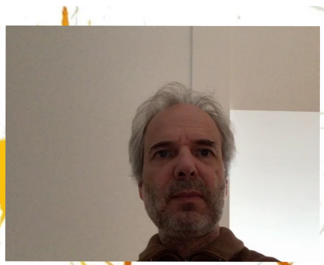


TIME SERIES ANALYSIS & RECURRENT NEURAL NETWORKS #12

- Sequential Variational Autoencoders
- Re-parameterization trick & "Stochastic Gradient Variational Bayes"

Main lecture: Daniel Durstewitz

Exercises: Leonard Bereska, Manuel Brenner, Daniel Kramer, Georgia Koppe

Heidelberg University

Generative models

$p_\theta(X, Z)$

$X = \{x_1 \dots x_T\}$, e.g. $x_t \in \mathbb{R}^N$

$p(z_t) = N(\underbrace{\phi(W z_{t-1} + h)}_{RNN}, \Sigma)$ latent process / DS

$p(x_t | z_t) = N(B z_t, \Gamma)$

EM: $p(Z|X)$

LDS: $q(Z|X) = p(Z|X)$

nonlin.: approx. EKF (Taylor expan.), Laplace $\Big\}\ O(T \cdot M^{2-3})$

partial filters: consistent estim. of $p(Z|X)$

→ big data?

→ Variational Inference (VI) & SGD

$p_\theta(X, Z)$ → max. likelihood

$\log p_\theta(X) = \log \int_Z p_\theta(X, Z)\, dZ$

$\geq \int_Z q(Z|X) \log \frac{p_\theta(X,Z)}{q(Z|X)}\, dZ =: ELBO[q, \theta]$

$= \log p_\theta(X) - KL[q(Z|X) \,\|\, p_\theta(Z|X)]$

$= E_q[\log \underbrace{p_\theta(X,Z)}_{p(X|Z)p(Z)}] - E_q[q(Z|X)]$ $= H[q]$

$= \underbrace{E_q[\log p_\theta(X|Z)]}_{\text{"fitting the obs."}} - \underbrace{KL[q(Z|X) \| p(Z)]}_{\text{"getting latent model right"}}$

$q^* = \arg\min_{q \in \mathcal{D}} KL[q(Z|X) \,\|\, p_\theta(Z|X)]$

→ Blei et al. (2018) arXiv

→ assume parameterized family $q_\xi(Z|X)$ parameterized variational density

$\{\xi^*, \theta^*\} = \arg\max_{\xi, \theta} ELBO[\xi, \theta]$

variational param. (ξ), generative model param. (θ)

- reasonably simple
- should be expressive!

Examples: $q_\xi(Z|X) = \sum_{k=1}^{K} \pi_k N(\mu_k, \Sigma_k)$, $\sum \pi_k = 1$

normalizing flows (Rezende & Mohamed 2015 ICML)

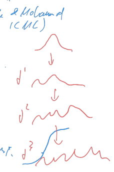
$y_0 \sim q_0 = N(0, I)$

$f^4(f^3(f^2(f^1(y_0)))) \dots$

$y_1 = f(y_0) \,|\, f$: smooth, invertible

$q_1(y_1) = q_0(y_0) \left| \det \frac{\partial f^{-1}}{\partial y_1} \right|$

$f = \sigma(\underbrace{W y_0 + h}_{\text{affine func.}})$ ← need to be invert.



Seq. Var. Auto-Encoders (seq-VAE)

→ Kingma & Welling (2014), Rezende et al. (2014)

terminology

generative model: $p_\theta(X, Z)$, $\theta = \{\theta_{lat}, \theta_{obs}\}$

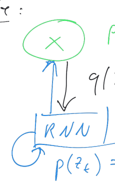
prior / latent model: $p_{\theta_{lat}}(Z)$

"decoder" model: $p_{\theta_{obs}}(X|Z)$

"encoder" model / "recognition": $q_\xi(Z|X)$ var. dens. / post. approx.

