Learning Partially Known Stochastic Dynamics with Empirical PAC Bayes

•••

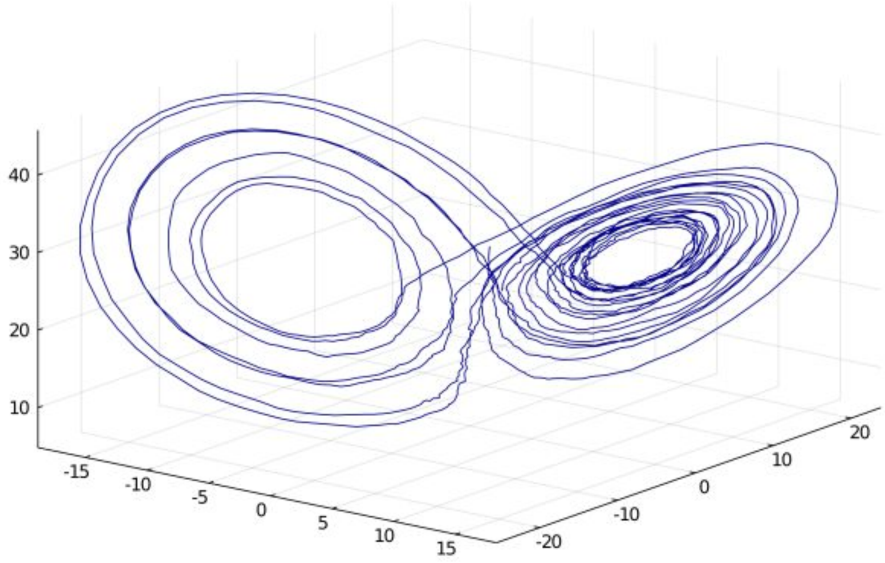
Manuel Haussmann Sebastian Gerwinn Andreas Look
Barbara Rakitsch Melih Kandemir

(Published at AISTATS 2021)

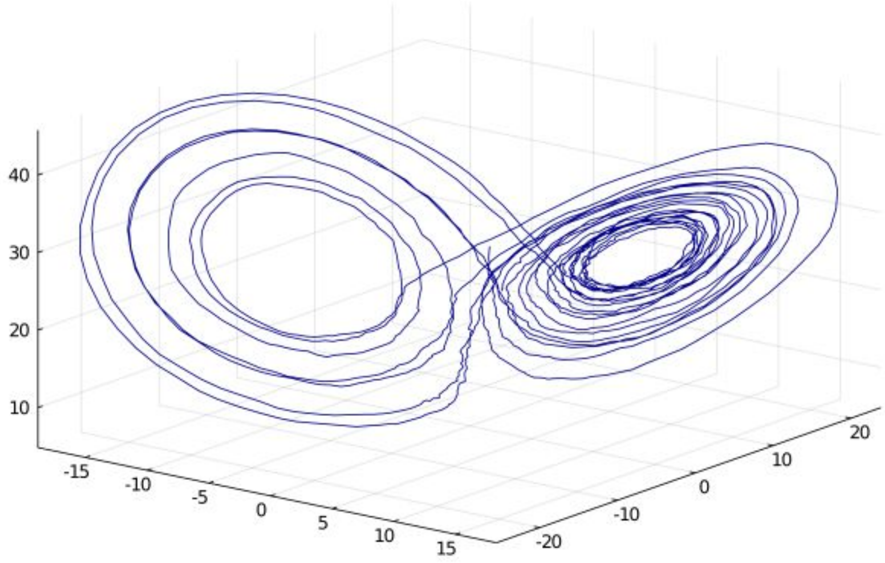
UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

BOSCH
Invented for life

Given observed measurements…



Given observed measurements…



… we want to learn the underlying dynamical system

Two complementary approaches

- The domain expert creates a system of differential equations
  - ✔ interpretable
  - ✔ can incorporate prior domain knowledge
  - ✔ few parameters
  - ✔ probably data efficient
  - ✗ not very flexible
  - ✗ what to do if only partial knowledge is available?

WHITE BOX ODE

$$d\mathbf{h}_t = r(\mathbf{h}_t, t)dt$$

Two complementary approaches

- The domain expert creates a system of differential equations
  - ✔ interpretable
  - ✔ can incorporate prior domain knowledge
  - ✔ few parameters
  - ✔ probably data efficient
  - ✗ not very flexible
  - ✗ what to do if only partial knowledge is available?

White Box ODE

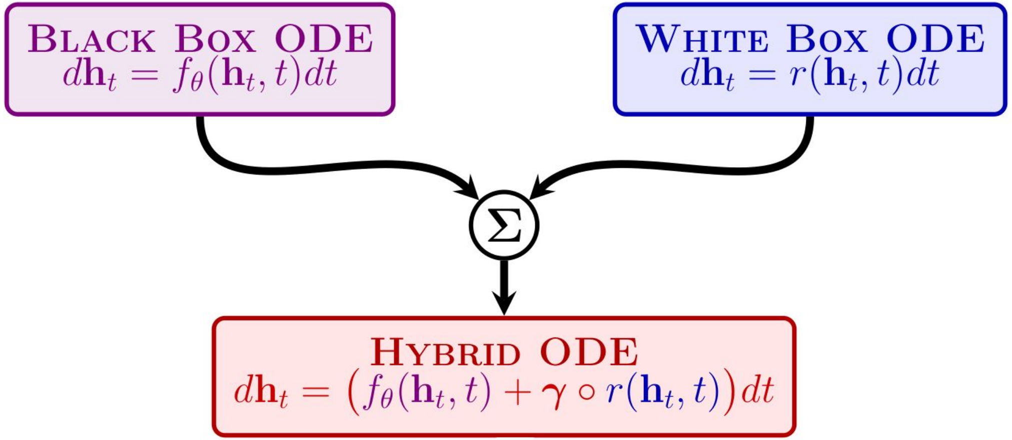
$$d\mathbf{h}_t = r(\mathbf{h}_t, t)dt$$

