaches (kriging, Gaussian Process Regression) are computationally expensive due to calculations with large covariance matrices.
- Constructing physically interpretable models for covariance kernels is not straightforward.

- Can we leverage ideas from statistical physics to improve modeling and prediction of space-time processes ?

Motivation

- Accurate prediction of space-time processes based on sparse data remains a challenge (geostatistics, machine learning).
- Covariance-based approaches (kriging, Gaussian Process Regression) are computationally expensive due to calculations with large covariance matrices.
- Constructing physically interpretable models for covariance kernels is not straightforward.

- Can we leverage ideas from statistical physics to improve modeling and prediction of space-time processes ?

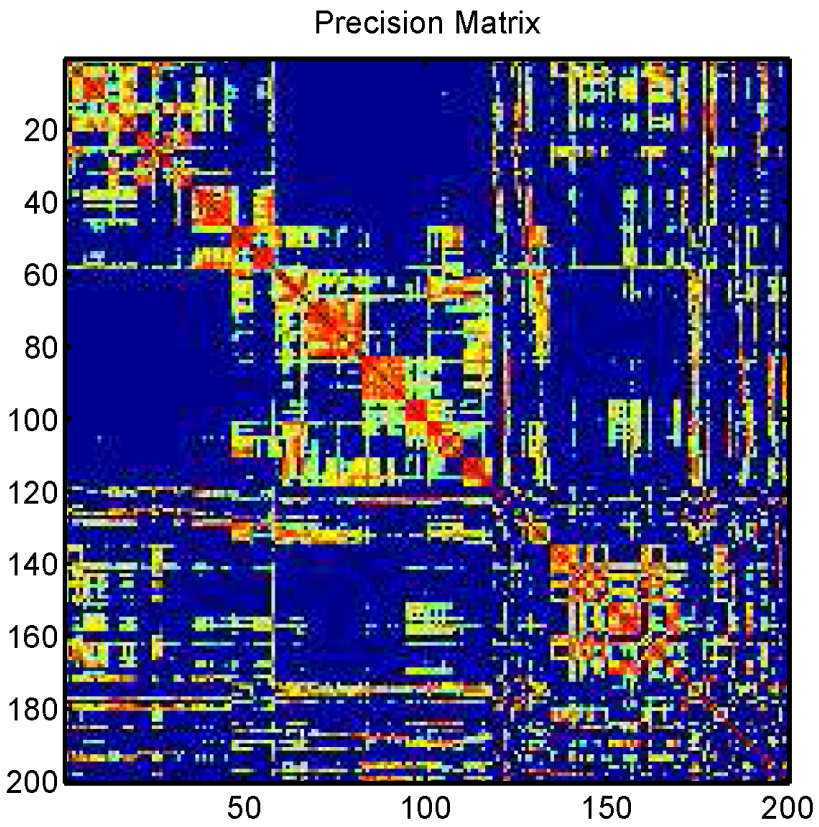
Motivation

- Accurate prediction of space-time processes based on sparse data remains a challenge (geostatistics, machine learning).
- Covariance-based approaches (kriging, Gaussian Process Regression) are computationally expensive due to calculations with large covariance matrices.
- Constructing physically interpretable models for covariance kernels is not straightforward.

- Can we leverage ideas from statistical physics to improve modeling and prediction of space-time processes ?

Motivation

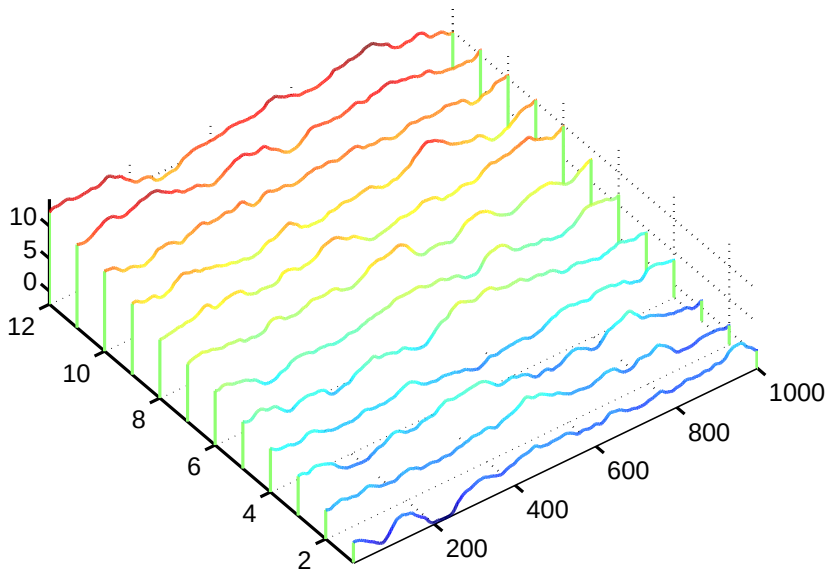
- Sparse models
- Local interactions $\Rightarrow$ precision "operator" formulations
- No covariance matrix inversion
- Interpretable models

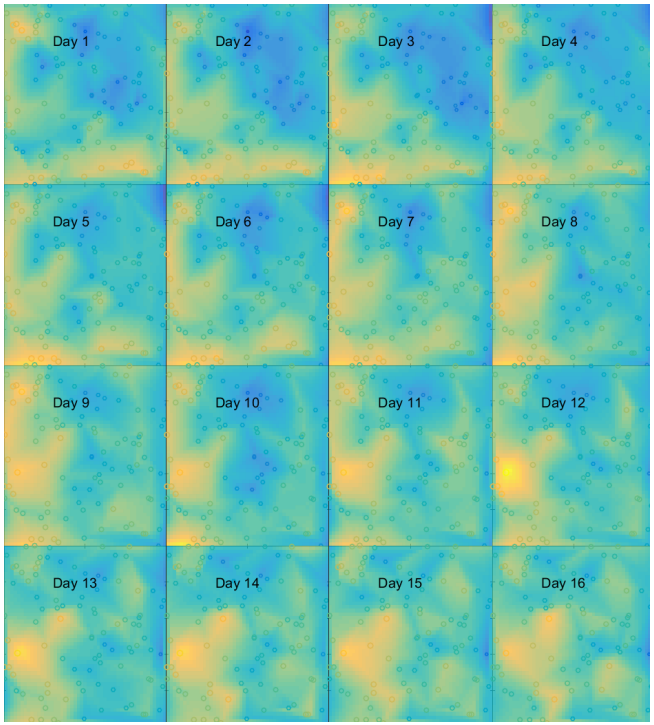


Definition of Gaussian Processes (GPs)

- Let $\mathbf{x} \in \mathbb{R}^d$ represent *input (feature vector)* and $z(\mathbf{x}) : \mathbb{R}^{\mathrm{d}} \to \mathbb{R}$ a scalar function.
- $m(\mathbf{x}) : \mathbb{R}^{\mathrm{d}} \to \mathbb{R}$ represents the expectation of $z(\mathbf{x})$.
- $C(\mathbf{x}, \mathbf{x}') : \mathbb{R}^{\mathrm{d}} \times \mathbb{R}^{\mathrm{d}} \to \mathbb{R}$ is a covariance kernel, i.e., a symmetric, non-negative definite function.
- A Gaussian process $z(\mathbf{x}) \sim \mathcal{GP}\left(m(\mathbf{x}), C(\mathbf{x}, \mathbf{x}')\right)$ is a prior over the functions $z(\mathbf{x})$:

$$z(\mathbf{x}) \mid \boldsymbol{\theta} \sim \mathcal{N}(\mathbf{m}, \mathbf{K}), \text{ where } m_i = m(\mathbf{x}_i),\ K_{i,j} = C(\mathbf{x}_i, \mathbf{x}_j),\ i, j = 1, \ldots, N$$





Definition of Gaussian Processes (GPs)

Gaussian processes are defined in a Bayesian framework in which observations help to update *a priori* beliefs.

GPs in Bayesian framework

- Observations: $y_i = z(\mathbf{x}_i) + \epsilon_i, \;\; \epsilon_i$: noise
- Observation model: $\mathbf{y} \mid \mathbf{z}, \phi \sim \prod_{i=1}^{N} p(y_i \mid z_i, \phi)$
- Gaussian Process prior: $z(\mathbf{x} \mid \boldsymbol{\theta}) \sim \mathcal{GP}\left(m(\mathbf{x}), C(\mathbf{x}, \mathbf{x}'); \boldsymbol{\theta}\right)$
- Hyperparameter prior: $\boldsymbol{\theta}, \phi \sim p_\theta(\boldsymbol{\theta})\, p_\phi(\phi)$

For Gaussian i.i.d noise with variance σ_n^2 and fixed hyperparameters:

Predictions at test point $\mathbf{x}_*$ are normally distributed with mean and variance given by (Rasmussen & Williams, 2006):

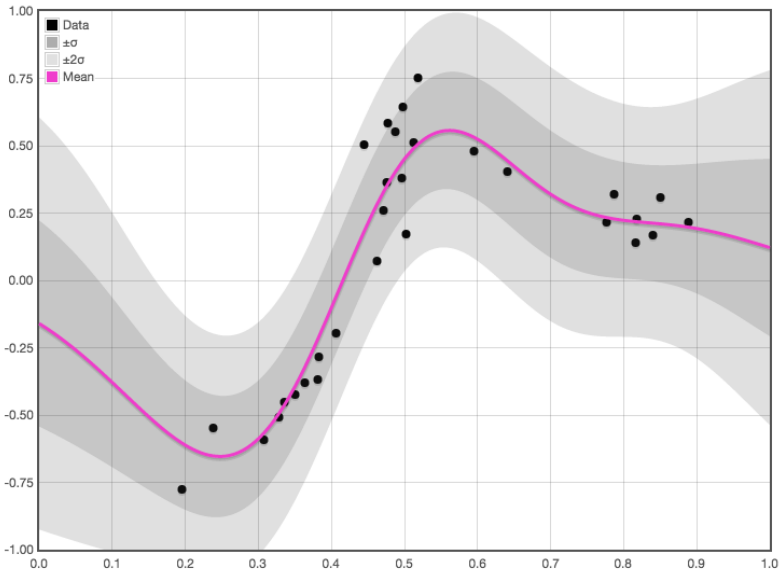
$$\mathbb{E}[z(\mathbf{x}_*)] = m(\mathbf{x}_*) + \mathbf{k}_*^\top \left(\mathbf{K} + \sigma_n^2 \mathbf{I}\right)^{-1} \mathbf{y}$$

$$\mathrm{Var}\left(z(\mathbf{x}_*)\right) = C(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^\top \left(\mathbf{K} + \sigma_n^2 \mathbf{I}\right)^{-1} \mathbf{k}_*$$

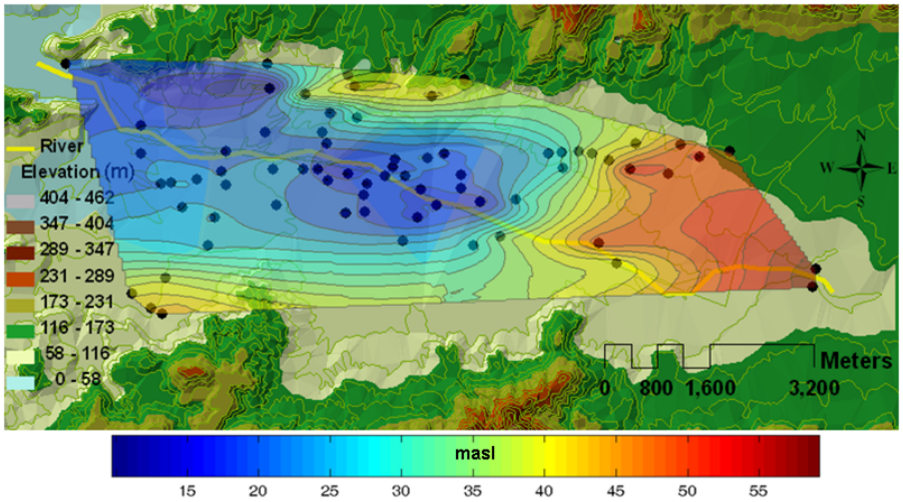
$$[\mathbf{K}]_{i,j} = C(\mathbf{x}_i, \mathbf{x}_j), \quad [\mathbf{k}_*]_i = C(\mathbf{x}_i, \mathbf{x}_*), \; i, j = 1, \ldots N.$$

Gaussian Processes vs Spatial (Geo-) Statistics

- Gaussian process
- Input: general features
- Kernel function
- White kernel: $C(t, t') = \sigma^2 \delta(t - t')$
- Gaussian process regression
- Generative model



- Gaussian random field
- Input: space/time coordinates
- Covariance function
- Nugget: $C(\mathbf{s}, \mathbf{s}') = \sigma^2 \delta(\mathbf{s} - \mathbf{s}')$
- Kriging
- Spatial simulation



Standard Decomposition of Earth Observation Data

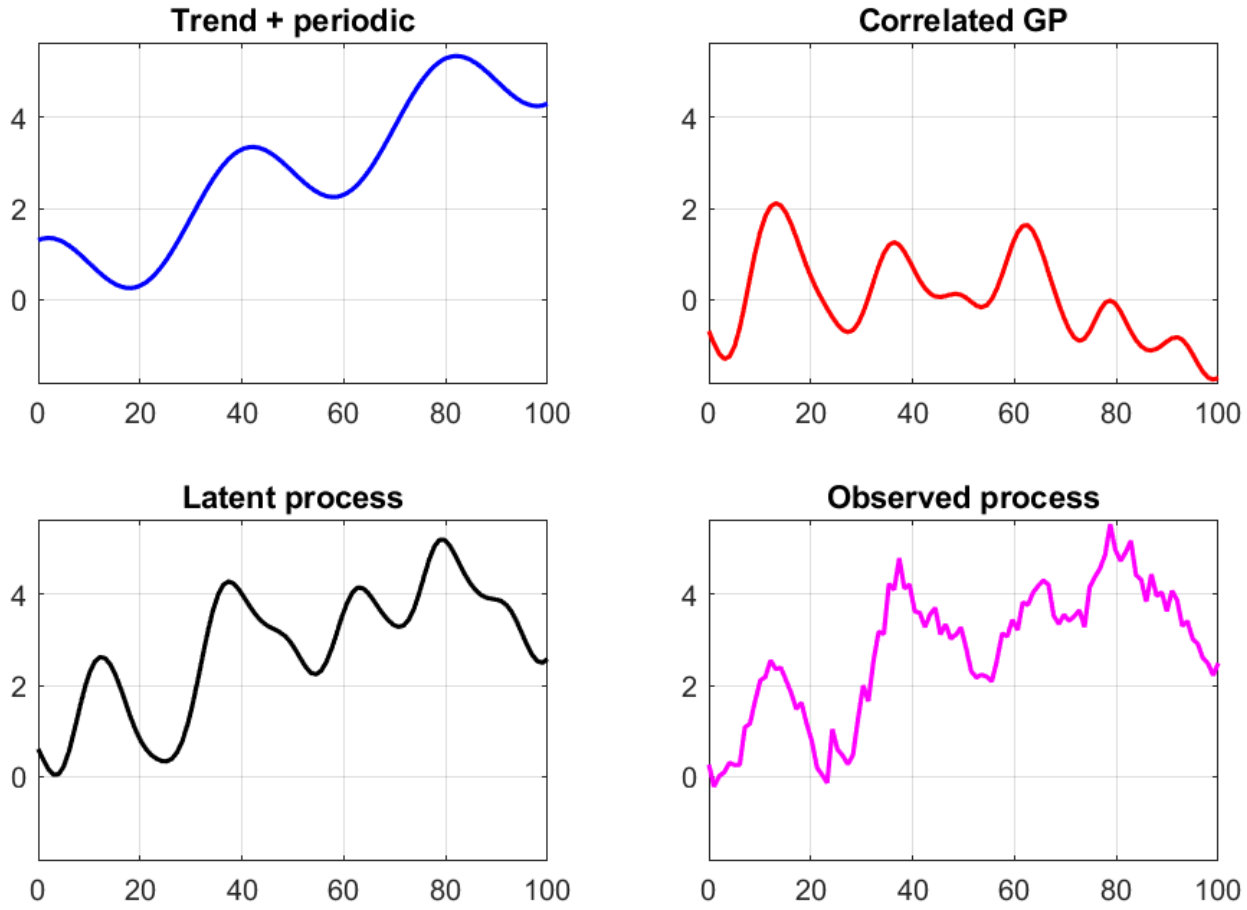
Trend-Fluctuation-Noise analysis of Observations

Latent model: $z(\mathbf{x};\omega) = m(\mathbf{x}) + S(\mathbf{x}) + z'(\mathbf{x};\omega)$

Observation model: $y(\mathbf{x};\omega) = z(\mathbf{x};\omega) + \epsilon(\mathbf{x};\omega)$

- $m(\mathbf{x})$: Trend (deterministic function for "large-scale" variations)
- $S(\mathbf{x})$: Periodic component $S(\mathbf{x} + \mathbf{x}_S) = S(\mathbf{x})$
- $z'(\mathbf{x};\omega)$: Correlated residual ("local" fluctuations)
- $\epsilon(\mathbf{x};\omega)$: Observation noise
- $m(\mathbf{x}) + S(\mathbf{x}) = \mathbb{E}[z(\mathbf{x};\omega)]$: Expectation over the ensemble of states
- For univariate earth observation data analyses $\mathbf{x} = (\mathbf{s}, t)$

Standard Decomposition of Earth Observation Data



Is this signal or noise?

