Variational Bayesian filtering and smoothing via low-dimensional transports

Massachusetts Institute of Technology

D. Bigoni A. Spantini
R. Baptista Y. Marzouk
Massachusetts Institute of Technology
✉: {dabi,spantini,rsb,ymarz}@mit.edu
🏠: http://transportmaps.mit.edu

Introduction

*We devise an algorithm for **nonlinear** Bayesian filtering and smoothing based on low-dimensional couplings between a tractable **reference** distribution ν_ρ and complex **target** distributions $\{\nu_\pi^k\}$ closely related to the lag-1 smoothing distributions of the data assimilation problem. To this end we we seek parametric maps $\{T^k\}$ pushing forward ν_ρ to $\{\nu_\pi^k\}$, denoted $T^k_\sharp \nu_\rho = \nu_\pi^k$, i.e., we look for:*

$$T^k : \mathbb{R}^d \to \mathbb{R}^d \quad \text{s.t.} \quad \nu_\rho(A) = \nu_\pi^k(T^k(A)) \ , \quad \forall A \in \sigma(\mathbb{R}^d) \ .$$

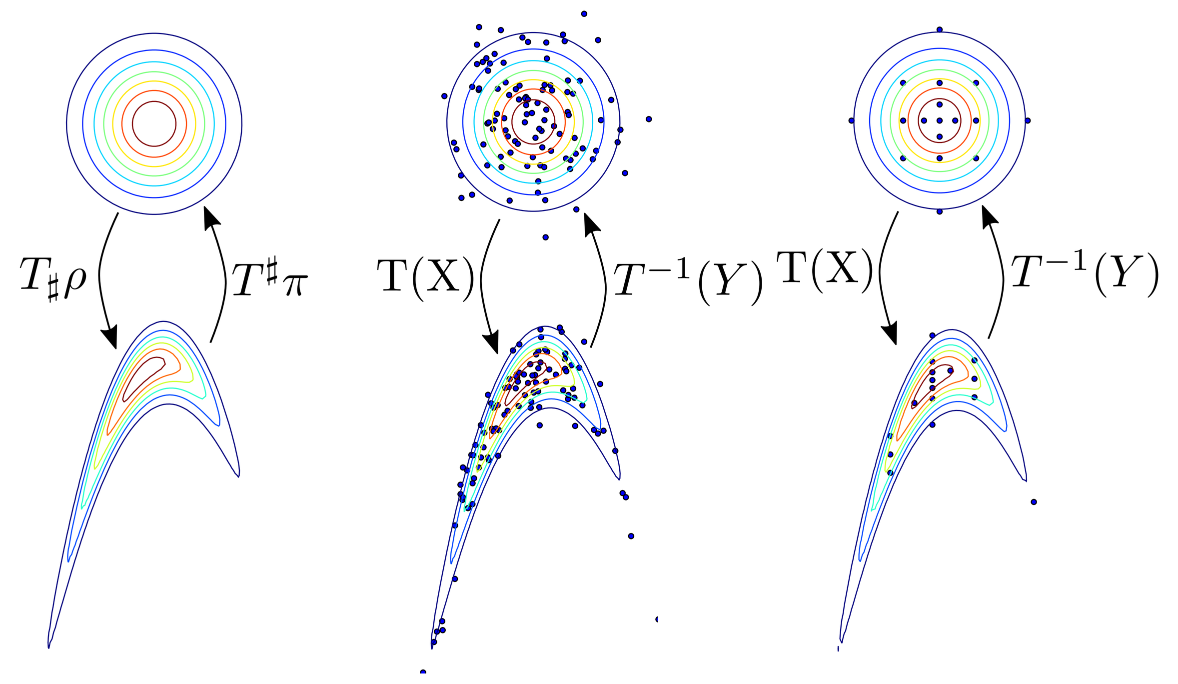
On top of characterizing the lag-1 smoothing distribution, such maps also deliver the solution to the full Bayesian problem, including all its filtering and smoothing marginals.

Measure transport

Let ρ be the density of ν_ρ and π be the density of ν_π.
The map T s.t. $T_\sharp \nu_\rho = \nu_\pi$ defines the following identities:

Pushforward: $T_\sharp \rho(\boldsymbol{x}) = \rho \circ T^{-1}(\boldsymbol{x}) \, |\nabla T^{-1}(\boldsymbol{x})| = \pi$,
Pullback: $T^\sharp \pi(\boldsymbol{x}) = \pi \circ T(\boldsymbol{x}) \, |\nabla T(\boldsymbol{x})| = \rho$,

and for $\mathbf{X} \sim \nu_\rho$, $T(\mathbf{X}) \sim \nu_\pi$.

