Computing transport maps by cumulant matching

Stephen Huan

Carnegie Mellon University
https://cgdct.moe

2024-09-17 Simons lightning talk

Overview

Fast inference in Gaussian processes

Towards general preconditioners

Transport by cumulant matching

Part 1: Fast inference in Gaussian processes

Collaborators



Joe Guinness,
Cornell



Matthias Katzfuß,
Texas A&M



Houman Owhadi,
Caltech



Florian Schäfer,
Gatech

The problem

Computations with large psd matrix Θ given entries $\Theta_{i,j}$

Ubiquitous in computational mathematics

- Kernel trick in machine learning
- Statistical inference in Gaussian processes on $\boldsymbol{y} \sim \mathcal{N}(\mathbf{0}, \Theta)$
- Second-order optimization, Newton's method

The problem

Computations with large psd matrix Θ given entries $\Theta_{i,j}$

Ubiquitous in computational mathematics

- Kernel trick in machine learning
- Statistical inference in Gaussian processes on $\boldsymbol{y} \sim \mathcal{N}(\mathbf{0}, \Theta)$
- Second-order optimization, Newton's method

Direct computation scales as $\mathcal{O}(N^3)$, limiting data size (10^4)

The problem

Computations with large psd matrix Θ given entries $\Theta_{i,j}$

Ubiquitous in computational mathematics

- Kernel trick in machine learning
- Statistical inference in Gaussian processes on $\boldsymbol{y} \sim \mathcal{N}(\mathbf{0}, \Theta)$
- Second-order optimization, Newton's method

Direct computation scales as $\mathcal{O}(N^3)$, limiting data size (10^4)

Seek *sparse* Cholesky factor for *dense* covariance matrix

Statistical Cholesky factorization

Cholesky factorization $\Leftrightarrow$ iterative conditioning of process

$$\mathbb{E}[\boldsymbol{y}_2 \mid \boldsymbol{y}_1] = \boldsymbol{\mu}_2 + \Theta_{2,1}\Theta_{1,1}^{-1}(\boldsymbol{y}_{\mathsf{Tr}} - \boldsymbol{\mu}_{\mathsf{Tr}})$$

$$\mathbb{C}\mathrm{ov}[\boldsymbol{y}_2 \mid \boldsymbol{y}_1] = \Theta_{2,2} - \Theta_{2,1}\Theta_{1,1}^{-1}\Theta_{1,2}$$

$$\mathrm{chol}(\Theta) = \begin{pmatrix} \mathsf{Id} & 0 \\ \Theta_{2,1}\Theta_{1,1}^{-1} & \mathsf{Id} \end{pmatrix} \begin{pmatrix} \mathrm{chol}(\Theta_{1,1}) & 0 \\ 0 & \mathrm{chol}(\Theta_{2,2} - \Theta_{2,1}\Theta_{1,1}^{-1}\Theta_{1,2}) \end{pmatrix}$$