$\boxed{X} \xrightarrow[q_\xi(Z|X)]{\text{"encode"}} \boxed{Z} \xrightarrow[p_{\theta_{obs}}(X|Z)]{\text{"decode"}} \boxed{X'}$

Example:



$p(x_t | z_t) = N[\mu_t^{(x)}, \Sigma_t^{(x)}]$

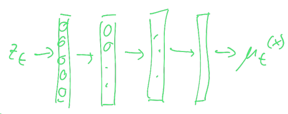
$q(z_t | x_t) = N[\mu_t^{(z)}, \Sigma_t^{(z)}]$

$p(z_t) = N[\mu_{t|t-1}, \Sigma_{t|t-1}]$

$\mu_{t|t-1} := \phi(W z_{t-1} + h)$, $\Sigma_{t|t-1} = \Sigma$

$\mu_t^{(x)} = B z_t$, $\Sigma_t^{(x)} = \Gamma$

$\mu_t^{(x)} = MLP[z_t]$, $CNN[z_t]$

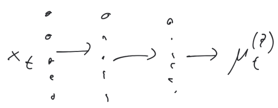


$\Sigma_t^{(x)}$ sym., positive-semidefinite

$\Sigma_t^{(x)} = diag[\mu_t^{(x)2}]$, $diag[\exp(\mu_t^{(x)})]$

$\Sigma_t^{(x)} = V V^T$, $V = MLP(z_t)$

$q_\xi(Z|X)$, $\mu_t^{(z)} = MLP(x_t)$



$ELBO[\xi, \theta] = E_{q(Z|X)}[\log p_\theta(X, Z)] + \underbrace{H_{q(Z|X)}[q(Z|X)]}_{\frac{MT}{2} + \frac{MT}{2}\ln(2\pi) + \frac{1}{2}\ln|\Sigma|}$ (latent states)

$= E_q[\log p_\theta(X|Z)] - \underbrace{KL[q_\xi(Z|X) \| p_\theta(Z)]}_{\frac{1}{2}\left[tr(\Sigma_p^{-1}\Sigma_q) + (\mu_p - \mu_q)^T \Sigma_p^{-1}(\mu_p - \mu_q) - MT + \ln\frac{|\Sigma_p|}{|\Sigma_q|}\right]}$

Example: Param. of cov. of posterior

Archer ... Paninski (2016) arXiv

$\Sigma_q = -H^{-1}$

$$H = \begin{array}{c|cccccc} & z_1 & z_2 & \dots & & & z_T \\ \hline z_1 & D_1 & B_1^T & & & & 0 \\ z_2 & B_1 & D_2 & B_2^T & & & \\ & & B_t & & & & \\ & & & & D_t & & \\ & & & & & \ddots & \\ & 0 & & & & & \ddots \end{array}$$

$\dfrac{z_t (\dots) z_{t-1}^T}{z_t (\dots) z_t^T}$

$D_t = MLP_{\xi_D}(x_t)$

$B_t = MLP_{\xi_B}(x_t, x_{t-1})$

$\to q_\xi(Z|X) = N(\mu^{(z)}, \Sigma^{(z)})$

Stochastic Gradient Variational Bayes

→ Rezende et al. (2014), Kingma & Welling (2014)

$$\mathrm{ELBO}[\xi,\theta] = E_q[\log p_\theta(x,z)] + H_q[q_\xi(z|x)]$$

$$= E_{q_\xi(z|x)}[\log p_\theta(x,z) - \log q_\xi(z|x)]$$

$$\approx \underset{\uparrow \, \#\text{samples}}{\frac{1}{L}} \sum_{l=1}^{L} [\log p_\theta(x,z^{(l)}) - \log q_\xi(z^{(l)}|x)]$$

$$z^{(l)} \sim q_\xi(z|x)$$

Grad-descent $\nabla_{\xi,\theta}\, \mathrm{ELBO}[\xi,\theta]$

$$\nabla_\xi \mathrm{ELBO} \neq \frac{1}{L} \sum \underbrace{\nabla_\xi \log p_\theta(x,z^{(l)})}_{0} - \nabla_\xi \log q_\xi(z^{(l)}|x)$$

