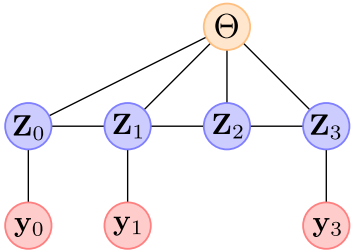
Variational Bayesian filtering and smoothing via low-dimensional transports

D. Bigoni (**dabi@mit.edu**), R. Baptista, A. Spantini, Y. Marzouk
Massachusetts Institute of Technology

AGU Fall meeting
Washington, DC – 12/11/2018

Sequential data assimilation

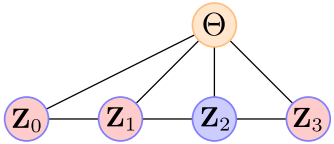


$$\pi\left(\Theta, \mathbf{Z}_\Lambda | \mathbf{y}_\Xi\right) \propto \mathcal{L}\left(\mathbf{y}_\Xi | \Theta, \mathbf{Z}_\Lambda\right) \pi\left(\Theta, \mathbf{Z}_\Lambda\right)$$

$$\mathcal{L}\left(\mathbf{y}_\Xi | \Theta, \mathbf{Z}_\Lambda\right) = \prod_{k \in \Xi} \mathcal{L}\left(\mathbf{y}_k | \Theta, \mathbf{Z}_k\right)$$

$$\pi\left(\Theta, \mathbf{Z}_\Lambda\right) = \pi\left(\Theta\right) \pi\left(\mathbf{Z}_0 | \Theta\right) \prod_{k \in \Lambda} \pi\left(\mathbf{Z}_k | \mathbf{Z}_{k-1}, \Theta\right)$$

Sequential data assimilation

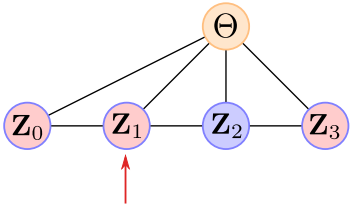


$$\pi\left(\Theta, \mathbf{Z}_\Lambda | \mathbf{y}_\Xi\right) \propto \mathcal{L}\left(\mathbf{y}_\Xi | \Theta, \mathbf{Z}_\Lambda\right) \pi\left(\Theta, \mathbf{Z}_\Lambda\right)$$

$$\mathcal{L}\left(\mathbf{y}_\Xi | \Theta, \mathbf{Z}_\Lambda\right) = \prod_{k \in \Xi} \mathcal{L}\left(\mathbf{y}_k | \Theta, \mathbf{Z}_k\right)$$

$$\pi\left(\Theta, \mathbf{Z}_\Lambda\right) = \pi\left(\Theta\right) \pi\left(\mathbf{Z}_0 | \Theta\right) \prod_{k \in \Lambda} \pi\left(\mathbf{Z}_k | \mathbf{Z}_{k-1}, \Theta\right)$$

Sequential data assimilation

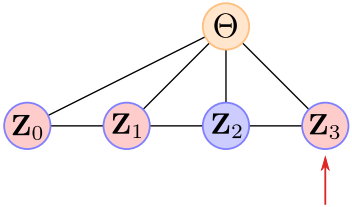


$$\pi\left(\mathbf{Z}_k|\mathbf{y}_\Xi\right) \propto \int \mathcal{L}\left(\mathbf{y}_\Xi|\Theta,\mathbf{Z}_\Lambda\right)\pi\left(\Theta,\mathbf{Z}_\Lambda\right)d\Theta\, d\mathbf{Z}_{j\neq k}$$

$$\mathcal{L}\left(\mathbf{y}_\Xi|\Theta,\mathbf{Z}_\Lambda\right) = \prod_{k\in\Xi}\mathcal{L}\left(\mathbf{y}_k|\Theta,\mathbf{Z}_k\right)$$

$$\pi\left(\Theta,\mathbf{Z}_\Lambda\right) = \pi\left(\Theta\right)\pi\left(\mathbf{Z}_0|\Theta\right)\prod_{k\in\Lambda}\pi\left(\mathbf{Z}_k|\mathbf{Z}_{k-1},\Theta\right)$$

Sequential data assimilation



$$\pi\left(\mathbf{Z}_k|\mathbf{y}_{j\leq k}\right) \propto \int \mathcal{L}\left(\mathbf{y}_{j\leq k}|\Theta, \mathbf{Z}_{j\leq k}\right) \pi\left(\Theta, \mathbf{Z}_{j\leq k}\right) d\Theta \, d\mathbf{Z}_{j<k}$$

$$\mathcal{L}\left(\mathbf{y}_\Xi|\Theta, \mathbf{Z}_\Lambda\right) = \prod_{k\in\Xi} \mathcal{L}\left(\mathbf{y}_k|\Theta, \mathbf{Z}_k\right)$$

$$\pi\left(\Theta, \mathbf{Z}_\Lambda\right) = \pi\left(\Theta\right) \pi\left(\mathbf{Z}_0|\Theta\right) \prod_{k\in\Lambda} \pi\left(\mathbf{Z}_k|\mathbf{Z}_{k-1}, \Theta\right)$$

Transport maps – Pullbacks [PB] and Pushforwards [PF]

- Distribution $\boldsymbol{\nu}_\rho$ with density $\rho : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$
- Distribution $\boldsymbol{\nu}_\pi$ with density $\pi : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$
- For $T : \mathbb{R}^d \to \mathbb{R}^d$ we define

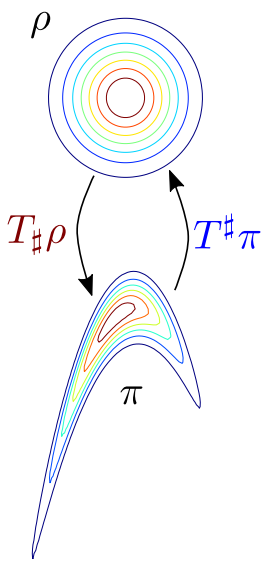
  PF $\quad T_\sharp \rho = \rho \circ T^{-1} |\nabla T^{-1}|$

  PB $\quad T^\sharp \pi = \pi \circ T |\nabla T|$

- We want T such that

  PF $\quad T_\sharp \rho = \pi$

  PB $\quad T^\sharp \pi = \rho$



Transport maps – Pullbacks [PB] and Pushforwards [PF]

- Distribution $\boldsymbol{\nu}_\rho$ with density $\rho : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$
- Distribution $\boldsymbol{\nu}_\pi$ with density $\pi : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$
- For $T : \mathbb{R}^d \to \mathbb{R}^d$ we define

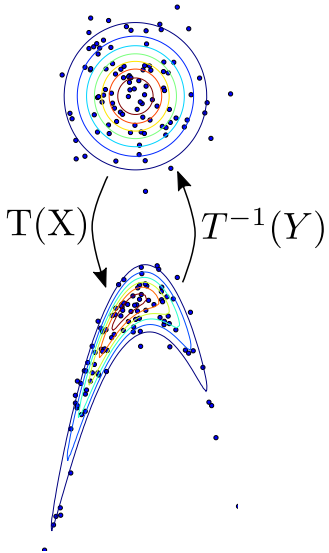
  PF $\quad T_\sharp \rho = \rho \circ T^{-1} |\nabla T^{-1}|$