"Today one person's noise may be someone else's signal". *Physics Today* 73(2): 42 (2020)

Table of Contents

1. Introduction
2. Boltzmann-Gibbs Gaussian Processes
3. Local Interactions and GMRFs
4. Boltzmann-Gibbs Processes and Smoothed Particle Hydrodynamics
5. Stochastic Local Interaction Model
6. SLI Interpolation & Applications
7. Connections with Deep Learning?
8. Concluding Remarks

Boltzmann-Gibbs Gaussian Processes

Langevin Equation

$Lz = \epsilon, \quad \epsilon$: GWN

L: Linear partial differential operator

$$\phi(\boldsymbol{t}) = E[\mathrm{e}^{\mathrm{i}\boldsymbol{t}\cdot(\boldsymbol{L}^{-1}\boldsymbol{\epsilon})}]$$

$$C(\mathbf{k}) = \frac{1}{L^{\dagger}(\mathbf{k})L(\mathbf{k})}$$

Boltzmann–Gibbs PDF

$p \sim \exp(-H(z))$

$H(z)$: Quadratic function

$$C(\mathbf{k}) = \frac{A}{\sum_{i=0}^{n} c_i \, \|\mathbf{k}\|^{2i}}, \qquad A > 0, \ \{c_i\}_{i=0}^{n} > 0$$

Rational SPD => Local Gaussian Processes

Whittle-Matérn case: L is a pseudo-differential operator. The SPD is a rational function only if $\boldsymbol{\nu} + \boldsymbol{d}/\boldsymbol{2} = \boldsymbol{n} \in \boldsymbol{N}$

Boltzmann-Gibbs "local" Gaussian processes

- Boltzmann-Gibbs probability density function (PDF)

$$p[z(\mathbf{x})] = Z^{-1}\,\mathrm{e}^{-\mathcal{H}[z(\mathbf{x})]}, \quad \mathcal{H}[z(\mathbf{x})]: \text{ energy functional}$$

- Z: partition function $Z = \int \mathcal{D}z(\mathbf{x})\,\mathrm{e}^{-\mathcal{H}[z(\mathbf{x})]}$
- The LAP-2 Gaussian model (Spartan random field-SSRF)[a]

$$\mathcal{H}_{\mathrm{LAP2}}[z(\mathbf{x})] = \frac{1}{2\eta_0\xi^d}\int \mathrm{d}\mathbf{x}\,\left\{z^2(\mathbf{x}) + \eta_1\,\xi^2\,[\nabla z(\mathbf{x})]^2 + \xi^4\,\left[\nabla^2 z(\mathbf{x})\right]^2\right\}$$

- The LAP-1 Gaussian model [b]

$$\mathcal{H}_{\mathrm{LAP1}}[z(\mathbf{x})] = \frac{1}{2}\int \mathrm{d}\mathbf{x}\,\left\{\alpha_0 z^2(\mathbf{x}) + \alpha_1\,[\nabla z(\mathbf{x})]^2\right\}$$

[a] Hristopulos, 2003. *SIAM J. Sci. Comput.*, 24(6), 2125–2162

[b] Allard, Hristopulos, Opitz (2021). *Elect. J. Stat.*, Vol. 15, 4085–4116.

Boltzmann-Gibbs "local" Gaussian processes

- Langevin equation for LAP-2 Gaussian processes (SSRFs)

$$\mathcal{L}\, z(\mathbf{s}) = \epsilon(\mathbf{s}), \text{ where } \epsilon(\cdot) \text{ is GWN (spatial Gaussian white noise)}$$

$$\mathcal{L} = \frac{1}{\sqrt{\eta_0 \xi^d}} \left[1 - \left(\sqrt{2+\eta_1} \right) \xi \, (-\triangle)^{1/2} + \xi^2 \triangle \right],$$

- $\triangle$: Laplacian operator, $(-\triangle)^{1/2}$: Half-Laplacian.
- LAP-2 (SSRF) Spectral density (**k** is the wavevector)

$$\tilde{C}(\mathbf{k}) \triangleq \frac{1}{\left|\tilde{\mathcal{L}}(\mathbf{k})\right|^2} = \frac{\eta_0 \xi^d}{1 + \eta_1 (\|\mathbf{k}\|\xi)^2 + (\|\mathbf{k}\|\xi)^4},$$

- $\tilde{\mathcal{L}}(\mathbf{k})$: Fourier image of $\mathcal{L}$

Boltzmann-Gibbs "local" Gaussian processes

- "Energy" of LAP-2 Gaussian process: $\mathcal{H}_{\mathrm{LAP2}} = \frac{1}{2} \int \mathrm{d}\mathbf{s} \int \mathrm{d}\mathbf{s}' \, z(\mathbf{s}') \, \mathbb{Q} \, z(\mathbf{s}) \, .$
- LAP-2 Precision Operator: $\mathbb{Q} = -\frac{1}{\eta_0 \xi^d} \left(\mathbf{1} - \eta_1 \, \xi^2 \triangle_{\mathbf{s}} + \xi^4 \triangle_{\mathbf{s}}^2 \right)$
- Covariance: Green's function for associated fourth-order PDE:

$$\mathbb{Q} C(\mathbf{s} - \mathbf{s}') = \delta(\mathbf{s} - \mathbf{s}')$$

- The covariance PDE is useful for Karhunen-Loéve expansions. [a]

[a] Hristopulos & Tsantili, *Prob. Engin. Mech.* 3:132–147 (2016).

Boltzmann-Gibbs "local" Gaussian processes

SPDEs, Covariance Functions and Green's functions

Dolph and Woodbury on the connection between covariance functions of processes governed by Langevin equations and the Green's function of associated differential equation [a]:

> *. . . our method is based on the intrinsically interesting relationship that exists between covariance functions of random processes generated by driving n-th order linear differential equations by so-called "pure noise" and the Green's function of a suitably defined self-adjoint equation of order 2n.*

[a] Dolph, C.L., Woodbury, M.A. *Trans. Am. Math. Soc.* 72(3), 519–550 (1952)

SSRF (LAP-2) covariance for time series

$$C(h) = \frac{\eta_0}{4}\, e^{-h\beta_2} \left[\frac{\cos(h\beta_1)}{\beta_2} + \frac{\sin(h\beta_1)}{\beta_1} \right], \quad |\eta_1| < 2$$

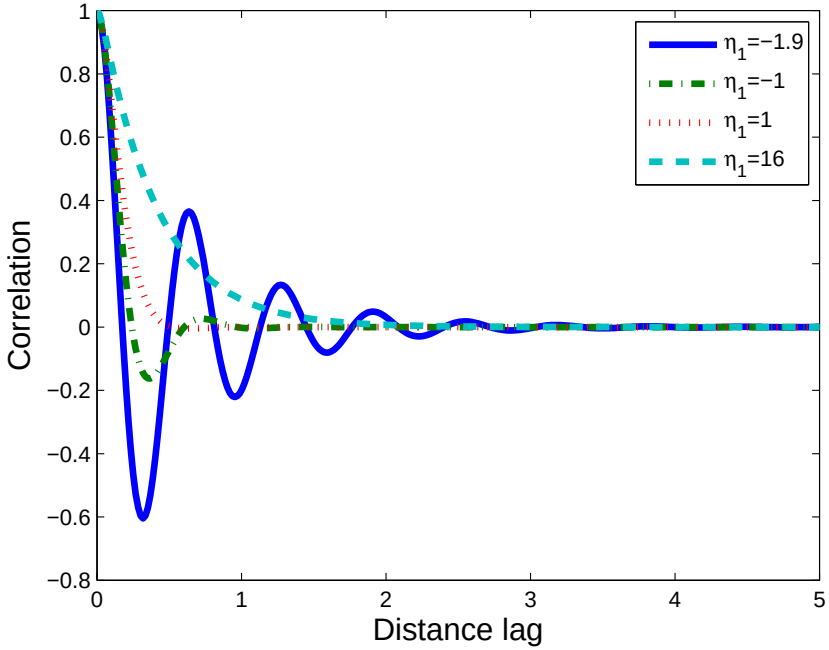
$$C(h) = \eta_0 \frac{(1+h)}{4\, e^{h}}, \quad \eta_1 = 2$$

$$C(h) = \frac{\eta_0}{2\,\Delta} \left(\frac{e^{-h\,\omega_1}}{\omega_1} - \frac{e^{-h\,\omega_2}}{\omega_2} \right), \quad \eta_1 > 2$$

Notation

- $h = |r|/\xi$: normalized lag
- $\beta_{1,2} = \left(\frac{|2 \mp \eta_1|}{4} \right)^{1/2}$
- $\omega_{1,2} = \left(\frac{|\eta_1 \mp \Delta|}{2} \right)^{1/2}$
- $\Delta = |\eta_1{}^2 - 4|^{\frac{1}{2}}$

d=1



Hristopulos & Elogne (2007), *IEEE Trans. Inform. Theory*, 53(12), 4667–4679

One-dimensional SSRF is a damped linear harmonic oscillator driven by Gaussian white noise

Langevin equation representation of LAP-2 process

$$\ddot{x}(t) + \Gamma\,\dot{x}(t) + \omega_0^2\,x(t) = \epsilon(t), \quad \epsilon(t)\text{: GWN}$$

$$C(h) = \frac{\eta_0}{4}\,\mathrm{e}^{-h\beta_2}\left[\frac{\cos(h\beta_1)}{\beta_2} + \frac{\sin(h\beta_1)}{\beta_1}\right], \; |\eta_1| < 2,$$ ■ Underdamping

$$C(h) = \eta_0\,\frac{(1+h)}{4\,\mathrm{e}^{h}}, \; \eta_1 = 2,$$ ■ Critical damping

$$C(h) = \frac{\eta_0}{2\,\Delta}\left(\frac{\mathrm{e}^{-h\,\omega_1}}{\omega_1} - \frac{\mathrm{e}^{-h\,\omega_2}}{\omega_2}\right), \; \eta_1 > 2$$ ■ Overdamping

SSRF (LAP-2) covariance functions for planar processes

$$C(h) = \frac{\eta_0 \, \Im\left[K_0(h\, z_+)\right]}{\pi\sqrt{4-{\eta_1}^2}}, \quad |\eta_1| < 2$$

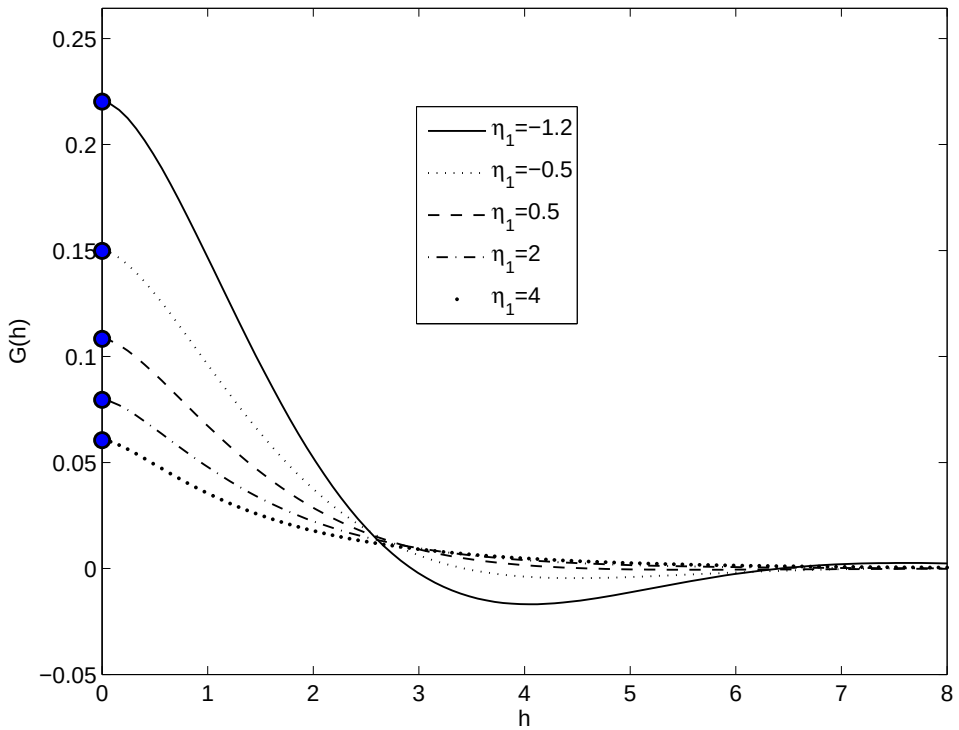
$$C(h) = \left(\frac{\eta_0\, h}{4\pi}\right) K_{-1}(h), \quad \eta_1 = 2$$

$$C(h) = \frac{\eta_0\left[K_0(h\, z_+) - K_0(h\, z_-)\right]}{2\pi\sqrt{{\eta_1}^2-4}}, \quad \eta_1 > 2$$

Notation

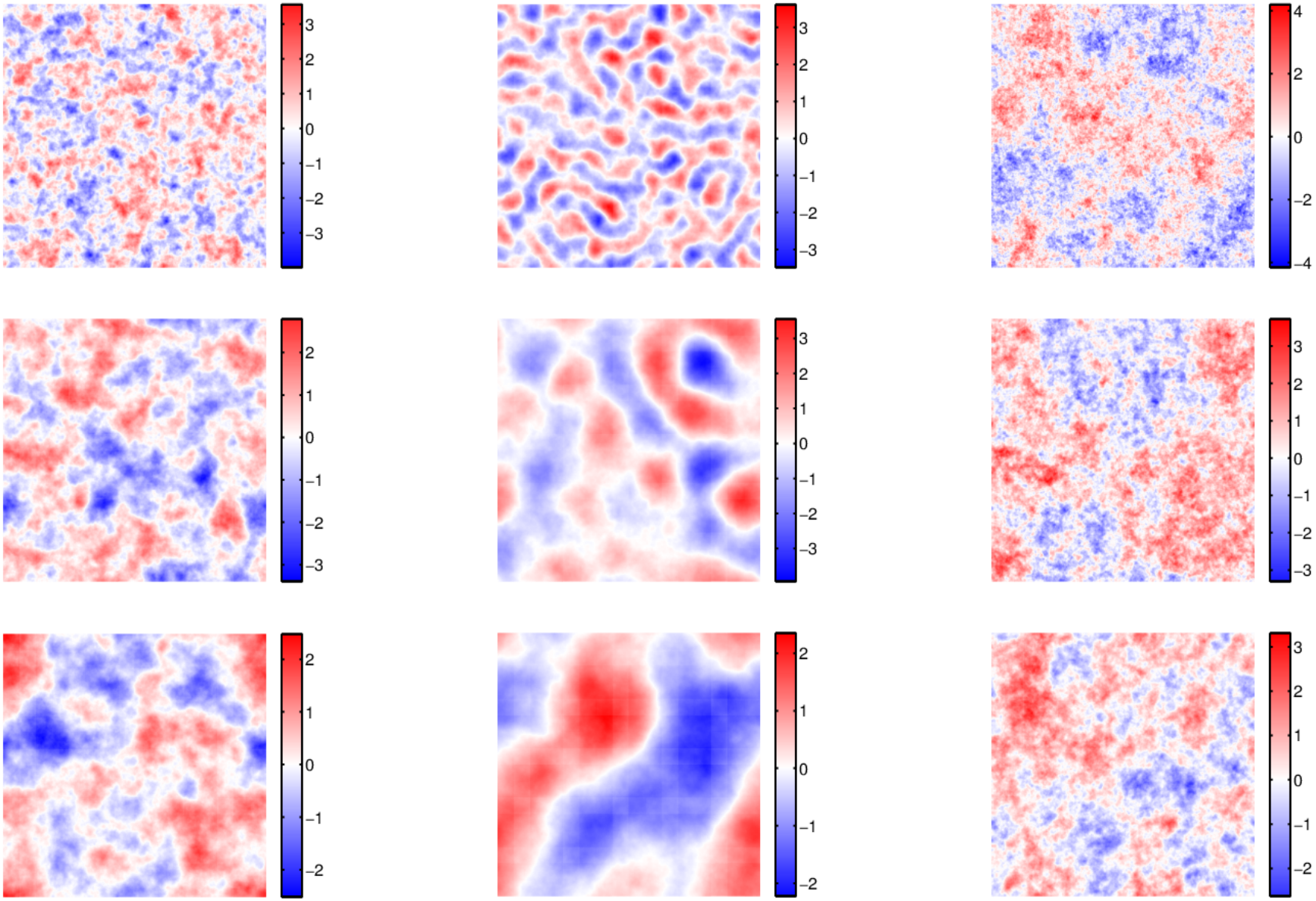
- $\Im$: Imaginary part
- $z_\pm = \sqrt{-t_\pm^*}$
- $t_\pm^* = \left(-\eta_1 \pm \sqrt{{\eta_1}^2-4}\right)/2$
- $K_\nu(z)$: modified Bessel function of the second kind and order ν

d=2



Hristopulos (2015), *Stoch. Environ. Res. Risk Assesss.*, 29(3), 739–754.

Examples of LAP-2 process realizations



Columns 1-3: $\eta_1 = 2, -1.95, 100$. Rows 1-3: $L/\xi = 50, 20, 10$.

