

Reconstructing Governing Equations of Influenza Virus Dynamics from Incomplete Measurements

Martin Steiger* Ion Victor Gosea** Daniel Rüdiger**
Peter Benner** Hans-Georg Brachtendorf* Udo Reichl**

* *University of Applied Sciences Upper Austria, Hagenberg im Mühlkreis, Austria*

** *Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany*

Abstract: Accurate dynamic models of virus infection processes are highly relevant for the optimization of vaccine production processes. However, fitting their parameters onto or even designing novel dynamic models from an existing set of biological measurement data is challenging as it is common that entire quantities are missing due to sub-optimal measurement setups. This work targets identifying virus dynamics models based on incomplete measurement data and domain knowledge. To this end, we unite sparse identification of non-linear dynamics (SINDy) with a commonly used approach to identify non-linear dynamical systems from incomplete measurements, namely an extended Kalman filter (EKF). This yields a hybrid model that is able to identify governing equations that describe the dynamic non-linear processes. The capabilities of this model are demonstrated on a set of incomplete artificial measurement data of an existing infection dynamics model.

Keywords: Non-Linear Dynamical Systems, Sparse Identification of Nonlinear Dynamics, Bioprocess Engineering, Signal Processing, Kalman Filters

1. INTRODUCTION AND PROBLEM SETUP

Mathematical models of virus infection are widely used for the optimization of biopharmaceutical production processes, to support the design of experiments and for the identification of antiviral treatment strategies as reviewed in Beauchemin and Handel (2011). Previously, they were applied to describe and analyze the infection dynamics for a large variety of virus species, e.g. HIV, hepatitis C virus, and SARS-COV2. For the description of influenza virus infection, various modeling approaches were used to great success, i.e. models covering the infection in humans as in Baccam et al. (2006), the host immune response as in Bocharov and Romanyukha (1994) and the development of antibiotic resistance as in Perelson et al. (2012). Here, we focus on influenza A virus replication dynamics in animal cells considered for vaccine production described by Heldt et al. (2012) and Rüdiger et al. (2024). The identification of the optimal model structure and size for the respective system depends on the available experimental data and is quite challenging. This is an inherent problem when using real biological data which are often plagued by a sparsity of measurement time points, large measurement error ranges, and missing initial conditions. Thus, it would be highly beneficial to develop a method to identify the underlying dynamics of virus infection processes directly from ill-defined experimental data. Although the underlying mechanisms of influenza A virus replication are a complex interplay of processes, the overall dynamics can be described by simple kinetics. To that end, we use a

reduced infection model that is defined as follows:

$$\begin{aligned}\dot{x}_1 &= -\text{Imp} \cdot x_1, & \dot{x}_2 &= \text{SynM} \cdot x_4 - \text{DegM} \cdot x_2 \\ \dot{x}_3 &= \text{SynC} \cdot x_4 x_6 - \text{DegRnp} \cdot x_3 \\ \dot{x}_4 &= \text{Imp} \cdot x_1 + \text{SynV} \cdot x_3 x_6 - \text{DegRnp} \cdot x_4 - \text{Exp} \cdot x_4 x_7 \\ \dot{x}_5 &= \text{Exp} \cdot x_4 x_7 - \text{DegRnp} \cdot x_5 - \text{Rel} \cdot x_5 \\ \dot{x}_6 &= \text{SynP} \cdot x_2 - (\text{SynC} x_4 + \text{SynV} x_3) x_6 \\ \dot{x}_7 &= \text{SynP} \cdot x_2, & \dot{x}_8 &= \text{Rel} \cdot x_5\end{aligned}\tag{1}$$

where x are the system states that describe the production of progeny virus particles following an infection and the rest resembles constant scalar system parameters. Initially, extracellular virions x_1 are imported to the cell nucleus with rate Imp and release viral genomes x_4 . These genomes produce messenger RNA x_2 in the nucleus with rate SynM, which is translated to two different viral proteins with rate SynP. Protein x_6 is involved in the replication of the viral genome and protein x_7 causes the export of viral genomes from the nucleus. An intermediate version of the genome x_3 is formed with rate SynC and is subsequently used to produce more viral genomes with rate SynV. Due to the accumulation of export proteins, the viral genomes are exported from the nucleus to the cytoplasm with rate Exp, which is accounted for in x_5 . Lastly, these cytosolic genomes are released from the cell with rate Rel. All genomic species can get degraded by cellular processes, which is accounted for by the rates DegM for messenger RNA and by DegRnp for all forms of the viral genome. We are simulating this system with the following parameter set [Imp = 10, SynM = 20, SynC = 0.9, SynV = 10, SynP = 1, DegM = 0.33, DegRnp = 0.09, Exp = 1, Rel