$$\nabla_\xi \int_z \underbrace{q_\xi(z|x)}_{=:q} [\log \underbrace{p_\theta(x,z)}_{=:p} - \log q_\xi(z|x)]\, dz$$

$$= \int_z \nabla_\xi [q \cdot \log p] - \nabla_\xi [q \cdot \log q]\, dz$$

$$[\nabla q] = q \nabla \log q$$

$$= \int_z q(\nabla \log q) \log p - q(\nabla \log q)\log q - \overbrace{q \nabla \log q}^{\nabla q}\, dz$$

$$\int \nabla q\, dz = \nabla \underbrace{\int q\, dz}_{1} = 0$$

$$= \int_z q(\nabla \log q)(\log p - \log q)\, dz$$

$$= E_q[(\nabla \log q)(\log p - \log q)]$$

score gradient

$$\approx \frac{1}{L} \sum_l (\nabla \log q)(\log p - \log q), \quad z \sim q$$

reparameterization trick

$$z^{(l)} = g(\xi, \varepsilon)$$

e.g. $z^{(l)} \sim N(\underbrace{\mu, \sigma^2}_{\xi})$

$$\Rightarrow z^{(l)} = \underbrace{\mu + \sigma}_{\xi} \varepsilon^{(l)}, \quad \varepsilon^{(l)} \sim N(0,1)$$

$$z^{(l)} = \mu + R\, \varepsilon^{(l)}, \quad \varepsilon^{(l)} \sim N(0, I)$$

$$\Sigma = R R^T$$

$$\mu = f(x) = \mathrm{MLP}(x) \text{ e.g.}$$

→ for all expon. family distrib.

→ Gumbel reparam.-trick for categ.-distrib.

$$\mathrm{ELBO}[\xi,\theta] \approx \frac{1}{L} \sum_{l=1}^{L} \log p_\theta(x, z^{(l)}) - \log q_\xi(z^{(l)}|x)$$

$$z^{(l)} \sim q_\xi(z|x)$$

$$= \frac{1}{L} \sum_{l=1}^{L} \log p_\theta(x, g(\xi, \varepsilon^{(l)})) - \log q_\xi(g(\xi, \varepsilon^{(l)}), x)$$

$$\varepsilon^{(l)} \sim N(0, I)$$

$$\rightarrow \nabla \mathrm{ELBO} = \nabla\left[\frac{1}{L} \sum \ldots\right]$$

in practice: $L \approx 1$

$$E_q[h(z)] = E_\varepsilon[h(g(\varepsilon))]$$

Proof:

I) Transformation theorem for func. of RV

Let ε be RV and $z = g(\varepsilon)$ an invertible (strictly monotonic) function

cumulative distrib.

$$F_z(z_0) = p(z \le z_0) = p(g(\varepsilon) \le z_0)$$

monoton. incr.

$$= p(g^{-1}(g(\varepsilon)) \le g^{-1}(z_0))$$

monoton. decr.: $(= p(\ldots \ge \ldots))$

$$= p(\varepsilon \le g^{-1}(z_0)) = F_\varepsilon(g^{-1}(z_0))$$

$$\Rightarrow \boxed{\frac{\partial F_z}{\partial z} = f_z(z) = f_\varepsilon(g^{-1}(z)) \cdot \left|\det \underbrace{\frac{\partial g^{-1}(z)}{\partial z}}_{\text{Jacobian}}\right|}$$

II) $$E_z[h(z)] = \int_z f_z(z)\, h(z)\, dz$$

$$= \int_\varepsilon f_\varepsilon(g^{-1}(z)) \cancel{\left|\frac{\partial g^{-1}(z)}{\partial z}\right|} h(z) \cancel{\left|\underbrace{\frac{\partial z}{\partial \varepsilon}}_{\partial z / \partial g^{-1}(z)}\right|} d\varepsilon$$

$$= \int_\varepsilon f_\varepsilon(\varepsilon)\, h(g(\varepsilon))\, d\varepsilon = E_\varepsilon[h(g(\varepsilon))]$$