SSRF (LAP-2) Covariance functions in $\mathbb{R}^3$

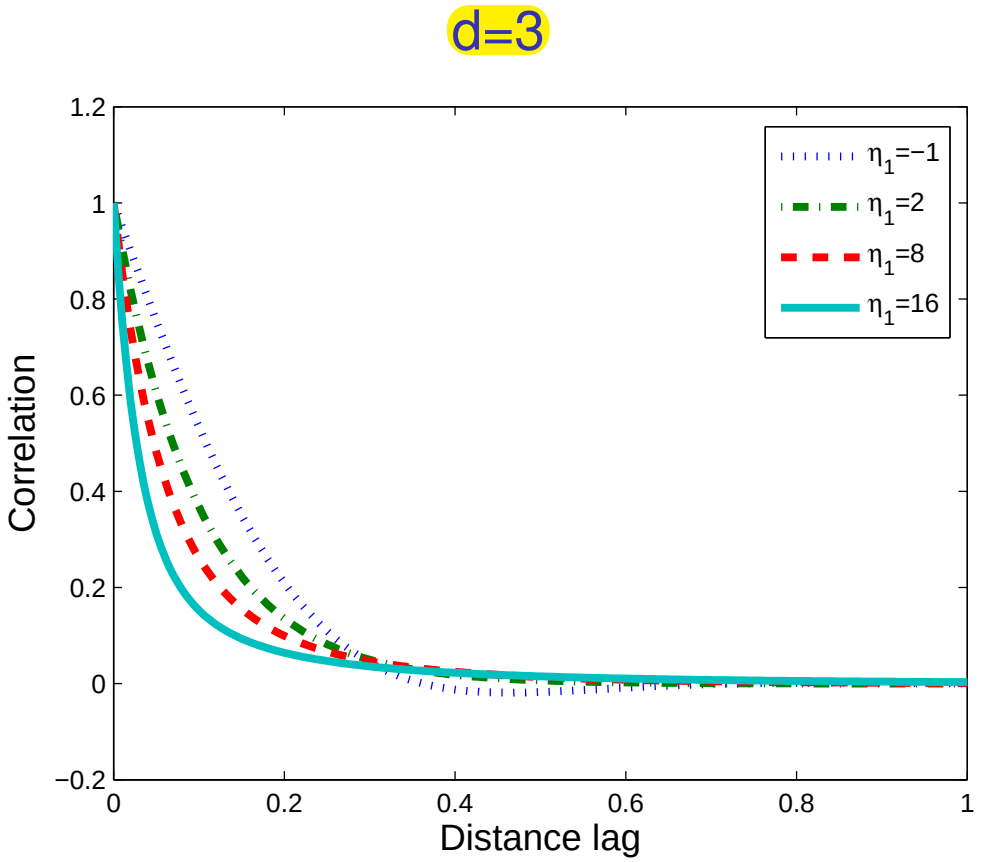
$$C(h) = \eta_0 \frac{e^{-h\beta_2}}{2\pi\Delta}\left[\frac{\sin(h\beta_1)}{h}\right], \quad |\eta_1| < 2$$

$$C(h) = \frac{\eta_0}{8\pi} e^{-h}, \quad \eta_1 = 2$$

$$C(h) = \frac{1}{4\pi\Delta}\left[\frac{e^{-h\omega_1} - e^{-h\omega_2}}{h(\omega_2 - \omega_1)}\right], \quad \eta_1 > 2$$

Notation

- $h = \|\mathbf{r}\|/\xi$,
- $\beta_{1,2} = \left(\frac{|2 \mp \eta_1|}{4}\right)^{1/2}$
- $\omega_{1,2} = \left(\frac{|\eta_1 \mp \Delta|}{2}\right)^{1/2}$
- $\Delta = |\eta_1{}^2 - 4|^{\frac{1}{2}}$



Hristopulos and Elogne, 2007, *IEEE Trans. Inform. Theory*, 53(12), 4667–4679

SSRF composite space-time covariance in $\mathbb{R}^2 \times T$

3D SSRF (LAP-2) Kernel

$$C(h) = \eta_0 \frac{e^{-h\beta_2}}{2\pi\Delta}\left[\frac{\sin(h\beta_1)}{h}\right], \quad |\eta_1| < 2$$

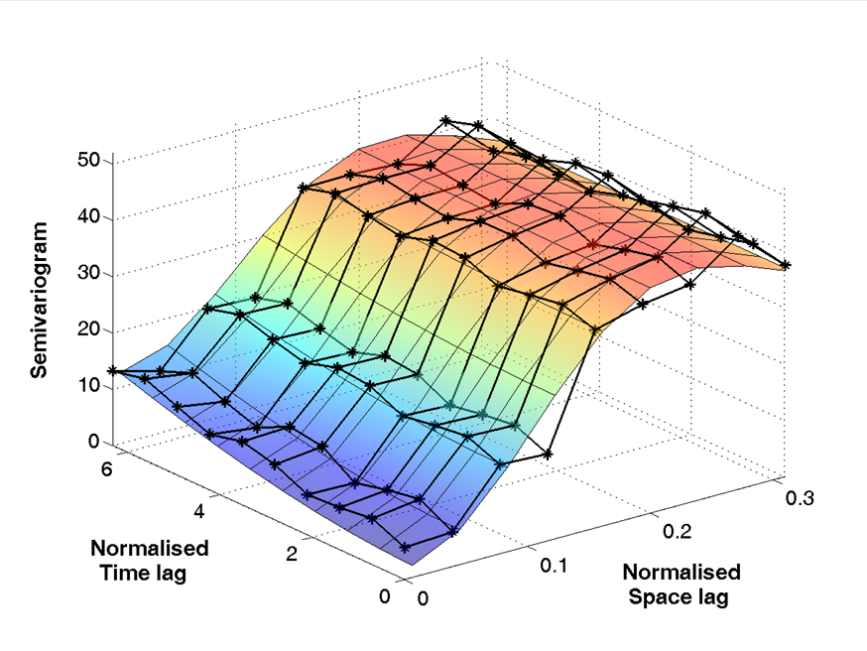
$$C(h) = \frac{\eta_0}{8\pi} e^{-h}, \quad \eta_1 = 2$$

$$C(h) = \frac{1}{4\pi\Delta}\left[\frac{e^{-h\omega_1} - e^{-h\omega_2}}{h(\omega_2 - \omega_1)}\right], \quad \eta_1 > 2$$

- $h = \|\mathbf{r}\|/\xi$,
- $\beta_{1,2} = \left(\frac{|2 \mp \eta_1|}{4}\right)^{1/2}$
- $\omega_{1,2} = \left(\frac{|\eta_1 \mp \Delta|}{2}\right)^{1/2}$
- $\Delta = |\eta_1{}^2 - 4|^{\frac{1}{2}}$

Composite space-time metric

$$h = \sqrt{\|\mathbf{r}\|^2/\xi^2 + \alpha\tau^2/\tau_c^2}$$



Varouchakis & Hristopulos, 2017. *Spatial Statistics*, https://doi.org/10.1016/j.spasta.2017.07.003.

LAP-1 Boltzmann-Gibbs Covariance Kernels

$$\mathcal{H}_{\mathrm{LAP1}}[z(\mathbf{x})] = \frac{1}{2}\int d\mathbf{x}\,\left\{\alpha_0 z^2(\mathbf{x}) + \alpha_1\,[\nabla z(\mathbf{x})]^2\right\}$$

Kernels with no cutoff

$$C_1(h) = \frac{\mathrm{e}^{-h}}{2\,\xi\alpha_0}\,,\; h = \|\mathbf{r}\|/\xi\,,$$

$$C_2(h) = \frac{K_0(h)}{2\pi\alpha_1}\,,\; \xi = \sqrt{\alpha_1/\alpha_0}\,,$$

$$C_3(h) = \frac{\mathrm{e}^{-h}}{4\pi\alpha_1 h\xi}\,,$$

$$C_4(h) = \frac{K_1(h)}{4\pi^2\alpha_1\xi^2 h}\,.$$

Regularized Kernels

- $K_\nu(\cdot)$: Modified Bessel functions of the second kind of order ν.
- $C_d(h)$ are generalized covariances in $d = 2, 3, 4$.
- Real-space regularization instead of momentum cutoff.
- The kernels $C_d^*(h) \triangleq C_d(h+\epsilon)$, where $\epsilon > 0$ are positive definite and non-singular.

Allard, Hristopulos, Opitz, *Elec. J. Stat.*, Vol. 15 (2021) 4085–4116

Table of Contents

1. Introduction
2. Boltzmann-Gibbs Gaussian Processes
3. Local Interactions and GMRFs
4. Boltzmann-Gibbs Processes and Smoothed Particle Hydrodynamics
5. Stochastic Local Interaction Model
6. SLI Interpolation & Applications
7. Connections with Deep Learning?
8. Concluding Remarks

B-G Gaussian Processes on Regular Lattices

- Derivatives $\approx$ finite differences $\Rightarrow$ Gauss-Markov random fields (GMRFs) [1]
- $\mathcal{H}_{\text{LAP1}}[z(\mathbf{s})] = \frac{1}{2}\int d\mathbf{s}\,\left\{\alpha_0 z^2(\mathbf{s}) + \alpha_1\,[\nabla z(\mathbf{s})]^2\right\}$ leads to two well-known lattice models in d dimensions[2]:
- the Besag model: $\alpha_0 = 0$, $\tau = \alpha_1/b^{-(d+2)}$, lattice step b

$$\hat{y}_p \mid \mathbf{y}_{-p} \sim \mathcal{N}\left(\frac{1}{|\mathrm{NB}_p|}\sum_{\tilde{p}\in\mathrm{NB}_p} y_{\tilde{p}},\ \frac{1}{|\mathrm{NB}_p|\,\tau}\right), \quad p = 1,\ldots,n$$

- the Leroux model with precision matrix

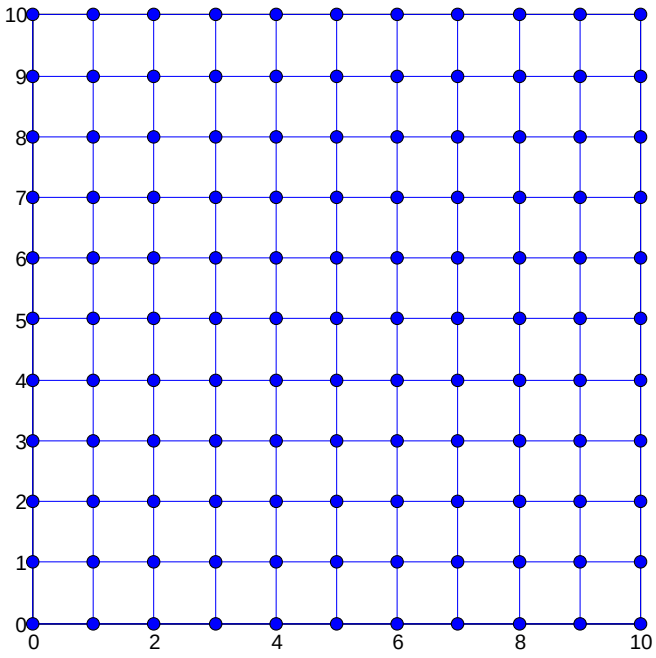
$$\mathbf{Q}_{\mathrm{L}} = \tau_{\mathrm{L}}\left[(1-\lambda)\,\mathbf{I}_n + \lambda\,\mathbf{Q}_{\mathrm{B}}(1)\right], \quad \lambda \in [0,1],$$

$\lambda = \frac{\alpha_1}{\alpha_0 b^2+\alpha_1}$, $\tau_{\mathrm{L}} = \frac{\alpha_0\, b^2+\alpha_1}{b^{d+2}}$, $\mathbf{Q}_{\mathrm{B}}(1)$: Besag precision matrix with $\tau = 1$.

[1] Rue and Held, *Gaussian Markov Random Fields*, Chapman and Hall/CRC, 2005
[2] Allard, Hristopulos, Opitz, *Elec. J. Stat.*, Vol. 15 (2021) 4085–4116

Anisotropic LAP-2 Processes on Square Lattices

$$\mathcal{H}_{LAP2}[z(\mathbf{s})] = \lambda \sum_{n=1}^{N} \left\{ [z(\mathbf{s}_n) - m]^2 + \sum_{i=1}^{d} c_{1,i} \left[\frac{z(\mathbf{s}_n + b\,\hat{\mathbf{e}}_i) - z(\mathbf{s}_n)}{b} \right]^2 + \sum_{i=1}^{d} c_{2,i} \left[\frac{z(\mathbf{s}_n + b\,\hat{\mathbf{e}}_i) - 2z(\mathbf{s}_n) + z(\mathbf{s}_n - b\,\hat{\mathbf{e}}_i)}{b^2} \right]^2 \right\}$$



Lattice LAP-2 energy with anisotropic couplings $c_{1,i}$ and $c_{2,i}$ [a]

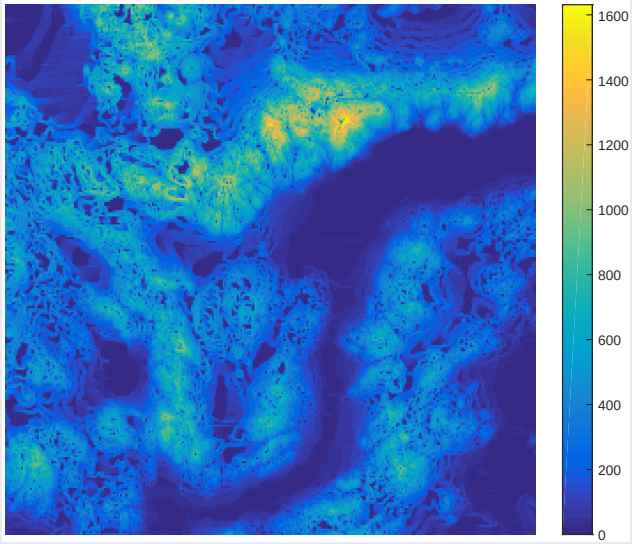
$\{\hat{\mathbf{e}}_i\}_{i=1}^{d}$: unit vectors in orthogonal lattice directions

b : lattice step

[a] Hristopulos. *SIAM J. Sci. Comp.*, 24(6):2125–2162, 2003

Reconstruction of Walker Lake Data (260 × 300 grid)

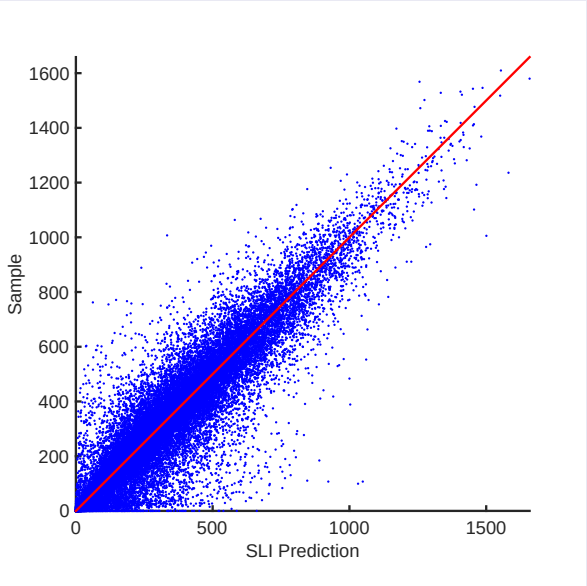
Anisotropic LAP-2 model based on 39 000 sampling points (50%)

