

# Variational Inference and Learning for Continuous-Time Nonlinear State-Space Models

**Antti Honkela, Markus Harva, Tapani Raiko, Juha Karhunen**

Adaptive Informatics Research Centre  
Helsinki University of Technology, Finland

May 28, 2008

# Outline

- Discrete and continuous-time nonlinear state-space models
- Variational inference for continuous-time models
- Demonstration with the Lorenz process

# Nonlinear dynamical systems

- Model data  $\mathbf{x}(t)$  with temporal dependencies
- Differential equation model for a continuous-time process:

$$\frac{d}{dt}\mathbf{x}(t) = \mathbf{g}(\mathbf{x}(t))$$

- Linear  $\mathbf{g}$  implies very restricted exponentially decaying dynamics, nonlinearity needed for interesting systems
- Sampling regularly at  $t = 1, \dots, T$  and integrating yields

$$\mathbf{x}(t + 1) = \phi_1(\mathbf{x}(t)),$$

a discrete-time difference equation

## Nonlinear state-space models (NSSMs)

- Instead of modelling the dynamics of the data  $\mathbf{x}$ , use a latent state-space  $\mathbf{s}$
- Discrete-time NSSM (Valpola & Karhunen, 2002):

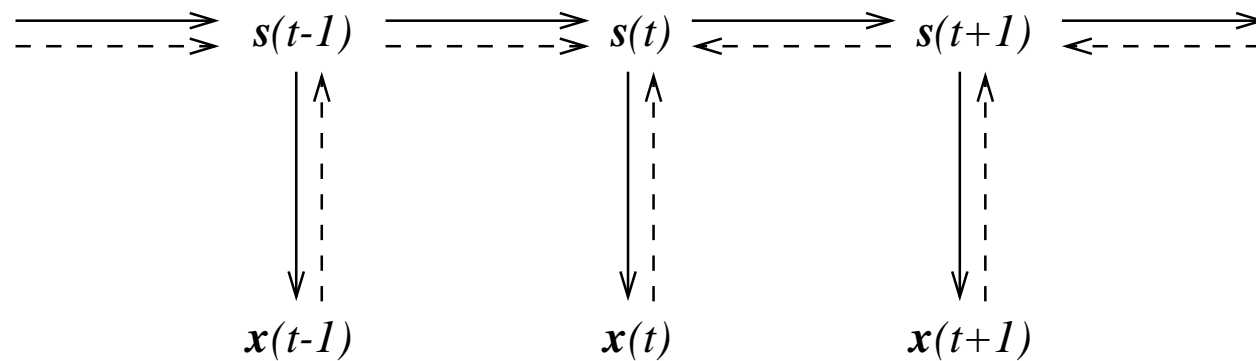
$$\begin{aligned}\mathbf{s}(t + 1) &= \mathbf{s}(t) + \mathbf{g}_{dt}(\mathbf{s}(t), \boldsymbol{\theta}_{\mathbf{g}}) + \mathbf{m}(t) \\ \mathbf{x}(t) &= \mathbf{f}(\mathbf{s}(t), \boldsymbol{\theta}_{\mathbf{f}}) + \mathbf{n}(t)\end{aligned}$$

- MLP networks used to model  $\mathbf{f}$  and  $\mathbf{g}$ :

$$\mathbf{f}(\mathbf{s}(t), \boldsymbol{\theta}_{\mathbf{f}}) = \mathbf{B} \tanh(\mathbf{A}\mathbf{s}(t) + \mathbf{a}) + \mathbf{b}$$

where  $\tanh$  is applied to each component separately

## The flow of information in the model



Information flow to sources  $s(t)$  in the nonlinear state-space model

# Variational learning

- Given data  $\mathbf{X}$ , estimate the sources  $\mathbf{S}$  and model parameters  $\boldsymbol{\theta}$
- Fit simple  $q(\mathbf{S}, \boldsymbol{\theta})$  to the posterior  $p(\mathbf{S}, \boldsymbol{\theta}|\mathbf{X})$
- Factorial approximation:

$$q(\mathbf{S}, \boldsymbol{\theta}) = q(\mathbf{S})q(\boldsymbol{\theta}) = q(\mathbf{S}) \prod_i q(\theta_i)$$

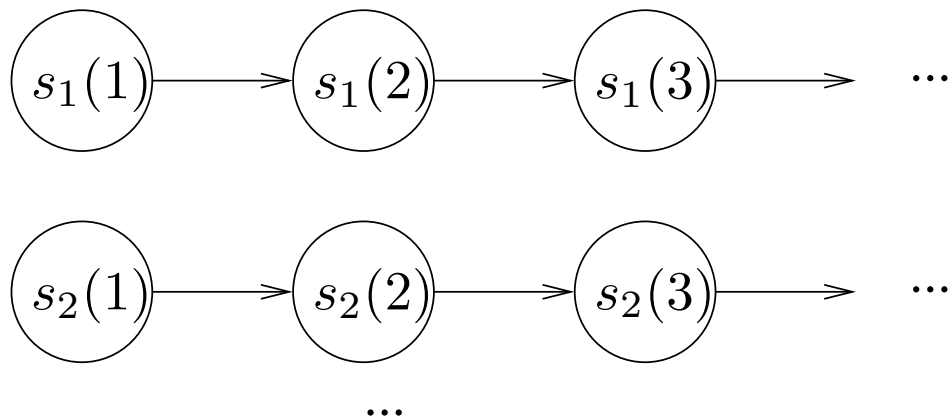
- Individual factors are scalar Gaussian distributions
- Optimize Kullback-Leibler divergence