$\square$

$$z = \mu + \sigma\varepsilon, \quad \varepsilon \sim N(0,1)$$

$$\Rightarrow \varepsilon = \frac{z - \mu}{\sigma}$$

Reconstructing dynamical systems

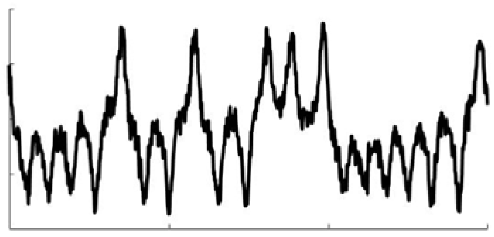
$p_\theta(x,z)$

$p(z_t) = \mathcal{N}(\phi(W z_{t-1} + h), \Sigma)$

Lorenz system

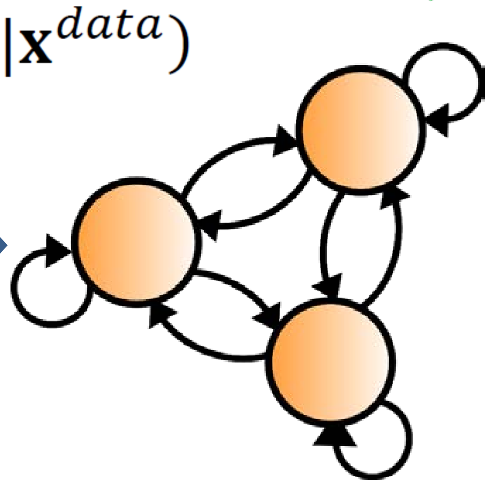
$$\mathbf{dx_1 / dt = s(x_2 - x_1)}$$
$$\mathbf{dx_2 / dt = rx_1 - x_2 - x_1x_3}$$
$$\mathbf{dx_3 / dt = x_1x_2 - bx_3}$$