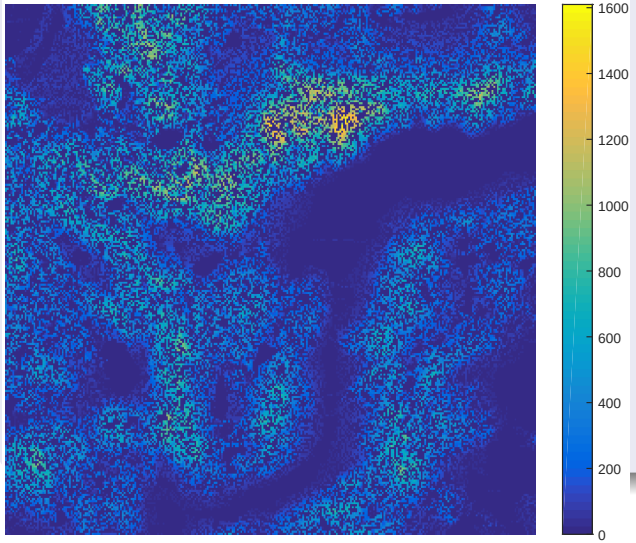


Left top: Full data set

Right top: SLI-based map

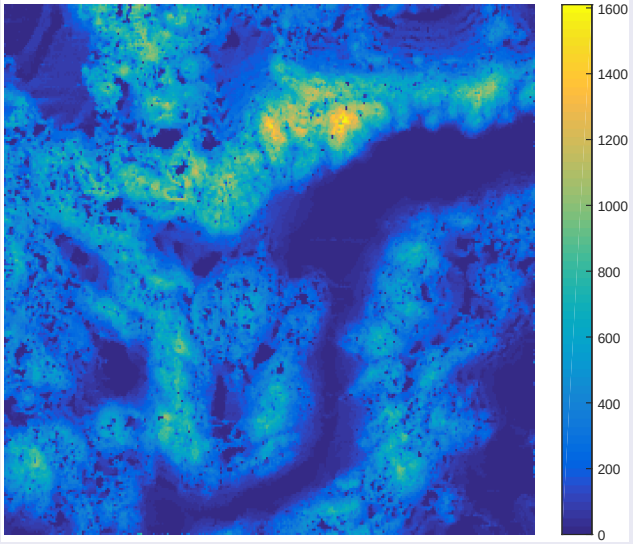
Below: Sample vs predictions





Left bottom: Sample

Right bottom: Prediction error



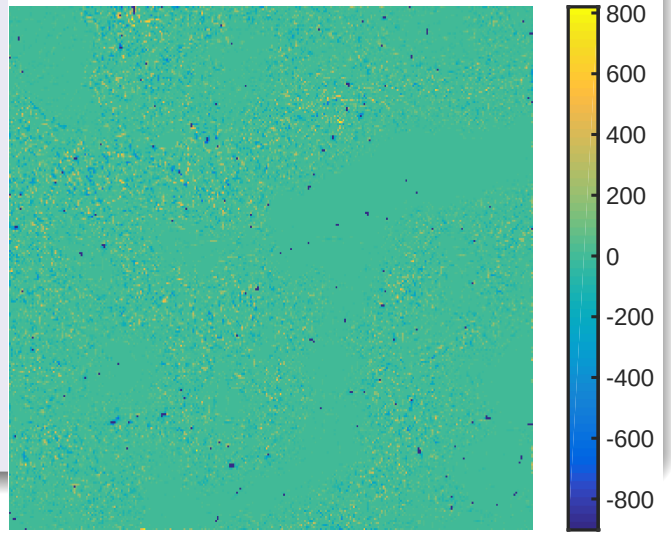
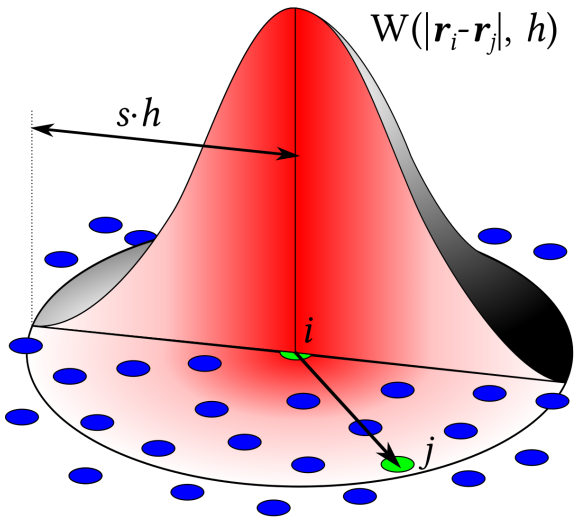


Table of Contents

1. Introduction
2. Boltzmann-Gibbs Gaussian Processes
3. Local Interactions and GMRFs
4. Boltzmann-Gibbs Processes and Smoothed Particle Hydrodynamics
5. Stochastic Local Interaction Model
6. SLI Interpolation & Applications
7. Connections with Deep Learning?
8. Concluding Remarks

Kernels Extend Interactions on Irregular Meshes

- How to extend the precision operator to unstructured grids?
- Smoothed-particle hydrodynamics (SPH)[1]



C.C. Attribution-Share Alike 4.0 International

SPH-field representation

1. $\mathbf{y} = (y_1, \ldots y_N)^\top$: vector of sample values at "particle" locations $\mathbb{S}_N = \{\mathbf{s}_n\}_{n=1}^N$.
2. $K(\cdot, \cdot; h)$: differentiable kernel function with bandwidth h.
3. Differentiable SPH representation:[2]

$$z^*(\mathbf{s}) = \sum_{n=1}^{N} K(\mathbf{s} - \mathbf{s}_n; h)\, y_n \,.$$

[a] R.A. Gingold; J.J. Monaghan (1977). *Mon. Not. R. Astron. Soc.* 181 (3): 375–89; L.B. Lucy (1977). *Astron. J.* 82: 1013–1024

[b] Hristopulos, *Theor. Probability and Math. Statist.* 107 (2022), 37-60

Kernels Extend Interactions on Irregular Meshes

- Then, the LAP-2 energy becomes $\mathcal{H}(\mathbf{y};\boldsymbol{\theta}) = \frac{1}{2}\,\mathbf{y}^\top\,\mathbf{Q}^*\,\mathbf{y}$
- LAP-2 precision matrix

$$\mathbf{Q}^* = \theta_0\boldsymbol{\Phi}_0(h) + \theta_1\boldsymbol{\Phi}_1(h) + \theta_2\boldsymbol{\Phi}_2(h).$$

- The interaction functions $\Phi_{\ell;,n,m}(h)$ between $\mathbf{s}_n$ and $\mathbf{s}_m$ are given by

$$\Phi_{\ell;n,m}(h) = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \mathrm{d}\mathbf{k}\;\mathrm{e}^{\jmath\mathbf{k}\cdot(\mathbf{s}_n-\mathbf{s}_m)}\,\|\mathbf{k}\|^{2\ell}\,\|\tilde{K}(\mathbf{k})\|^2\,,\ \ell = 0,1,2\,.$$

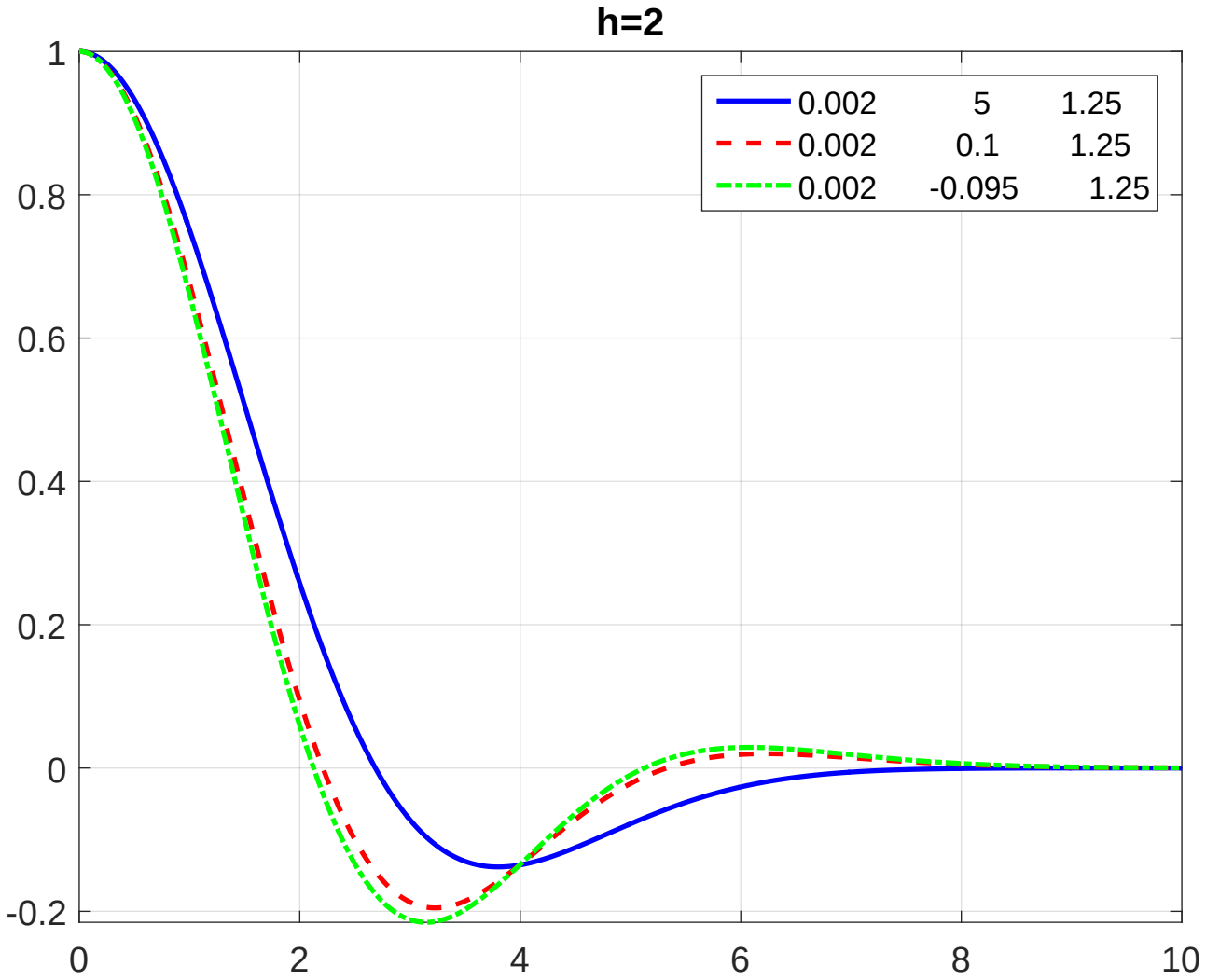
- FT of the LAP-2 precision function *(for Gaussian smoothing kernel)*

$$\tilde{Q}^*(k;\boldsymbol{\theta},h) = \mathrm{e}^{-k^2h^2/2}\,\left(\theta_0 + \theta_1 k^2 + \theta_2 k^4\right),\ \text{where } k = \|\mathbf{k}\|\,,$$

- Real-space SPH-LAP2 Precision function (r is the spatial lag):

$$Q^*(r;\boldsymbol{\theta},h) = \frac{\mathrm{e}^{-r^2/2h^2}}{\left(h\sqrt{2\pi}\right)^d}\left\{\theta_0 - \frac{\theta_1}{h^2}\left(\frac{r^2}{h^2} - d\right) + \frac{\theta_2}{h^4}\left[\frac{r^4}{h^4} - 2(d+2)\frac{r^2}{h^2} + d(d+2)\right]\right\}.$$

Kernels Extend Interactions on Irregular Meshes



SPH-LAP2 precision function $Q^*(r;\boldsymbol{\theta},h)$ for $d=2$ an $h=2$.

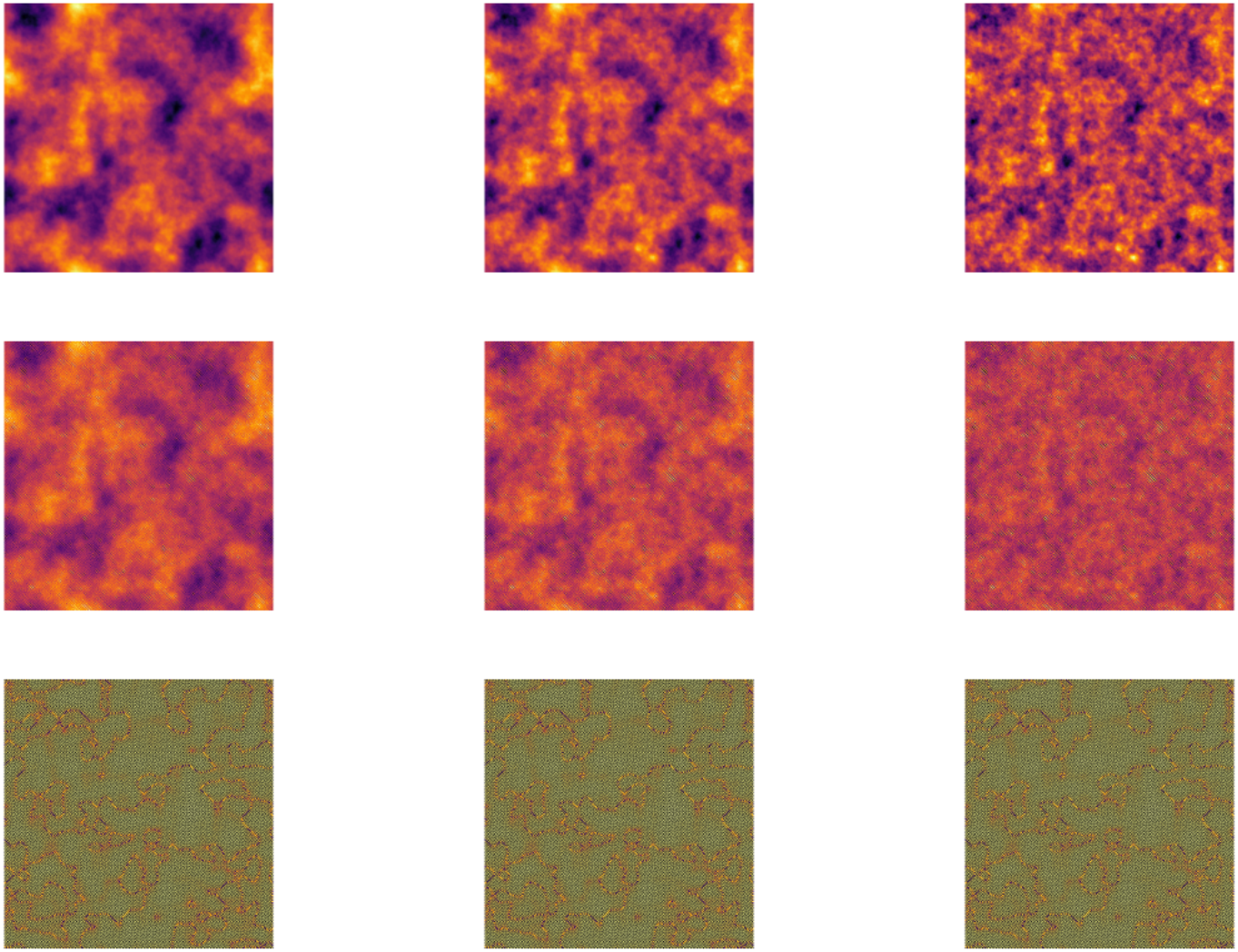
$\boldsymbol{\theta}$: $(2\,10^{-3}, 5, 1.25)$

$\boldsymbol{\theta}$: $(2\,10^{-3}, 0.1, 1.25)$

$\boldsymbol{\theta}$: $(0.002, -0.095, 1.25)$

All curves are normalized by $Q^*(0;\boldsymbol{\theta},h=2)$.

Kernels Extend Interactions on Irregular Meshes



Simulations of SPH-LAP2 on 256×256 grid.

$\boldsymbol{\theta} = \frac{1}{4\pi}(\xi^{-2}, 2, \xi^{2})^{\top}$.

Left: $\xi = 20$

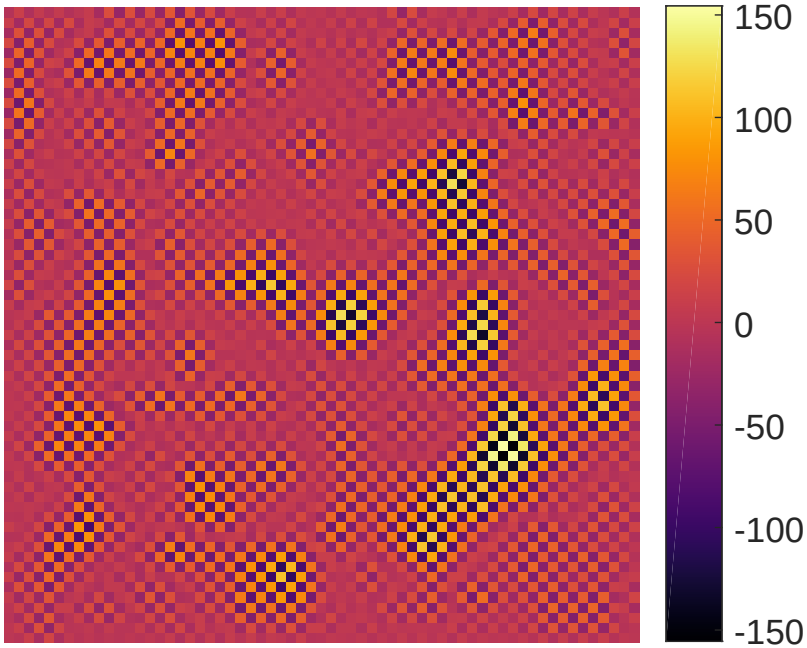
Middle: $\xi = 10$

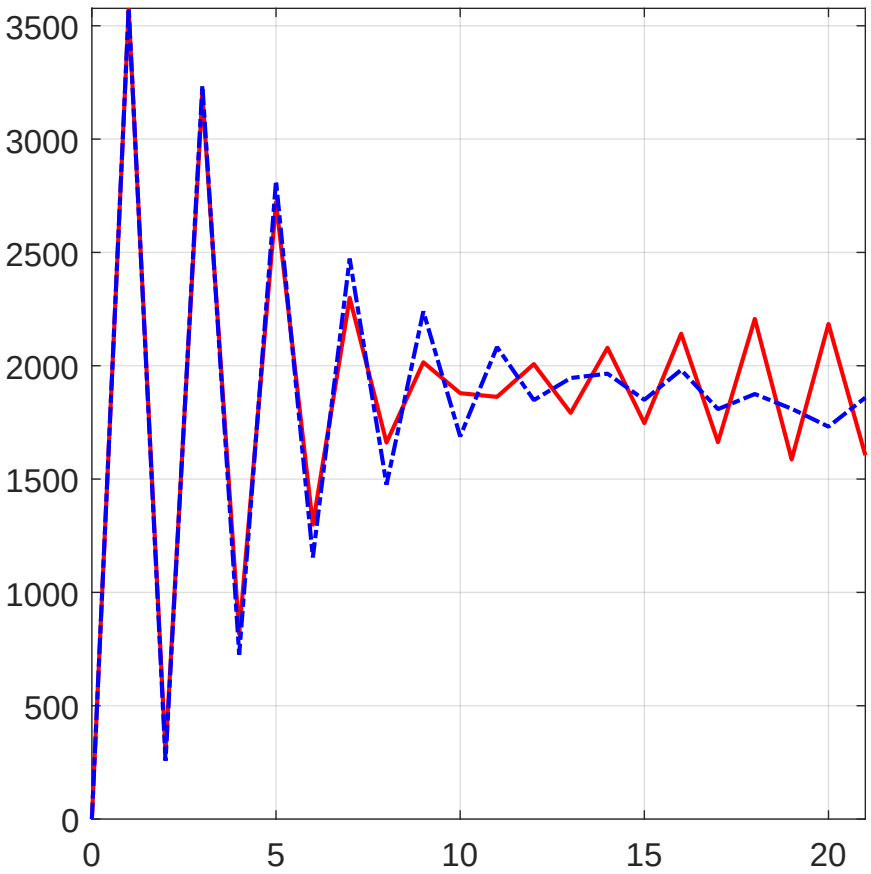
Right: $\xi = 5$

Top: $h = 0.05$

Middle: $h = 0.8$ Bottom: $h = 1.5$

Kernels Extend Interactions on Irregular Meshes





Simulated SPH-LAP2 Gaussian process with Gaussian kernel on 64×64 grid (left) and structure function (right) in orthogonal directions. $\boldsymbol{\theta} = \frac{1}{4\pi}(\xi^{-2}, 2, \xi^{2})^{\top}$ with $\xi = 20$.

