Variational Bayes for Continuous-Time Nonlinear State-Space Models

Antti Honkela, Matti Tornio, and Tapani Raiko

Adaptive Informatics Research Centre

Helsinki University of Technology, Finland

December 9, 2006

Outline

- Discrete and continuous-time nonlinear state-space models
- Variational inference for continuous-time models
- State inference methods
- Experiments

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 $t = 1, \ldots, T$ and integrating yields

$$\mathbf{x}(t+1) = \boldsymbol{\phi}_1(\mathbf{x}(t)),$$

  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):

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

- MLP networks used to model $\mathbf{f}$ and $\mathbf{g}$
- Variational inference as an extension to nonlinear factor analysis (Lappalainen (Valpola) & Honkela, 2000)

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):

$$\mathbf{s}(t+1) = \mathbf{s}(t) + \mathbf{g}'_{dt}(\mathbf{s}(t), \boldsymbol{\theta}_{\mathbf{g}}) + \mathbf{m}(t)$$
$$\mathbf{x}(t) = \mathbf{f}(\mathbf{s}(t), \boldsymbol{\theta}_{\mathbf{f}}) + \mathbf{n}(t)$$

- MLP networks used to model $\mathbf{f}$ and $\mathbf{g}$
- Variational inference as an extension to nonlinear factor analysis (Lappalainen (Valpola) & Honkela, 2000)

Variational inference for the NSSM

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

- Linearise the nonlinearities (Honkela & Valpola, NIPS 2004)
  - Like Taylor approximation, but use quadratures instead of derivatives to estimate global behaviour
- Gradient-based minimisation of the free energy with respect to the variational parameters of the Gaussian posterior approximation

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?
– Processes with different time scales very challenging

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)*

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

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

Stochastic Differential Equations

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

- 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:

$$d\mathbf{s}(t) = \mathbf{g}(\mathbf{s}(t), \boldsymbol{\theta}_{\mathbf{g}})\, dt + \sqrt{\boldsymbol{\Sigma}}\, dW(t)$$
$$\mathbf{x}(t_i) = \mathbf{f}(\mathbf{s}(t_i), \boldsymbol{\theta}_{\mathbf{f}}) + \mathbf{n}(t_i)$$

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

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

- 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

State inference

- How to estimate the sequence of dependent state values $\mathbf{S}$?
- Traditional solution: extended/unscented (variational) Kalman filter
  - Potentially unstable with long sequences
  - Not an exact minimum of the free energy
- Solution of Valpola & Karhunen (2002): minimise the free energy ignoring dependencies
  - Provably stable and convergent but slow algorithm

Faster state inference

- General principle: take into account relevant dependencies to minimise free energy more efficiently
- One heuristic: instead of partial derivatives, use total derivatives of the free energy

$$\frac{d\mathcal{C}}{d\overline{\mathbf{s}}(t)} = \sum_{\tau=1}^{T} \frac{\partial \mathcal{C}}{\partial \overline{\mathbf{s}}(\tau)} \frac{\partial \overline{\mathbf{s}}(\tau)}{\partial \overline{\mathbf{s}}(t)}.$$

- Solve the optimal mean assuming the linearisation and evaluate

$$\frac{\partial \overline{\mathbf{s}}(\tau)}{\partial \overline{\mathbf{s}}(t)} \approx \frac{\partial \overline{\mathbf{s}}^{\mathsf{opt}}(\tau)}{\partial \overline{\mathbf{s}}^{\mathsf{opt}}(t)}, \quad \tau \in \{t-1, t+1\}$$

- Total derivatives can now be evaluated using chain rule and dynamic programming