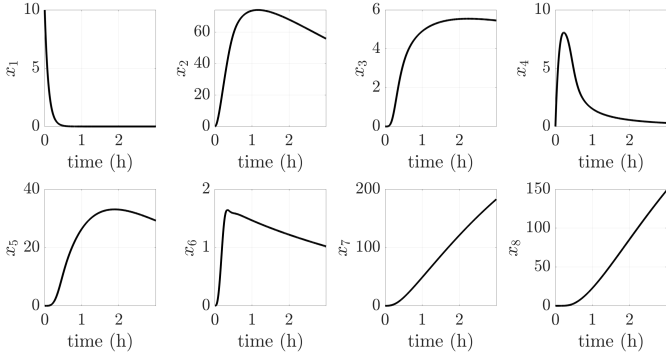


Fig. 1. Behavior of an influenza virus in a cell culture from the reduced model in (1)

$= 2]$ and the initial condition $x(0) = [10, 0, 0, 0, 0, 0, 0, 0]$, which forms a set of measurement data $X = (x(t_n))_{\ell=0}^{N-1}$, $x(t_n) \in \mathbb{R}^{K_x}$ in a possibly variable grid of N collocation points t_n . A visualization of the resulting trajectories can be found in Figure 1. Various methods exist to reconstruct behavioral models of ODEs from measurement data. As an introductory example, we apply sparse identification of non-linear dynamics (SINDy) by Brunton et al. (2016) as it yields a highly interpretable closed-form solution that can be easily reused. For this purpose, we introduce a function library that contains a set of M unique ansatz functions

$$\mathcal{A} = \{a_m : \mathbb{R}^{K_x} \rightarrow \mathbb{R}, m = 1 \dots M\} \quad (2)$$

This ansatz function library is usually highly diverse and contains a variety of memoryless functions such as polynomials, trigonometric, or exponential terms. The core idea of SINDy is to model the mapping f of $\dot{x} = f(x)$ via a weighted sum of said ansatz functions in the library \mathcal{A} which yields the following linear regression problem

$$Y = \dot{X} = \mathcal{A}(X)\Xi, \quad \hat{\Xi} = \underset{\Xi}{\operatorname{argmin}}(\|Y - \mathcal{A}(X)\Xi\|_2) \quad (3)$$

Problem (3) can be solved with standard methods such as SVD, however, Brunton et al. (2016) suggest including a regularization to promote sparsity in Ξ

$$\hat{\Xi} = \underset{\Xi}{\operatorname{argmin}}(\|Y - \mathcal{A}(X)\Xi\|_2 + \lambda\|\Xi\|_p) \quad (4)$$

The selection of p is crucial for the problem solution as e.g. $p = 0$ yields a mixed-integer optimization problem that requires dedicated solvers. For this work, we set $p = 1$. Using sequential thresholded linear least squares (STLSQ) in this case as described by Brunton et al. (2016) is beneficial. However, applying SINDy by itself is not feasible for identifying the underlying dynamics of virus infection processes, since typically measurements are missing for non-observable quantities. Hence we can not solve (4) and therefore, we introduce a way to extend SINDy for such scenarios in the next section.

2. RECONSTRUCTION FROM INCOMPLETE MEASUREMENT DATA

A fundamental problem of the SINDy approach is the reliance on complete measurement data. In Section 1, we introduced a reduced system that models influenza virus infection dynamics. Under laboratory conditions, however, we can not measure all system states and therefore SINDy can not be applied. For the purpose of reconstructing

governing equations from incomplete measurements, we introduce a Kalman filter SINDy hybrid that can not only deal with incomplete measurements but also yields the actual underlying system equation similar to SINDy. The proposed model is largely based on the so-called extended Kalman filter (EKF), which is suitable for non-linear system identification. To use this, we switch from common non-linear ODEs to non-linear difference equations, as such models are commonly used to describe discrete systems. Applying the following methods to a continuous system is trivial.

2.1 Extended Kalman Filter Basics

In this Section, the EKF and corresponding derivations are presented. It summarizes the work of Terejanu et al. (2008) with slight changes in notation. We use the following non-linear system

$$\begin{aligned} x_n &= f(x_{n-1}) + q_{n-1}, \quad 0 \leq n \leq N, \\ y_n &= g(x_n) + r_n, \end{aligned} \quad (5)$$

where $x_n \in \mathbb{R}^{K_x}$ is the system state at timestamp n , $y_n \in \mathbb{R}^{K_y}$ is the system output, $q_n \in \mathbb{R}^{K_x}$ is the process and $r_n \in \mathbb{R}^{K_y}$ the measurement noise. Suppose we can measure y , but not the system states x . One application of Kalman filters is to reconstruct (possibly non-observable) system states from output measurements alone. For this we require the initial state $x(0) \equiv x_0$ of the system with known mean $\mu_0 = E[x_0]$ and covariance $P_0 = E[(x_0 - \mu_0)(x_0 - \mu_0)^T]$. Furthermore, we require an uncertainty model of the process that manifests in q . In certain cases with extensive domain knowledge, we may derive such from physical laws, but for now, we set this quantity to be unknown. An estimation procedure for q (respectively its covariance matrix Q) is introduced in Section 2.3. As for the further noise characteristics we set