Table of Contents

1. Introduction
2. Boltzmann-Gibbs Gaussian Processes
3. Local Interactions and GMRFs
4. Boltzmann-Gibbs Processes and Smoothed Particle Hydrodynamics
5. Stochastic Local Interaction Model
6. SLI Interpolation & Applications
7. Connections with Deep Learning?
8. Concluding Remarks

Stochastic Local Interaction Model (SLI)

- $\mathbf{y} \triangleq (y_1, y_2, \ldots, y_N)^\top$: data at the space-time point set $\{(\mathbf{s}_i, t_i), (\mathbf{s}_i, t_i)\}_{i=1}^N$.
- Boltzmann-Gibbs probability density function (PDF)

$$z(\mathbf{y}; \boldsymbol{\theta}) = Z^{-1}(\boldsymbol{\theta})\, \mathrm{e}^{-\mathcal{H}(\mathbf{y};\boldsymbol{\theta})}\,.$$

- For convenience, regular sampling in time with step δt will be assumed, i.e., $t_n = n\delta t$.
- $\mathcal{H}(\mathbf{y}; \boldsymbol{\theta})$: "energy" (interaction) function.
- $Z(\boldsymbol{\theta})$: Partition function (normalization factor).

SLI versus the "Classical" Model

- Gaussian SLI for the zero-mean case ($\mathbf{y} \in \mathbb{R}^N$, $\mathbf{m} = \mathbf{0}$)

$$\mathcal{H}(\mathbf{y}; \theta) = \frac{1}{2}\mathbf{y}^\top \mathbf{Q}\,\mathbf{y}.$$

- $\mathbf{Q}$ is the "interaction matrix".
- The model is well-defined if $\mathcal{H}(\mathbf{y}; \theta) \geq 0$ for all $\mathbf{y}$.
- Note that if $\mathbf{Q} = \mathbf{C}^{-1}$, where $\mathbf{C}$ is a covariance matrix, the Gaussian SLI has the same form as the standard geostatistical model.

SLI versus the "Classical" Model

- Gaussian SLI for the zero-mean case ($\mathbf{y} \in \mathbb{R}^N$, $\mathbf{m} = \mathbf{0}$)

$$\mathcal{H}(\mathbf{y};\theta) = \frac{1}{2}\mathbf{y}^\top \mathbf{Q}\,\mathbf{y}.$$

- $\mathbf{Q}$ is the "interaction matrix".
- The model is well-defined if $\mathcal{H}(\mathbf{y};\theta) \geq 0$ for all $\mathbf{y}$.
- Note that if $\mathbf{Q} = \mathbf{C}^{-1}$, where $\mathbf{C}$ is a covariance matrix, the Gaussian SLI has the same form as the standard geostatistical model.

SLI versus the "Classical" Model

- Gaussian SLI for the zero-mean case ($\mathbf{y} \in \mathbb{R}^N$, $\mathbf{m} = \mathbf{0}$)

$$\mathcal{H}(\mathbf{y};\theta) = \frac{1}{2}\mathbf{y}^\top \mathbf{Q}\,\mathbf{y}.$$

- $\mathbf{Q}$ is the "interaction matrix".
- The model is well-defined if $\mathcal{H}(\mathbf{y};\boldsymbol{\theta}) \geq 0$ for all $\mathbf{y}$.
- Note that if $\mathbf{Q} = \mathbf{C}^{-1}$, where $\mathbf{C}$ is a covariance matrix, the Gaussian SLI has the same form as the standard geostatistical model.

SLI versus the "Classical" Model

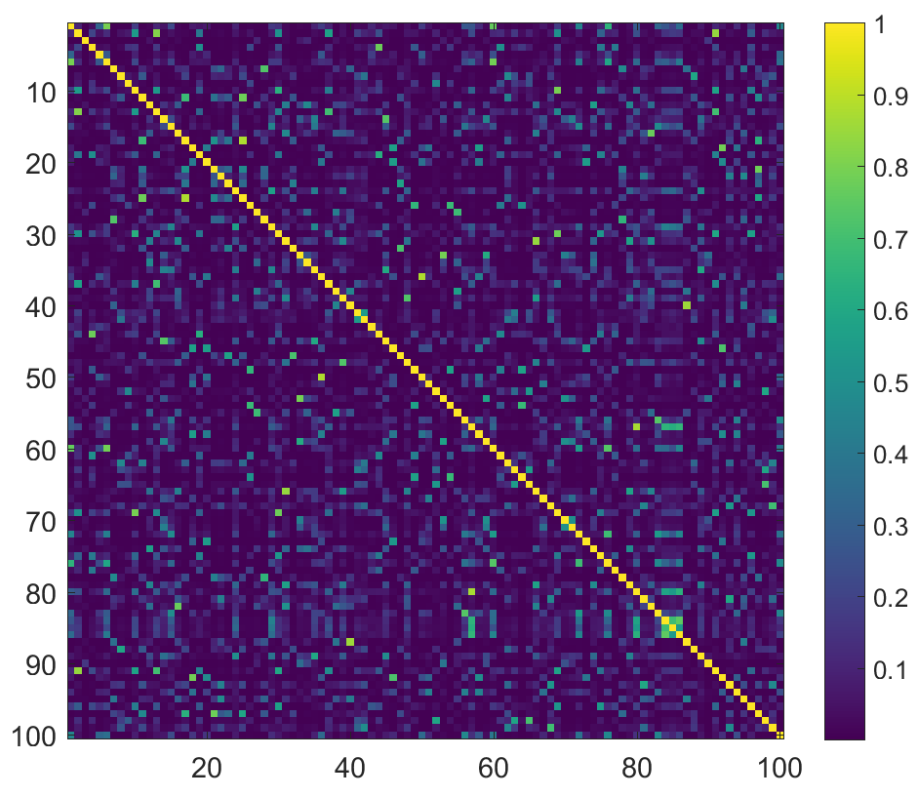
- Gaussian SLI for the zero-mean case ($\mathbf{y} \in \mathbb{R}^N$, $\mathbf{m} = \mathbf{0}$)

$$\mathcal{H}(\mathbf{y};\theta) = \frac{1}{2}\mathbf{y}^\top \mathbf{Q}\,\mathbf{y}.$$

- $\mathbf{Q}$ is the "interaction matrix".
- The model is well-defined if $\mathcal{H}(\mathbf{y};\theta) \geq 0$ for all $\mathbf{y}$.
- Note that if $\mathbf{Q} = \mathbf{C}^{-1}$, where $\mathbf{C}$ is a covariance matrix, the Gaussian SLI has the same form as the standard geostatistical model.

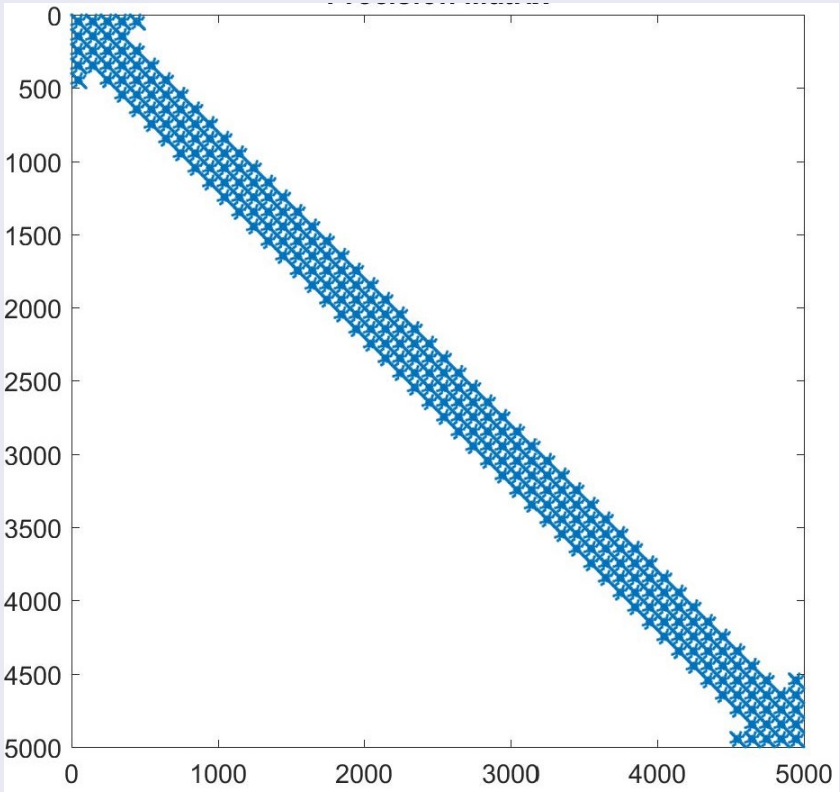
However, there is a difference ...

Precision Matrix of Exponential Covariance



Dense

Sparse Precision Matrix from SLI Model



Sparse

What is a Permissible Gaussian SLI?

- Recall the energy of the Gaussian SLI

$$\mathcal{H}(\mathbf{y};\boldsymbol{\theta}) = \frac{1}{2}\left(\mathbf{y}-\mathbf{m}\right)^{\top}\mathbf{Q}\left(\mathbf{y}-\mathbf{m}\right) = \frac{1}{2}\mathbf{y}'^{\top}\mathbf{Q}\,\mathbf{y}',$$

- where $\mathbf{Q}$ is a sparse, local interaction matrix.
- SLI permissibility: $\mathbf{Q}$ non-negative definite $\equiv$ $\mathbf{C}$ non-negative definite.
- This is ensured if $\mathcal{H}(\mathbf{y};\boldsymbol{\theta}) \geq 0$ for all $\mathbf{y}' \in \mathbb{R}^N$.
- Hence, for Gaussian SLI, permissibility is equivalent to a non-negative energy, which is easy to guarantee by construction.

Defining local space-time neighborhoods

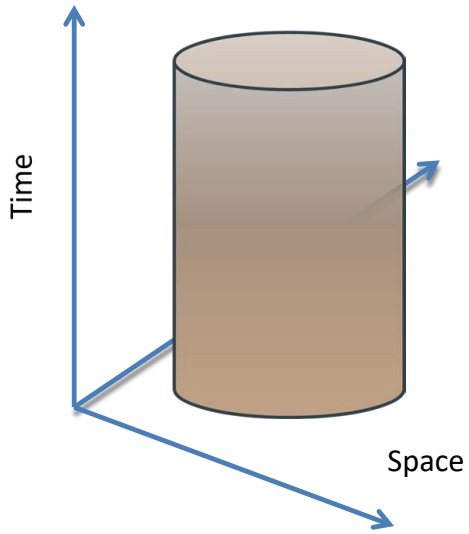
Let $K(\mathbf{u}) = K(\|\mathbf{u}\|)$, where $\mathbf{u} \in \mathbb{R}^{\mathrm{d}}$ be a radial, non-negative, bounded, and compactly supported (or exponentially decaying) kernel function.

Separable space-time structure

$$w_{n,m} = \underbrace{K\left(\frac{\|\mathbf{r}_{n,m}\|}{h_{s,n}}\right) K\left(\frac{|\tau_{n,m}|}{h_{t,n}}\right)}_{\text{Neighborhood selection}} \underbrace{f(\|\mathbf{r}_{n,m}\|)}_{\text{Strength of interaction}}$$

Normalized weights: $\tilde{w}_{n,m} = \dfrac{w_{n,m}}{\sum_{n=1}^{N} \sum_{m=1}^{N} w_{n,m}}$

Key factor: Automate bandwidth estimation to reduce number of parameters



SLI Model for Space-Time Data

- LAP-1-like model of space-time energy function

$$\mathcal{H}(\mathbf{y};\boldsymbol{\theta}) = \frac{1}{2\,\lambda}\left[\frac{(\mathbf{y} - m\,\mathbf{l})^2}{N} + c_1\,\mathcal{S}_1(\mathbf{y};\mathbf{h}_s,\mathbf{h}_t)\right], \quad \lambda, c_1 > 0$$

- $\mathcal{S}_1(\mathbf{y};\mathbf{h}_s,\mathbf{h}_t)$ contains average squared increments based on kernel functions

$$\mathcal{S}_1(\mathbf{y};\mathbf{h}_s,\mathbf{h}_t) = \left\langle\, (y_n - y_k)^2 \,\right\rangle,$$

$$\left\langle\, (y_n - y_k)^2 \,\right\rangle = \sum_{n=1}^{N}\sum_{k=1}^{N} \tilde{w}_{n,k}(\mathbf{r}_{n,k},\tau_{n,k};h_{s,n},h_{t,n})\,(y_n - y_k)^2\,.$$

- Local space bandwidth: $\mathbf{h}_s = (h_{s,1}, h_{s,2}, \ldots, h_{s,n})^\top$.
- Local time bandwidth: $\mathbf{h}_t = (h_{t,1}, h_{t,2}, \ldots, h_{t,n})^\top$.

SLI Model Hyperparameters and Parameters

Hyperparameters (selected by user)

- Number (and type) of space-time trend basis functions
- Number of harmonics and respective periods
- Order of nearest neighbors in space and time (e.g., k_s, k_t)
- Kernel function(s) used (e.g., Epanechnikov, spherical, etc.)

Parameters (estimated with Maximum Likelihood)

- Trend and seasonality coefficients, e.g., $m_n = \sum_{\ell=1}^{L} b_\ell \, f_\ell(\mathbf{s}_n, t_n)$
- *Short-range space-time correlations* (e.g., λ, c_1, μ_s , μ_t, ...)
- *Parameters determining correlations of seasonal components*

Kernel-based weights

The weights $w_{n,m}, \; n, m = 1, \ldots N$ are determined from the kernel function and k_s, k_t, μ_s , μ_t.

Table of Contents

1. Introduction
2. Boltzmann-Gibbs Gaussian Processes
3. Local Interactions and GMRFs
4. Boltzmann-Gibbs Processes and Smoothed Particle Hydrodynamics
5. Stochastic Local Interaction Model
6. SLI Interpolation & Applications
7. Connections with Deep Learning?
8. Concluding Remarks

SLI Mode Predictor for Interpolation of Missing Data

- The prime ($'$) indicates a fluctuation around m
- Energy functional with *unobserved point* $\mathbf{x}_p$:

$$\mathcal{H}(\mathbf{y}\,, z_p; \boldsymbol{\theta}) = \frac{1}{2} \begin{bmatrix} \mathbf{y}'^{\top} & z_p' \end{bmatrix} \begin{bmatrix} \mathbf{Q}_{S,S} & \mathbf{Q}_{S,P} \\ \mathbf{Q}_{P,S} & Q_{p,p} \end{bmatrix} \begin{bmatrix} \mathbf{y}' \\ z_p' \end{bmatrix}$$

- Predictive mean: $\hat{z}_p = \arg\min_{z_p} \hat{\mathcal{H}}(\mathbf{y}, z_p; \boldsymbol{\theta}^*)$
- The predictor is a linear equation that can be evaluated in $\mathcal{O}(N)$ time:

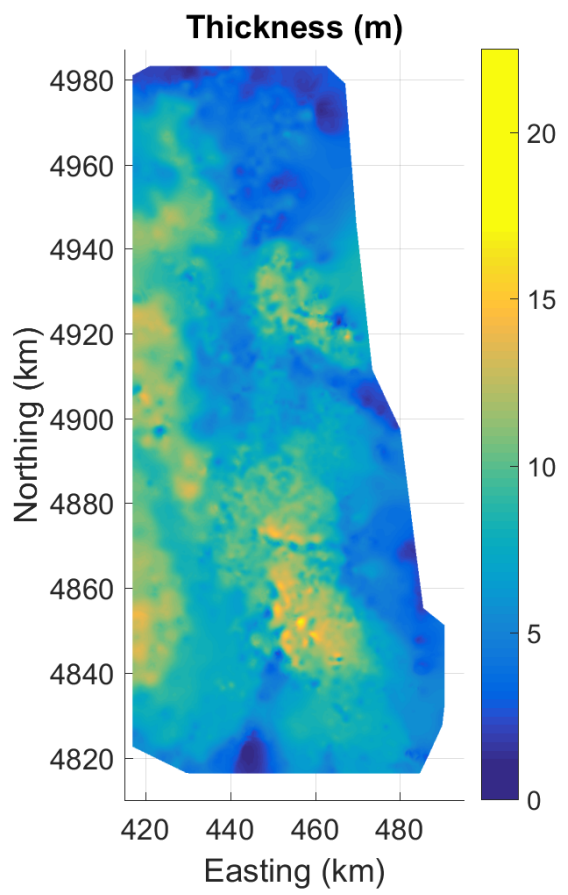
$$\hat{z}_p = m - \frac{1}{Q_{p,p}(\boldsymbol{\theta}^*)} \sum_{n=1}^{N} Q_{p,n}(\boldsymbol{\theta}^*)\,(y_n - m)$$

- Predictive variance: $\hat{\sigma}_p^2 = \frac{1}{Q_{p,p}}$.