X



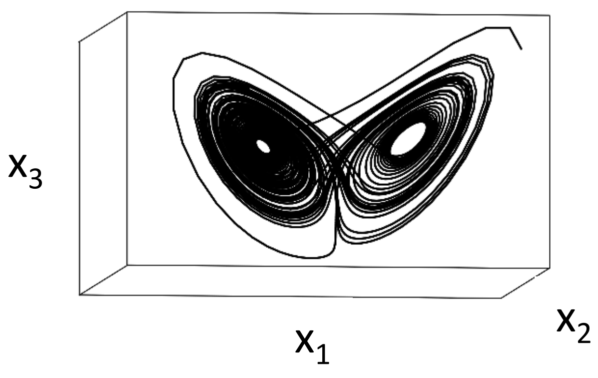
Inference

$p_{inf}(\mathbf{z}, \boldsymbol{\theta} | \mathbf{x}^{data})$



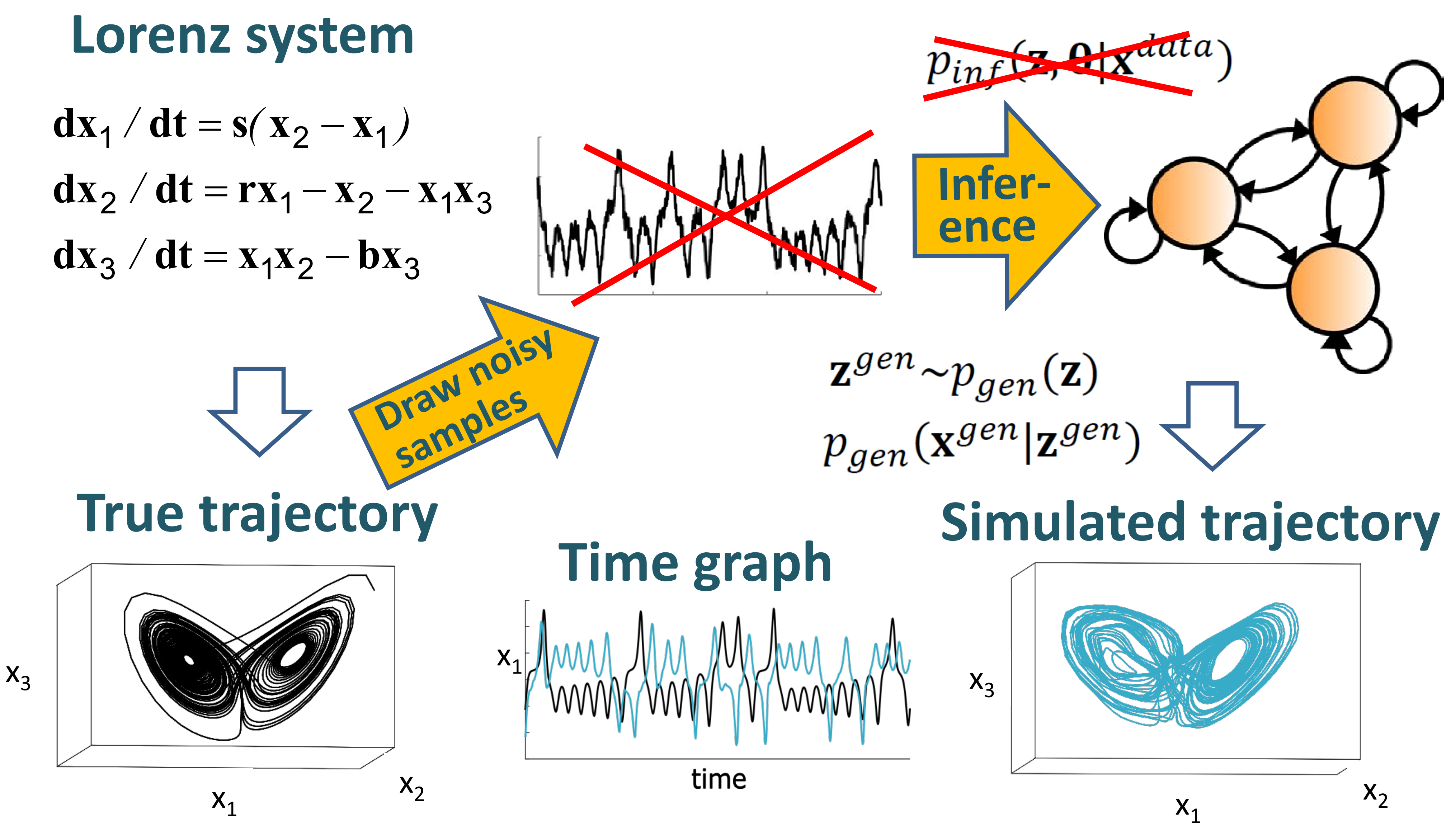
Draw noisy samples

True trajectory

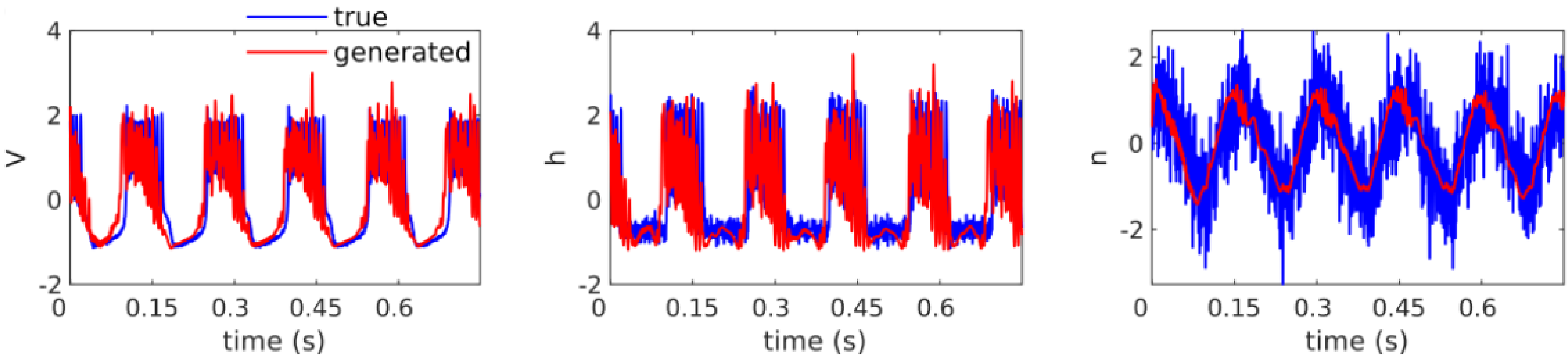


Koppe et al. (2019), *PLoS Comp Biol*

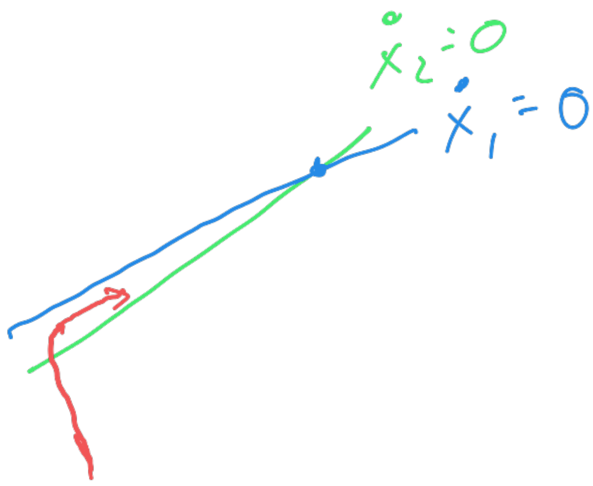
Reconstructing dynamical systems



Koppe et al. (2019), *PLoS Comp Biol*