$$\begin{aligned} E[q_n] &= 0, \quad E[q_n q_n^T] = Q_n, \quad E[q_i q_j^T] = 0 \quad \forall i \neq j, \\ E[q_n x_0^T] &= 0 \quad \forall n, \quad E[r_n] = 0, \quad E[r_n r_n^T] = R_n, \\ E[r_i r_j^T] &= 0 \quad \forall i \neq j, \quad E[r_n x_0^T] = 0 \quad \forall n, \quad E[q_i r_j^T] \quad \forall i, j \end{aligned} \quad (6)$$

meaning both noise quantities are temporally and mutually uncorrelated (white noise) with zero mean and furthermore uncorrelated with the initial system state x_0 . Regarding the vectorial functions $f : \mathbb{R}^{K_x \times 1} \rightarrow \mathbb{R}^{K_x \times 1}$, $g : \mathbb{R}^{K_x \times 1} \rightarrow \mathbb{R}^{K_y \times 1}$, we require at least C^1 smoothness in the given domain. Kalman filtering essentially consists of two steps

- (1) a *model forecast* produces estimates of the current state variable and its associated uncertainties before considering the current observation (a-priori);
- (2) a *model correction* updates the current state variable (a-posteriori) to maximize the probability of our observations y given the a-priori estimation from the model forecast

In the following we derive the EKF equations whereby quantities with a minus exponent (e.g. x_n^-) are a-priori values calculated before considering the current observation and quantities with a plus exponent (e.g. x_n^+) are a-posteriori values calculated with respect to the current observation. A central aspect of the EKF derivation is the C^1 assumption for the non-linearities, as we can apply Taylor series expansion. We assume that we received an

optimal estimation in the last time step given the observation $x_{n-1}^+ = E[x_{n-1}|y_{n-1}]$ with the corresponding state covariance P_{n-1}^+ . The predicted current state value of our EKF model reads

$$x_n^- = E[x_n|y_{n-1}] = E[f(x_{n-1})|y_{n-1}] \quad (7)$$

neglecting q_{n-1} due to being uncorrelated to the system state. Expanding f about x_{n-1}^+ yields

$$f(x_{n-1}) \equiv f(x_{n-1}^+) + \nabla f(x_{n-1}^+)(x_{n-1} - x_{n-1}^+) + h.o.t \quad (8)$$

where ∇f is the Jacobian of f . As is common with Taylor series expansion, the higher order terms (*h.o.t*) are neglected. Therefore, from (7)

$$x_n^- \approx f(x_{n-1}^+) + \nabla f(x_{n-1}^+)E[e_{n-1}|y_{n-1}] \quad (9)$$

with $e_{n-1} \equiv x_{n-1} - x_{n-1}^+$ being the residual of previously estimated and true system state. $E[e_{n-1}|y_{n-1}] = 0$ since we assumed an optimal estimate. The corresponding forecast error and covariance are

$$\epsilon_n^- \equiv x_n - x_n^- \approx \nabla f(x_{n-1}^+)\epsilon_{n-1} + w_{n-1} \quad (10)$$

$$P_n^- \equiv E[e_n^- (e_n^-)^T] = \nabla f(x_{n-1}^+)P_{n-1}^+ \nabla f^T(x_{n-1}^+) + Q_{n-1}$$

In the *model forecast step*, we evaluated our underlying process model and gained knowledge about the forecast value x_n^- and covariance P_n^- . In the model correction, we aim to utilize this information to approximate the best state estimate x_n^+ . Terejanu et al. (2008) mentions the work of Lewis et al. (2006) regarding one possible approach to get an unbiased estimate x_n^+ of x_n (in least-squares sense)

$$x_n^+ = a + K_n z_k \quad (11)$$

From the unbiasedness condition follows

$$\begin{aligned} 0 &= E[x_n - x_n^+ | y_n] \quad (12) \\ &= E[(x_n^- + \epsilon_n) - (a + K_n g(x_n) + K_n v_n) | y_n] \\ &= x_n^- - a - K_n E[g(x_n) | y_n] \\ a &= x_n^- - K_n E[g(x_n) | y_n] \end{aligned}$$