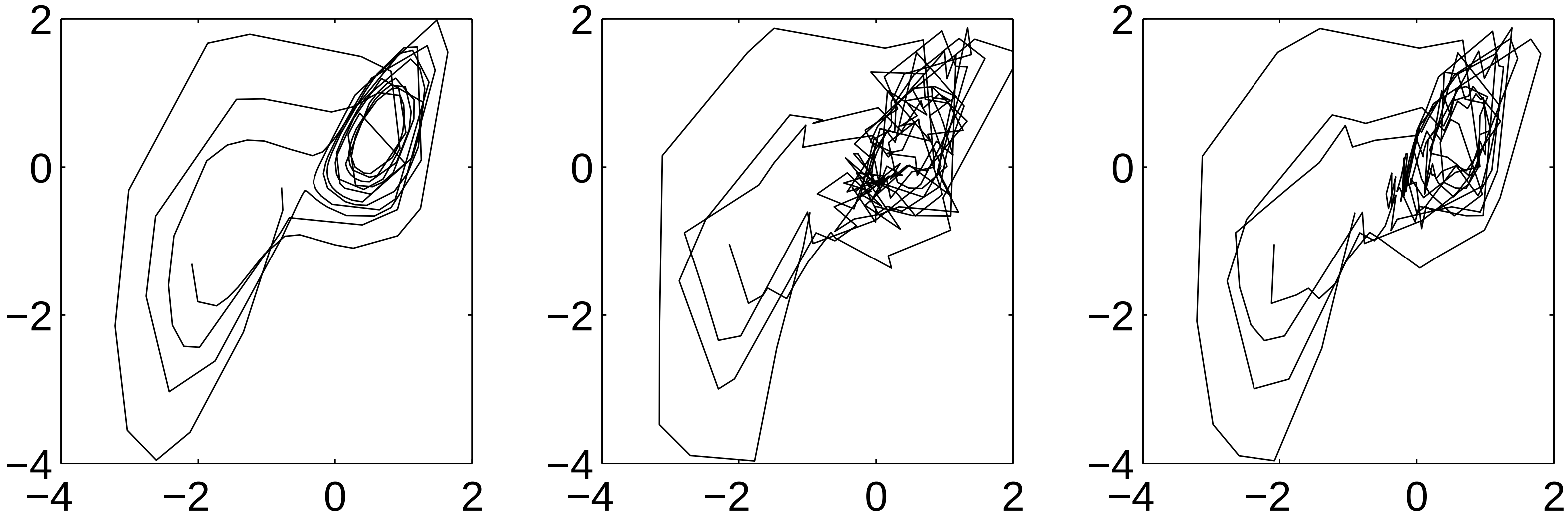
Experiment: Continuous-time NSSM

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

$$
\begin{aligned}
\frac{dz_1}{dt} &= \sigma(z_1 - z_2) \\
\frac{dz_2}{dt} &= \rho z_1 - z_2 - z_1 z_3 \\
\frac{dz_3}{dt} &= z_1 z_2 - \beta z_3
\end{aligned}
$$