  PB $\quad T^\sharp \pi = \pi \circ T |\nabla T|$

- We want T such that

  PF For $X \sim \boldsymbol{\nu}_\rho$, $T(X) \sim \boldsymbol{\nu}_\pi$

  PB For $Y \sim \boldsymbol{\nu}_\pi$, $T^{-1}(Y) \sim \boldsymbol{\nu}_\rho$



Transport maps – Pullbacks [PB] and Pushforwards [PF]

- Distribution $\boldsymbol{\nu}_\rho$ with density $\rho : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$
- Distribution $\boldsymbol{\nu}_\pi$ with density $\pi : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$
- For $T : \mathbb{R}^d \to \mathbb{R}^d$ we define

  PF $\quad T_\sharp \rho = \rho \circ T^{-1} |\nabla T^{-1}|$

  PB $\quad T^\sharp \pi = \pi \circ T |\nabla T|$

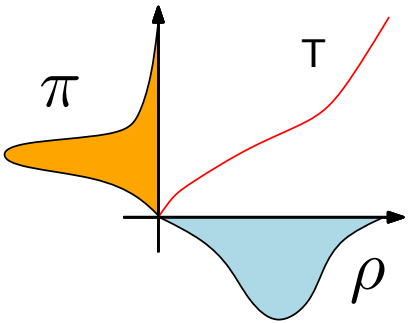
- We want T such that

  PF For $X \sim \boldsymbol{\nu}_\rho$, $T(X) \sim \boldsymbol{\nu}_\pi$

  PB For $Y \sim \boldsymbol{\nu}_\pi$, $T^{-1}(Y) \sim \boldsymbol{\nu}_\rho$

Knothe-Rosenblatt rearrangement

$\forall\ \boldsymbol{\nu}_\rho, \boldsymbol{\nu}_\pi$ Lebesgue absolutely continuous
$\exists$ a **triangular monotone** map s.t. $T_\sharp \rho = \pi$



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

Minimize Kullback-Leibler divergence to find optimal map [ElMoselhy2012]

$$T^\star = \arg\min_{T \in \mathcal{T}_>} D_{\mathrm{KL}}(T_\sharp \rho \| \pi) = \arg\min_{T \in \mathcal{T}_>} \mathbb{E}_\rho \left[\log \frac{\rho}{T^\sharp \pi} \right]$$

+ **Gradient-based unconstrained optimization** (if gradients are available)
+ We can **explore π in parallel**
+ We can **generate i.i.d. samples** from $T_\sharp^\star \rho \propto \pi$ **in parallel**
+ We can **assess convergence**: $D_{\mathrm{KL}}(T_\sharp \rho \| \pi) \approx \frac{1}{2} \mathbb{V}_\rho \left[\log \frac{\rho}{T^\sharp \pi} \right]$
+ The map can be used as a **preconditioner** for other unbiased methods
− We need to **approximate d functions of up to d variables!**

Minimize Kullback-Leibler divergence to find optimal map [ElMoselhy2012]

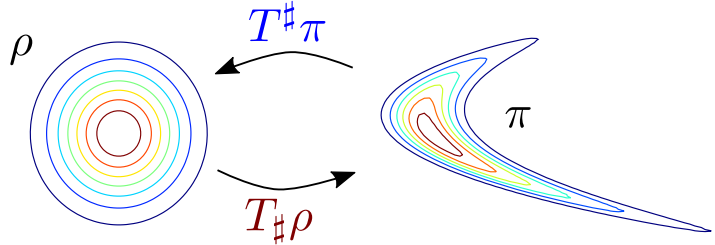
$$T^\star = \operatorname*{arg\,min}_{T \in \mathcal{T}_>} D_{\mathrm{KL}}(T_\sharp \rho \| \pi) = \operatorname*{arg\,min}_{T \in \mathcal{T}_>} \mathbb{E}_\rho \left[\log \frac{\rho}{T^\sharp \pi} \right]$$

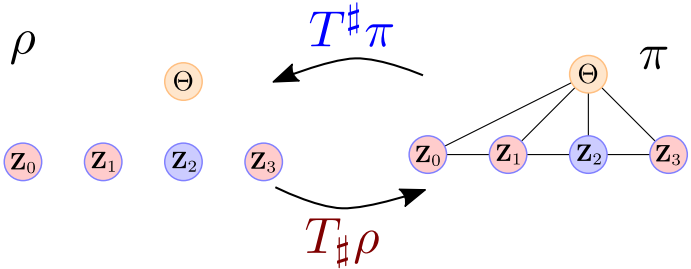
+ **Gradient-based unconstrained optimization** (if gradients are available)
+ We can **explore π in parallel**
+ We can **generate i.i.d. samples** from $T_\sharp^\star \rho \propto \pi$ **in parallel**
+ We can **assess convergence**: $D_{\mathrm{KL}}(T_\sharp \rho \| \pi) \approx \frac{1}{2} \mathbb{V}_\rho \left[\log \frac{\rho}{T^\sharp \pi} \right]$
+ The map can be used as a **preconditioner** for other unbiased methods
− We need to **approximate d functions of up to d variables!**

Sources of low-dimensional structure

- Smoothness [B., Spantini, Marzouk]
- Marginal independence [B., Spantini, Marzouk]
- Conditional independence [Spantini2018]
- Low-rank structure [B., Zahm, Spantini, Marzouk]

How to remove conditional dependencies?





Removing conditional dependencies – step 1

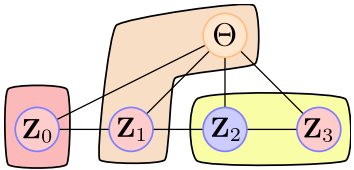
Find the map

$$\mathfrak{M}_0(\boldsymbol{\theta}, \mathbf{z}_0, \mathbf{z}_1) = \begin{bmatrix} \mathfrak{M}_0^{\Theta}(\boldsymbol{\theta}) \\ \mathfrak{M}_0^{0}(\boldsymbol{\theta}, \mathbf{z}_0, \mathbf{z}_1) \\ \mathfrak{M}_0^{1}(\boldsymbol{\theta}, \mathbf{z}_1) \end{bmatrix}$$

pushing forward $\mathcal{N}(0, \mathbf{I})$ to the **first Markov component** of $\pi\left(\Theta, \mathbf{Z}_\Lambda \middle| \mathbf{y}_\Xi\right)$:

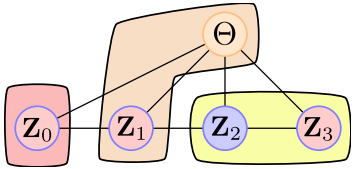
$$\pi^0(\Theta, \mathbf{Z}_0, \mathbf{Z}_1) := \mathcal{L}(\mathbf{y}_0|\Theta, \mathbf{Z}_0)\mathcal{L}(\mathbf{y}_1|\Theta, \mathbf{Z}_1)\pi(\mathbf{Z}_1|\Theta, \mathbf{Z}_0)\pi(\mathbf{Z}_0|\Theta)\pi(\Theta)$$

π :



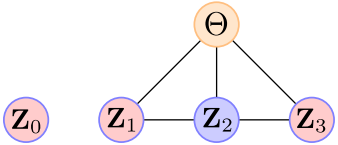
Removing conditional dependencies – step 1