Now we can substitute (12) in (11) to get

$$x_n^+ = x_n^- + K_n (y_n - E[g(x_n) | y_n]) \quad (13)$$

Following the same Taylor series expansion idea from the model forecast step, expanding g around x_n^- yields

$$g(x_n) \equiv g(x_n^-) + \nabla g(x_n^-)(x_n - x_n^-) + h.o.t \quad (14)$$

where ∇g is the Jacobian of the observation non-linearity g . Taking the conditional expected value w.r.t y_n of both sides and once again neglecting the higher-order terms yields

$$E[g(x_n) | y_n] \approx g(x_n^-) + \nabla g(x_n^-)E[x_n - x_n^- | y_n] \quad (15)$$

with $E[x_n - x_n^- | y_n] = 0$. Substituting in (13)

$$x_n^+ \approx x_n^- + K_n (y_n - g(x_n^-)) \quad (16)$$

Now we can also calculate the a-posteriori state error (with the derivations of Akhlaghi et al. (2017))

$$\begin{aligned} \epsilon_n^+ &\equiv x_n - x_n^+ \quad (17) \\ &= f(x_{n-1}) + w_{n-1} - x_n^- - K_n (y_n - g(x_n^-)) \\ &\approx (I - K_n \nabla g(x_n^-)) \nabla f(x_{n-1}^+) \epsilon_{n-1}^+ \\ &\quad + (I - K_n \nabla g(x_n^-)) w_{n-1} - K_n v_n \end{aligned}$$

and the a-posteriori state covariance estimate

$$\begin{aligned} P_n^+ &= E[\epsilon_n^+ (\epsilon_n^+)^T] \quad (18) \\ &= P_n^- - K_n \nabla g(x_n^-) P_n^- - P_n^- \nabla g^T(x_n^-) K_n^T \\ &\quad + K_n \nabla g(x_n^-) P_n^- \nabla g^T(x_n^-) K_n^T + K_n R_n K_n^T \end{aligned}$$

The only remaining unknown is the so-called Kalman gain K_n . As with the conventional (linear) Kalman filter, it can be calculated by minimizing the trace of P_n^+ ($\text{tr}(P_n^+)$) with regard to K_n :

$$\begin{aligned} 0 &= \frac{\partial \text{tr}(P_n^+)}{\partial K_n} \quad (19) \\ &= -(\nabla g(x_n^-) P_n^-)^T - P_n^- \nabla g^T(x_n^-) \\ &\quad + 2K_n \nabla g(x_n^-) P_n^- \nabla g^T(x_n^-) + 2K_n R_n \\ &\Leftrightarrow K_n = P_n^- \nabla g^T(x_n^-) (\nabla g(x_n^-) P_n^- \nabla g^T(x_n^-) + R_n)^{-1} \end{aligned}$$

where the trace is the sum of diagonal elements of P_n^+ and represents the total variance (uncertainty) of the a-posteriori state estimates. Finally substituting into (18) yields

$$P_n^+ = (I - K_n \nabla g(x_n^-)) P_n^- \quad (20)$$

A summary of the EKF update procedure can be found in Figure 2.

2.2 Including SINDy using State Augmentation

To include the SINDy approach into the EKF, we can simply replace the non-linearities f and g with corresponding weighted sums of ansatz functions

$$\begin{aligned} x_n &= \mathcal{A}_f(x_{n-1}) \Xi^{(f)} + q_{n-1} \quad (21) \\ y_n &= \mathcal{A}_g(x_n) \Xi^{(g)} + r_n \end{aligned}$$

where

- \mathcal{A}_f is the process non-linearity with M_f ansatz functions $a_m^{(f)} : \mathbb{R}^{K_x \times 1} \rightarrow \mathbb{R}$
- $\Xi^{(f)} \in \mathbb{R}^{M_x \times K_x}$ is the sparse process weight matrix
- \mathcal{A}_g is the observation non-linearity with M_g ansatz functions $a_m^{(g)} : \mathbb{R}^{K_x \times 1} \rightarrow \mathbb{R}$
- $\Xi^{(g)} \in \mathbb{R}^{M_y \times K_y}$ is the sparse observation weight matrix

Evaluating the non-linearities at x_n produces a row vector that contains the results of the individual ansatz functions such as $\mathcal{A}_f(x_n) \equiv [a_1^{(f)}(x_n), \dots, a_{M_f}^{(f)}(x_n)]$ and $\mathcal{A}_g(x_n) \equiv [a_1^{(g)}(x_n), \dots, a_{M_g}^{(g)}(x_n)]$. We require that each ansatz function in \mathcal{A}_f and \mathcal{A}_g is at least C^1 smooth. Our goal is to identify a system characterized by the process and observation weight matrices $\Xi^{(f)}$ and $\Xi^{(g)}$ that approximates our given observations y_n in the least squares sense. To simplify this task, we may assume that we know $\Xi^{(g)}$ based on domain knowledge. However $\Xi^{(f)}$ remains unknown and therefore, we define an augmented system state

$$\tilde{x}_n = [x_n^T, \xi^{(f)}]^T \quad (22)$$

which also contains the vectorization of $\Xi^{(f)}$, namely $\xi^{(f)}$. Since $\Xi^{(f)}$ is a constant weight matrix, we can define a new EKF model with an augmented state

