Reconstruction of complex space-time dynamics using historical datasets: Analog HMM

Redouane Lguensat

-

In collaboration with:

R. Fablet, P. Ailliot, P. Tandeo

Institut Mines-Telecom
Signal & Communications Department

May 26, 2015

Overview

Data Driven methods for Dynamic Systems

Objective

Solving the following problem, which is of key interest in many dynamic data-driven applications
Given:

- Partial and noisy observations of a complex dynamical system
- Historical datasets about that system

Develop a reconstruction method of that system without having access to the equations of motion and without using a parametric model.

Non-Parametric Analog Data Assimilation

- Methods based on stochastic filtering:
  - Analog Ensemble Kalman Filter and Smoother (Tandeo et al 2014)
  - Particle Filter based reformulation.
- Method based on Hidden Markov Models:
  - Discrete reformulation: Analog Forward Backward Algorithm

Overview

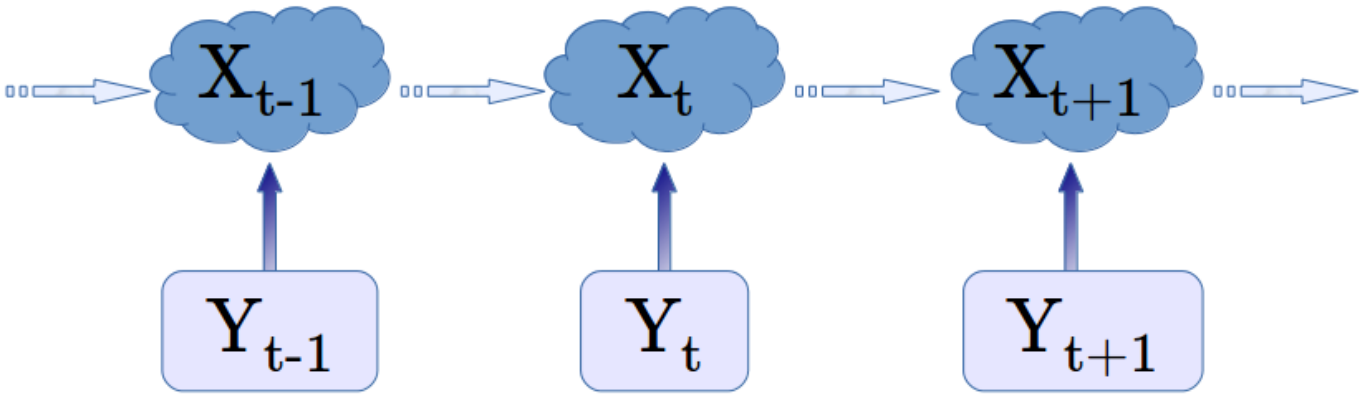


Figure : General architecture of a HMM: The random variable X_t is the hidden state at time t. The random variable Y_t is the corresponding observation (or measurement) at time t.

HMM parameters

We note by $\Lambda = (A, B, \pi_1)$ the parameters of the HMM where:

- Assuming that $P(X_t|X_{t-1})$ is independent of time t, the definition of the time independent **transition matrix** is given by:

$$A = \{a_{ij}\} = P(X_t = j|X_{t-1} = i)$$

- The **initial state distribution** (i.e. when $t = 1$) is given by:

$$\pi_1 = \{\pi_i\} \text{ where } \pi_i = P(X_1 = i)$$

- The **observation matrix** (called also the emission matrix) gives the probability of a certain observation at time t for state j and it is expressed as:

$$B = \{b_j(Y_t)\} \text{ where } b_j(Y_t) = P(Y_t|X_t = j)$$

Inference problem

Task

Given a foreknowledge of an HMM parameters and an observation sequence, compute the posterior marginals $P(X_t \mid Y_{1:T})$ of all hidden state variables where $Y_{1:T} = Y_1, \cdots, Y_T$ and $T > t$.