Schmidt et al. (2021) ICLR

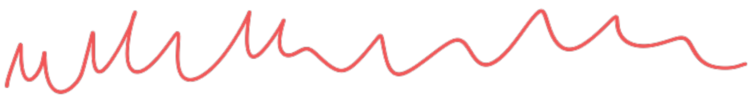


Inferring PLRNN from fMRI data

process model

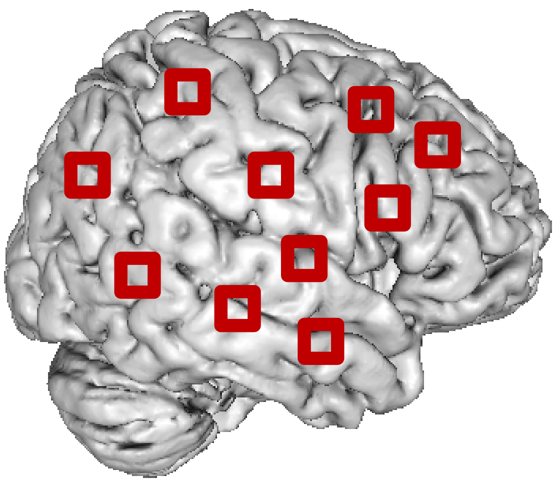
$$\mathbf{z}_t = \mathbf{A}\mathbf{z}_{t-1} + \mathbf{W}\phi(\mathbf{z}_{t-1}) + \mathbf{h} + \mathbf{C}\mathbf{u}_t + \boldsymbol{\varepsilon}_t$$

$$\boldsymbol{\varepsilon}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma})$$