$$\begin{aligned} \tilde{x}_n &= \begin{bmatrix} x_n \\ \xi^{(f)} \end{bmatrix} = \begin{bmatrix} \mathcal{A}_f(x_{n-1}) \Xi_{n-1}^{(f)} \\ \xi_{n-1}^{(f)} \end{bmatrix} + \tilde{q}_{n-1} \quad (23) \\ y_n &= \mathcal{A}_g(x_n) \Xi^{(g)} + r_n \end{aligned}$$

where $\tilde{q}_{n-1} = [q_{n-1}^T, \omega]$ and ω is the expected uncertainty of the process weight matrix coefficients. Note that $\xi_n^{(f)}$ has no direct effect on the observation y_n and is therefore neglected in the second equation. The process Jacobian has to be updated to reflect the augmented state as well

$$\nabla f(\tilde{x}_n) = \begin{bmatrix} \nabla \mathcal{A}_f(x_n) \Xi_n^{(f)} & \mathcal{A}_f(x_n) \otimes I \\ 0 & I \end{bmatrix} \quad (24)$$

where $\nabla \mathcal{A}_f(x_n)$ is the Jacobian of the process non-linearity (our SINDy model) w.r.t x_n and $\nabla f(\tilde{x}_n)$ represents the Jacobian of the augmented process equation. We can employ a similar process in the calculation of $\nabla g(\tilde{x}_n)$, but in most cases, a dedicated output non-linearity is not required. We can write

$$y_n = Gx_n + v_n \quad (25)$$

where G is an observability matrix indicating which states in x_n appear in y_n . The calculation of the corresponding Jacobian $\nabla g(\tilde{x}_n)$ is trivial.

2.3 Estimating Process/Observation Covariance Matrices

Another challenge of the introduced SINDy-augmented EKF is the selection/estimation of the process noise covariance Q_n , which also implicitly includes the SINDy parameter uncertainty. One common approach for linear Kalman filters is to update Q_n and the measurement covariance matrix R_n at each step according to Mehra (1970). For EKFs, Akhlaghi et al. (2017) provide an innovation/residual-based estimation approach which is referred to as *adaptive* EKF (AEKF), whereby innovation d_n is the difference between observations and a-priori state estimations contrary to residuals e_n using a-posteriori estimations

$$\begin{aligned} d_n &= y_n - g(\tilde{x}_n^-) && \text{(innovation)} && (26) \\ e_n &= y_n - g(\tilde{x}_n^+) && \text{(residual)} \end{aligned}$$

Since covariance matrices are positive definite, Akhlaghi et al. (2017) uses a residual-based approach to estimate R_n as introduced by Wang (1999)

$$\begin{aligned} S_n &= E[e_n e_n^T] = E[v_n v_n^T] - \nabla g(\tilde{x}_n^-) P_n^- \nabla g^T(\tilde{x}_n^-) && (27) \\ R_n &= E[v_n v_n^T] = S_n + \nabla g(\tilde{x}_n^-) P_n^- \nabla g^T(\tilde{x}_n^-) \end{aligned}$$

We can estimate R_n using an exponential window with a forgetting factor $0 < \alpha \leq 1$ and $\beta = 1 - \alpha$. This yields

$$R_n = \alpha R_{n-1} + \beta (e_n e_n^T + \nabla g(\tilde{x}_n^-) P_n^- \nabla g^T(\tilde{x}_n^-)) \quad (28)$$

A similar principle can be applied regarding the process noise covariance matrix Q_n using the innovation. We can estimate the process noise using the a-posteriori state estimation as in Akhlaghi et al. (2017)

$$\begin{aligned} \hat{q}_{n-1} &= \tilde{x}_n^+ - \tilde{f}(\tilde{x}_{n-1}^+) && (29) \\ &= \tilde{x}_n^+ - \tilde{x}_n^- = K_n [y_n - g(\tilde{x}_n^-)] = K_n d_n \end{aligned}$$

and therefore

$$\begin{aligned} Q_{n-1} &\approx E[\hat{q}_{n-1} \hat{q}_{n-1}^T] && (30) \\ &= E[K_n (d_n d_n^T) K_n^T] = K_n E[d_n d_n^T] K_n^T \end{aligned}$$

We can determine $E[d_n d_n^T]$ using the same exponential window as with the process noise covariance

$$Q_n = \alpha Q_{n-1} + \beta (K_n d_n d_n^T K_n^T) \quad (31)$$

Combining the principles of covariance matrix estimation with our SINDy augmenting approach forms an algorithm that we call *augmented adaptive* EKF (AAEKF). The