π :



$$T_0(\boldsymbol{\theta}, \mathbf{z}_\Lambda) = \begin{bmatrix} \mathfrak{M}_0(\boldsymbol{\theta}, \mathbf{z}_0, \mathbf{z}_1) \\ \mathbf{z}_2 \\ \mathbf{z}_3 \end{bmatrix}$$

$T_0^\sharp \pi$:



Removing conditional dependencies – step 2

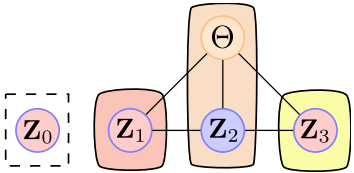
Find the map

$$\mathfrak{M}_1(\boldsymbol{\theta}, \mathbf{z}_1, \mathbf{z}_2) = \begin{bmatrix} \mathfrak{M}_1^{\Theta}(\boldsymbol{\theta}) \\ \mathfrak{M}_1^{0}(\boldsymbol{\theta}, \mathbf{z}_1, \mathbf{z}_2) \\ \mathfrak{M}_1^{1}(\boldsymbol{\theta}, \mathbf{z}_2) \end{bmatrix}$$

pushing $\mathcal{N}(0, \mathbf{I})$ to the **second Markov component** of $T_0^\sharp \pi$:

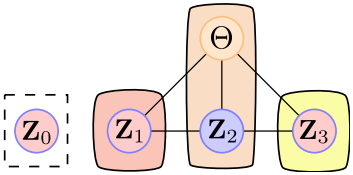
$$\pi^1(\Theta, \mathbf{Z}_1, \mathbf{Z}_2) := \eta(\Theta, \mathbf{Z}_1)\, \pi\left(\mathbf{Z}_2 \middle| \Theta', \mathbf{Z}_1'\right) \quad \text{(missing observation)}$$
$$\Theta' := \mathfrak{M}_0^{\Theta}(\Theta)\,, \quad \mathbf{Z}_1' := \mathfrak{M}_0^{1}(\Theta, \mathbf{Z}_1)$$

$T_0^\sharp \pi$:



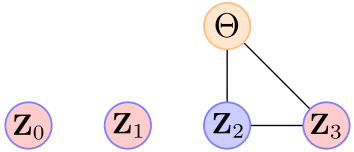
Removing conditional dependencies – step 2

$T_0^\sharp \pi$:



$$T_1(\boldsymbol{\theta}, \mathbf{z}_\Lambda) = \begin{bmatrix} \mathfrak{M}_1^\Theta(\boldsymbol{\theta}) \\ \mathbf{z}_0 \\ \mathfrak{M}_1^0(\boldsymbol{\theta}, \mathbf{z}_1, \mathbf{z}_2) \\ \mathfrak{M}_1^1(\boldsymbol{\theta}, \mathbf{z}_2) \\ \mathbf{z}_3 \end{bmatrix}$$

$T_1^\sharp T_0^\sharp \pi$:



Removing conditional dependencies – step 3

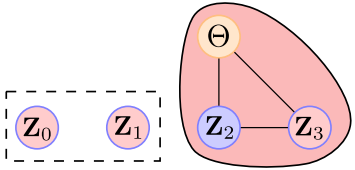
Find the map

$$\mathfrak{M}_2(\boldsymbol{\theta}, \mathbf{z}_2, \mathbf{z}_3) = \begin{bmatrix} \mathfrak{M}_2^{\Theta}(\boldsymbol{\theta}) \\ \mathfrak{M}_2^{0}(\boldsymbol{\theta}, \mathbf{z}_2, \mathbf{z}_3) \\ \mathfrak{M}_2^{1}(\boldsymbol{\theta}, \mathbf{z}_3) \end{bmatrix}$$

pushing $\mathcal{N}(0, \mathbf{I})$ to the **third Markov component** of $T_1^\sharp T_0^\sharp \pi$:

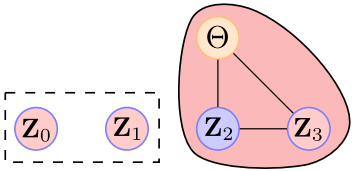
$$\pi^2\left(\Theta, \mathbf{Z}_2, \mathbf{Z}_3\right) := \eta(\Theta, \mathbf{Z}_2)\, \mathcal{L}\left(\mathbf{y}_3 \,\middle|\, \Theta'', \mathbf{Z}_3\right) \pi\left(\mathbf{Z}_3 \,\middle|\, \Theta'', \mathbf{Z}_2''\right)$$
$$\Theta'' := \mathfrak{M}_0^{\Theta} \circ \mathfrak{M}_1^{\Theta}(\Theta)\,, \quad \mathbf{Z}_1'' := \mathfrak{M}_1^{1}(\Theta, \mathbf{Z}_2)$$

$T_1^\sharp T_0^\sharp \pi$:



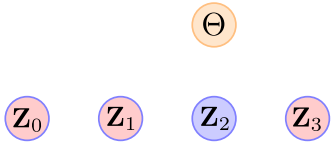
Removing conditional dependencies – step 3

$T_1^\sharp T_0^\sharp \pi$:



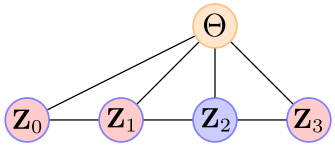
$$T_2(\boldsymbol{\theta}, \mathbf{z}_\Lambda) = \begin{bmatrix} \mathfrak{M}_2^\Theta(\boldsymbol{\theta}) \\ \mathbf{z}_{0:1} \\ \mathfrak{M}_2^0(\boldsymbol{\theta}, \mathbf{z}_2, \mathbf{z}_3) \\ \mathfrak{M}_2^1(\boldsymbol{\theta}, \mathbf{z}_3) \end{bmatrix}$$

$T_2^\sharp T_1^\sharp T_0^\sharp \pi$:



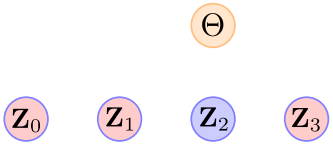
Removing conditional dependencies

π :



$$\mathfrak{T} = \begin{bmatrix} \mathfrak{M}_0^{\Theta}(\boldsymbol{\theta}) \\ \mathfrak{M}_0^{0}(\boldsymbol{\theta}, \mathbf{z}_0, \mathbf{z}_1) \\ \mathfrak{M}_0^{1}(\boldsymbol{\theta}, \mathbf{z}_1) \\ \mathbf{z}_2 \\ \mathbf{z}_3 \end{bmatrix} \circ \begin{bmatrix} \mathfrak{M}_1^{\Theta}(\boldsymbol{\theta}) \\ \mathbf{z}_0 \\ \mathfrak{M}_1^{0}(\boldsymbol{\theta}, \mathbf{z}_1, \mathbf{z}_2) \\ \mathfrak{M}_1^{1}(\boldsymbol{\theta}, \mathbf{z}_2) \\ \mathbf{z}_3 \end{bmatrix} \circ \begin{bmatrix} \mathfrak{M}_1^{\Theta}(\boldsymbol{\theta}) \\ \mathbf{z}_0 \\ \mathbf{z}_1 \\ \mathfrak{M}_2^{0}(\boldsymbol{\theta}, \mathbf{z}_2, \mathbf{z}_3) \\ \mathfrak{M}_2^{1}(\boldsymbol{\theta}, \mathbf{z}_3) \end{bmatrix}$$