- The data scientist relies on deep learning and creates a neural ODE
  - ✔ very flexible
  - ✔ can adapt itself to complex relationship
  - ✗ lots of parameters
  - ✗ data hungry

Black Box ODE

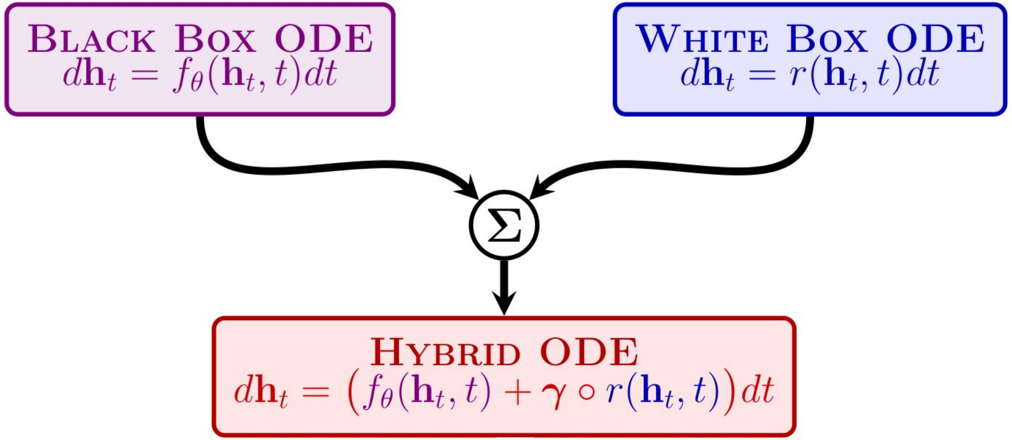
$$d\mathbf{h}_t = f_\theta(\mathbf{h}_t, t)dt$$

Can we combine the two?



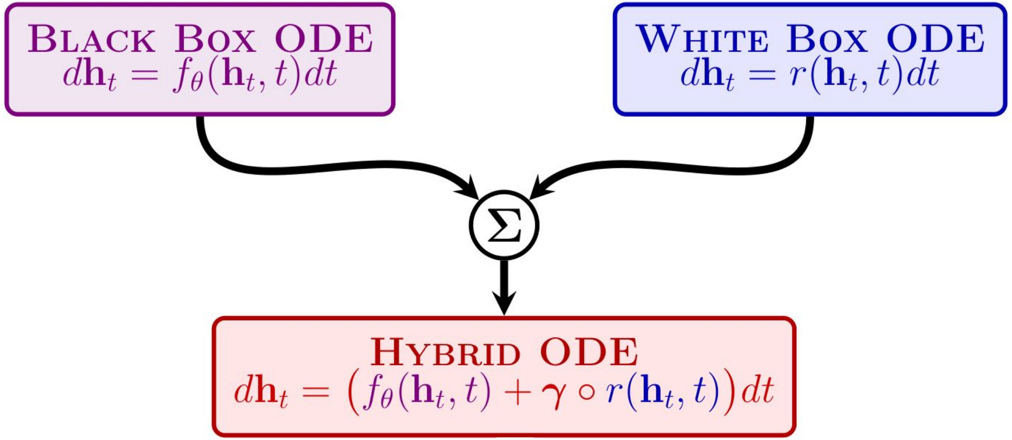
- Get the best of both worlds with in a hybrid model (with $\gamma \in \{0, 1\}^D$)

Can we combine the two?



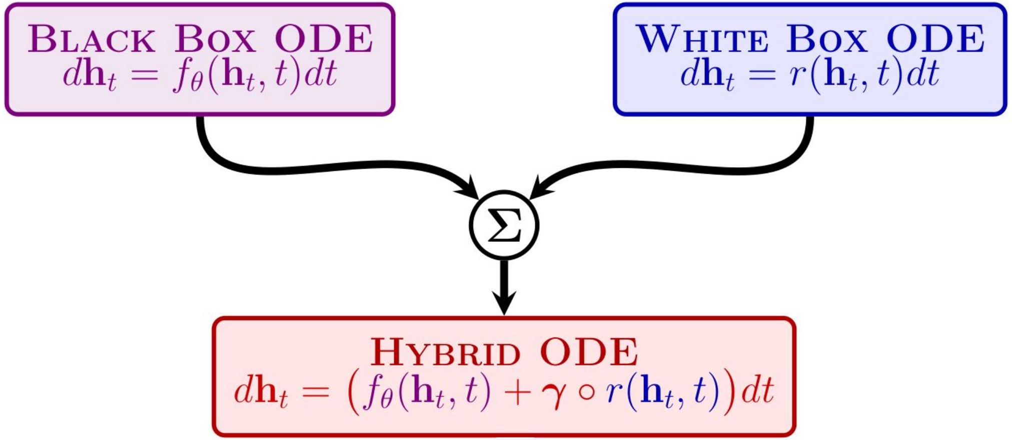
- Get the best of both worlds with in a hybrid model (with $\gamma \in \{0, 1\}^D$)
- Can we go one step further?