$\twoheadrightarrow$ This is called *smoothing* in stochastic filtering.
$\twoheadrightarrow$ In HMM framework this can be solved using the Forward Backward Algorithm.

- Given $\Lambda = (A, B, \pi_1)$ we are interested in evaluating $\gamma_t(i) = P(X_t = i | Y_{1:T})$:
- Let consider the *forward* variable $\alpha_t(i)$ and the *backward* variable $\beta_t(i)$ defined as:

$$\alpha_t(i) = P(Y_{1:t}, X_t = i) \quad \beta_t(i) = P(Y_{t+1:T} | X_t = i)$$

- We show that:

$$\alpha_t(j) = [\Sigma_{i=1}^{Q} \alpha_{t-1}(i).a_{ij}] b_j(Y_t)$$
$$\beta_t(j) = \Sigma_{i=1}^{Q} b_i(Y_{t+1}).a_{ji}.\beta_{t+1}(i)$$
$$\gamma_t(i) = \frac{\alpha_t(i).\beta_t(i)}{\Sigma_{i=1}^{Q} \alpha_t(i).\beta_t(i)}$$

Solution of the problem

Task

Given a foreknowledge of an HMM parameters and an observation sequence, compute the posterior marginals $P(X_t \mid Y_{1:T})$ of all hidden state variables where $Y_{1:T} = Y_1, \cdots, Y_T$ and $T > t$.

Choosing the most likely state for the system at time t is straightforward by taking the index of the state with the larger probability value:

$$X_t = \arg \max_{i=1 \cdots Q} \gamma_t(i)$$

Motivation

- Discrete reformulation of the method used in the Analog Ensemble Kalman Filter
- States are taken from the database
- Easy to implement

Idea

- Catalog of historical trajectories

Analogues (t)	Successeurs (t+dt)
$(-0.3268, +3.2644, +25.5134)$	$(+0.0131, +3.2278, +24.8371)$
$(+0.0131, +3.2278, +24.8371)$	$(+0.3177, +3.2017, +24.1889)$
$\vdots$	$\vdots$
$(-2.7587, -4.5007, +19.1790)$	$(-2.9344, -4.7112, +18.8037)$
$(-2.9344, -4.7112, +18.8037)$	$(-3.1147, -4.9464, +18.4530)$

- States of the HMM are historical values in the database (the analogs)
- Transition matrix is a sparse matrix where for every state there is only K possible transitions