$\mathfrak{T}^\sharp\pi$:



A framework for sequential inference

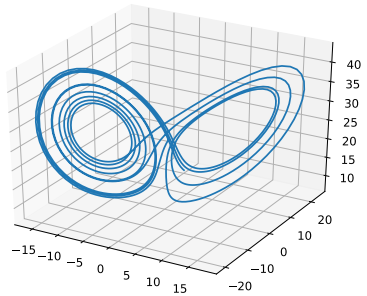
- **Variational Bayesian algorithms** for joint state-parameter inference
- Yelds **filtering**, **smoothing**, **full posterior** via composition of maps
- Constant cost per assimilation/prediction step
- Map parametrization *(basis, sparsity)* governs **complexity/accuracy trade-off**

Lorenz 63

Dynamics

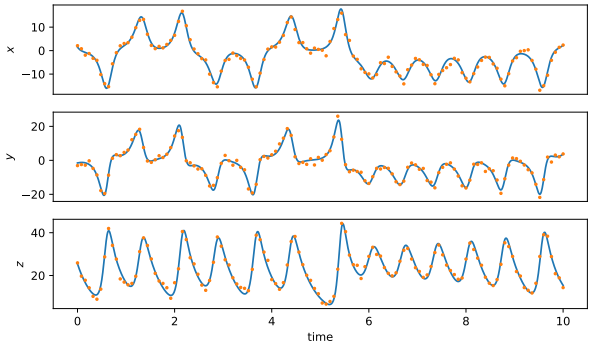
$$\begin{cases} \dot{x} = \sigma(y - x) \\ \dot{y} = x(\rho - z) - y \\ \dot{z} = xy - \beta z \end{cases}$$

$$\sigma = 10,\ \beta = 8/3,\ \rho = 28$$

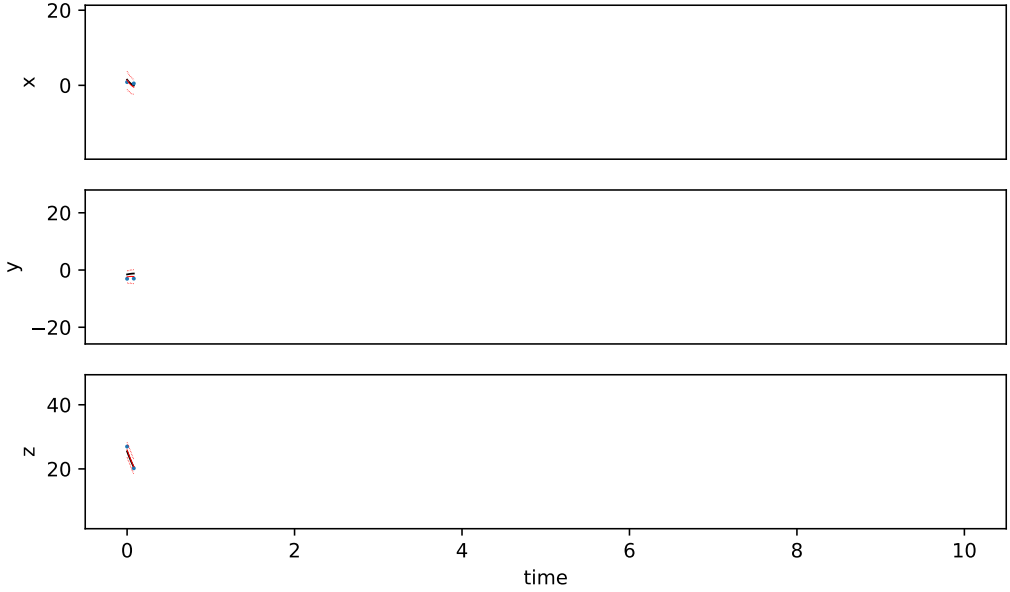


Assimilation model

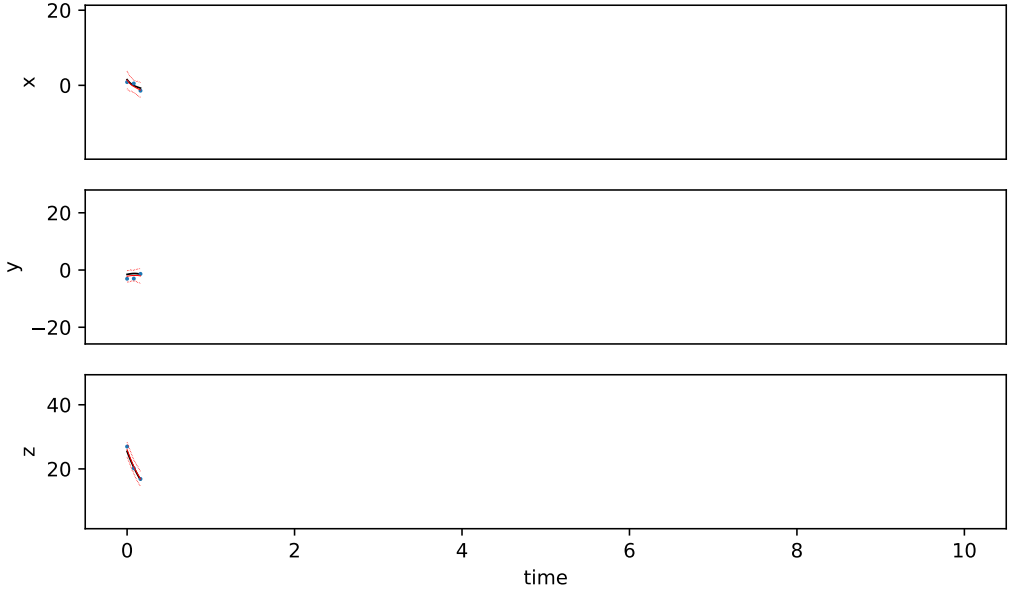
$$\mathbf{u}_{k+1} = \mathbf{u}_k + 0.01 \cdot \mathcal{F}(\mathbf{u}_k) + \mathcal{N}(0, 10^{-1} \cdot \mathbf{I})$$
$$\mathbf{h}_k = \mathcal{O}(\mathbf{u}_k) + \mathcal{N}(0, 2 \cdot \mathbf{I}) \,, \quad k = 0, 8, 16, \ldots$$