Can we combine the two?



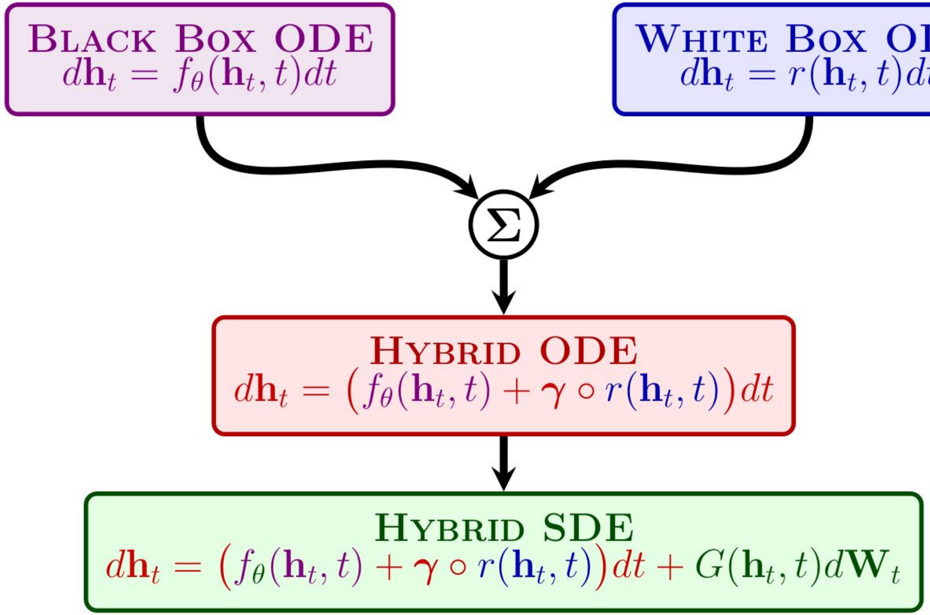
- Get the best of both worlds with in a hybrid model (with $\gamma \in \{0, 1\}^D$)
- Can we go one step further?
  - Model parameter (epistemic) uncertainty by switching to a BNN

Can we combine the two?



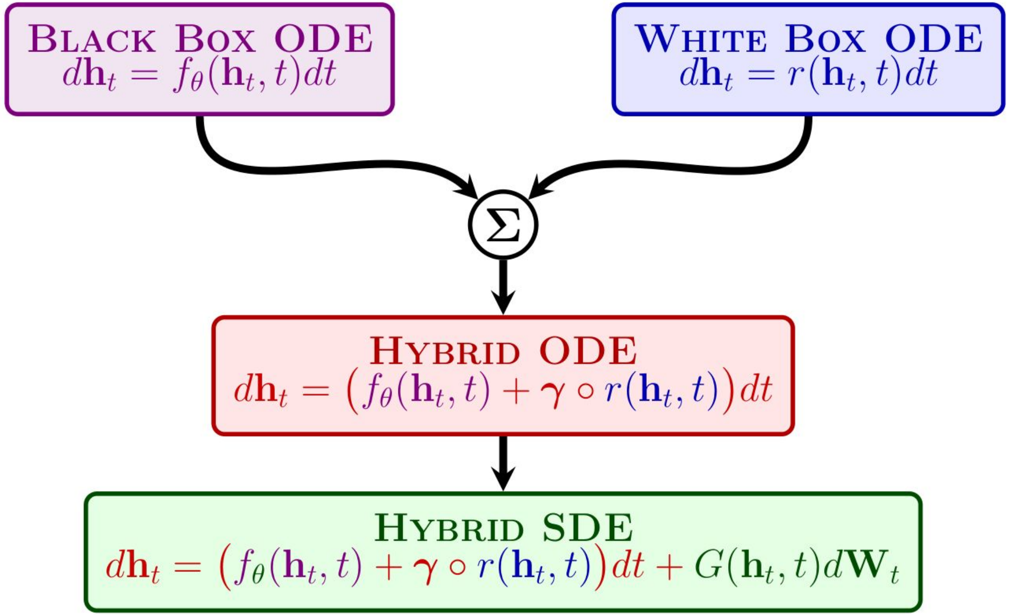
- Get the best of both worlds with in a hybrid model (with $\gamma \in \{0,1\}^D$)
- Can we go one step further?
  - Model parameter (epistemic) uncertainty by switching to a BNN
  - Model (aleatoric) uncertainty by switching to an SDE

The complete pipeline



✔ flexibility & adaptability ✔ domain knowledge ✔ uncertainty

The complete pipeline



✔ flexibility & adaptability ✔ domain knowledge ✔ uncertainty

Great, but... How do I actually learn such a thing?