Statistical Cholesky factorization

Cholesky factorization $\Leftrightarrow$ iterative conditioning of process

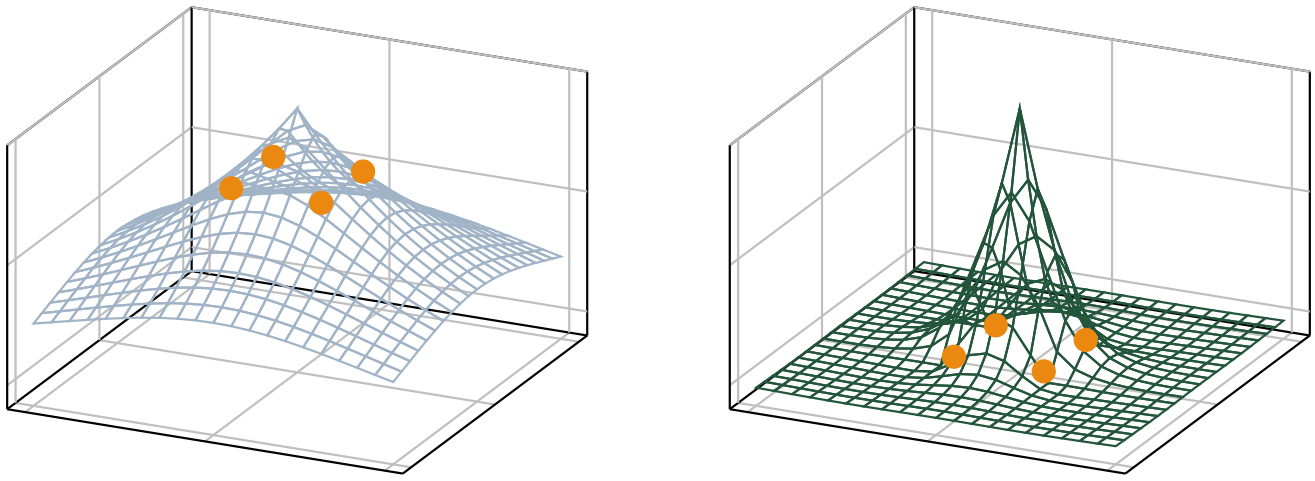
$$\mathbb{E}[\boldsymbol{y}_2 \mid \boldsymbol{y}_1] = \boldsymbol{\mu}_2 + \Theta_{2,1}\Theta_{1,1}^{-1}(\boldsymbol{y}_{\mathsf{Tr}} - \boldsymbol{\mu}_{\mathsf{Tr}})$$

$$\mathbb{C}\mathrm{ov}[\boldsymbol{y}_2 \mid \boldsymbol{y}_1] = \Theta_{2,2} - \Theta_{2,1}\Theta_{1,1}^{-1}\Theta_{1,2}$$

$$\mathrm{chol}(\Theta) = \begin{pmatrix} \mathsf{Id} & 0 \\ \Theta_{2,1}\Theta_{1,1}^{-1} & \mathsf{Id} \end{pmatrix} \begin{pmatrix} \mathrm{chol}(\Theta_{1,1}) & 0 \\ 0 & \mathrm{chol}(\Theta_{2,2} - \Theta_{2,1}\Theta_{1,1}^{-1}\Theta_{1,2}) \end{pmatrix}$$

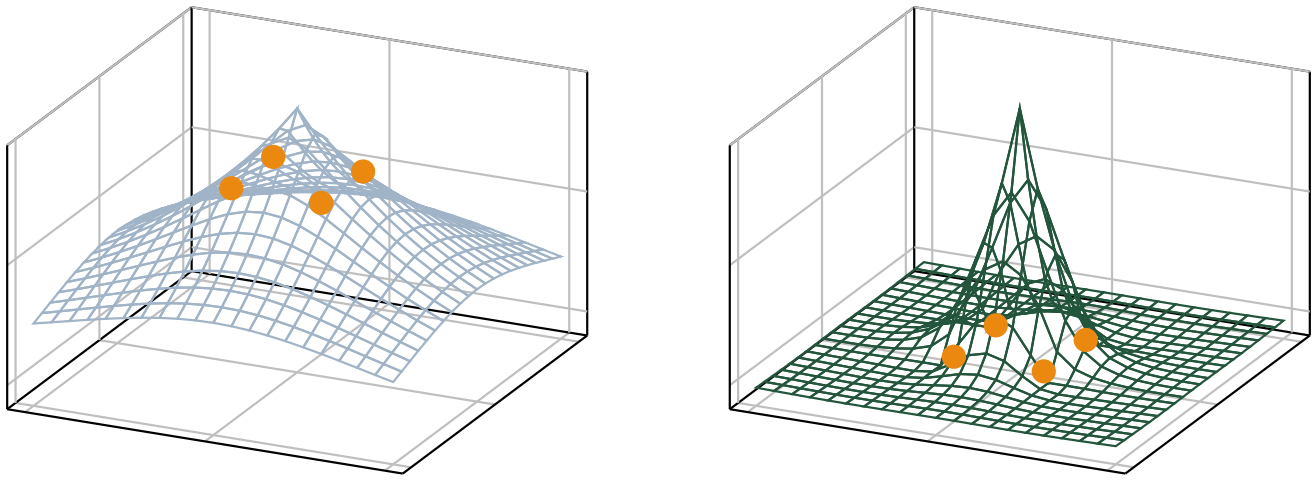
Conditional (near)-independence $\Leftrightarrow$ (approximate) sparsity