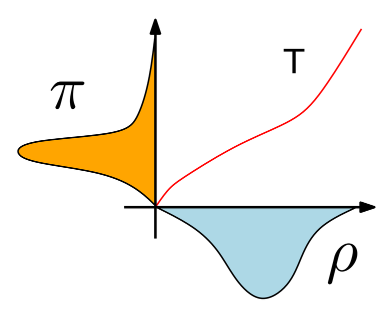


(a) Densities (b) Monte Carlo (c) Sparse grid

Knothe-Rosenblatt rearrangement

For any ν_ρ, ν_π Lebesgue absolutely continuous there exists a **triangular monotone** map $T \in \mathcal{T}_>$ s.t. $T(d\nu_\rho) = d\nu_\pi$

$$T(\boldsymbol{x}) = \begin{bmatrix} T_1(x_1) \\ T_2(x_1, x_2) \\ \vdots \\ T_d(x_1, x_2, \ldots, x_d) \end{bmatrix}$$



The transport map framework

The transportation problem is cast as a minimization problem in terms of the **Kullback-Leibler divergence** [1, 2, 3]:

$$T^\star = \underset{T \in \mathcal{T}_>}{\arg\min} \, \mathcal{D}_{\mathrm{KL}}(T_\sharp \nu_\rho \| \nu_\pi) = \underset{T \in \mathcal{T}_>}{\arg\min} \, \mathbb{E}_\rho \left[-\log T^\sharp \tilde{\pi} \right] \ .$$

We approximate $\mathbb{E}_\rho$ using the quadrature $\mathcal{Q}_q$:

$$T^\star \approx T^\star_q = \underset{T \in \mathcal{T}_>}{\arg\min} \, \mathcal{Q}_q \left(-\log T^\sharp \tilde{\pi} \right) \ .$$

The elements in $\mathcal{T}_>$ are defined by their components $T^{(i)}$,

$$T^{(i)}(x_{1:i}) = c^{(i)}(x_{1:i-1}) + \int_0^{x_i} \left(h^{(i)}(x_{1:i-1}, t) \right)^2 + \varepsilon \, dt \ .$$

Let $\mathcal{T}^k_> \subset \mathcal{T}_>$, with $n_k = \dim(\mathcal{T}^k_>) < \dim(\mathcal{T}_>) = \infty$.
This leads to the further approximation:

$$T^\star_q \approx T^\star_{q,k} = \underset{T \in \mathcal{T}^k_>}{\arg\min} \, \mathcal{Q}_q \left(-\log T^\sharp \tilde{\pi} \right) \ .$$

The following **variance diagnostic** is a global convergence criterion for the approximation of $T^\star$ in the space $\mathcal{T}_>$:

$$\mathcal{D}_{\mathrm{KL}}(T_\sharp \nu_\rho \| \nu_\pi) \approx \frac{1}{2} \mathbb{V}\left[\log \frac{\rho}{T^\sharp \tilde{\pi}} \right] \quad \text{as} \quad T \to T^\star$$

[1] T. a. El Moselhy et al. "Bayesian inference with optimal maps". In: *Journal of Computational Physics* 231.23 (Oct. 2012).
[2] Y. Marzouk et al. "Sampling via Measure Transport: An Introduction". In: *Handbook of Uncertainty Quantification*. Ed. by R. G. Ghanem et al. Cham: Springer International Publishing, 2016. arXiv: 1602.05023.
[3] D. Bigoni et al. "On the computation of monotone transports". In: *preprint* (2019).
[4] A. Spantini et al. "Inference via low-dimensional couplings". In: *Journal of Machine Learning Research* 19.66 (2018), pp. 1–71. arXiv: 1703.06131.