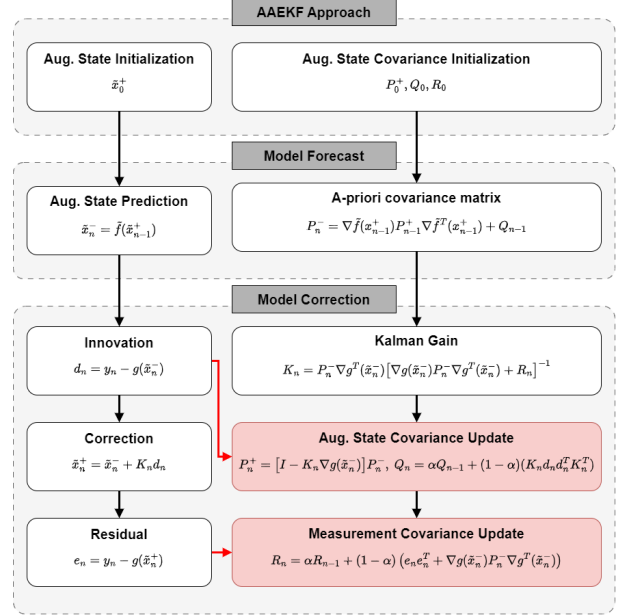


Fig. 2. Flowchart of the proposed augmented AAEKF, inspired by Akhlaghi et al. (2017)

flowchart for a single iteration of the AAEKF is displayed in Figure 2.

2.4 Introducing Sparsity into AAEKF

AAEKF using SINDy models does however lack a crucial aspect in its current form: the added augmentation state $\xi^{(f)}$ is not sparse. To solve this, we introduce a sparsity criterion for the process equation into AAEKF, which leads to more compact system approximations. One way to include sparsity is to modify the Kalman gain accordingly by adding a regularization term to (19)

$$0 = \frac{\partial \text{tr}(P_n^+) + \lambda \|x_n^+\|_p}{\partial K_n} \quad (32)$$

where $x_n^+ \approx x_n^- + K_n (y_n - g(x_n^-))$ as in (16). Similar to (19) we need to solve (32) for K_n , which is trivial for $p = 2$ but only moderately useful to facilitate sparsity in x_n^+ as it penalizes the sum of squares

$$\begin{aligned} K_n &= (P_n^- \nabla g^T(x_n^-) + \lambda x_n^- d_n^T) && (33) \\ &\cdot (\nabla g(x_n^-) P_n^- \nabla g^T(x_n^-) + R_n + \lambda d_n d_n^T)^{-1} \end{aligned}$$

where $d_n = y_n - g(x_n^-)$ once again resembles the model innovation. Selecting $p = 1$ as commonly used for SINDy by Brunton et al. (2016) yields

$$0 = \frac{\partial \text{tr}(P_n^+)}{\partial K_n} + \lambda \cdot \text{sign}(x_n^- + K_n d_n) d_n^T \quad (34)$$

which can not be directly solved for K_n . There exist different methods to handle such problems, but we can also formulate an algorithm similar to *sequential thresholded linear least squares* (STLSQ) as introduced by Brunton et al. (2016). In Algorithm 1 we initially solve the non-sparse regression problem and systematically remove elements in the a-posteriori covariance matrix P_n^- to update the Kalman gain estimations until a sufficient degree of sparsity is reached.

Algorithm 1 Adapted STLSQ for AAEKF

- 1: $B = P_n^- \nabla g(\tilde{x}_n^-)$, $A = (\nabla g(x_n^-) P_n^- \nabla g^T(x_n^-) + R_n)^{-1}$
 - 2: Compute the non-sparse initial solution $K = BA^{-1}$
 - 3: **while** $\xi_n^+ \in x_n^+ = x_n^- + K_n d_n$ is non-sparse **do**
 - 4: find non-zeros in ξ_n^+ and ξ_n^- smaller than λ
 - 5: set the app. rows and columns in P_n^- to zero
 - 6: update A and B
 - 7: re-calculate the Kalman gain $K = BA^{-1}$
 - 8: **end while**
 - 9: calculate the posterior state $\tilde{x}_n^+ = \tilde{x}_n^- + K_n d_n$
 - 10: set all non-zeros $\xi_n^+ \in \tilde{x}_n^+$ smaller than λ to zero
-

3. PARALLELS TO EXISTING LITERATURE

Literature research indicates there are other publications concerned with fusing some aspects of SINDy with the capabilities of EKF. To highlight the novelty of this work, we are therefore adding context on the main differences, most notably w.r.t the works of Stevens-Haas et al. (2024) and Rosafalco et al. (2024). Especially similarities to the method of Stevens-Haas et al. (2024) are minor, as it is concerned with making SINDy more robust to noise by using Kalman smoothing in the data assimilation step to compensate the noise amplification in numerical derivatives. The proposed method in this work however utilizes EKFs to perform the sparse regression step of SINDy itself and therefore covers a different anchor point. In Rosafalco et al. (2024) multiple algorithms are presented, e.g. time delay embedding approaches, but the focus in the context of this work lies on the EKF-related part. The underlying idea of utilizing a SINDy approach for the process and observation equations as in (21) and co-estimating the corresponding weight coefficients alongside the actual system state can be considered equivalent. However, Rosafalco et al. (2024) does not account for estimating the process and observation covariance matrices Q, R which makes the proposed method harder to parameterize. Additionally no direct sparsity mechanism has been introduced into the EKF-SINDy approach (online phase) of Rosafalco et al. (2024). For AAEKFs both points are covered.