Screening effect



Conditional on points near a point of interest,
far away points are almost independent [Stein 2002]

Screening effect



Conditional on points near a point of interest,
far away points are almost independent [Stein 2002]

Suggests space-covering ordering [Guinness 2018] and
selecting nearby points for the sparsity pattern

Kullback-Leibler minimization

Compute entries by minimizing Kullback-Leibler divergence

$$L \coloneqq \underset{\hat{L} \in \mathcal{S}}{\operatorname{argmin}} \, \mathbb{D}_{\mathrm{KL}}\Big(\mathcal{N}(\mathbf{0}, \Theta) \,\Big\|\, \mathcal{N}(\mathbf{0}, (\hat{L}\hat{L}^{\top})^{-1})\Big)$$

Kullback-Leibler minimization

Compute entries by minimizing Kullback-Leibler divergence

$$L \coloneqq \underset{\hat{L} \in \mathcal{S}}{\operatorname{argmin}} \, \mathbb{D}_{\mathrm{KL}}\left(\mathcal{N}(\mathbf{0}, \Theta) \,\middle\|\, \mathcal{N}(\mathbf{0}, (\hat{L}\hat{L}^{\top})^{-1})\right)$$

Efficient and embarrassingly parallel closed-form solution

$$L_{s_i,i} = \frac{\Theta^{-1}_{s_i,s_i} \boldsymbol{e}_1}{\sqrt{\boldsymbol{e}_1^{\top} \Theta^{-1}_{s_i,s_i} \boldsymbol{e}_1}}$$

Kullback-Leibler minimization

Compute entries by minimizing Kullback-Leibler divergence

$$L \coloneqq \underset{\hat{L}\in\mathcal{S}}{\operatorname{argmin}}\ \mathbb{D}_{\mathrm{KL}}\left(\mathcal{N}(\mathbf{0},\Theta)\ \middle\|\ \mathcal{N}(\mathbf{0},(\hat{L}\hat{L}^{\top})^{-1})\right)$$

Efficient and embarrassingly parallel closed-form solution

$$L_{s_i,i} = \frac{\Theta^{-1}_{s_i,s_i}\boldsymbol{e}_1}{\sqrt{\boldsymbol{e}_1^{\top}\Theta^{-1}_{s_i,s_i}\boldsymbol{e}_1}}$$

Achieves state of the art ε-accuracy in time complexity $\mathcal{O}\left(N\log^{2d}\left(\frac{N}{\varepsilon}\right)\right)$ with $\mathcal{O}\left(N\log^{d}\left(\frac{N}{\varepsilon}\right)\right)$ nonzero entries [Schäfer, Katzfuss, and Owhadi 2021]

KL-minimization, revisited

Plug optimal L back into the KL divergence

$$\mathbb{D}_{\mathrm{KL}}\left(\Theta \,\middle\|\, (LL^\top)^{-1}\right) = \sum_{i=1}^{N} \left[\log\left(\Theta_{i,i|s_i\setminus\{i\}}\right) - \log\left(\Theta_{i,i|i+1:}\right)\right]$$

KL-minimization, revisited

Plug optimal L back into the KL divergence

$$\mathbb{D}_{\mathrm{KL}}\left(\Theta \,\middle\|\, (LL^\top)^{-1}\right) = \sum_{i=1}^{N} \left[\log\left(\Theta_{i,i|s_i\setminus\{i\}}\right) - \log\left(\Theta_{i,i|i+1:}\right)\right]$$