$$\begin{aligned} D(q(\mathbf{S}, \boldsymbol{\theta})||p(\mathbf{S}, \boldsymbol{\theta}|\mathbf{X})) &= \int q(\mathbf{S}, \boldsymbol{\theta}) \log \frac{q(\mathbf{S}, \boldsymbol{\theta})}{p(\mathbf{S}, \boldsymbol{\theta}|\mathbf{X})} d\mathbf{S}d\boldsymbol{\theta} \\ &= E_q \left[ \log \frac{q(\mathbf{S}, \boldsymbol{\theta})}{p(\mathbf{S}, \boldsymbol{\theta}|\mathbf{X})} \right] \end{aligned}$$

# The source model

- The approximation  $q(\mathbf{S})$  for the sources is not fully factorial but contains temporal dependencies

$$q(\mathbf{S}) = \prod_i \left( q(s_i(1)) \prod_t q(s_i(t+1)|s_i(t)) \right)$$



## Learning the NSSM

- Linearise the nonlinearities (Honkela & Valpola, NIPS 2004)
  - Like Taylor approximation, but use quadratures instead of derivatives to estimate global behaviour
- For many variables, the optimal value (assuming others are fixed) can be solved easily
- Gradient-based iteration is used for the weights of the MLPs and the sources  $s(t)$



## Learning the NSSM, continued

- The update algorithm resembles back-propagation, but
  - Inputs and some of the outputs of the MLPs are latent
  - All unknown variables have a distribution of values
  - The cost function is based on Kullback-Leibler divergence
- Source updates are local in time, no global smoothing
  - In time, iterative updates work globally
  - New inference algorithms under development (using global smoothing and a natural gradient)

## Discrete-time models: pros and cons

- + Relatively simple and efficient methods for learning and inference
- What happens between the samples?
- Uneven sampling, missing time points?

## Continuous-time NSSM

- Instead of a discrete-time map, use a differential equation to model state evolution
- Introducing the noise makes this a *stochastic differential equation (SDE)*

$$ds(t) = \mathbf{g}(s(t)) dt + \sqrt{\Sigma} dW(t),$$

where  $dW$  is the differential of a Wiener process (Brownian motion)

# Stochastic Differential Equations

Comparison:

$$\mathbf{s}(t + 1) = \mathbf{s}(t) + \mathbf{g}_{dt}(\mathbf{s}(t), \boldsymbol{\theta}_{\mathbf{g}}) + \mathbf{m}(t)$$

$$d\mathbf{s}(t) = \mathbf{g}(\mathbf{s}(t)) dt + \sqrt{\boldsymbol{\Sigma}} dW(t)$$

- Intuitively: deterministic drift + stochastic part
- The solution is a continuous-time stochastic process with Markov property
- Sampling methods similar to numerical solution methods of ODEs

## Continuous-time NSSM

- Assume data  $\mathbf{X} = \{\mathbf{x}(t_i) | i = 1, \dots, N\}$ , introduce latent variables for the states  $\mathbf{S} = \{\mathbf{s}(t_i) | i = 1, \dots, N\}$
- Continuous-time NSSM equations:

$$d\mathbf{s}(t) = \mathbf{g}(\mathbf{s}(t), \boldsymbol{\theta}_g) dt + \sqrt{\boldsymbol{\Sigma}} dW(t)$$

$$\mathbf{x}(t_i) = \mathbf{f}(\mathbf{s}(t_i), \boldsymbol{\theta}_f) + \mathbf{n}(t_i)$$

- Because of the Markov property of  $\mathbf{s}(t)$ , need the dynamics only to evaluate  $p(\mathbf{s}(t_{i+1}) | \mathbf{s}(t_i))$

# Approximations

- How to evaluate  $p(\mathbf{s}(t_{i+1})|\mathbf{s}(t_i))$ ?
- Derive differential equations for the mean and covariance of a Gaussian process satisfying the same SDE by linearising  $\mathbf{g}$  about the current mean:

$$\frac{d}{dt}\boldsymbol{\mu}(t) = \langle \mathbf{g}(\boldsymbol{\mu}(t)) \rangle$$

$$\frac{d}{dt}\mathbf{P}(t) = \langle \mathbf{G}(\boldsymbol{\mu}(t)) \rangle \mathbf{P}^T(t) + \mathbf{P}(t) \langle \mathbf{G}^T(\boldsymbol{\mu}(t)) \rangle + \boldsymbol{\Sigma}$$