The sequential variational algorithm

Standard data assimilation problems can be formalized as follows: for the index sets $\Lambda \supset \Xi$,

$$\pi\left(\mathbf{Z}_\Lambda | \mathbf{y}_\Xi\right) \propto \prod_{k \in \Xi} \mathcal{L}\left(\mathbf{y}_k | \mathbf{Z}_k\right) \prod_{k \in \Lambda} \pi\left(\mathbf{Z}_k | \mathbf{Z}_{k-1}\right) \pi\left(\mathbf{Z}_0\right) \ .$$

Let us assume that at each step $k \in \Lambda$ the map $\mathfrak{M}^1_{k-1}(\mathbf{z})$ such that $(\mathfrak{M}^1_{k-1})_\sharp \rho(\mathbf{z}_k) = \pi\left(\mathbf{z}_k | \mathbf{y}_{j \in \Xi, \, j \leq k}\right)$ is available, and we are able to characterize the map $\mathfrak{M}_k$ such that

$$(\mathfrak{M}_k)_\sharp \rho(\mathbf{z}_k, \mathbf{z}_{k+1}) = \begin{bmatrix} \mathfrak{M}^1_{k-1}(\mathbf{z}_k) \\ \mathbf{z}_{k+1} \end{bmatrix}^\sharp \pi(\mathbf{z}_k, \mathbf{z}_{k+1} | \mathbf{y}_{j \leq k+1}) \ , \quad \text{where} \quad \mathfrak{M}_k(\mathbf{z}_k, \mathbf{z}_{k+1}) = \begin{bmatrix} \mathfrak{M}^0_k(\mathbf{z}_k, \mathbf{z}_{k+1}) \\ \mathfrak{M}^1_k(\mathbf{z}_{k+1}) \end{bmatrix} \ .$$

Then the following hold true [4]:

Filtering: $(\mathfrak{M}^1_k)_\sharp \rho(\mathbf{z}_{k+1}) = \pi\left(\mathbf{z}_{k+1} | \mathbf{y}_{j \in \Xi, \, j \leq k+1}\right)$
Full solution: $(\mathfrak{T}_k)_\sharp \rho(\mathbf{z}_{0:k+1}) = \pi\left(\mathbf{z}_{0:k+1} | \mathbf{y}_{\Xi \leq k+1}\right)$ where,

$$\mathfrak{T}_k := \begin{bmatrix} \mathfrak{M}_0(\mathbf{z}_0, \mathbf{z}_1) \\ \mathbf{z}_2 \\ \mathbf{z}_3 \\ \vdots \\ \mathbf{z}_{k+1} \end{bmatrix} \circ \begin{bmatrix} \mathbf{z}_0 \\ \mathfrak{M}_1(\mathbf{z}_1, \mathbf{z}_2) \\ \mathbf{z}_3 \\ \vdots \\ \mathbf{z}_{k+1} \end{bmatrix} \circ \cdots \circ \begin{bmatrix} \mathbf{z}_0 \\ \vdots \\ \mathbf{z}_{k-2} \\ \mathbf{z}_{k-1} \\ \mathfrak{M}_k(\mathbf{z}_k, \mathbf{z}_{k+1}) \end{bmatrix}$$

Preconditioning the transport problem.

At step k one needs to solve the problem

$$\mathfrak{M}_k = \underset{\mathfrak{M} \in \mathcal{T}_>}{\arg\min} \, \mathcal{D}_{\mathrm{KL}} \left(\mathfrak{M}_\sharp \rho(\mathbf{z}_k, \mathbf{z}_{k+1}) \middle\| \begin{bmatrix} \mathfrak{M}^1_{k-1}(\mathbf{z}_k) \\ \mathbf{z}_{k+1} \end{bmatrix}^\sharp \pi(\mathbf{z}_k, \mathbf{z}_{k+1} | \mathbf{y}_{j \leq k+1}) \right) \ ,$$

where the distribution on the right is $\widetilde{\pi}^k(\mathbf{z}_k, \mathbf{z}_{k+1}) := \rho(\mathbf{z}_k) \, \pi_{\mathbf{Z}_{k+1}|\mathbf{Z}_k}(\mathbf{z}_{k+1} | \mathfrak{M}^1_{k-1}(\mathbf{z}_k)) \, \mathcal{L}(\mathbf{y}_{k+1} | \mathbf{z}_{k+1})$.
For many dynamical system a reasonable *ansatz* for the next filtering distribution is the latest filtering itself. This observation leads to the preconditioned, i.e. **Gaussianized**, problem

$$\widetilde{\mathfrak{M}}_k = \underset{\mathfrak{M} \in \mathcal{T}_>}{\arg\min} \, \mathcal{D}_{\mathrm{KL}} \left(\mathfrak{M}_\sharp \rho(\mathbf{z}_k, \mathbf{z}_{k+1}) \middle\| \begin{bmatrix} \mathbf{z}_k \\ \mathfrak{M}^1_{k-1}(\mathbf{z}_{k+1}) \end{bmatrix}^\sharp \widetilde{\pi}^k(\mathbf{z}_k, \mathbf{z}_{k+1}) \right) \ ,$$