KL is accumulated error over independent regression problems

KL-minimization, revisited

Plug optimal L back into the KL divergence

$$\mathbb{D}_{\mathrm{KL}}\left(\Theta \,\middle\|\, (LL^\top)^{-1}\right) = \sum_{i=1}^{N} \left[\log\left(\Theta_{i,i|s_i\setminus\{i\}}\right) - \log\left(\Theta_{i,i|i+1:}\right)\right]$$

KL is accumulated error over independent regression problems

Goal: minimize posterior variance of i-th prediction point by selecting training points s_i *most informative* to that point

KL-minimization, revisited

Plug optimal L back into the KL divergence

$$\mathbb{D}_{\mathrm{KL}}\left(\Theta \,\middle\|\, (LL^\top)^{-1}\right) = \sum_{i=1}^{N} \left[\log\left(\Theta_{i,i|s_i\setminus\{i\}}\right) - \log\left(\Theta_{i,i|i+1:}\right)\right]$$

KL is accumulated error over independent regression problems

Goal: minimize posterior variance of i-th prediction point by selecting training points s_i *most informative* to that point

Variance $\Leftrightarrow$ mutual information $\Leftrightarrow$ mean squared error

$$\mathbb{H}[\boldsymbol{y}_{\mathsf{Pr}}] = \mathbb{H}[\boldsymbol{y}_{\mathsf{Pr}} \mid \boldsymbol{y}_{\mathsf{Tr}}] + \mathbb{I}[\boldsymbol{y}_{\mathsf{Pr}}; \boldsymbol{y}_{\mathsf{Tr}}]$$
$$\mathbb{V}\mathrm{ar}[\boldsymbol{y}_{\mathsf{Pr}}] = \mathbb{E}[\mathbb{V}\mathrm{ar}[\boldsymbol{y}_{\mathsf{Pr}} \mid \boldsymbol{y}_{\mathsf{Tr}}]] + \mathbb{V}\mathrm{ar}[\mathbb{E}[\boldsymbol{y}_{\mathsf{Pr}} \mid \boldsymbol{y}_{\mathsf{Tr}}]]$$

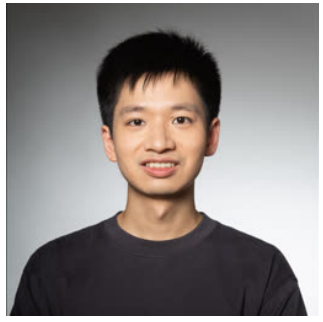
https://cgdct.moe/projects/cholesky/

Part 2: Towards general preconditioners

Collaborators



Greg Lavrentiadis,
Caltech



Yifan Chen,
Caltech



Pau Batlle,
Caltech



Ethan Epperly,
Caltech



Houman Owhadi,
Caltech



Florian Schäfer,
Gatech

Sparse versus low rank

Sparsity doesn't scale to high dimension

RPCholesky [Chen et al. 2023] for Nyström approximation

"All high dimensional data is low rank"

Best-of-both-worlds approximation?