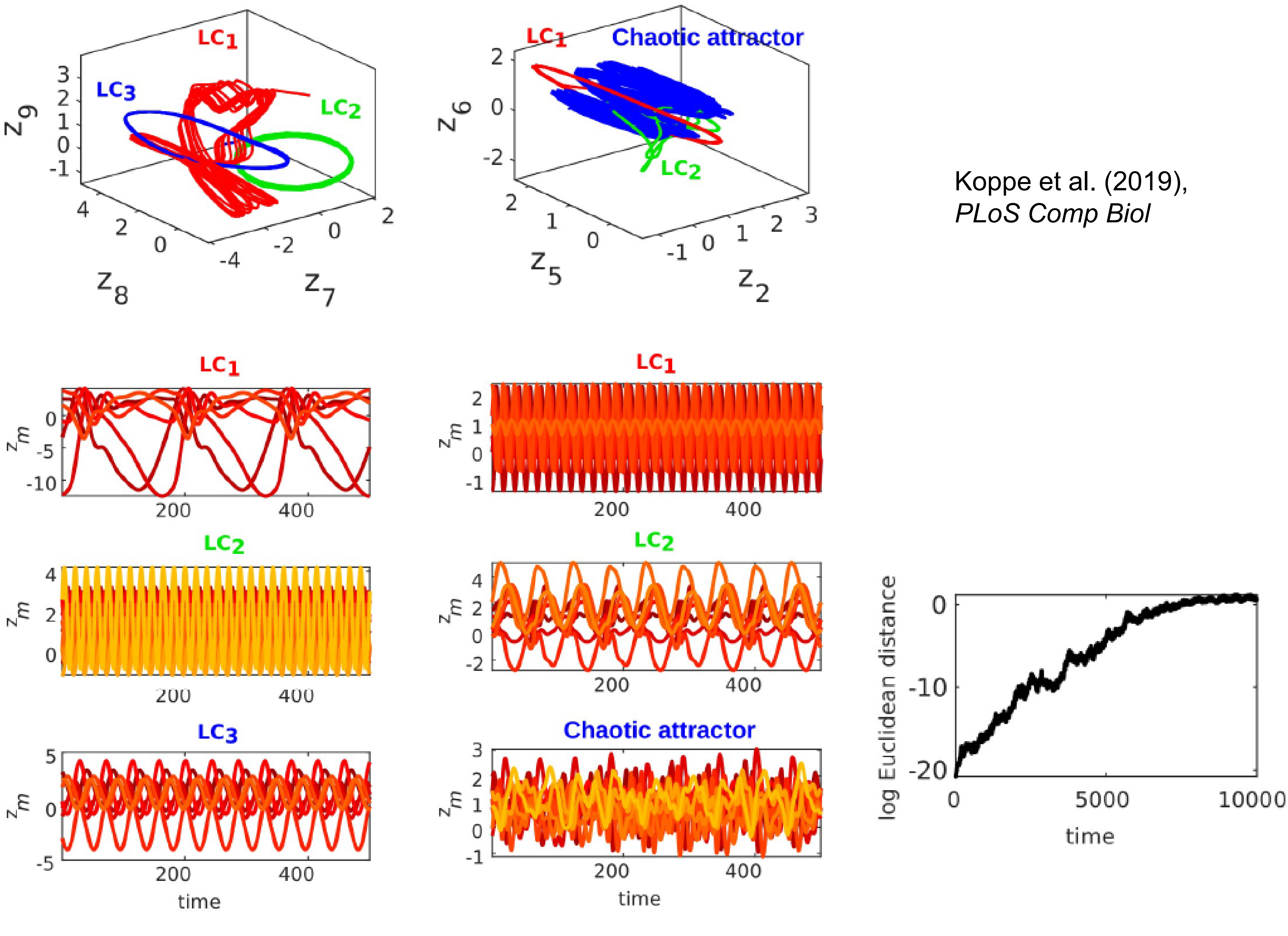
fMRI observation model

$$\mathbf{x}_t \mid \mathbf{z}_t \sim \mathrm{N}(\mathbf{B}[hrf * \mathbf{z}_{t-\tau:t}] + \mathbf{M}\mathbf{r}_t, \boldsymbol{\Gamma})$$