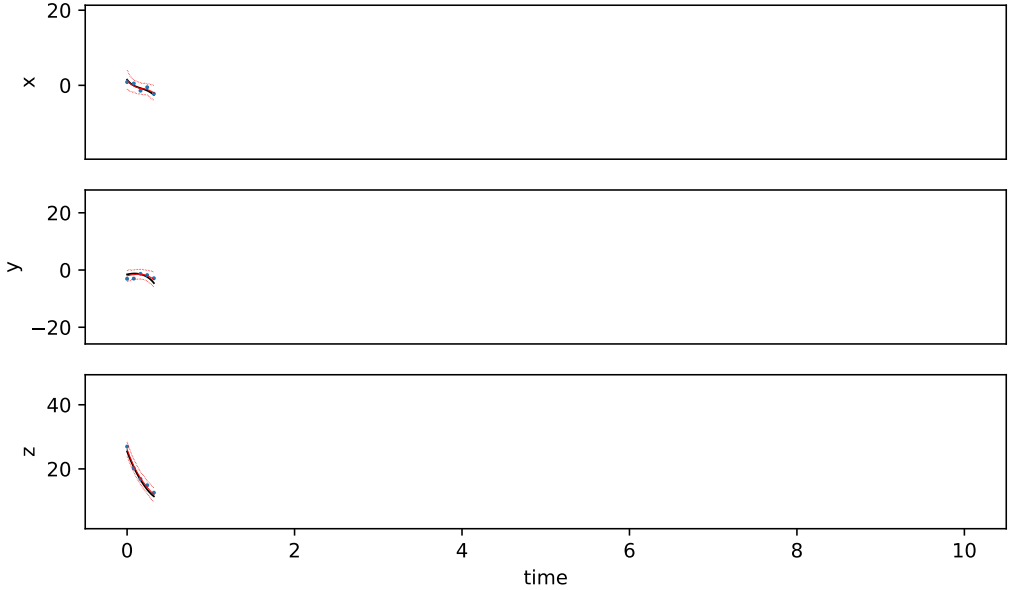
Smoothing



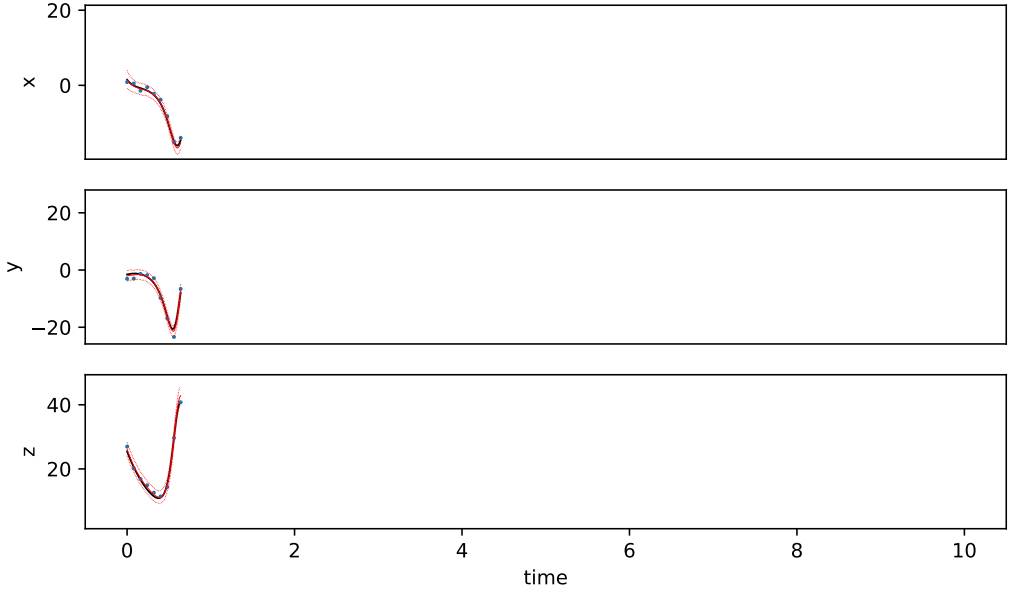
Smoothing



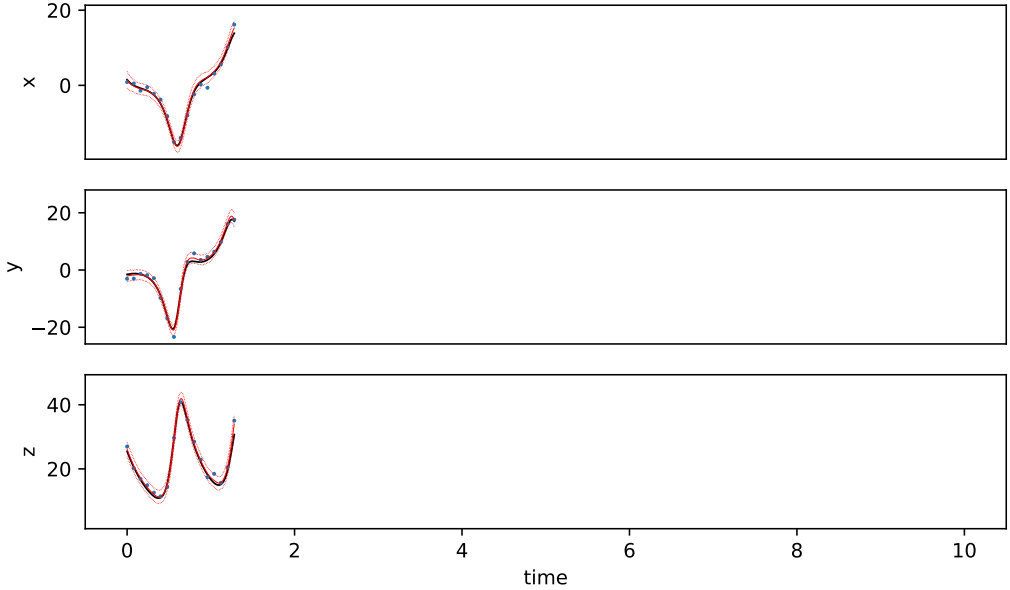
Smoothing



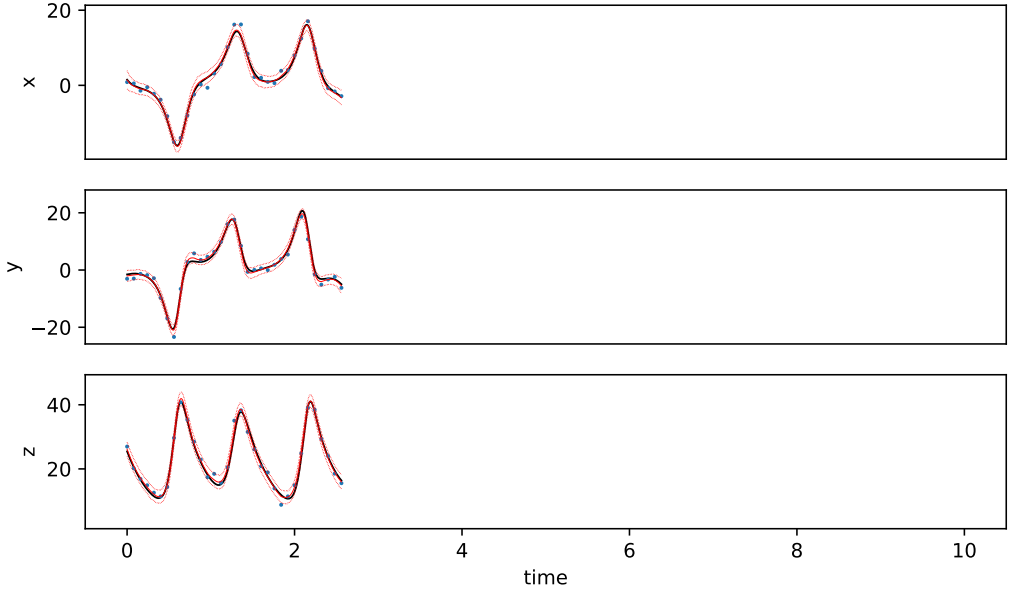
Smoothing



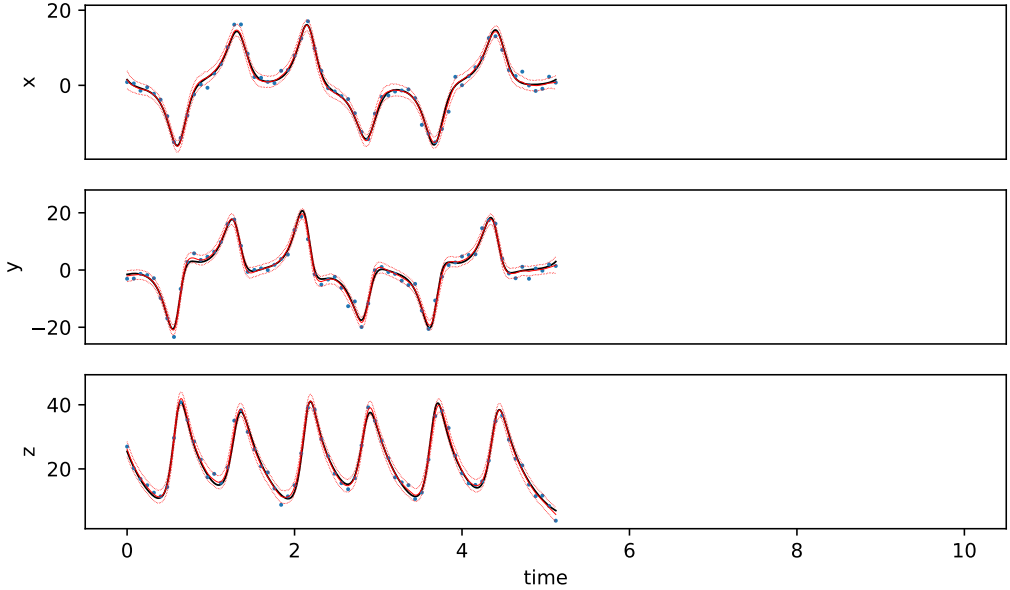
Smoothing



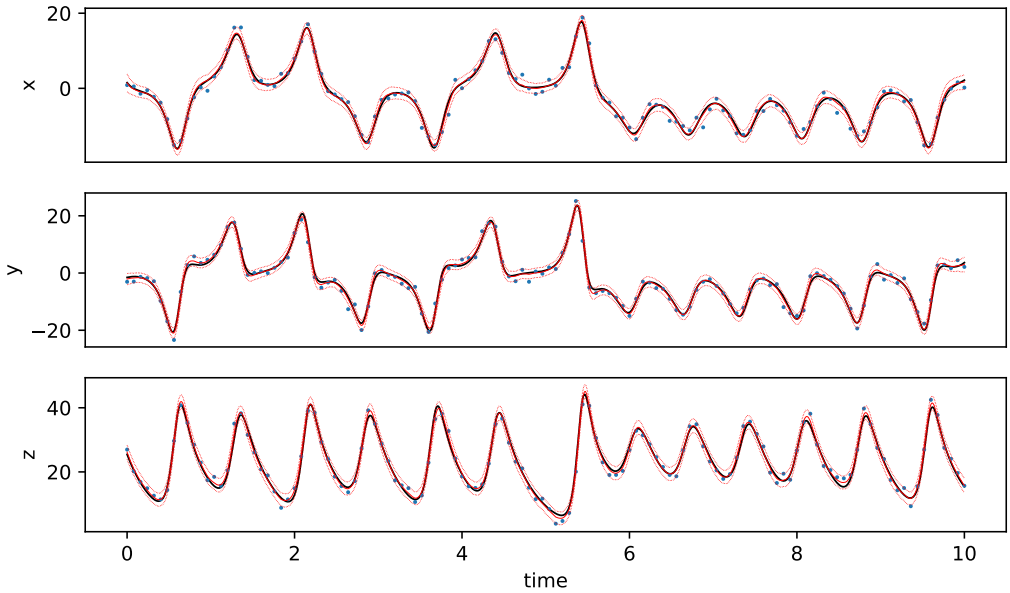
Smoothing



Smoothing