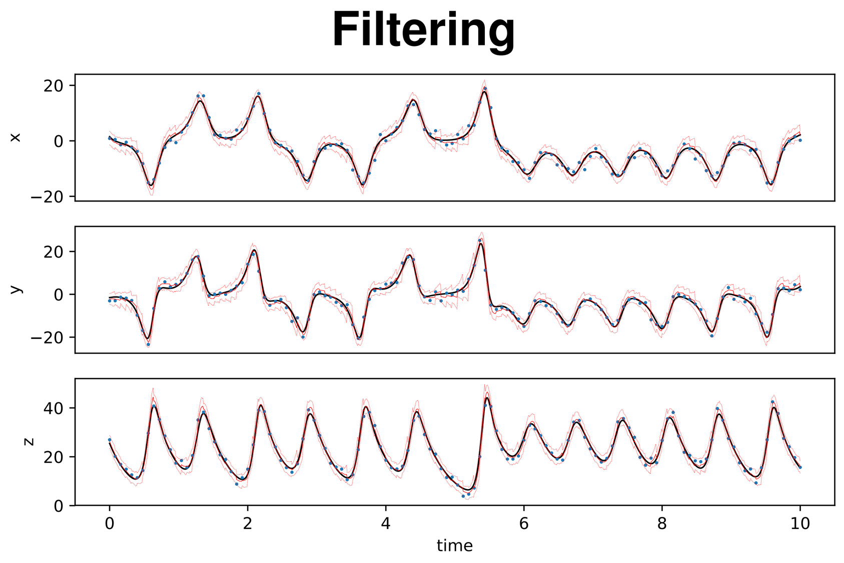
We then define $\mathfrak{M}_k(\mathbf{z}_k, \mathbf{z}_{k+1}) := \begin{bmatrix} \mathbf{z}_k \\ \mathfrak{M}^1_{k-1}(\mathbf{z}_{k+1}) \end{bmatrix} \circ \begin{bmatrix} \widetilde{\mathfrak{M}}^0_k(\mathbf{z}_k, \mathbf{z}_{k+1}) \\ \widetilde{\mathfrak{M}}^1_k(\mathbf{z}_{k+1}) \end{bmatrix}$.

One inexpensive **regression** step can be used to avoid the recursion of the filtering maps in the algorithm:

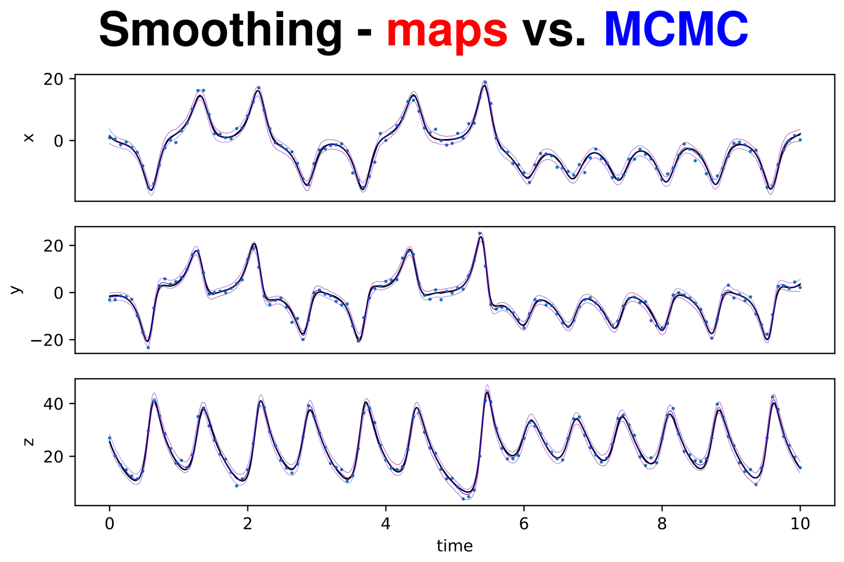
$$\mathfrak{M}^1_k(\mathbf{z}_{k+1}) = \underset{\mathfrak{M}^1 \in \mathcal{T}_>}{\arg\min} \left\| \mathfrak{M}^1(\mathbf{z}_{k+1}) - \mathfrak{M}^1_{k-1} \circ \widetilde{\mathfrak{M}}^1_k(\mathbf{z}_{k+1}) \right\|$$