Reformulation - Transition model

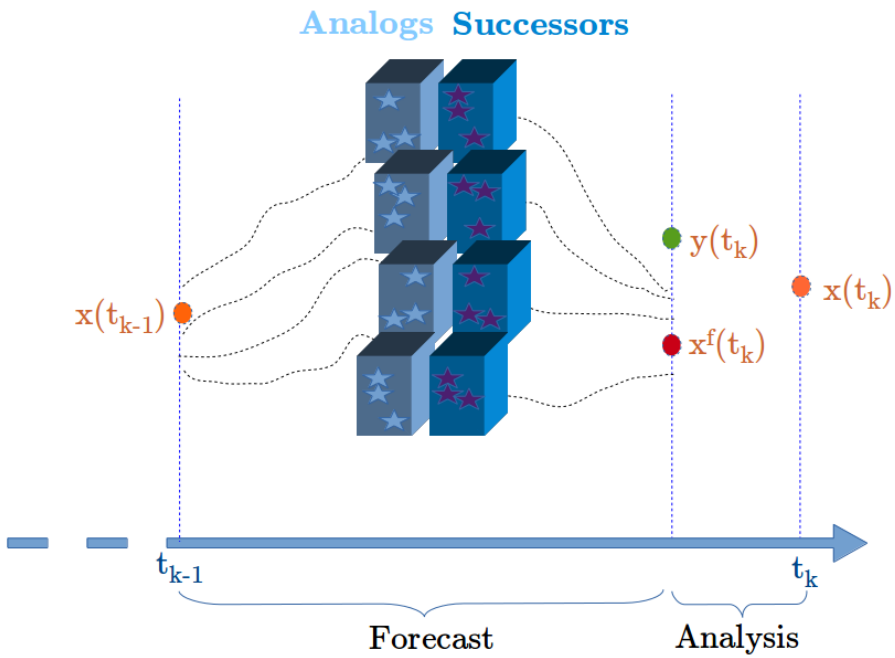


Figure : Stochastic filtering

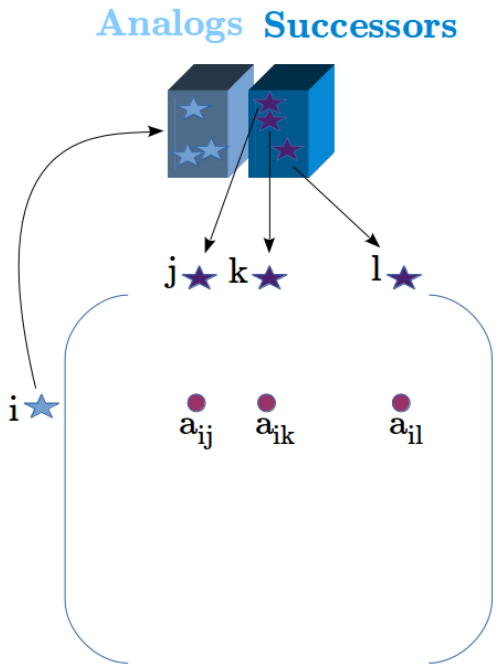


Figure : Analog HMM

Reformulation - Observation model

Matrix B expresses the observation model $P(Y_t|X_t)$

- evaluating the matrix for all the Q states is very heavy computationally.
- storing a $Q \times T$ matrix when a big catalog is used (very large Q) is not always feasible.

When using the Analog Forward-Backward algorithm we will at each time step $t \in 1, \cdots, T$ evaluate $b_j(Y_t)$ only in the possible states that the variable can take at time t.

Optimizing memory and execution time

t=1

$$\alpha_1(i) = \pi_i b_i(Y_1) \quad \forall i \in 1:1:Q$$

For $t = 2:1:T$

- Select states from α_{t-1} with non-null probability
- Determinine possible transitions from these states