4. INFLUENZA DYNAMICS RECONSTRUCTION

In the following section, we apply our Kalman filter hybrid model (AAEKF) to identify a realistic influenza virus model containing only observations of measurable quantities. To promote the stability of the identified system we utilize the integral formulation in combination with an appropriate ODE solver

$$x(t_n) = x(t_{n-1}) + \int_{t_{n-1}}^{t_n} f(x(t)) dt \quad (35)$$

whereby $x(t_n) \equiv x_n$ for all collocation points. This formulation has the decisive advantage, that fewer measurements are required compared to the discrete variant and furthermore we can detect instabilities during the Kalman filter update process by evaluating the ODE solver output. The Kalman filter flow from Figure 2 holds nonetheless. Under realistic conditions resulting from common experimental setups, only x_2, x_3, x_8 , the sum of x_4, x_5 and the initial value of x_1 are directly measurable. To generate artificial observations, we solve (1) for $x(0) = [10, 1, 1, 1, 1, 1, 1]$ in the interval $t = [0, 5]$ and select $N = 100$ equidistant

collocation points along the trajectories. As we introduced mechanisms to estimate Q, R along the identification procedure, we initialize those to be positive definite diagonal matrices of random values around $\mu = 10^{-3}$. This is mainly due to the reason that we do not consider explicit process or observation noise in our artificially generated observations, yet. Our sparsification parameter is $\lambda = 1e-4$ and $\alpha = 0.99$. From domain knowledge (see Equation (1)) we know that a linear-quadratic ansatz function library is sufficient to describe the virus dynamics. As a comparative measure we employ different figures of merit (FOM), namely the maximum approximation error, the mean squared error (MSE) and its root variants (RMSE, NRMSE). As evident in Figure 3 and numerically backed by Table 1, the approximated trajectories (red, dashed) for $x_2, x_3, x_4 + x_5$ and x_8 match the available observations (black) well, whereby the a-priori estimations of the states x_n^- from the Kalman filter are displayed. However, we could only approximate the sum of x_4, x_5 well which is due to the fact that we don't have separated observations for these variables. In Figure 4, the remaining unobservable system state approximations for x_1, x_6, x_7 are displayed including the confidence intervals resulting from the corresponding values within the state covariance matrix P_n^- at the end of the Kalman filter update process. As visible in Figure 4, the reconstructed unobservable system states are partially negative, which is not possible in actual infection dynamics. However, this can not be prevented during the approximation process without constraints, which is currently under investigation. A promising approach includes reformulating the minimization process of the a-posteriori state covariance matrix P_n^+ into a quadratic programming (QP) problem and applying constraints onto the a-posteriori estimation x_n^+ of the system state. However, the identified coefficients for the chosen SINDy ansatz function library of at most quadratic polynomials (see Figure 5) bears little resemblance to the reference system in (1). This is to be expected since we are missing crucial information due to the incomplete state measurements. What can be noticed is the significant sparsity of the Ξ coefficients over all sub-equations due to applying Algorithm 1.

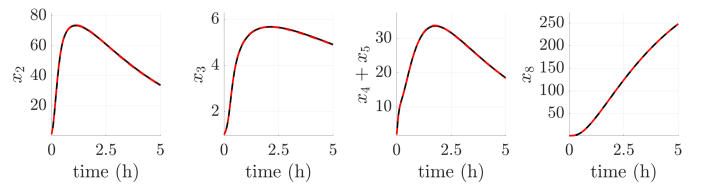


Fig. 3. Approximation of x_n^- (red) for the observed influenza virus behavior ($x_2, x_3, x_4 + x_5, x_8$) (black)

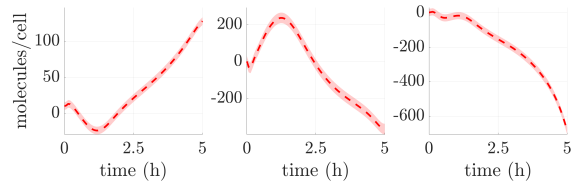


Fig. 4. Approximation of x_n^- for the hidden influenza virus behavior (x_1, x_6, x_7)

ACKNOWLEDGEMENTS

This work is part of the project SparseRF, which has been funded by the Austrian Research Promotion Agency (FFG) and the state of Upper Austria under grant 40338