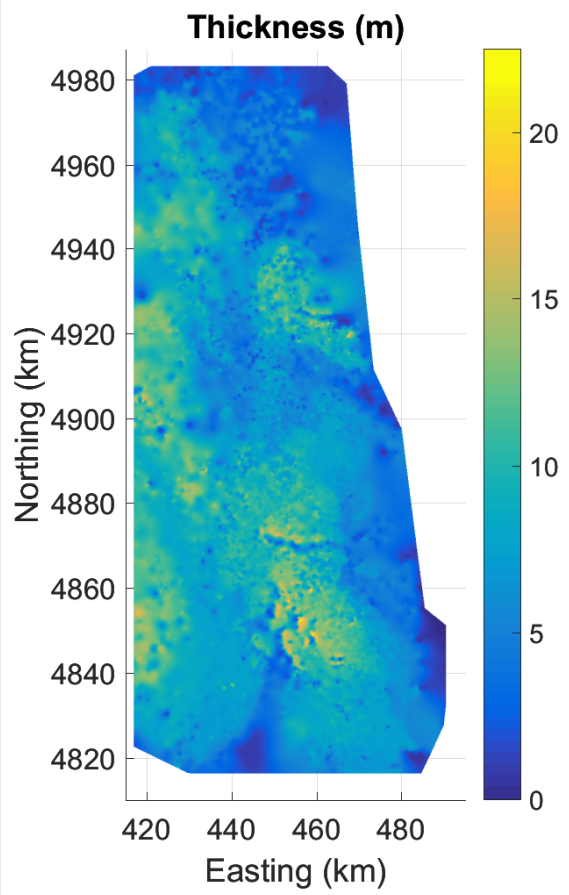
Hristopulos, 2015, Hristopulos and Agou, 2019

SLI-Kriging Comparison: Coal Data ($\approx$ 11K samples)

SLI Map



Ordinary Kriging Map



Map grid: 653×1388 cells (size $\approx 115 \times 135\text{m}^2$); Estimated reserves: 74.1×10^9 m^3

SLI-Kriging Comparison: Coal Data ($\approx$ 11K samples)

Table: LOO-CV measures of coal thickness based on (i) ordinary kriging with the spherical variogram model and (ii) SLI with spherical kernel and neighbor order $K_s = 3$. MAE: mean absolute error; MaxAE: Maximum absolute error; MARE: mean absolute relative error; RMSE: root mean square error; RMSRE: root mean square relative error; R: Pearson's correlation coefficient.

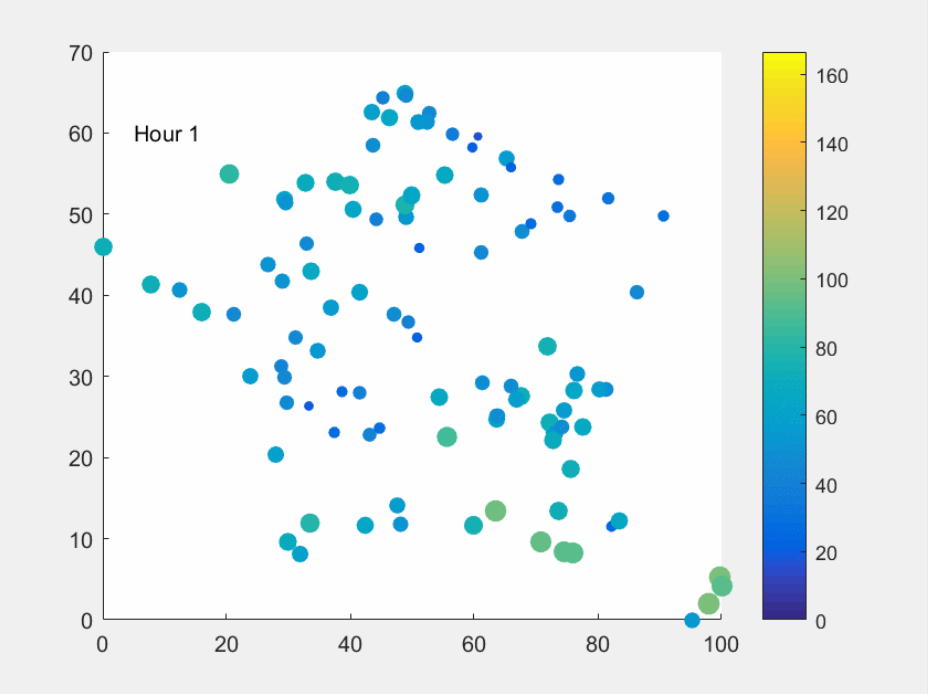
Method	ME (m)	MAE (m)	MARE	RMSE (m)	RMSRE	MaxAE (m)	R(%)
OK (9.1 km)	4.0224×10^{-4}	1.1788	0.2480	1.8067	0.8312	16.7286	80.27%
OK (14.6 km)	4.0573×10^{-4}	1.1787	0.2480	1.8066	0.8313	16.7290	80.27%
SLI ($K_s = 3$)	5.6420×10^{-4}	1.1199	0.2480	1.7033	0.8527	14.8781	82.23%

OK search radius of $r_s = 14.6$ km is used to estimate at the map grid nodes. The SLI predictor is $\approx 3 - 25$ times faster than OK ($\approx$ 12 min with SLI versus $\approx$ 42 min for OK with $r_s = 9.1$ km and $\approx$ 5 hr with $r_s = 14.6$ km).

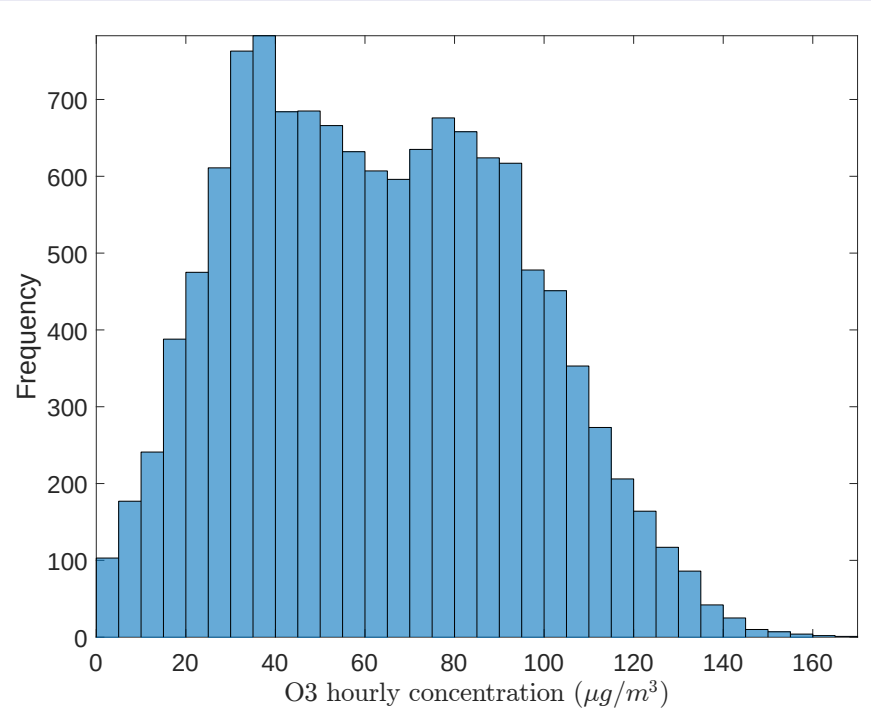
Hristopulos et al., *Math. Geosci.*, 53, 1907–1949 (2021).

Space-time Interpolation of Ozone Data

Scatter plot of ST data



Histogram of ozone data

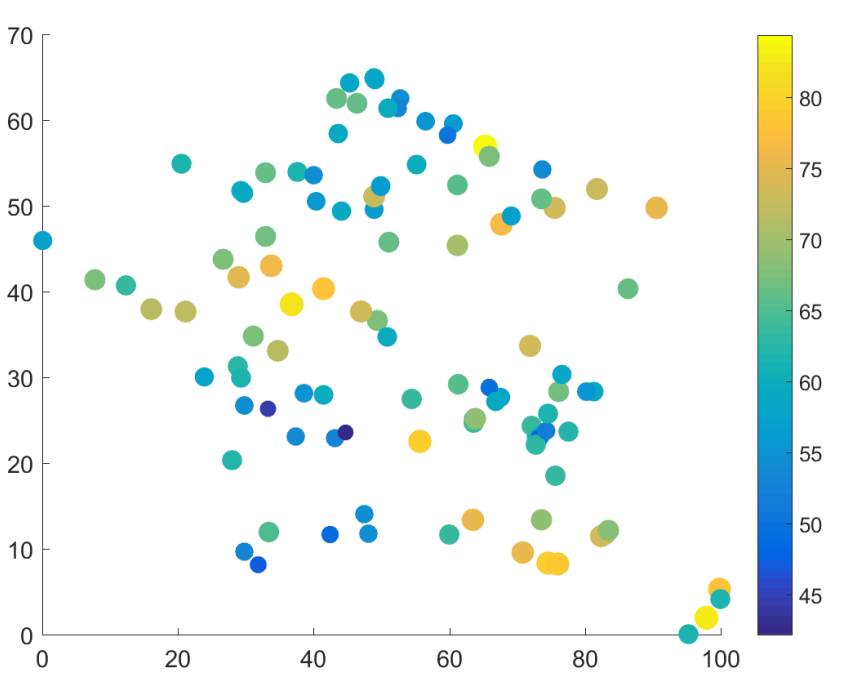


Data Summary: 5 days (120 hours), 107 stations, 12,840 sampling points.

Data downloaded from the French GEOD'AIR database: www.prevair.org.

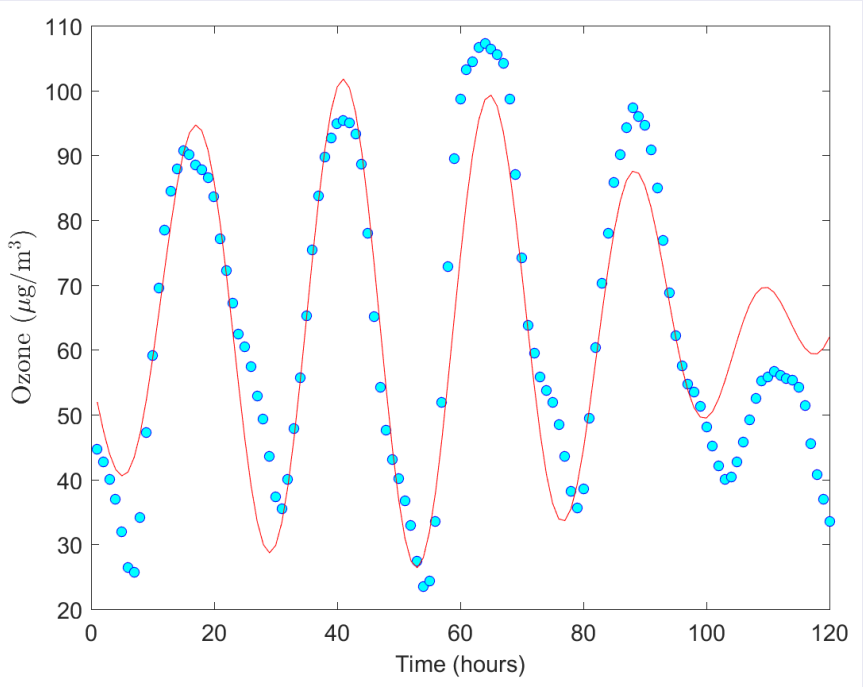
Space-time Interpolation of Ozone Data

Temporal average



No discernible spatial trend.

Spatial average

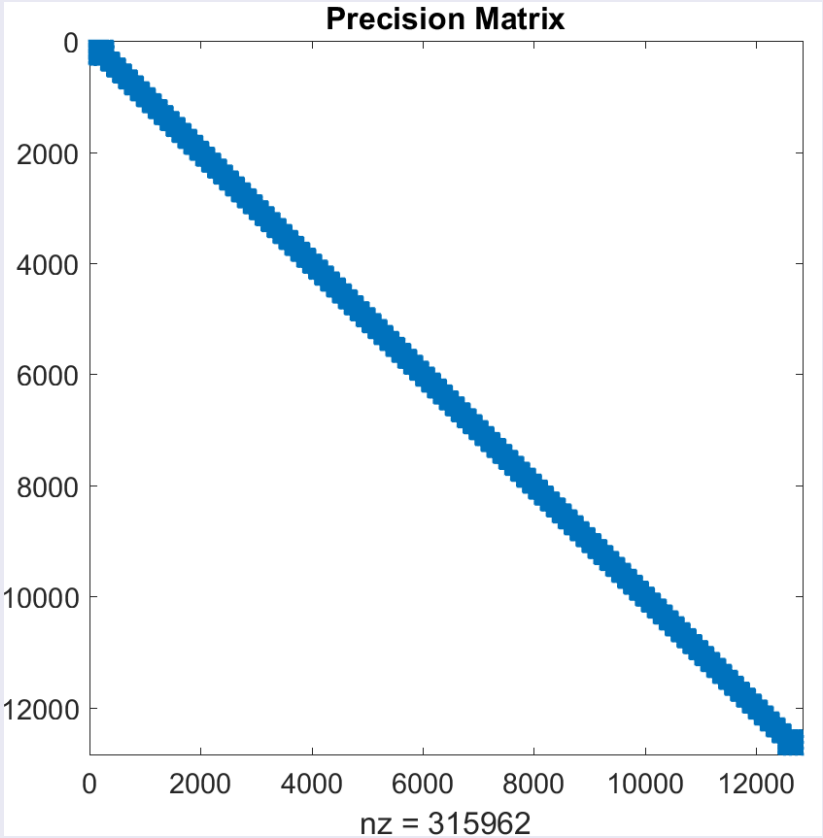


Clear daily periodicity.

$$m(t) = b_1 + \left(b_2 + b_3 t + b_4 t^2\right) \cos\left(2\pi t/24\right) + \left(b_5 + b_6 t + b_7 t^2\right) \sin\left(2\pi t/24\right)$$

Space-time Interpolation of Ozone Data

SLI Precision matrix **Q** obtained by MLE



How sparse is it?

$$\text{Sparsity index} = \frac{n(Q \neq 0)}{N} \approx 0.19\%.$$

Computational speed

MLE: $\approx$ 10 minutes (12840 data points).

Mapping time: $\approx$ 20 minutes for $200 \times 200 \times 120$ grid (4.8×10^6 points).

Cross-Validation Results

Table: Leave one-slice out cross validation using SLI with Epanechnikov spatial and temporal kernels with $K_t = 3$ and two K_s values.

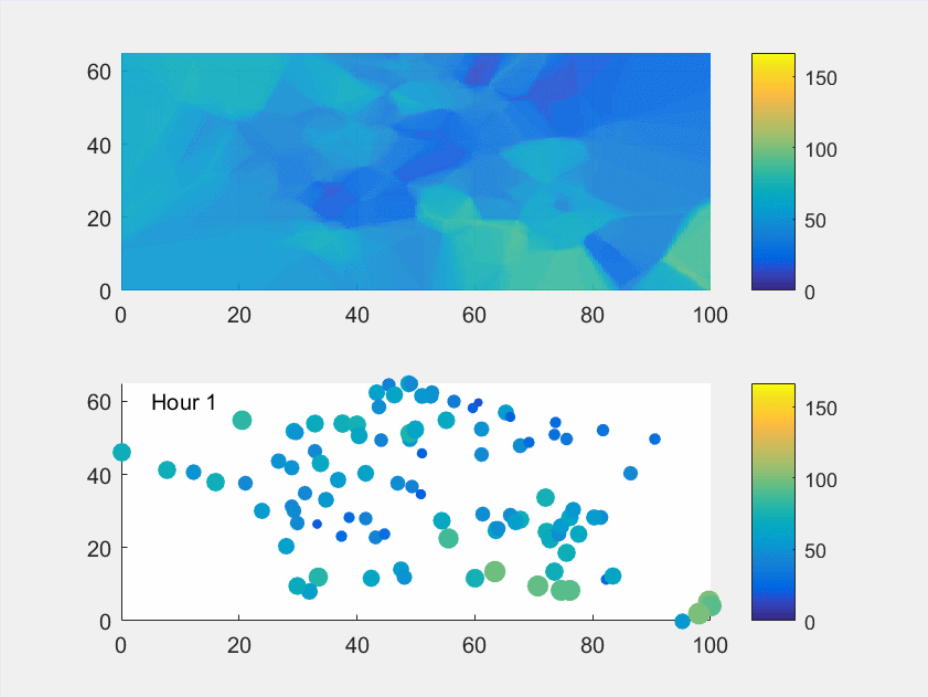
ME: mean error; MAE: mean absolute error; RMSE: root mean square error; R: Pearson's correlation coefficient.