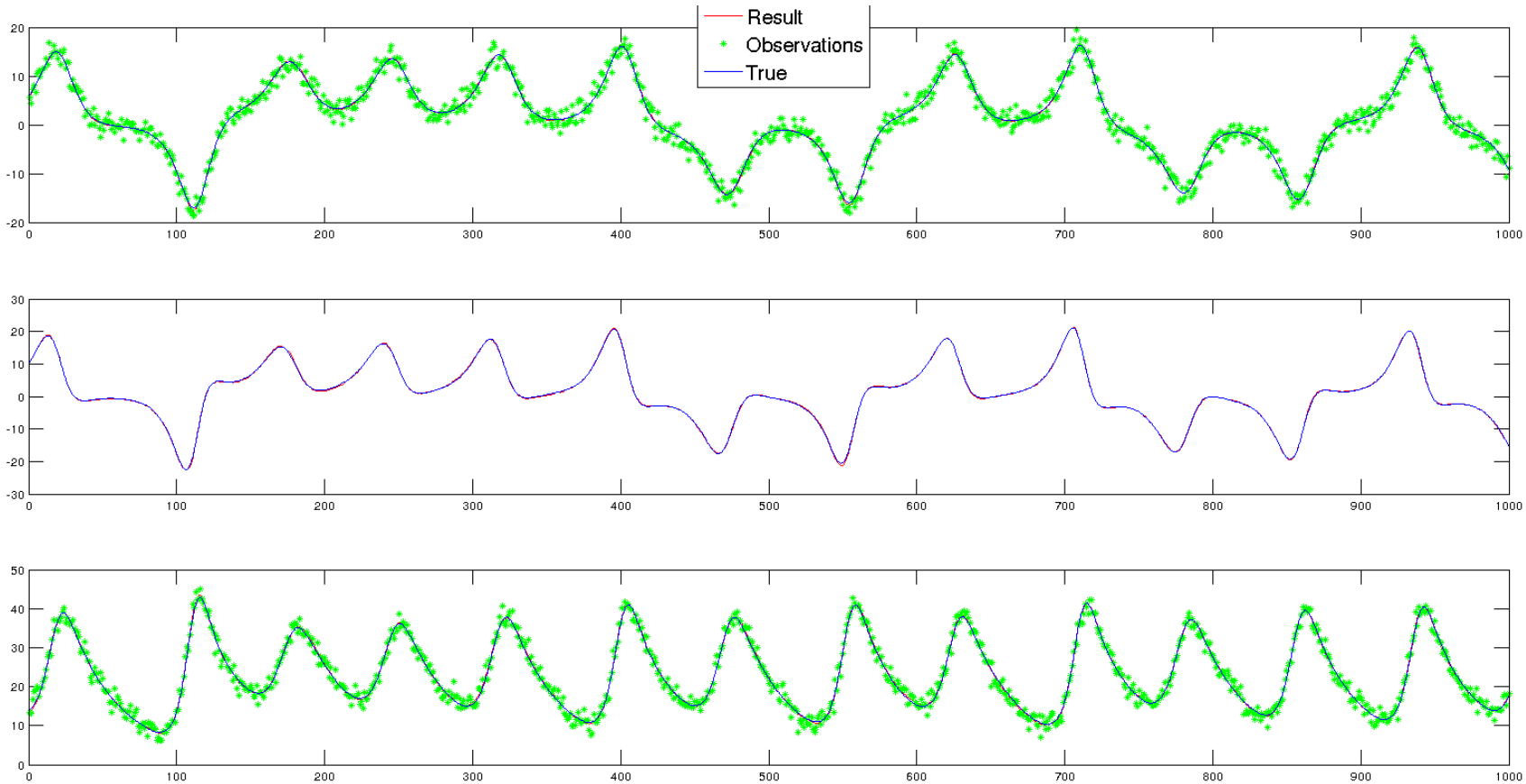
For $j \in possibletrans$

Evaluate $b_j(Y_t)$