Lorenz 63

Filtering



Smoothing - maps vs. MCMC



Dynamics

$$\begin{cases} \dot{x} = 10(y - x) \\ \dot{y} = x(28 - z) - y \\ \dot{z} = xy - \frac{8}{3} z \end{cases}$$

Assimilation model

$$\mathbf{u}_{k+1} = \mathbf{u}_k + 0.01 \cdot \mathcal{F}(\mathbf{u}_k) + \mathcal{N}(0, 10^{-1} \cdot \mathbf{I})$$
$$\mathbf{h}_k = \mathcal{O}(\mathbf{u}_k) + \mathcal{N}(0, 2 \cdot \mathbf{I})$$
$$k = 0, 8, 16, \ldots$$

	filtering		smoothing	
	Order 1	Order 3	Order 1	Order 3
Avg. RMSE	7.39×10^{-1}	7.46×10^{-1}	4.82×10^{-1}	4.86×10^{-1}
Med. RMSE	6.49×10^{-1}	6.61×10^{-1}	4.18×10^{-1}	4.23×10^{-1}
Var. RMSE	1.51×10^{-1}	1.55×10^{-1}	5.62×10^{-2}	5.78×10^{-2}

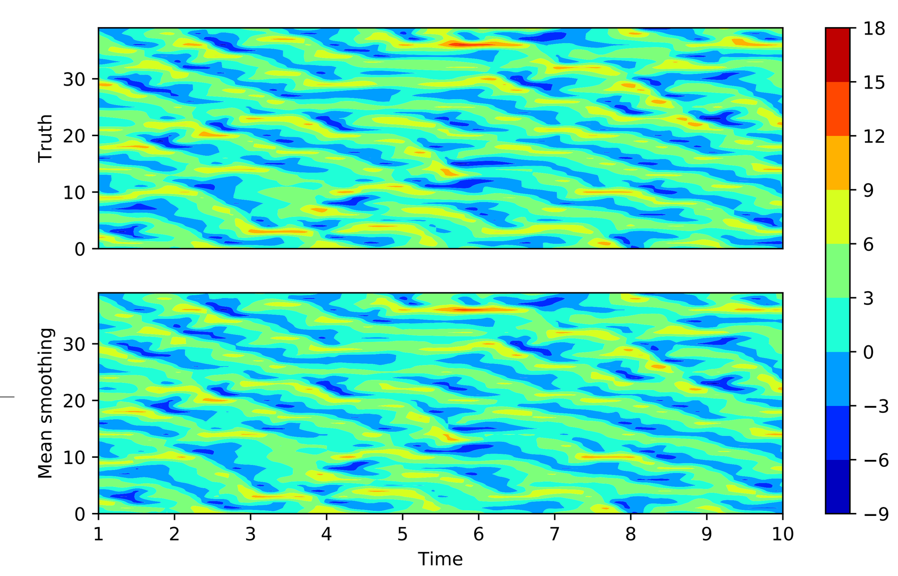
Posterior accuracy ($d = 3000$)	Order 1	Order 3
$\mathbb{V}[\log \frac{\rho}{T^\sharp \pi}]$	8.58×10^{-1}	2.11×10^{-1}
Metropolis indep. A/R	33.7%	62.9%
Metropolis indep. ESS (worst)	0.68%	12.54%

Lorenz 96 – model I

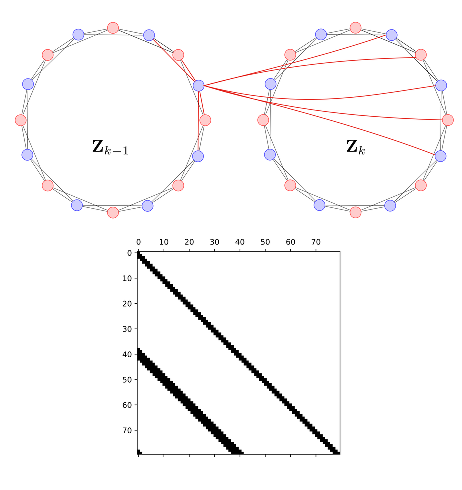
$$\frac{d\mathbf{Z}_j}{dt} = (\mathbf{Z}_{j+1} - \mathbf{Z}_{j-2}) \, \mathbf{Z}_{j-1} - \mathbf{Z}_j + F$$
$j \in \{1, \ldots, 40\}$, $F = 8$ (chaotic)