Automated orderings and sparsity pattern

Faster algorithms for optimal transport

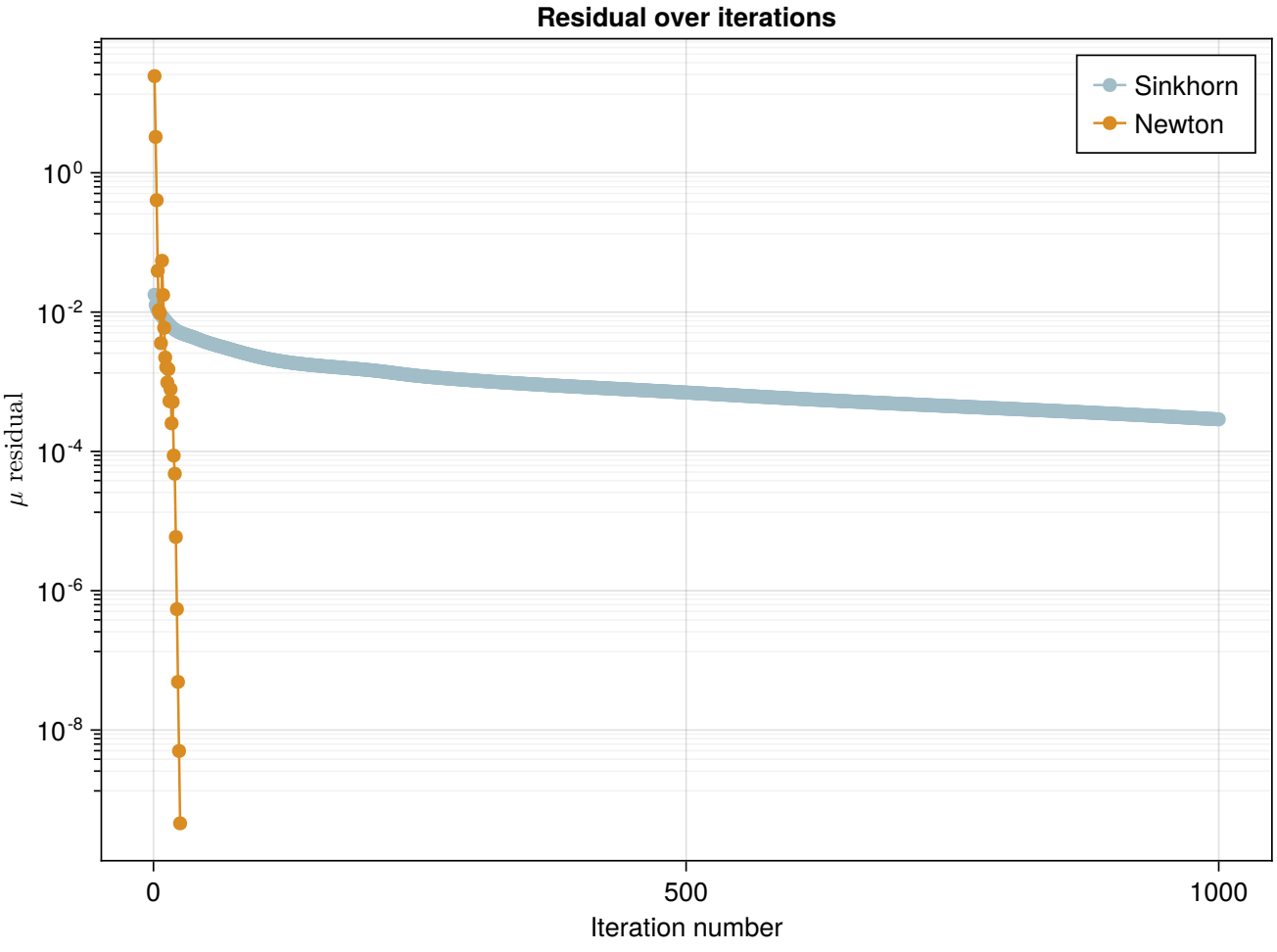


Figure: Newton's method versus Sinkhorn method.

https://kolesky.cgdct.moe/

Part 3: Transport by cumulant matching

Collaborators



Andreas Robertson,
Sandia



Youssef Marzouk,
MIT



Florian Schäfer,
Gatech

Transport maps

Random variable X with density π (possibly only samples)

Transport maps

Random variable X with density π (possibly only samples)

Let Z be a reference random variable with density η

Pick Z easy to work with, e.g. $\mathcal{N}(\mathbf{0}, \mathsf{Id})$

Transport maps

Random variable X with density π (possibly only samples)

Let Z be a reference random variable with density η

Pick Z easy to work with, e.g. $\mathcal{N}(\mathbf{0}, \mathsf{Id})$

Find map T such that given samples $\boldsymbol{z}_i \sim Z$,
the transformed samples satisfy $T(\boldsymbol{z}_i) \sim X$