Step 1: Euler-Murayama discretization (see e.g. Särkkä and Solin, 2019)

Problem 1: How to deal with the SDE?

Hybrid SDE

$$d\mathbf{h}_t = \big(f_\theta(\mathbf{h}_t, t) + \boldsymbol{\gamma} \circ r(\mathbf{h}_t, t)\big)dt + G(\mathbf{h}_t, t)d\mathbf{W}_t$$

- W_t denotes a Wiener process, G gives the diffusion dynamics
- Discretization of the SDE into K steps gives us

$$\mathbf{h}_{t_{k+1}} = \mathbf{h}_{t_k} + \big(f_\theta(\mathbf{h}_{t_k}, t_k) + \boldsymbol{\gamma} \circ r(\mathbf{h}_{t_k}, t_k)\big)\Delta t_k + G(\mathbf{h}_{t_k}, t_k)\Delta W_k$$

$$\Delta W_k \sim \mathcal{N}(0, \Delta t_k \mathbb{1}_P) \qquad \Delta t_k := t_{k+1} - t_k$$

Step 1.5: Design a generative pipeline

- Our generative pipeline is then given as

$$
\begin{aligned}
\theta &\sim p_\phi(\theta) \\
\mathbf{h}_0 &\sim p(\mathbf{h}_0) \\
\mathbf{h}_{k+1}|\mathbf{h}_k, \theta &\sim \mathcal{N}\big(\mathbf{h}_{k+1}|\mathbf{h}_k + d(\mathbf{h}_k, t_k)\Delta t_k, \Sigma_k\big), \qquad k = 0, \dots, K-1 \\
&\text{where} \quad d(\mathbf{h}_k, t_k) := f_\theta(\mathbf{h}_k, t_k) + \boldsymbol{\gamma} \circ r(\mathbf{h}_k, t_k) \\
&\qquad\qquad \Sigma_k := G(\mathbf{h}_k, t_k)G(\mathbf{h}_k, t_k)^\top \Delta t_k \\
&\qquad\qquad \Delta t_k := t_{k+1} - t_k \\
\mathbf{y}_k|\mathbf{h}_k &\sim p(\mathbf{y}_k|\mathbf{h}_k), \qquad k = 1, \dots, K
\end{aligned}
$$

Where we have observed N trajectories $\mathcal{D} = \{\mathbf{Y}_1, \dots, \mathbf{Y}_N\}$ with $\mathbf{Y}_n = \{\mathbf{y}_1^n, \dots, \mathbf{y}_K^n\}$

- Classical approach: Infer a posterior over local and global parameters

$$
p(\theta, \mathbf{H}_1, \dots, \mathbf{H}_N|\mathcal{D})
$$

Step 2: Empirical Bayes / Type II Maximum Likelihood

- Problem 2: Infer a posterior over local and global parameters

$$p(\theta, \mathbf{H}_1, \ldots, \mathbf{H}_N | \mathcal{D})$$

- How?
  - MCMC? Becomes very quickly too expensive computationally
  - Variational Inference? Becomes too restrictive due to strong independence assumptions

- Instead, Type II Maximum Likelihood

$$\hat{\phi} = \arg\max_{\phi} \int p(\mathcal{D}|\mathbf{H}) p(\mathbf{H}|\theta) p_{\phi}(\theta) \, \mathrm{d}\mathbf{H} \, \mathrm{d}\theta$$

$$\approx \arg\max_{\phi} \log \left(\frac{1}{S} \sum_{s=1}^{S} p(\mathcal{D}|\mathbf{H}^s) \right) \qquad \text{where } \mathbf{H}^s \sim p(\mathbf{H}|\theta^s), \; \theta^s \sim p_{\phi}(\theta)$$

Step 3: PAC-based regularization

- Problem 3: We now have a tractable approach, but we lost a lot of regularization
- A very quick primer on PAC-Bayes (see e.g. McAllester, 1999,2003)
    - A risk of a hypothesis h given a loss l is $R(h) = \mathbb{E}_x \left[l(x, h(x)) \right]$
    - The empirical counterpart is given as $R_{\mathcal{D}}(h) = \frac{1}{|D|} \sum_{x \in \mathcal{D}} l(x, h(x))$
    - Given a distribution over hypothesis Q and a prior distribution P

$$\mathbb{P}\Big(\forall Q : \mathbb{E}_{h \sim Q} \left[R(h) \right] \leq \mathbb{E}_{h \sim Q} \left[R_{\mathcal{D}}(h) \right] + \mathcal{C}(P, Q, \delta, N) \Big) \geq 1 - \delta$$

Step 3: PAC-based regularization

We place distributions over the hybrid and a prior process

$$Q_{0\to T}(\mathbf{h}_{0\to T}, \theta) = p_{\text{hyb}}(\mathbf{h}_{0\to T}|\theta)p_{\phi}(\theta)$$
$$P_{0\to T}(\mathbf{h}_{0\to T}, \theta) = p_{\text{pri}}(\mathbf{h}_{0\to T})p_{\text{pri}}(\theta)$$

and define the risk as

$$R(H) \triangleq \mathbb{E}_{\boldsymbol{Y}_k \sim \mathfrak{G}(t)} \left[1 - \frac{1}{\overline{B}_K} \prod_{k=1}^{K} p(\boldsymbol{y}_k | \boldsymbol{h}_k) \right]$$

$$\overline{B}_K \triangleq \max_{\boldsymbol{y}_k, \boldsymbol{h}_k} \prod_{k=1}^{K} p(\boldsymbol{y}_k | \boldsymbol{h}_k) \leq \left(\max_{\boldsymbol{y}_k, \boldsymbol{h}_k} p(\boldsymbol{y}_k | \boldsymbol{h}_k) \right)^K$$