K_s	ME	MAE	RMSE	R
3	0.251	8.54	11.19	0.94
8	0.018	11.40	14.53	0.89

Space-Time Maps with SLI

SLI-generated maps with Epanechnikov kernels on $200 \times 200 \times 120$ space-time grid: Fast!

$K_s = K_t = 3$



$K_s = 8, K_t = 3$

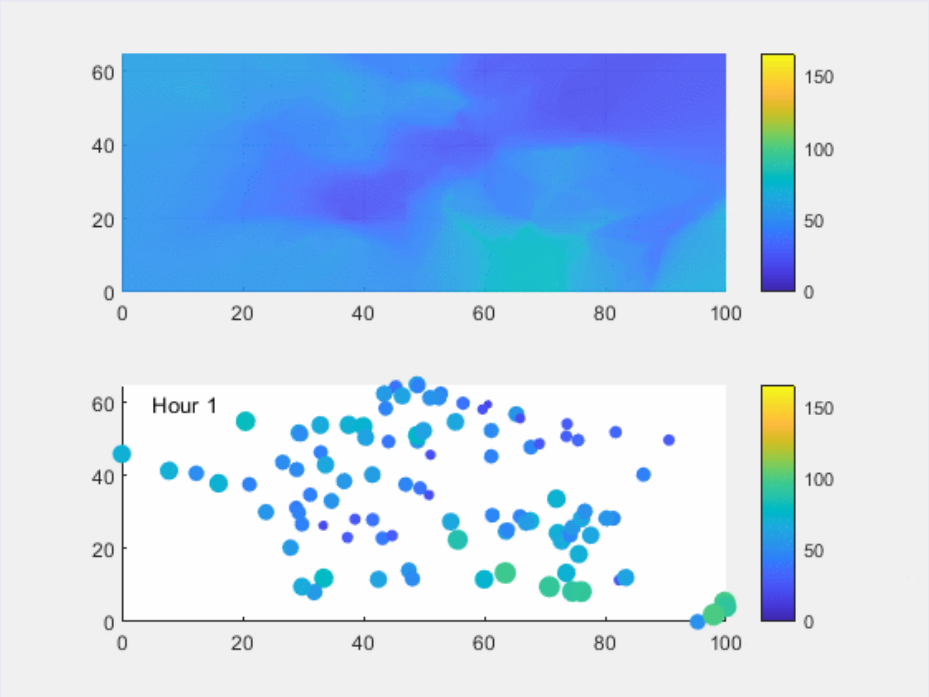


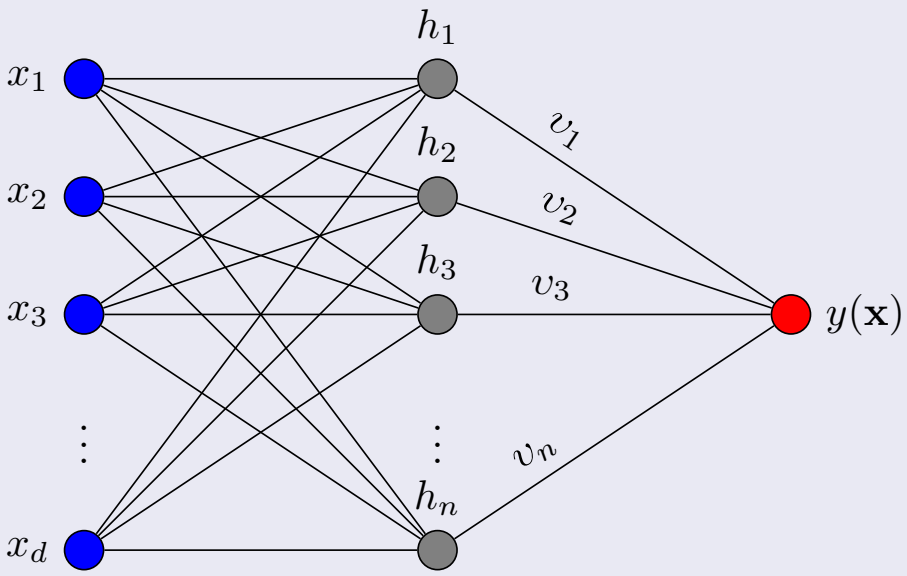
Table of Contents

1. Introduction
2. Boltzmann-Gibbs Gaussian Processes
3. Local Interactions and GMRFs
4. Boltzmann-Gibbs Processes and Smoothed Particle Hydrodynamics
5. Stochastic Local Interaction Model
6. SLI Interpolation & Applications
7. Connections with Deep Learning?
8. Concluding Remarks

Connections with Deep Learning?

Infinitely wide, single-layer ANN ⇔ Gaussian process

- Single-layer, feed-forward Bayesian artificial neural network ANN
- Infinite number of hidden units (i.e., infinitely wide ANN).
- *Independent, identically distributed* priors over the parameters.
- These ANN are equivalent to GPs.
- The kernel of the equivalent GP can be obtained in closed form (neural network covariance).

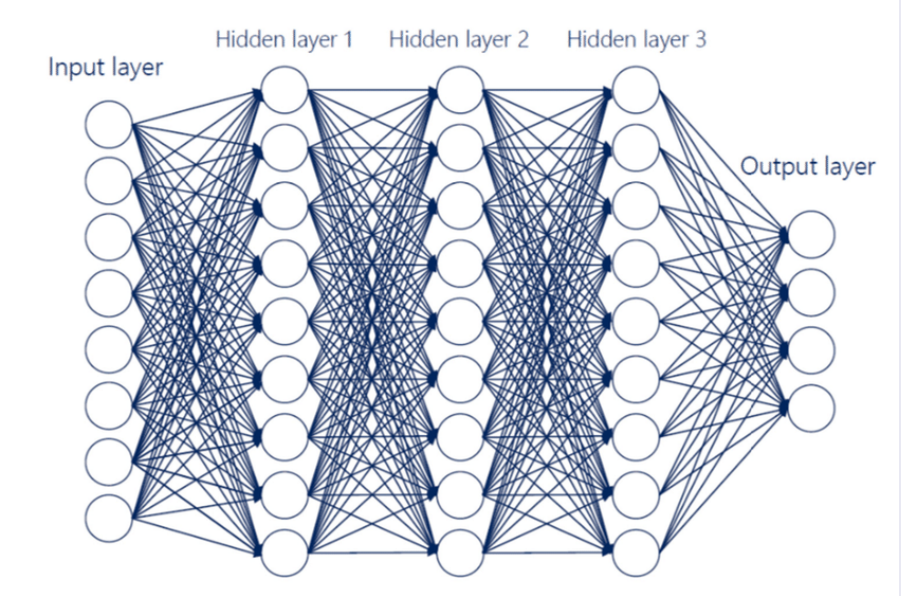


Neal R. M., *Bayesian Learning for Neural Networks*, Vol. 118, Springer, 1996.

Connections with Deep Learning?

Deep, infinitely wide ANN $\Leftrightarrow$ Gaussian process

- The GP kernel of wide ANNs with finite number of layers is obtained computationally. Lee, J., *et al.*, 2017. Deep neural networks as Gaussian processes. arXiv:1711.00165.
- The output of a (residual) convolutional NN in the limit of infinite depth is equivalent to a GP (using an appropriate prior over the weights and biases).
- The equivalent kernel can be computed exactly.
- Deep, infinitely wide NN are also equivalent to GPs.



Garriga-Alonso, A., Rasmussen, C.E. and Aitchison, L., 2018. Deep convolutional networks as shallow Gaussian processes. arXiv:1808.05587

Table of Contents

1. Introduction
2. Boltzmann-Gibbs Gaussian Processes
3. Local Interactions and GMRFs
4. Boltzmann-Gibbs Processes and Smoothed Particle Hydrodynamics
5. Stochastic Local Interaction Model
6. SLI Interpolation & Applications
7. Connections with Deep Learning?
8. Concluding Remarks

Concluding Remarks

- Boltzmann-Gibss Gaussian processes with local interactions are powerful tools for space-time analysis of geo-referenced data.
- Local precision operators: small and beautiful.
- Kernel functions extend interactions to scattered grids: SLI and SPH formalisms.
- SLI provides flexible and fast space-time interpolation based on designed precision (interaction) matrices.
- The SLI model is well-defined for any dimensionality of input space and distance metric.

Concluding Remarks

Future Research

- Investigate interaction strength functions to mitigate kernel boundary effects.
- Construct kernels that use different distance measures (e.g., Lagrangian distances, tele-connections, etc.).
- Develop multivariate extensions of the SLI model.
- Derive new spatiotemporal covariance kernels with interpretable parameters.
- Connections with deep learning NNs.

For more information ...

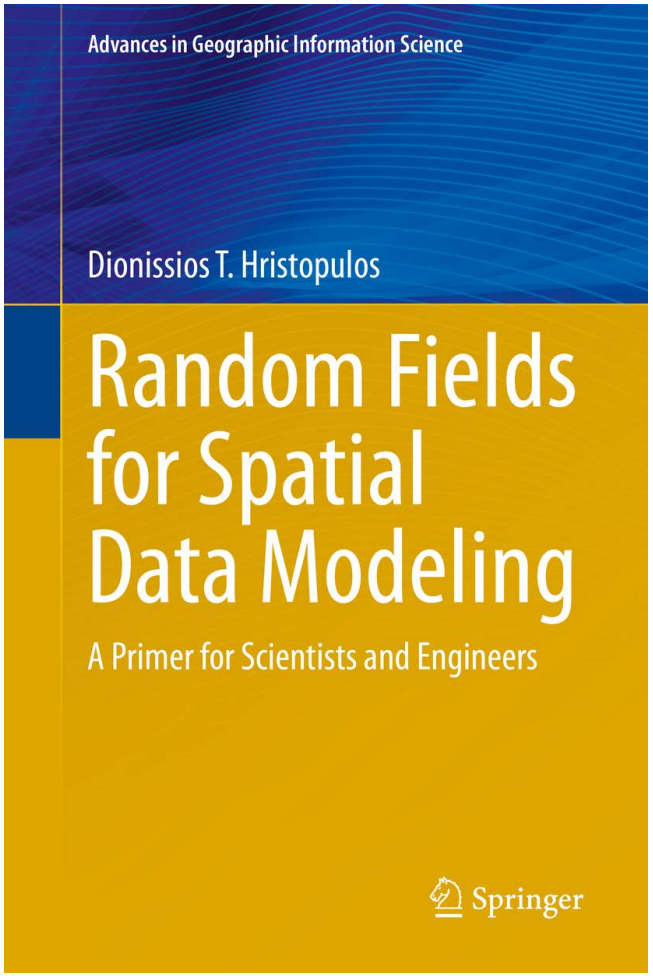
D. T. Hristopulos (2015). “Stochastic Local Interaction (SLI) model: Bridging machine learning and geostatistics.” *Computers & Geosciences* 85, 26-37

D. T. Hristopulos and V. Agou (2020). “Stochastic local interaction model with sparse precision matrix for space-time interpolation,” *Spatial Statistics*, Volume 40, 100403 (2020). 10.1016/j.spasta.2019.100403

D. T. Hristopulos, “Random Fields for Spatial Data Modeling: A Primer for Scientists and Engineers,” *Springer*, Dordrecht (2020), 867pp. https://link.springer.com/book/10.1007/978-94-024-1918-4

Hristopulos, D.T., Pavlides, A., Agou, V.D., and Gkafa, P., “Stochastic Local Interaction Model: An Alternative to Kriging for Massive Datasets.” *Mathematical Geosciences*, 53, 1907–1949 (2021). 10.1007/s11004-021-09957-7

Hristopulos, D.T., “Boltzmann–Gibbs Random Fields with Mesh-free Precision Operators Based on Smoothed Particle Hydrodynamics.” *Theor. Probability and Math. Statist.* 107, 37-60 (2022). 10.1090/tpms/1180



SLI Precision Matrix Formulation

Quadratic Form of SLI Energy Functional [HA19]

$$\mathcal{H}(\mathbf{y};\theta) = \frac{1}{2}(\mathbf{y} - m\mathbf{l})^\top \mathbf{Q}(\theta')\,(\mathbf{y} - m\mathbf{l}) \tag{1}$$

Precision matrix (sparse, explicit) using only the first two terms

$$\mathbf{Q}(\theta') = \frac{1}{\lambda}\left\{\frac{\mathbf{I}}{N} + c_1\,\mathbf{Q}_1(\mathbf{h}_s,\mathbf{h}_t)\right\},\ \theta' = (\lambda, c_1, \mu_s, \mu_t)^\top \tag{2}$$

"Gradient" Precision sub-matrix based on normalized-weights

$$[\mathbf{Q}_1(\mathbf{h}_s,\mathbf{h}_t)]_{i,j} = -\tilde{w}_{i,j}(\mathbf{h}_s,\mathbf{h}_t) - \tilde{w}_{j,i}(\mathbf{h}_s,\mathbf{h}_t) + \delta_{i,j}\sum_{n=1}^{N}\left[\tilde{w}_{i,n}(\mathbf{h}_s,\mathbf{h}_t) + \tilde{w}_{n,i}(\mathbf{h}_s,\mathbf{h}_t)\right] \tag{3}$$

Normalized kernel weights: $\tilde{w}_{i,j}(\mathbf{h}_s,\mathbf{h}_t)$

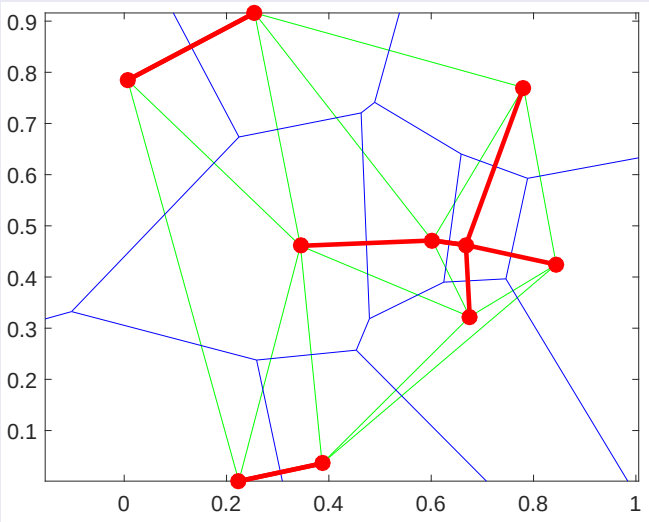
Kernel Bandwidth Selection & Parameter Estimation

Bandwidth estimation algorithm

1. $h_{s,n} = \mu_s \; D_{n,[k]}(\mathbb{S}_N)$
2. $D_{n,[k]}(\mathbb{S}_N)$: distance between $\mathbf{s}_n$ and its k-nearest neighbor in $\mathbb{S}_N$
3. $\mu_s > 1$: global bandwidth parameter inferred by MLE
4. $k \in \{2, 3, 4, 5\}$: Neighbor-order hyper-parameter

SLI weights are fully determined by selecting hyper-parameters (kernel function, k_s, k_t) and ML estimation of parameters μ_s, μ_t

Nearest-neighbor distances



Same idea for temporal neighbors:
$h_{t,n} = \mu_t \, (k_t - 1) \, \delta t$, where δt is the time step

Weight Matrix for Separable S-T Kernel Structure

Sampling pattern: N_s locations and N_t times

$$\mathbf{K}_s = \begin{bmatrix} K\left(\frac{\|\mathbf{r}_{1,1}\|}{h_{s,1}}\right) & \dots & K\left(\frac{\|\mathbf{r}_{1,N_s}\|}{h_{s,1}}\right) \\ \vdots & \vdots & \vdots \\ K\left(\frac{\|\mathbf{r}_{N_s,1}\|}{h_{s,N_s}}\right) & \dots & K\left(\frac{\|\mathbf{r}_{N_s,N_s}\|}{h_{s,N_s}}\right) \end{bmatrix}, \mathbf{K}_t = \begin{bmatrix} K\left(\frac{\|\tau_{1,1}\|}{h_{t,1}}\right) & \dots & K\left(\frac{\|\tau_{1,N_t}\|}{h_{t,1}}\right) \\ \vdots & \vdots & \vdots \\ K\left(\frac{\|\tau_{N_t,1}\|}{h_{t,N_t}}\right) & \dots & K\left(\frac{\|\tau_{N_t,N_t}\|}{h_{t,N_t}}\right) \end{bmatrix}$$

Kernel weight matrix: $\mathbf{W} = \mathbf{K}_s \otimes \mathbf{K}_t$. For *non-negative-valued* kernel functions the *normalized-weight matrix* is

$$\widetilde{\mathbf{W}} = \frac{\mathbf{W}}{\|\mathbf{W}\|_1}, \tag{4}$$

where $\boxed{\|\mathbf{W}\|_1 = \sum_{k=1}^{N} \sum_{l=1}^{N} |W_{k,l}| \text{ is the entry-wise } L_1 \text{ norm.}}$

Squared Increment for Separable Space-Time Kernels

The kernel-averaged square gradient is given by

$$\langle (y_n - y_k)^2 \rangle = \left\| \left[(\mathbf{y} \otimes \mathbf{1}) - (\mathbf{y} \otimes \mathbf{1})^\top \right] \circ \mathbf{U} \circ \left[(\mathbf{y} \otimes \mathbf{1}) - (\mathbf{y} \otimes \mathbf{1})^\top \right] \right\|_1$$

where $\mathbf{1} = (1, \ldots, 1)^\top$ is the $N \times 1$ vector of ones and $\circ$ denotes the *Hadamard product*: $[\mathbf{A} \circ \mathbf{B}]_{i,j} = A_{i,j} B_{i,j}$.