Transport maps

Random variable X with density π (possibly only samples)

Let Z be a reference random variable with density η

Pick Z easy to work with, e.g. $\mathcal{N}(\mathbf{0}, \mathsf{Id})$

Find map T such that given samples $\boldsymbol{z}_i \sim Z$,
the transformed samples satisfy $T(\boldsymbol{z}_i) \sim X$

VAE, GAN, normalizing flow, score-based/diffusion models

Knothe-Rosenblatt transport

Restrict to *triangular* maps in the sense that

$$T(\boldsymbol{x}) = \begin{bmatrix} T^1(x_1) \\ T^2(x_1, x_2) \\ \vdots \\ T^d(x_1, \ldots, x_d) \end{bmatrix}$$

and control the sign of the diagonal by enforcing $\nabla T \succ 0$

Unique map T called the Knothe-Rosenblatt rearrangement

Knothe-Rosenblatt transport

Restrict to *triangular* maps in the sense that

$$T(\boldsymbol{x}) = \begin{bmatrix} T^1(x_1) \\ T^2(x_1, x_2) \\ \vdots \\ T^d(x_1, \ldots, x_d) \end{bmatrix}$$

and control the sign of the diagonal by enforcing $\nabla T \succ 0$

Unique map T called the Knothe-Rosenblatt rearrangement

Cholesky factor L pushes $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathsf{Id})$ to $L\boldsymbol{z} + \boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu}, \Theta)$

Computational benefits

Encodes conditioning by construction [Marzouk et al. 2016]

Computational benefits

Encodes conditioning by construction [Marzouk et al. 2016]

Fast eigenvalues, determinant, inverse

Computational benefits

Encodes conditioning by construction [Marzouk et al. 2016]

Fast eigenvalues, determinant, inverse

Inherits sparsity from conditional independence

[Spantini, Bigoni, and Marzouk 2018]

A preview

cdf_X^{-1} is a transport from $U \sim \mathrm{Unif}(0,1)$ to 1-d X

Natural generalization to n-d with conditional maps

(cf. sampling from copula)

Möbius inversion, Cornish-Fisher expansion, Edgeworth series

https://misc.cgdct.moe/papers/undergrad_thesis.pdf

References I

- Chen, Yifan et al. (Feb. 2023). *Randomly Pivoted Cholesky: Practical Approximation of a Kernel Matrix with Few Entry Evaluations*. DOI: 10.48550/arXiv.2207.06503. arXiv: 2207.06503 [cs, math, stat].
- Guinness, Joseph (Oct. 2018). "Permutation and Grouping Methods for Sharpening Gaussian Process Approximations". In: *Technometrics* 60.4, pp. 415–429. ISSN: 0040-1706, 1537-2723. DOI: 10.1080/00401706.2018.1437476. arXiv: 1609.05372 [stat].
- Marzouk, Youssef et al. (2016). "An Introduction to Sampling via Measure Transport". In: *arXiv:1602.05023 [math, stat]*, pp. 1–41. DOI: 10.1007/978-3-319-11259-6_23-1. arXiv: 1602.05023 [math, stat].

References II

- Schäfer, Florian, Matthias Katzfuss, and Houman Owhadi (Oct. 2021). "Sparse Cholesky Factorization by Kullback-Leibler Minimization". In: *arXiv:2004.14455 [cs, math, stat]*. arXiv: 2004.14455 [cs, math, stat].
- Spantini, Alessio, Daniele Bigoni, and Youssef Marzouk (Jan. 2018). "Inference via Low-Dimensional Couplings". In: *The Journal of Machine Learning Research* 19.1, pp. 2639–2709. ISSN: 1532-4435.
- Stein, Michael L. (Feb. 2002). "The Screening Effect in Kriging". In: *The Annals of Statistics* 30.1, pp. 298–323. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/aos/1015362194.