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

Experiment: Continuous-time NSSM

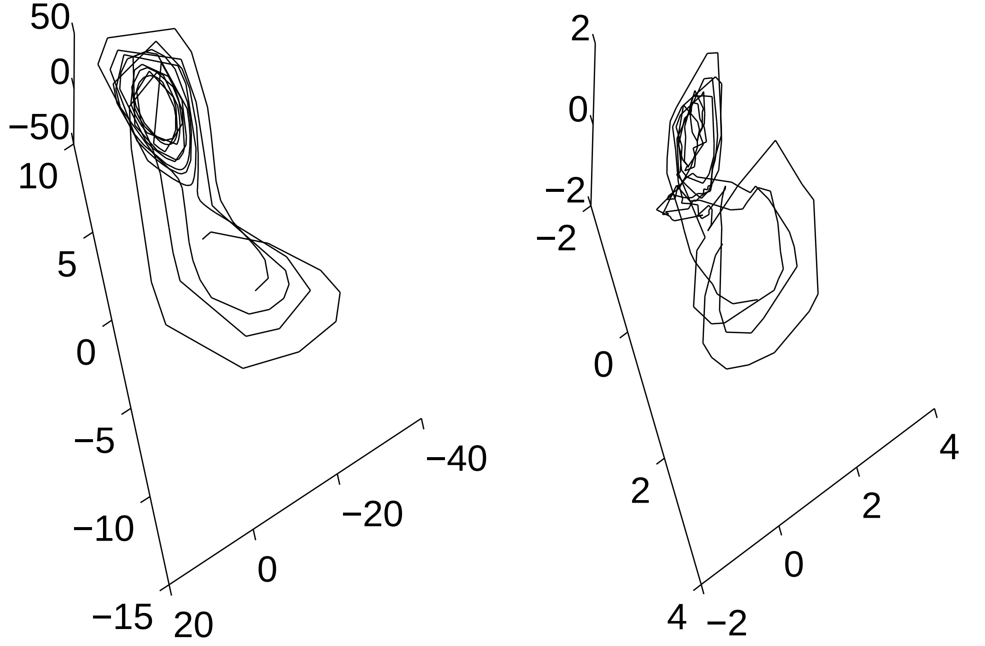


Left: The original data set without noise.

Middle: The noisy data set used in the experiment.

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

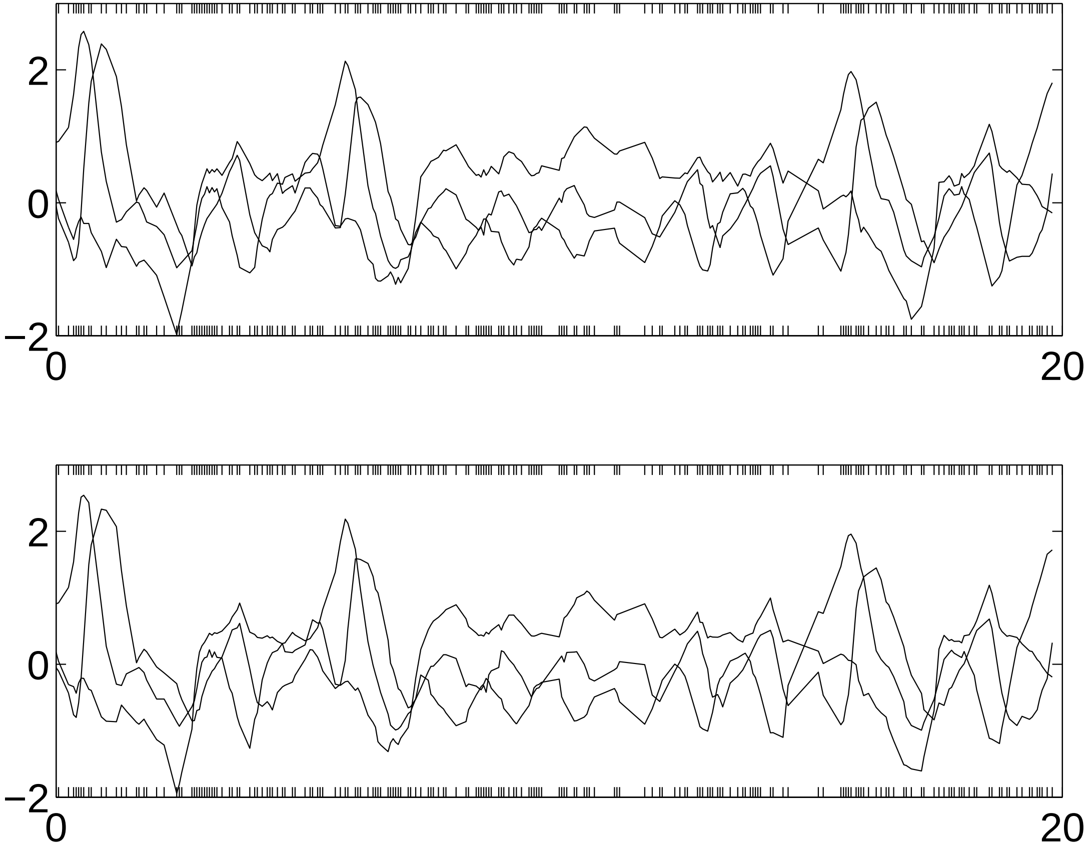
Experiment: Continuous-time NSSM



Left: The original three-dimensional Lorenz process without noise.