Koppe et al. (2019), *PLoS Comp Biol*

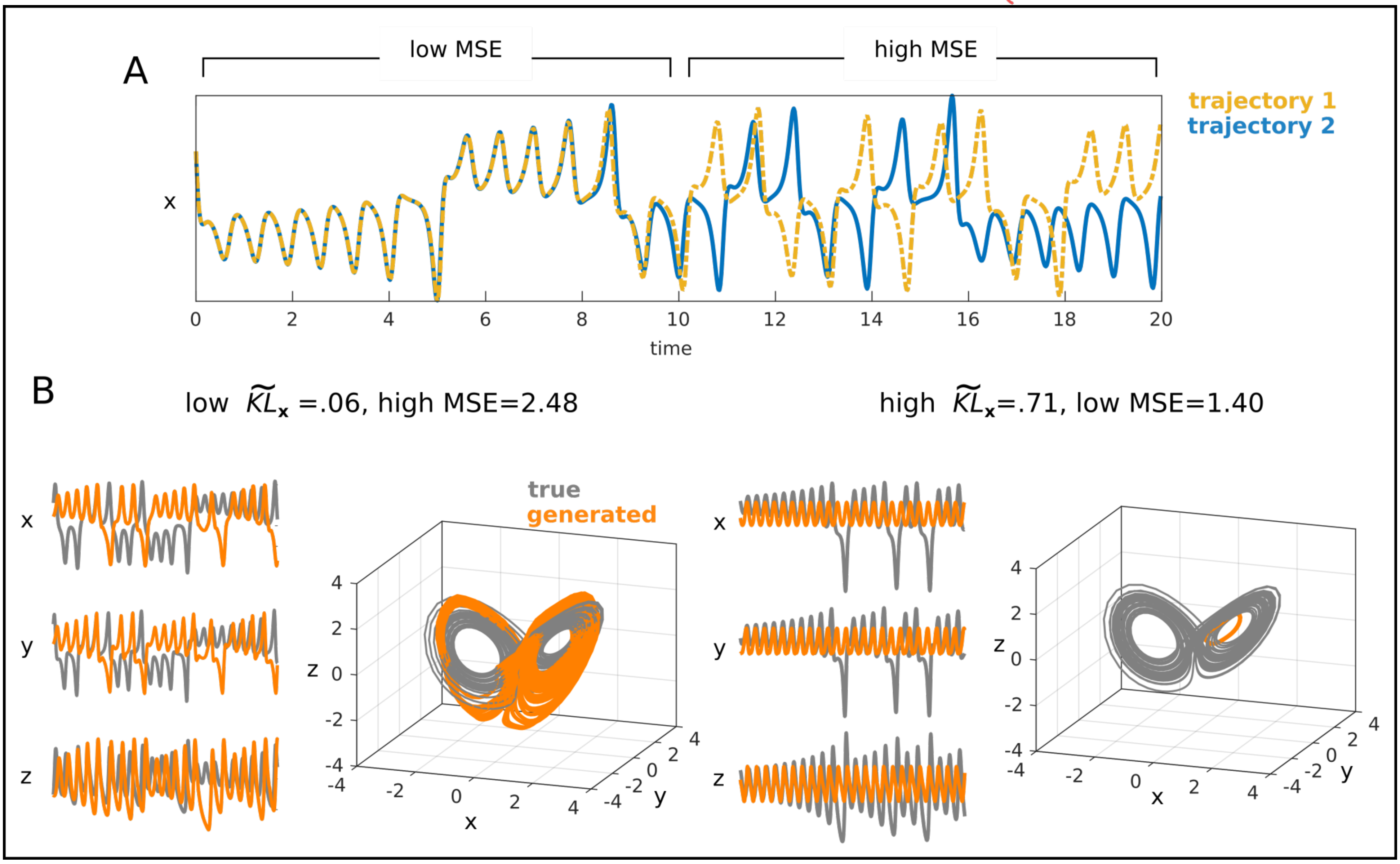
Nonlinear phenomena from fMRI data



Koppe et al. (2019),
PLoS Comp Biol

Simple ahead prediction errors may be meaningless

$\sum_t (x_t - \hat{x}_t)^T \Sigma^{-1} (\ldots)$



Koppe et al. (2019), *PLoS Comp Biol*