- Solve these numerically using an Euler method
- Expected statistics of  $\mathbf{g}$  and its Jacobian  $\mathbf{G}$  evaluated using the global linearisation (Honkela & Valpola, NIPS 2004)

## Variational continuous-time NSSM

- The resulting learning method for continuous-time NSSM is mainly rather similar to discrete-time variant
- Main conceptual difference: process noise ( $\mathbf{m}(t)$ ) is generated by the SDE, not just i.i.d. Gaussian
- Possibility to study the process between the samples by introducing new time points without observations
- Scales well: Computational complexity linear w.r.t. number of connections in the model

# Experiment: Continuous-time NSSM

- Proof-of-concept experiment: learning a Lorenz process

$$\frac{dz_1}{dt} = \sigma(z_1 - z_2)$$

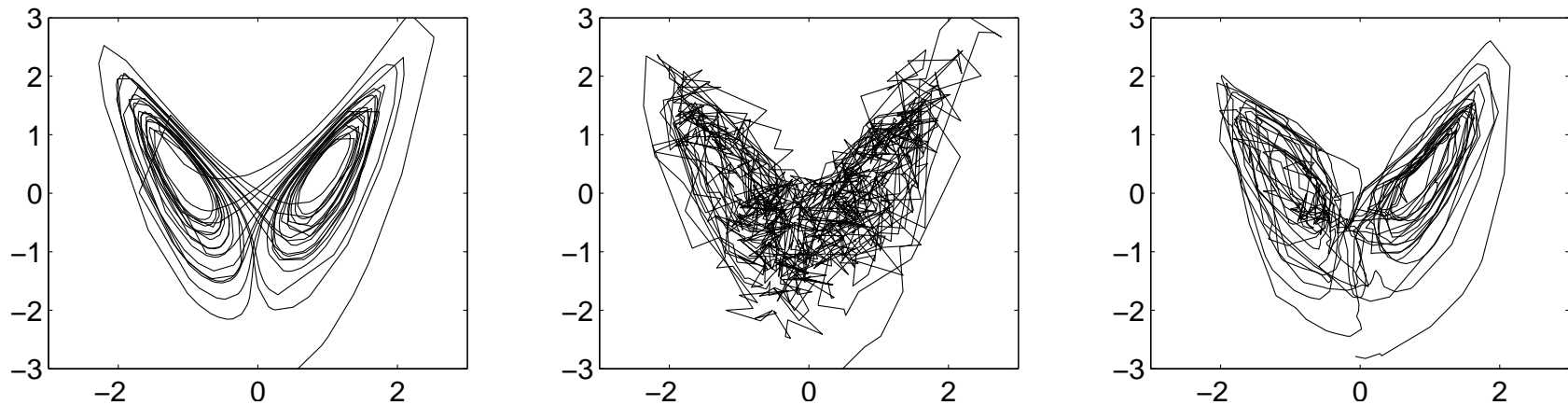
$$\frac{dz_2}{dt} = \rho z_1 - z_2 - z_1 z_3$$

$$\frac{dz_3}{dt} = z_1 z_2 - \beta z_3$$

- Two observations, relatively high observation noise level
- No process noise
- 1000 unevenly sampled data points



## Experiment: Continuous-time NSSM



Left: The original data set without noise.

Middle: The noisy data set used in the experiment.

Right: The reconstruction of the data set by the model.

## Comparison to (Archambeau & al. 2008)

- Archambeau & al. study variational inference on continuous-time models based on measuring KL-divergence between the actual paths
- Our approach is based on reducing the continuous-time inference essentially to a discrete-time problem
  - Learning the model becomes a solved problem
  - System noise covariance does not have to be identical to that of the prior process

## Discussion

- Nonlinearities are clearly needed to model dynamical systems
- Continuous time opens new possibilities
  - Maybe help with different time scales?
- Proof-of-concept continuous-time nonlinear state-space model
- Interesting application potential in modelling gene transcription regulation (Mjolsness 2007)
- Future work: heteroscedastic variations of the basic model — Model observation noise or innovation process variance

## References (1/2)

- A. Honkela, M. Tornio, and T. Raiko. Variational Bayes for continuous-time nonlinear state-space models. In NIPS\*2006 Workshop on Dynamical Systems, Stochastic Processes and Bayesian Inference, Whistler, B.C., Canada, 2006.
- H. Valpola and J. Karhunen. An unsupervised ensemble learning method for nonlinear dynamic state-space models. *Neural Computation*, 14(11):2647–2692, 2002.
- A. Honkela and H. Valpola. Unsupervised variational Bayesian learning of nonlinear models. In L. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 593–600. MIT Press, Cambridge, MA, USA, 2005.

## References (2/2)

- C. Archambeau, M. Opper, Y. Shen, D. Cornford, and J. Shawe-Taylor. Variational inference for diffusion processes. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 17–24. MIT Press, Cambridge, MA, 2008.
- E. Mjolsness. On cooperative quasi-equilibrium models of transcriptio of *Bioinformatics and Computational Biology*, 5(2b):467–490, 2007.