Step 3: PAC-based regularization

We place distributions over the hybrid and a prior process

$$Q_{0\to T}(\mathbf{h}_{0\to T}, \theta) = p_{\text{hyb}}(\mathbf{h}_{0\to T}|\theta)p_\phi(\theta)$$
$$P_{0\to T}(\mathbf{h}_{0\to T}, \theta) = p_{\text{pri}}(\mathbf{h}_{0\to T})p_{\text{pri}}(\theta)$$

and can then show (Theorem 1) that

$$\text{with } \mathbb{P} \geq 1-\delta \qquad \mathbb{E}_{H\sim Q_{0\to T}}\left[R(H)\right] \leq \mathbb{E}_{H\sim Q_{0\to T}}\left[R_{\mathcal{D}}(H)\right] + \mathcal{C}_\delta(Q_{0\to T}, P_{0\to T})$$

Ultimate objective is then

$$\hat{\phi} := \arg\max_{\phi} \underbrace{\frac{1}{SN}\sum_{n=1}^{N}\sum_{s=1}^{S}\sum_{k=1}^{K}\log\left(p(\mathbf{y}_k^n|\mathbf{h}_k^{n,s})\right)}_{\text{approx marginal likelihood}} + \underbrace{\sqrt{\frac{\text{KL}\left(Q_{0\to T} \parallel P_{0\to T}\right) + \log(4\sqrt{N}/\delta)}{2N}}}_{\text{regularizer}}$$

Theoretical Summary

- Step 0: Design a hybrid SDE model that allows for knowledge and ignorance
- Step 1: Turn into a probabilistic model via discretization
- Step 2: Empirical Bayes/Type II ML to get a principled, tractable objective (v1.0)
- Step 3: PAC-Bayes to get an improved principled objective (v2.0)

Results: Lorenz Attractor as an SDE

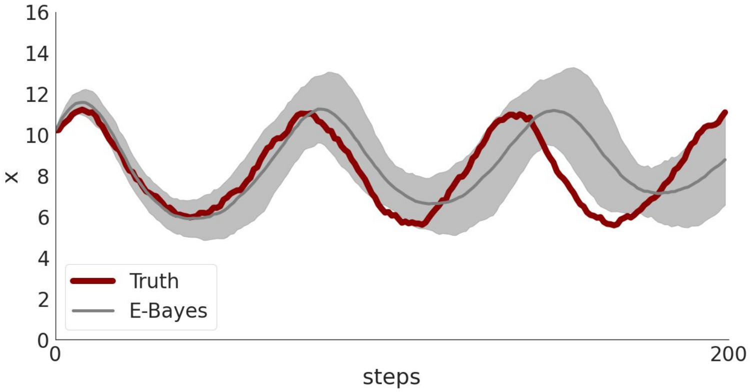
The underlying SDE

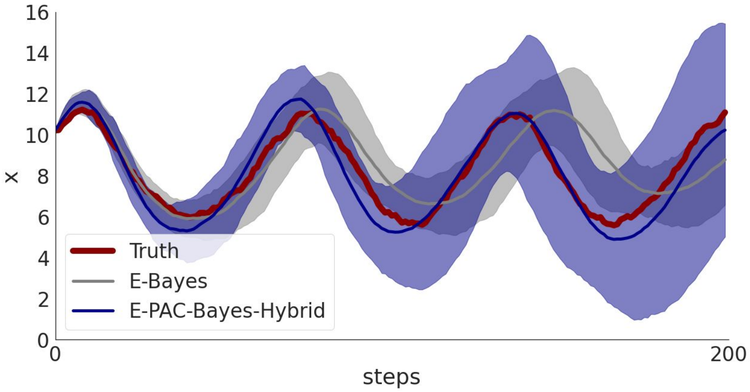
$$
\begin{aligned}
dx_t &= \zeta(y_t - x_t)\,dt + dW_t, \\
dy_t &= \big(x_t(\kappa - z_t) - y_t\big)\,dt + dW_t, \\
dz_t &= (x_t y_t - \rho z_t)\,dt + dW_t,
\end{aligned}
$$