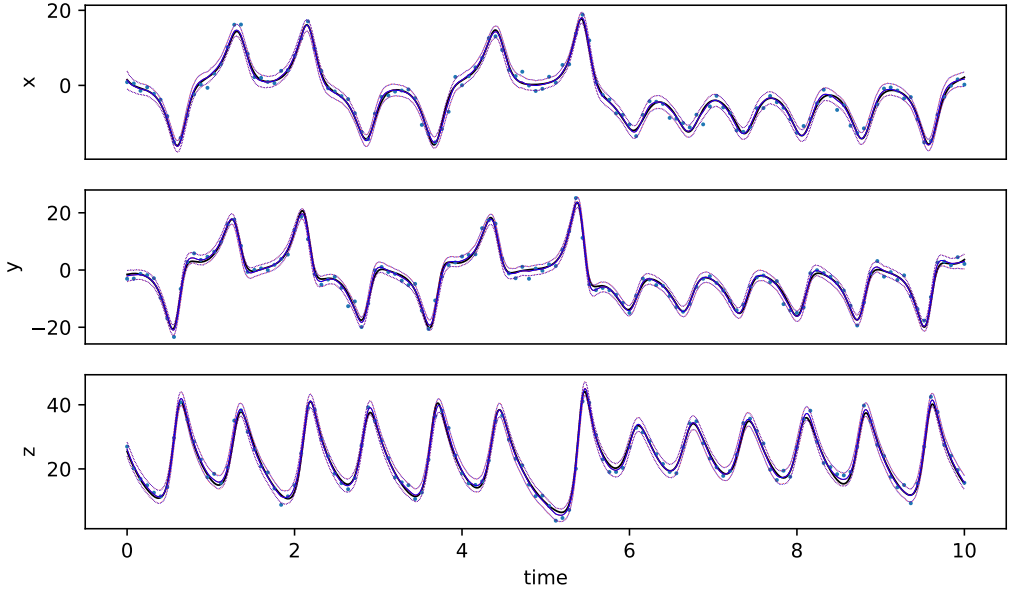


Smoothing



		filtering		smoothing	
		Order 1	Order 3	Order 1	Order 3
at obs.	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}
incl. pred.	Avg. RMSE	8.37×10^{-1}	8.42×10^{-1}	4.67×10^{-1}	4.72×10^{-1}
	Med. RMSE	7.54×10^{-1}	7.50×10^{-1}	4.11×10^{-1}	4.14×10^{-1}
	Var. RMSE	2.23×10^{-1}	2.25×10^{-1}	5.53×10^{-2}	5.72×10^{-2}