The computational complexity of the above operations is $\mathcal{O}(N^2)$. However, more efficient algorithms can be designed based on the fact that $\mathbf{U}$ is a *sparse matrix*.

SLI Likelihood

- SLI negative log-likelihood

$$\mathrm{NLL} = \frac{1}{2}(\mathbf{y} - m\mathbf{l})^\top \mathbf{Q}(\boldsymbol{\theta}')\,(\mathbf{y} - m\mathbf{l}) - \frac{1}{2}\ln|\mathbf{Q}(\boldsymbol{\theta}')| + \frac{N}{2}\ln 2\pi. \tag{5}$$

- Simplify: $\mathbf{Q}(\boldsymbol{\theta}') = \mathbf{Q}_r(\boldsymbol{\theta}_1)/\lambda$, where $\boldsymbol{\theta}_1 = (c_1, \mu_s, \mu_t)^\top$
- NLL with explicit scale dependence (drop last term)

$$\mathrm{NLL} = \frac{1}{2\lambda}(\mathbf{y} - m\mathbf{l})^\top \mathbf{Q}_r(\boldsymbol{\theta}_1)\,(\mathbf{y} - m\mathbf{l}) - \frac{1}{2}\ln|\mathbf{Q}_r(\boldsymbol{\theta}_1)| + \frac{N}{2}\ln\lambda \tag{6}$$

- Minimize with respect to λ:

$$\lambda^* = \frac{(\mathbf{y} - m\mathbf{l})^\top \mathbf{Q}_r(\boldsymbol{\theta}_1)\,(\mathbf{y} - m\mathbf{l})}{N} \tag{7}$$

SLI Likelihood

- "New" NLL after replacing λ with λ^* (constant term dropped):

$$\mathrm{NLL} = \frac{N}{2} \ln \left[\frac{(\mathbf{y} - m\mathbf{I})^\top \mathbf{Q}_r(\boldsymbol{\theta}_1)\,(\mathbf{y} - m\mathbf{I})}{N} \right] - \frac{1}{2} \ln |\mathbf{Q}_r(\boldsymbol{\theta}_1)| \tag{8}$$

- $\mathbf{Q}_r(\boldsymbol{\theta}_1)$ is the sparse precision matrix obtained from (2) and (3) by setting $\lambda = 1$
- The NLL function should be minimized with respect to m, c_1, μ_s, μ_t under the constraint that c_1, μ_s, μ_t are positive:

$$(m^*, c_1^*, \mu_s^*, \mu_t^*) = \arg\min_{\mathrm{m,c_1,\mu_s,\mu_t}} \mathrm{NLL} \tag{9}$$

- After optimal values for these parameters are determined, they are inserted in (7) to obtain λ^*.

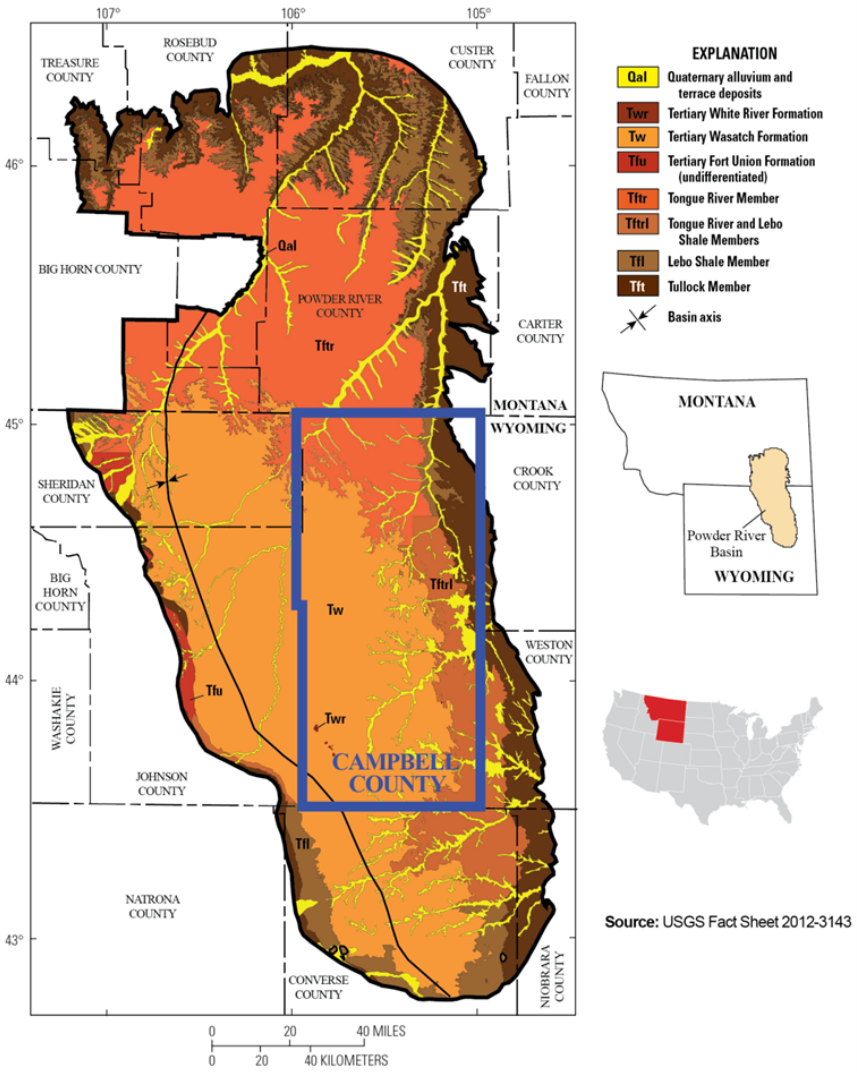
Steps for parameter estimation

1. Set the order of space and time near neighbours, i.e., $k_s = k_t = 3$ and select a kernel function (e.g., quadratic kernel).
2. Based on k_s and k_t find the respective near neighbor distances $D_{n,[k_s]}(\mathbb{S}_N)$ for each point in the sample set $\mathbb{S}_N$.
3. Initialize the parameters c_1, m, μ_s, μ_t: e.g., $\mu_s^{(0)}, \mu_t^{(0)} = 3$, $m^{(0)} = \bar{x}$ (sample average), $c_1 = 100$
4. Use iterative constrained optimization method to optimize the NLL given by Eq. (8)
   - The bandwidths $\{h_{s,n}, h_{t,n}\}_{n=1}^{N}$ at each iteration i are determined according to $h_{s,n}^{(i)} = \mu_s^{(i)} D_{n,[k_s]}(\mathbb{S}_N)$ and $h_{t,n}^{(i)} = \mu_t^{(i)} (k_t - 1)\, \delta t$ (for regular sampling in time)
   - The precision matrix $\mathbf{Q}_r(\boldsymbol{\theta}_1)$, where $\boldsymbol{\theta}_1 = (c_1, \mu_s, \mu_t)^\top$ is calculated based on Eqs. (3) (with $\lambda = 1$) and (4).
5. Question: Is the global optimum $c_1 = 0$? What does this mean for the model?

Study Area in Powder River Basin

Features of dataset

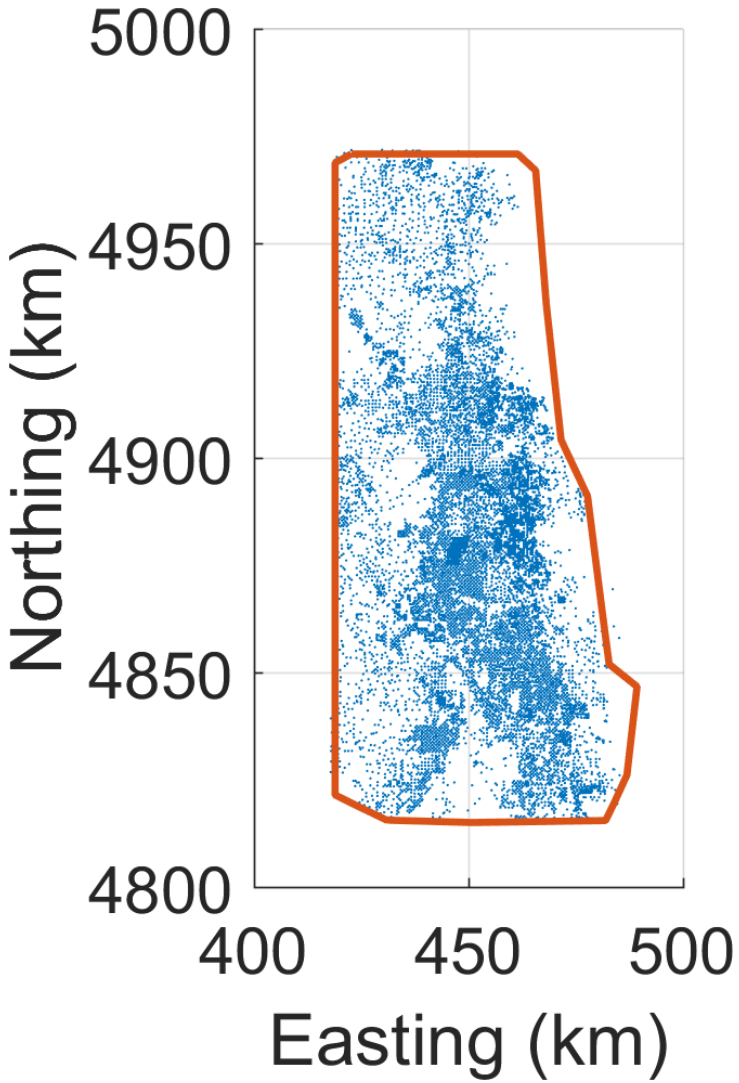
- Study area: Campbell county in the Wyoming side of the Powder River Basin.
- Area: $9,500\ \text{km}^2$
- Data: 11,416 drill-holes (including the two largest coal mines in the world)
- Coal Beds: Roland Coal to Canyon Coal, i.e., the upper part of the Fort Union Formation



Study Area in Powder River Basin

Features of dataset

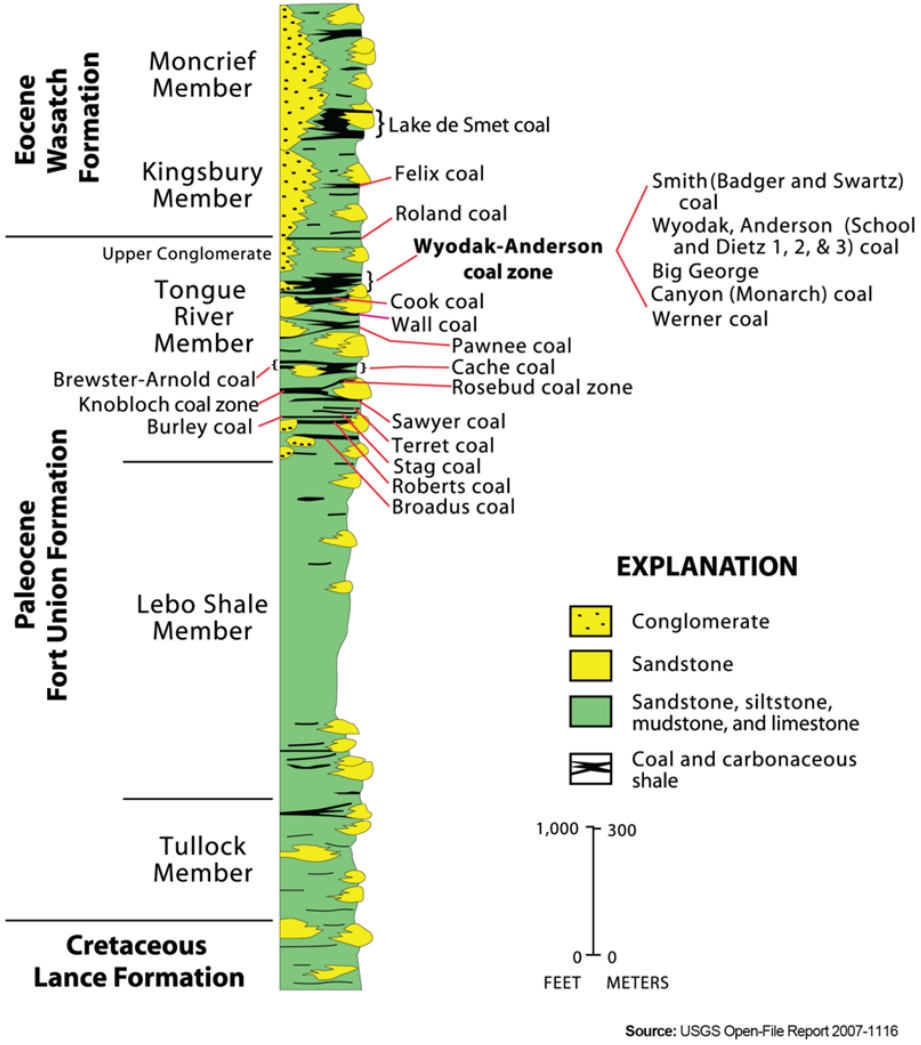
- Study area: Campbell county in the Wyoming side of the Powder River Basin.
- Area: $9,500$ km^2
- Data: 11,416 drill-holes (including the two largest coal mines in the world)
- Coal Beds: Roland Coal to Canyon Coal, i.e., the upper part of the Fort Union Formation



Study Area in Powder River Basin

Features of dataset

- Study area: Campbell county in the Wyoming side of the Powder River Basin.
- Area: $9,500\ km^2$
- Data: 11,416 drill-holes (including the two largest coal mines in the world)
- Coal Beds: Roland Coal to Canyon Coal, i.e., the upper part of the Fort Union Formation



Exploratory Data Analysis

Mean (m)	Median (m)	Max (m)	Min (m)	STD (m)	Skewness
7.83	7.71	22.67	0.33	2.99	0.13

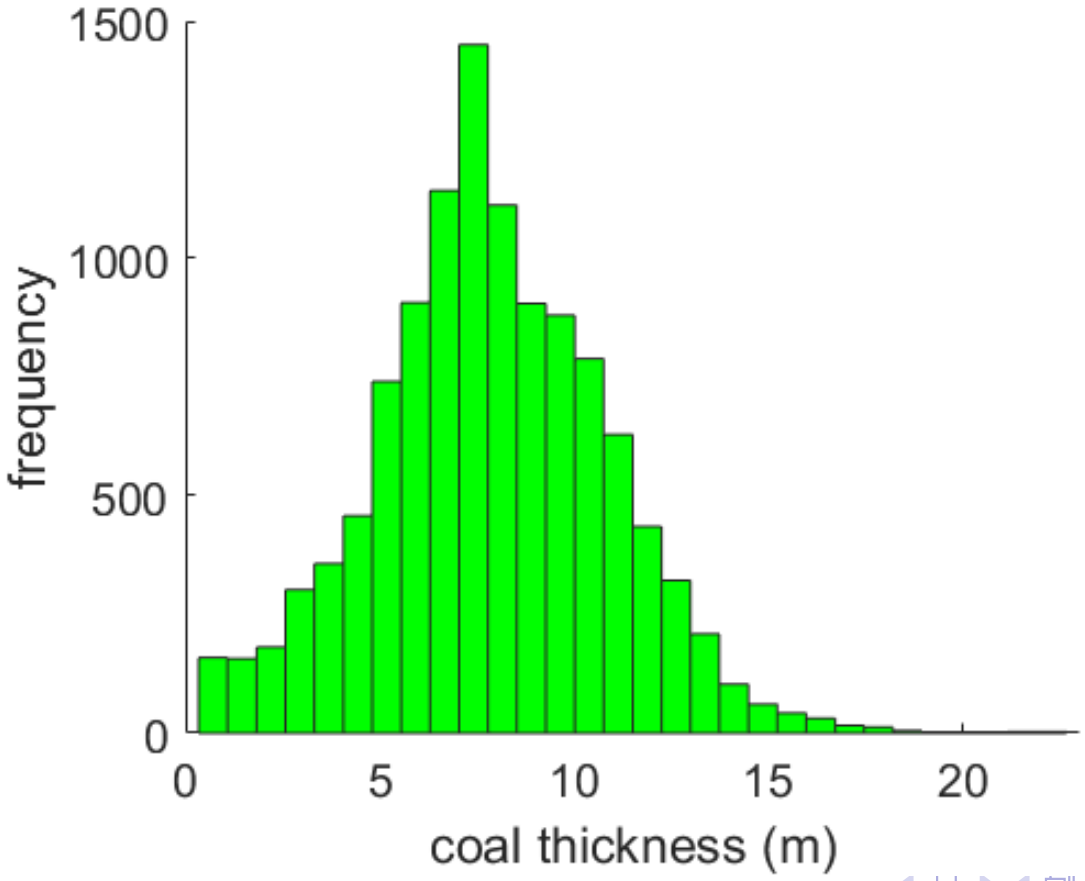



Table: Statistics of the coal thickness data at the 11,416 drill-hole locations