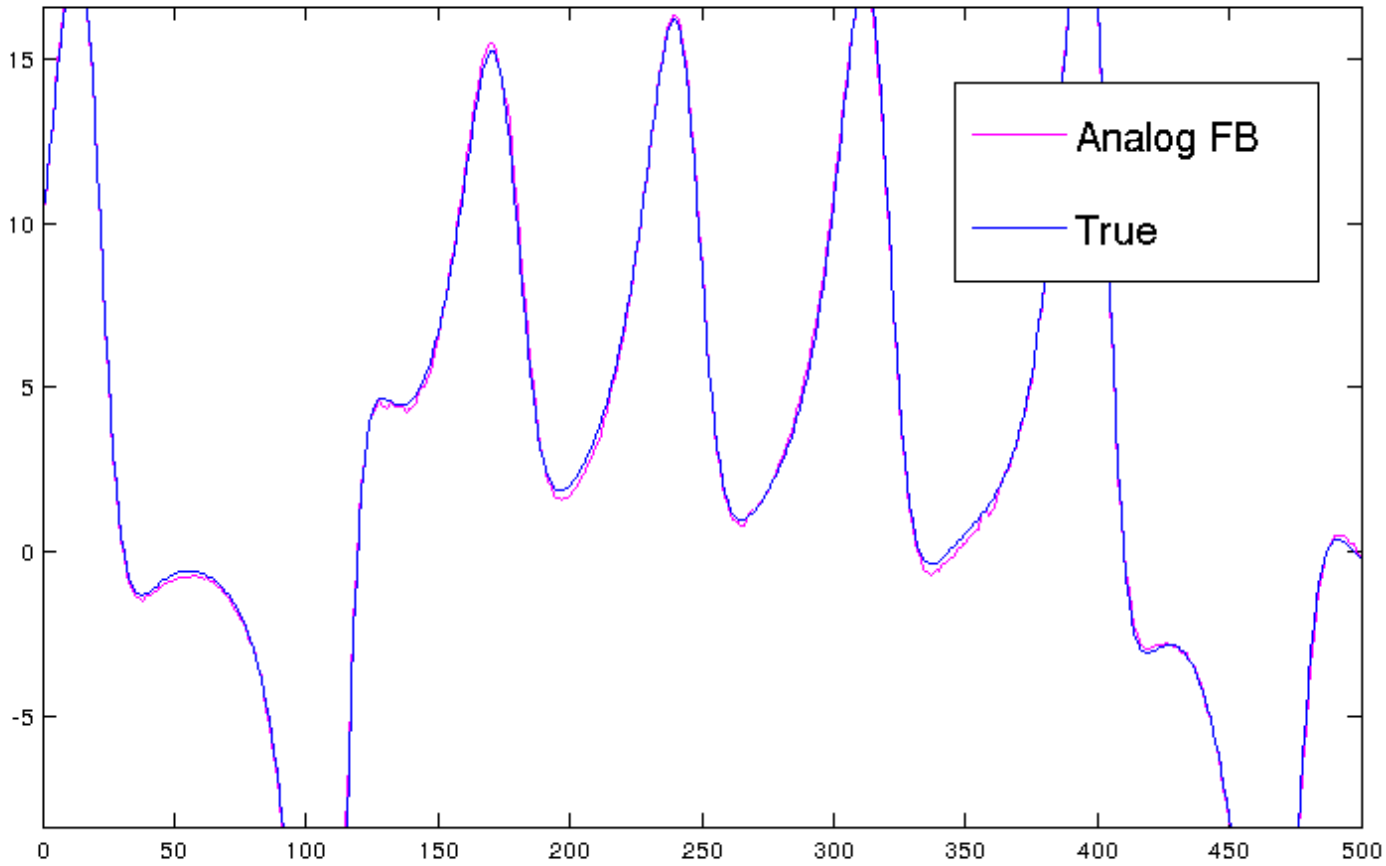
$$\alpha_t(j) = [\Sigma_{i=1}^{Q} \alpha_{t-1}(i).a_{ij}] b_j(Y_t)$$