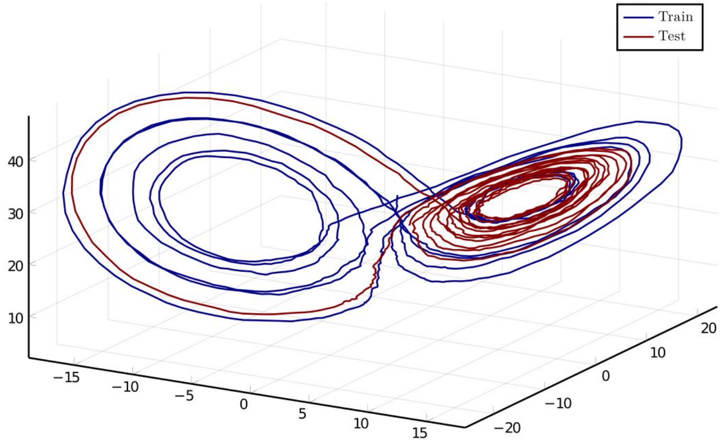
Prior Knowledge	**Model**	**Test MSE**
None	(i)	$29.20_{\pm 0.19}$
	(ii)	$29.05_{\pm 0.23}$
$\gamma = [1, 0, 0], \quad \zeta \sim \mathcal{N}(10, 1)$	(iii)	$27.58_{\pm 0.17}$
	(iv)	$27.42_{\pm 0.16}$
$\gamma = [0, 1, 0], \quad \kappa \sim \mathcal{N}(28, 1)$	(iii)	$15.87_{\pm 0.46}$
	(iv)	$15.06_{\pm 0.35}$
$\gamma = [0, 0, 1], \quad \rho \sim \mathcal{N}(2.67, 1)$	(iii)	$27.82_{\pm 0.26}$
	(iv)	$28.37_{\pm 0.21}$
$\gamma = [1, 1, 1],$ $(\zeta, \kappa, \rho)^\top \sim \mathcal{N}\big((10, 28, 2.67)^\top, \mathbb{1}_3\big)$	(v)	$16.40_{\pm 2.31}$

(i) Empirical Bayes (EB) based without prior knowledge
(ii) EB with PAC regularization
(iii) EB with prior knowledge
(iv) EB with prior knowledge and PAC regularization
(v) A model with full prior knowledge over all three equations