Assimilation model

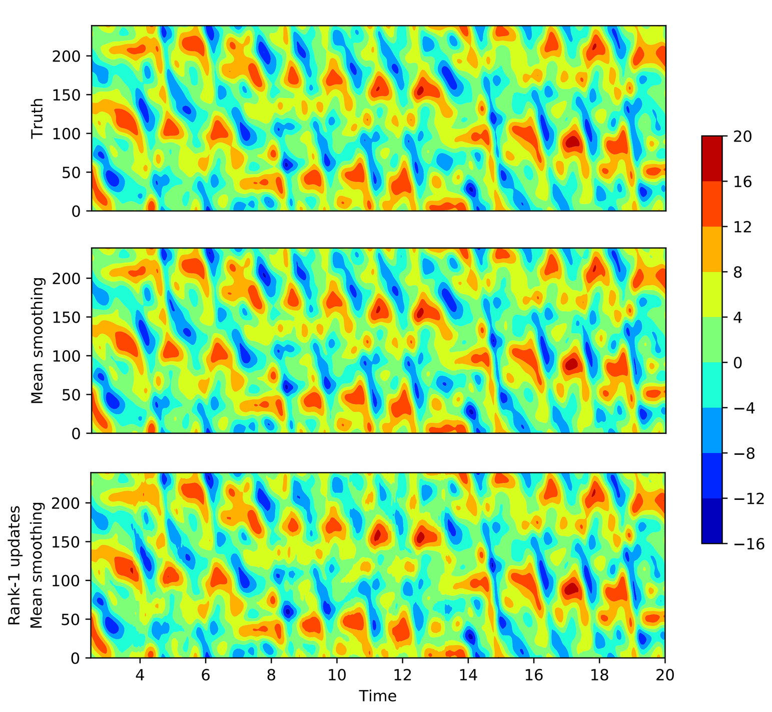
$$\mathbf{z}_{k+1} = \mathbf{z}_k + 0.01 \cdot \mathcal{F}(\mathbf{z}_k) + \mathcal{N}(0, 10^{-1} \cdot \mathbf{I})$$
$$\mathbf{h}_i = \mathcal{O}(\mathbf{z}_i) + \mathcal{N}(0, 0.5 \cdot \mathbf{I}) \ , \ i = 0, 10, \ldots$$



Map sparsity



Lorenz 96 – model II



$$\frac{d\mathbf{X}_n}{dt} = \sum_{j=-J}^{J} \sum_{i=-J}^{J} \left(X_{n-K+j-i} X_{n+K+j} - X_{n-2K-i} X_{n-K-j} \right) / K^2 - X_n + F$$

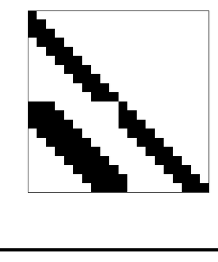
$\mathbf{X} \in \mathbb{R}^d$, $d = 240$, $K = 33$, $J = (K-1)/2$, $F = 14$ (chaotic)

Assimilation model (Reduced model including 99.9% of total power)

$$\mathbf{b}_{k+1} = \mathbf{b}_k + 0.025 \cdot \mathcal{K}^{-1} \circ \mathcal{F} \circ \mathcal{K}(\mathbf{b}_k) + \mathcal{K}^{-1} \mathcal{N}(0, 10^{-1} \cdot \mathbf{I})$$
$$\mathbf{h}_i = \mathcal{O}(\mathcal{K}(\mathbf{b}_i)) + \mathcal{N}(0, 10^{-2} \cdot \mathbf{I}) \ , \ i = 0, 4, \ldots \quad \mathcal{K}(\mathbf{b}) = \boldsymbol{\mu} + \mathbf{K}\mathbf{b}$$
where $\mathbb{V}[\mathbf{X}] = \mathbf{U} \boldsymbol{\Lambda} \mathbf{U}^\top$, $\mathbf{K} := \mathbf{U}_{1:r} \boldsymbol{\Lambda}^{1/2}_{1:r}$ and $r = 21$

including predictions	filtering	smoothing
Avg. RMSE	1.00	7.32×10^{-1}
Med. RMSE	8.27×10^{-1}	6.23×10^{-1}
Var. RMSE	5.63×10^{-1}	3.08×10^{-1}

Map sparsity



This work was supported by the Air Force Office of Scientific Research, Computational Mathematics program, and by the US Department of Energy, Office of Advanced Scientific Computing Research.