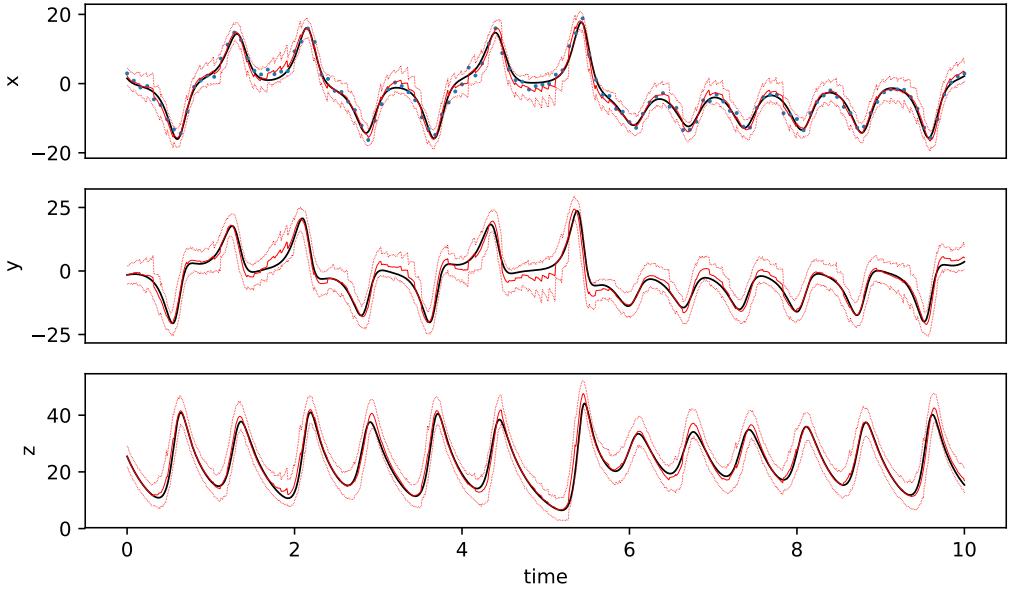
Smoothing – maps vs. MCMC



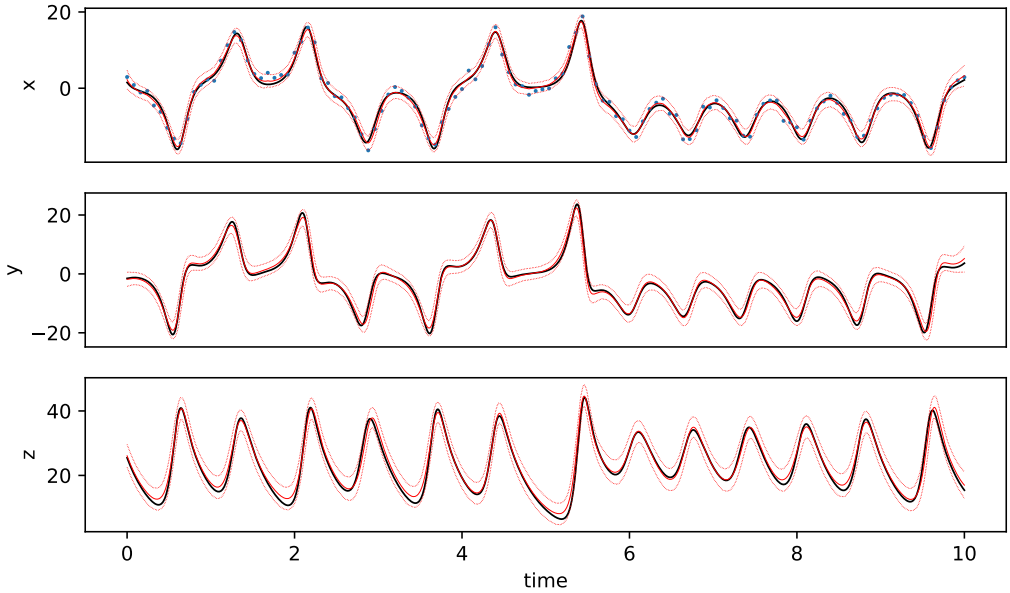
		filtering		smoothing	
		Order 1	Order 3	Order 1	Order 3
at obs.	Avg. RMSE	7.39×10^{-1}	7.46×10^{-1}	4.82×10^{-1}	4.86×10^{-1}
	Med. RMSE	6.49×10^{-1}	6.61×10^{-1}	4.18×10^{-1}	4.23×10^{-1}
	Var. RMSE	1.51×10^{-1}	1.55×10^{-1}	5.62×10^{-2}	5.78×10^{-2}

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%

Filtering



Smoothing

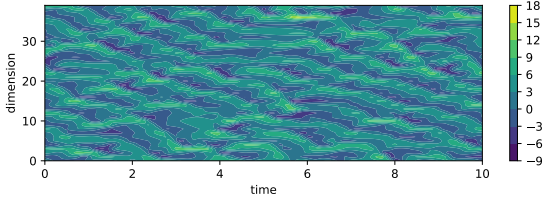


Lorenz 96

Dynamics

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

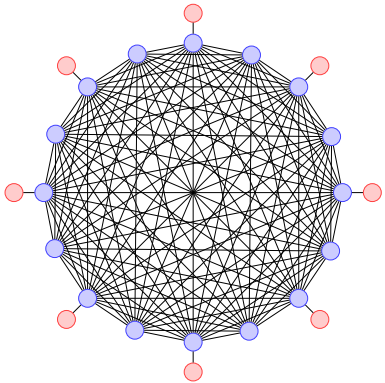
$$j \in \{1, \dots, 40\}\,, \quad F = 8 \text{ (chaotic)}$$



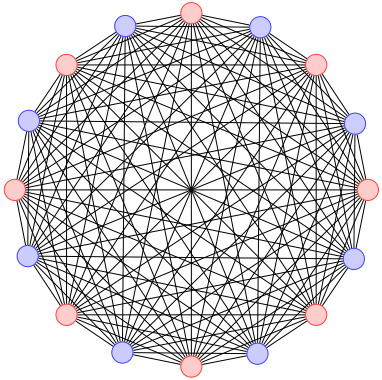
Assimilation model

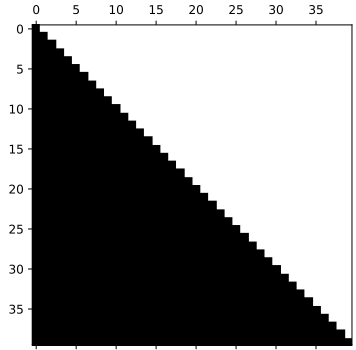
$$\mathbf{u}_{k+1} = \mathbf{u}_k + 0.01 \cdot \mathcal{F}(\mathbf{u}_k) + \mathcal{N}(0, 10^{-1} \cdot \mathbf{I})$$

$$\mathbf{h}_k = \mathcal{O}(\mathbf{u}_k) + \mathcal{N}(0, 0.5 \cdot \mathbf{I})\,,\ k = 0, 10, \dots$$