Right: The three-dimensional latent state-space of the model.

Experiment: Continuous-time NSSM

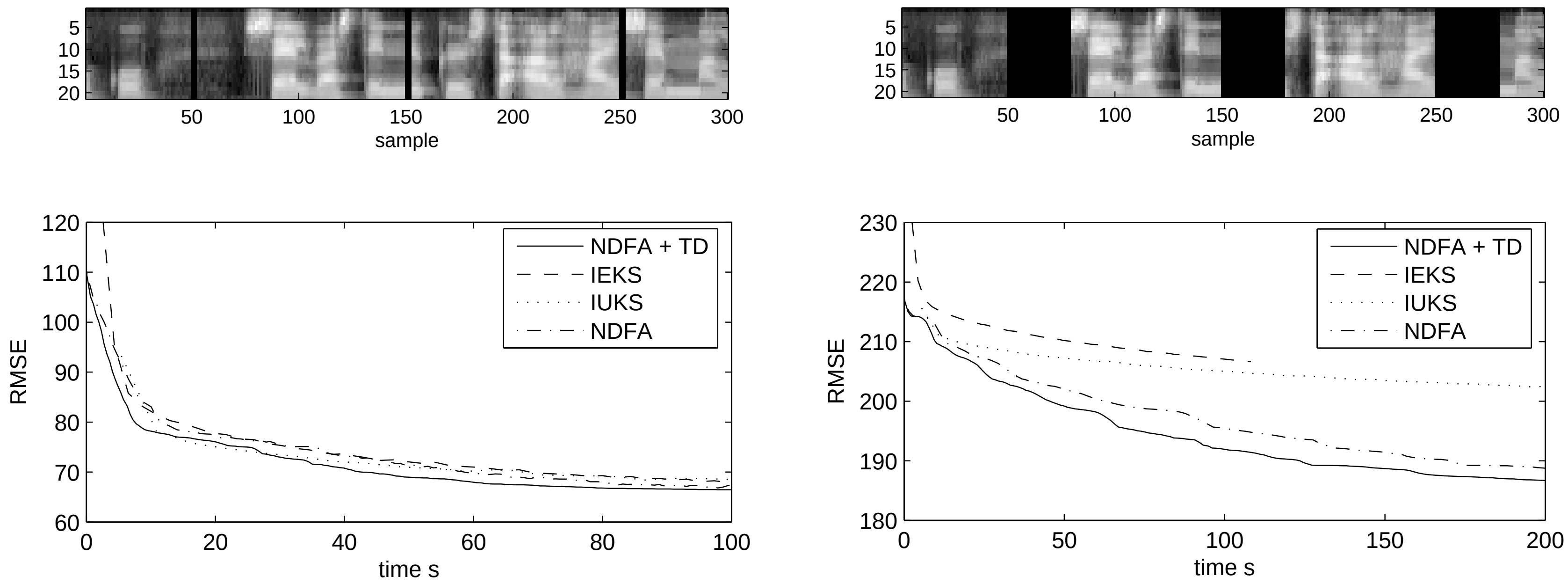


Top: The latent state values. Bottom: The values predicted from the previous time step.

Experiment: State inference

- Data: 21-dimensional spectrograms of continuous human speech
- 10 000 samples to learn the dynamics, 1 200 for testing
- Learn a discrete-time NSSM with 7 hidden states
- Task: reconstruct gaps of 3 or 30 samples in observations
- Compare state inference between proposed method, iterated extended Kalman smoother (IEKS) and iterated unscented Kalman smoother (IUKS)

Experiment: State inference



Conclusion

- 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
- State inference in nonlinear models (for learning)