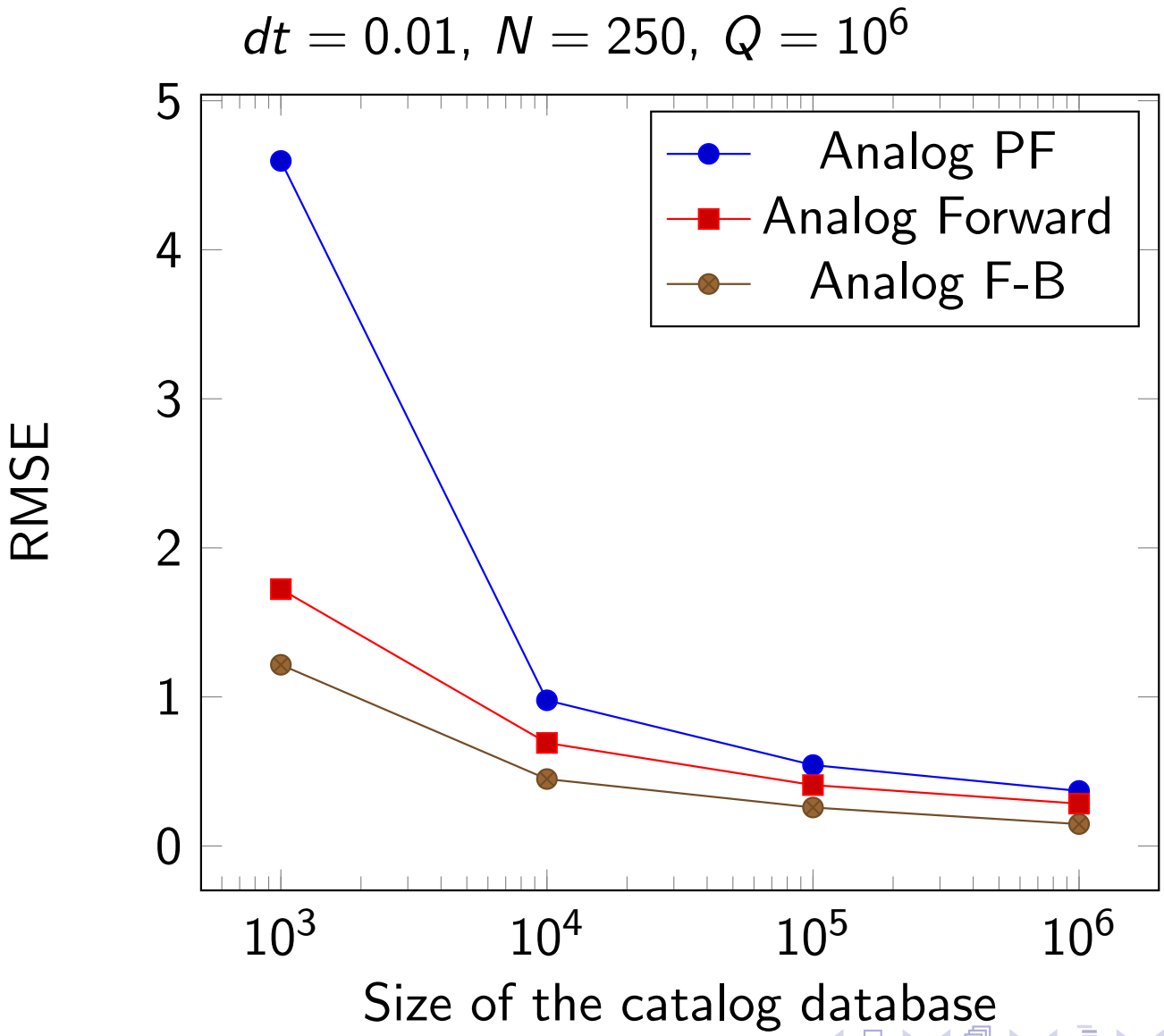
- Only consider the N most likely states and set the rest to zero

$$dt = 0.01,\ N = 250,\ Q = 10^6$$



$dt = 0.01,\ N = 250,\ Q = 10^6$




$dt = 0.01,\ N = 250,\ Q = 10^6$
RMSE
Size of the catalog database
Analog PF
Analog Forward
Analog F-B
0
1
2
3
4
5
10^3
10^4
10^5
10^6

Perspectives

- $dt_{obs} > 1 \Longrightarrow$ Long analogs
- Application on SST real data
- Perspectives on other scientific fields - Big Data.

References

- Evensen, G. (2003). The ensemble Kalman filter: Theoretical formulation and practical implementation. Ocean dynamics, 53(4), 343-367.
- Tandeo, P., Ailliot, P., Fablet, R., Ruiz, J., Rousseau, F., and Chapron, B. (2014). The Analog Ensemble Kalman Filter and Smoother. Climate informatics, Boulder, Colorado, US.
- Tandeo, P. et al. Combining analog method and ensemble data assimilation: application to the Lorenz-63 chaotic system. In book: Machine Learning and Data Mining Approaches to Climate Science, Chapter: Machine Learning Methods, Editors: Springer
- Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2), 257-286.

The End