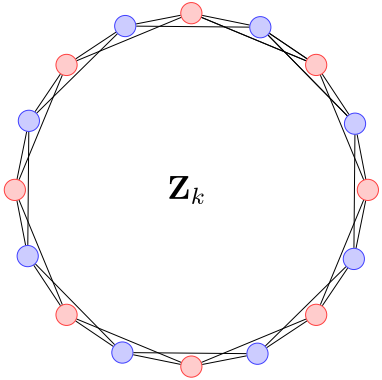


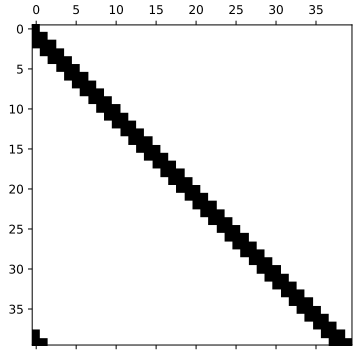
Map sparsity ansatz



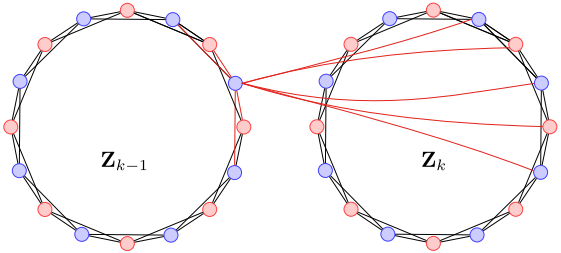


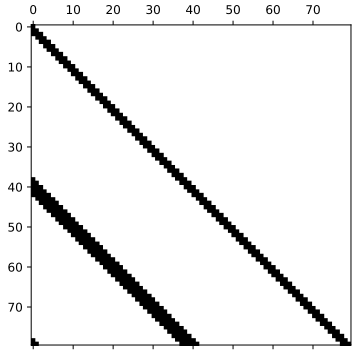
Map sparsity ansatz



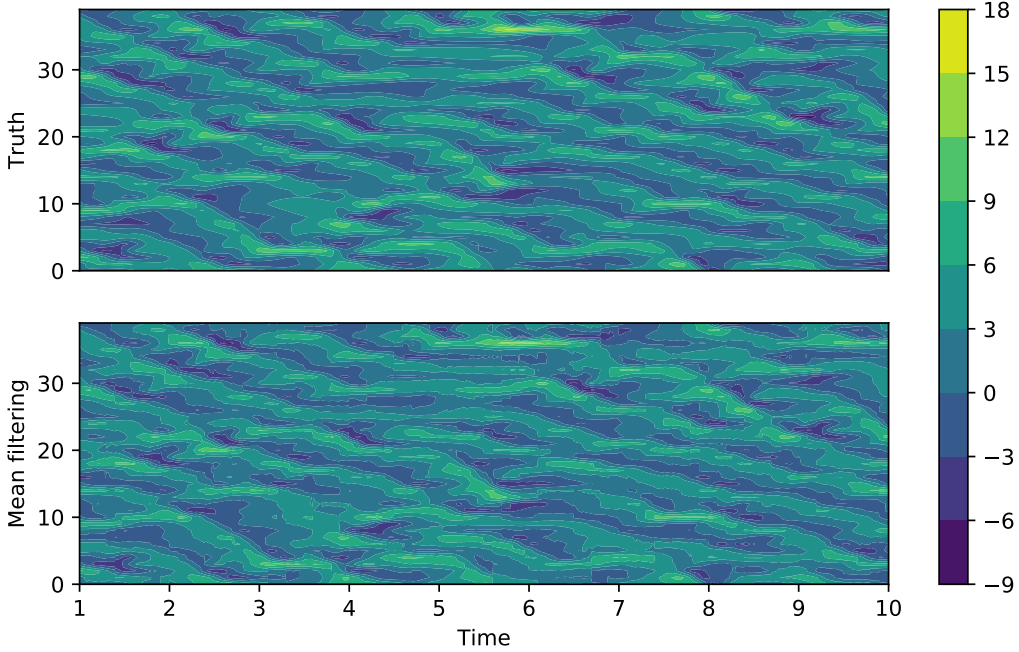


Map sparsity ansatz

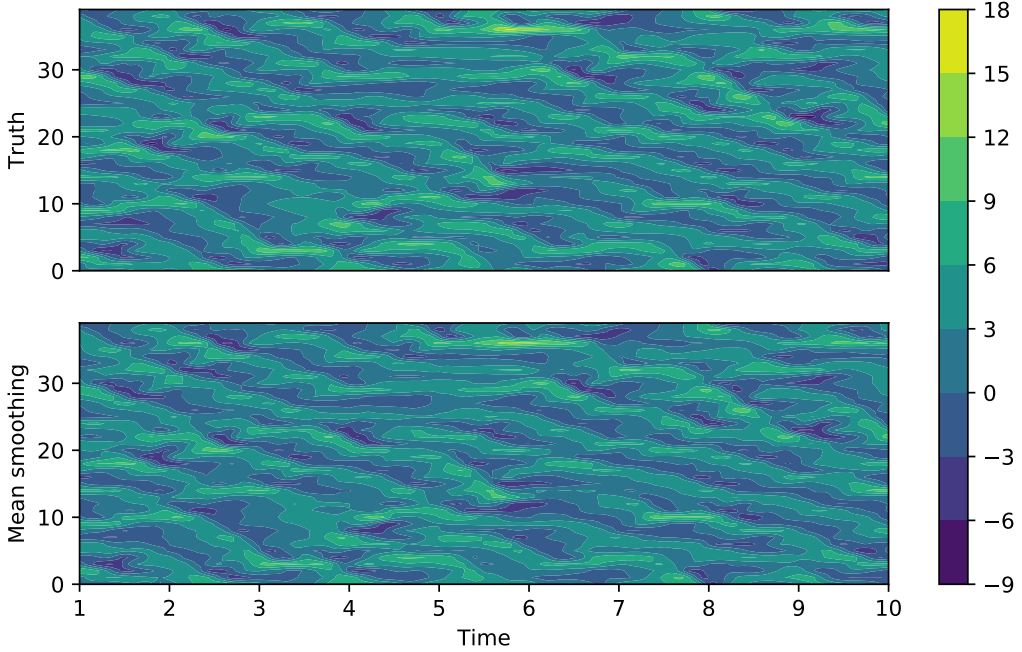




Filtering


Truth
Mean filtering
0
10
20
30
1
2
3
4
5
6
7
8
9
10
Time
18
15
12
9
6
3
0
−3
−6
−9

Smoothing


Truth
Mean smoothing
0
10
20
30
1
2
3
4
5
6
7
8
9
10
Time
18
15
12
9
6
3
0
−3
−6
−9

		Order 1	
		Filtering	Smoothing ($d = 4 \times 10^4$)
at obs.	Avg. RMSE	8.35×10^{-1}	7.40×10^{-1}
	Med. RMSE	8.11×10^{-1}	7.29×10^{-1}
	Var. RMSE	3.26×10^{-2}	2.42×10^{-2}
incl. pred.	Avg. RMSE	9.20×10^{-1}	7.34×10^{-1}
	Med. RMSE	9.01×10^{-1}	7.25×10^{-1}
	Var. RMSE	4.21×10^{-2}	2.51×10^{-2}

Key contributions

Algorithms for filtering, smoothing, and parameters estimation
via **deterministic couplings** and **optimization**.

Contact: Daniele Bigoni – **dabi@mit.edu**

Software: `https://transportmaps.mit.edu`

References:

Bigoni et al. "Adaptive construction of measure transports for Bayesian inference"
Spantini et al. "Inference via low-dimensional couplings"
Marzouk et al. "An introduction to sampling via measure transport"
Parno et al. "Transport map accelerated Markov chain Monte Carlo"
El Moselhy et al. "Bayesian inference with optimal maps"

Thanks to:

U.S. DEPARTMENT OF ENERGY

DARPA