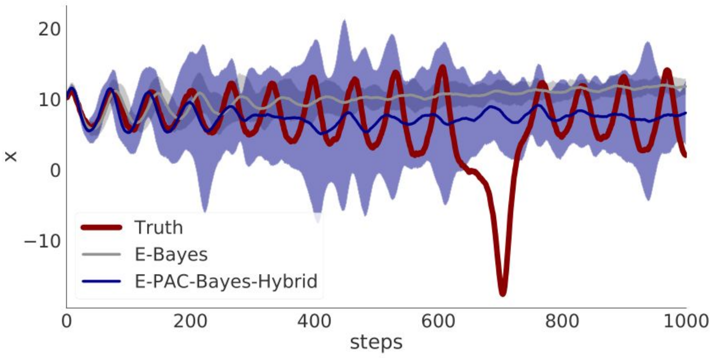
Results: Lorenz Attractor as an SDE

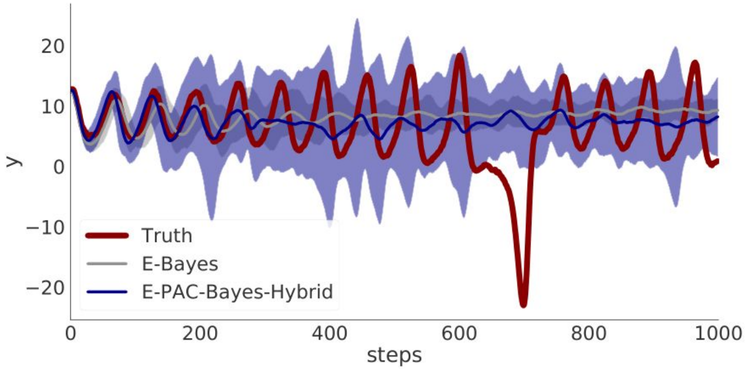


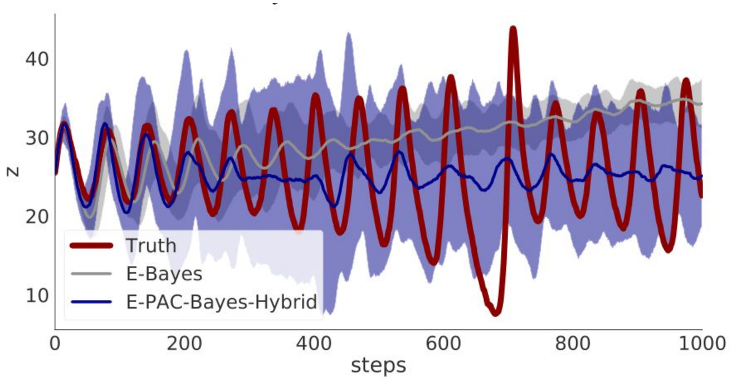




Results: Lorenz Attractor as an SDE

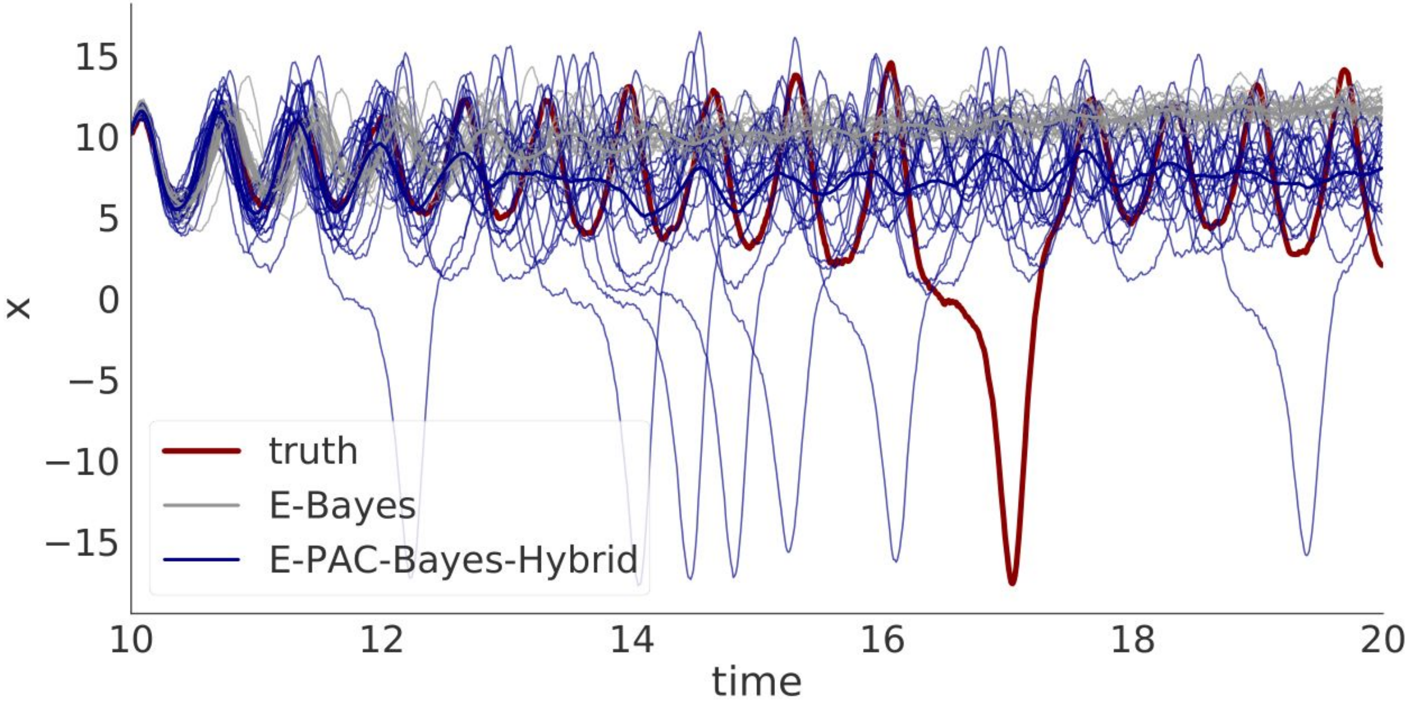




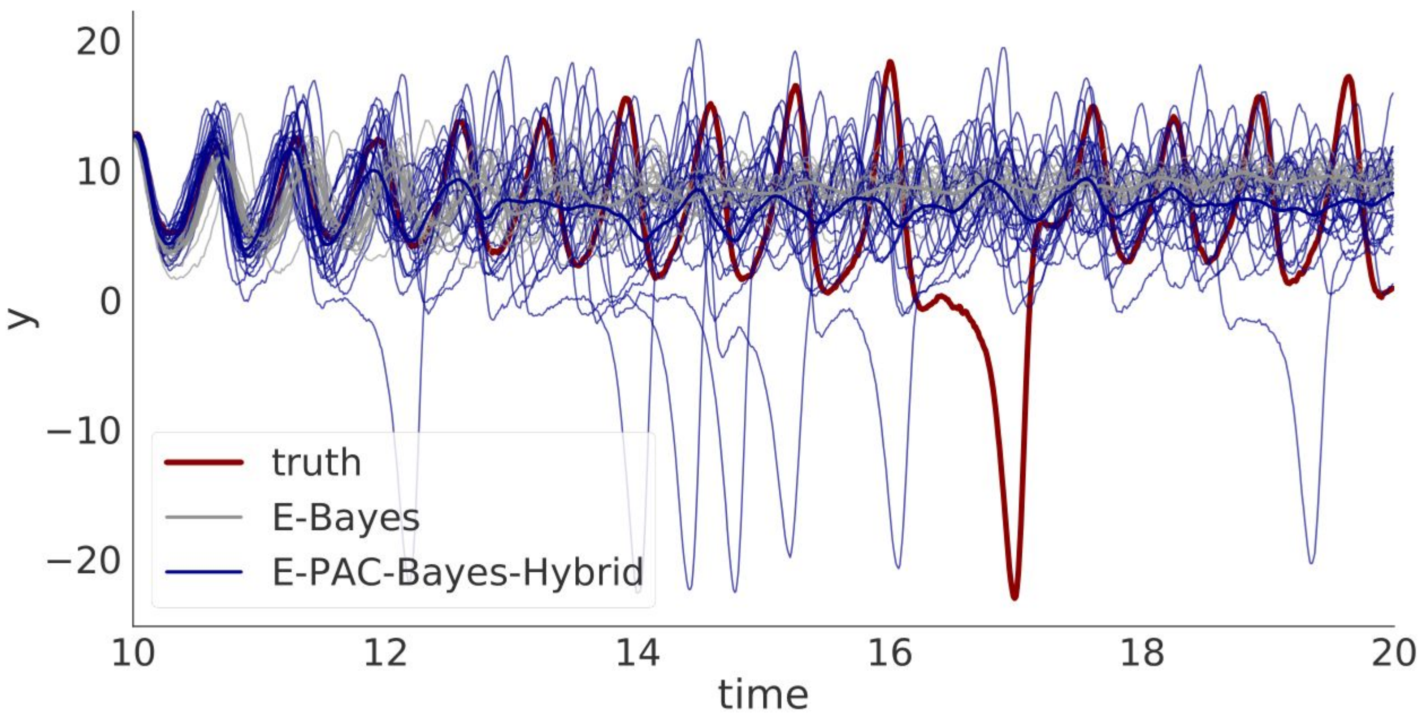


- Noisy knowledge over dZ
- Std over 21 trajectories

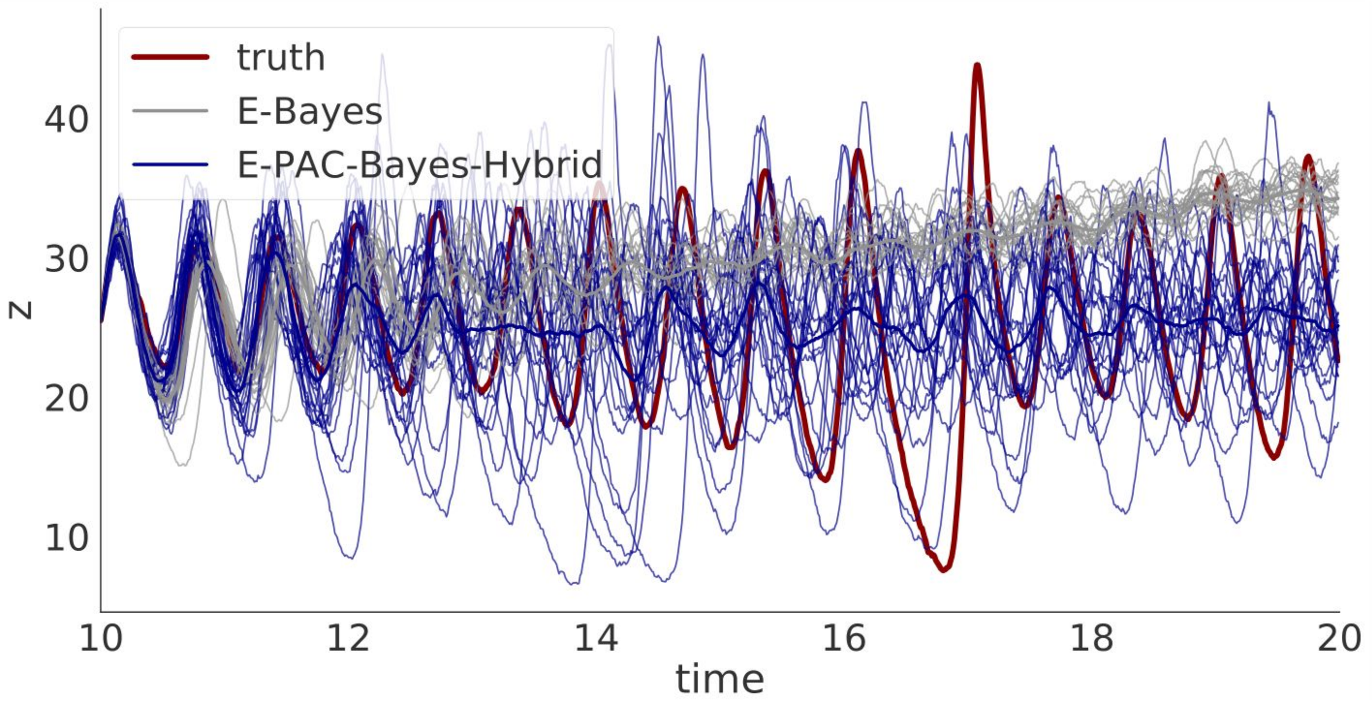
Results: Lorenz Attractor as an SDE



Results: Lorenz Attractor as an SDE



Results: Lorenz Attractor as an SDE



Results: CMU Motion Capture Data (Sequences of human walking dynamics)

Method	**Test MSE**	**Test NLL**
DTSBN-S (Gan et al., 2015)	34.86±0.02	Not Applicable
npODE (Heinonen et al., 2018)	22.96	Not Applicable
Neural-ODE (Chen et al., 2018a)	22.49±0.88	Not Applicable
ODE^2VAE (Yildiz et al., 2019)	10.06±1.40	Not Reported
ODE^2VAE-KL (Yildiz et al., 2019)	8.09±1.95	Not Reported
D-BNN (SGLD) (Look and Kandemir, 2019)	13.89±2.56	747.92±58.49
D-BNN (VI) (Hegde et al., 2019)	9.05±2.05	452.47±102.59
(i) E-Bayes	8.68±1.56	433.76±77.78
(ii) E-PAC-Bayes	9.17±1.20	489.82±67.06
(iii) E-Bayes-Hybrid	9.25±1.99	462.82±99.61
(iv) E-PAC-Bayes-Hybrid	**7.84±1.41**	**415.38±80.37**

Thank you for listening!

Please see arxiv.org/abs/2006.09914 for further details

Feel free to contact me at manuel.haussmann@iwr.uni-heidelberg.de with further questions