REFERENCES

- Akhlaghi, S., Zhou, N., and Huang, Z. (2017). Adaptive adjustment of noise covariance in Kalman filter for dynamic state estimation. In *2017 IEEE Power & Energy Society General Meeting*, 1–5. IEEE.
- Baccam, P., Beauchemin, C., Macken, C.A., et al. (2006). Kinetics of influenza a virus infection in humans. *J Virol*, 80(15), 7590–9.
- Beauchemin, C.A. and Handel, A. (2011). A review of mathematical models of influenza a infections within a host or cell culture: lessons learned and challenges ahead. *BMC Public Health*, 11 Suppl 1(Suppl 1), S7.
- Bocharov, G.A. and Romanyukha, A.A. (1994). Mathematical model of antiviral immune response III. Influenza a virus infection. *J Theor Biol*, 167(4), 323–60.
- Brunton, S.L., Proctor, J.L., and Kutz, J.N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15), 3932–3937.
- Heldt, F.S., Frensing, T., and Reichl, U. (2012). Modeling the intracellular dynamics of influenza virus replication to understand the control of viral RNA synthesis. *J Virol*, 86(15), 7806–17.
- Lewis, J.M., Lakshmivarahan, S., and Dhall, S. (2006). *Dynamic Data Assimilation: A Least Squares Approach*. Encyclopedia of Mathematics and its Applications. Cambridge University Press.
- Mehra, R. (1970). On the identification of variances and adaptive Kalman filtering. *IEEE Transactions on Automatic Control*, 15(2), 175–184.
- Perelson, A.S., Rong, L., and Hayden, F.G. (2012). Combination antiviral therapy for influenza: predictions from modeling of human infections. *J Infect Dis*, 205(11), 1642–1645.
- Rosafalco, L., Conti, P., Manzoni, A., Mariani, S., and Frangi, A. (2024). EKF-sindy: Empowering the extended kalman filter with sparse identification of nonlinear dynamics. *Computer Methods in Applied Mechanics and Engineering*, 431, 117264. doi: <https://doi.org/10.1016/j.cma.2024.117264>. URL <https://www.sciencedirect.com/science/article/pii/S0045782524005206>.
- Rüdiger, D., Piasecka, J., Kuchler, J., et al. (2024). Mathematical model calibrated to in vitro data predicts mechanisms of antiviral action of the influenza defective interfering particle "OP7". *iScience*, 27(4), 109421.
- Stevens-Haas, J.M., Bhangale, Y., Nathan Kutz, J., and Aravkin, A. (2024). Learning nonlinear dynamics using kalman smoothing. *IEEE Access*, 12, 138564–138574. doi:10.1109/ACCESS.2024.3465390.
- Terejanu, G.A. et al. (2008). Extended Kalman filter tutorial. *University at Buffalo*, 27.
- Wang, J. (1999). Stochastic modeling for real-time kinematic GPS/GLONASS positioning. *Navigation*, 46(4), 297–305.

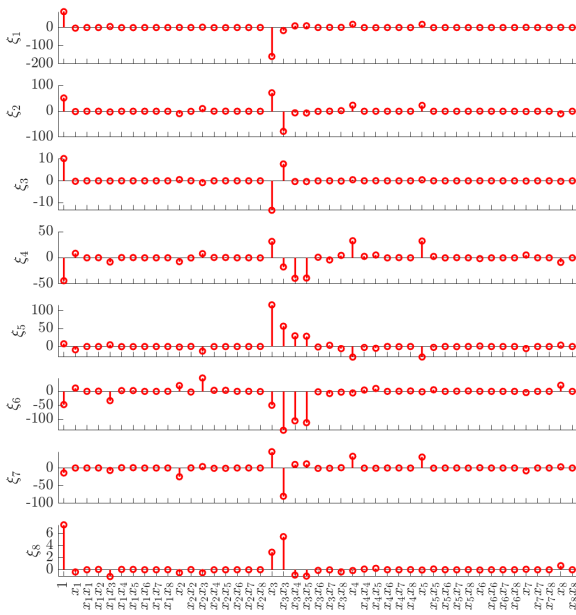


Fig. 5. SINDy library coefficients for each Influenza dynamics sub-system (x_1, \dots, x_8)

State	FOM			
	max	MSE	RMSE	NRMSE
x_2	5.2e-2	4.8e-2	2.2e-1	3.0e-3
x_3	2.1e-2	2.1e-4	1.5e-2	3.1e-3
x_4	1.0e3	1.8e5	4.3e2	4.9e1
x_5	1.0e3	1.8e5	4.3e2	1.3e1
$x_4 + x_5$	1.7e-1	7.8e-3	8.8e-2	2.8e-3
x_8	1.3e-1	4.1e-3	6.4e-2	2.6e-4

Table 1. Influenza dynamics model FOM

5. CONCLUSION AND REMARKS

We introduced a novel extended Kalman filter SINDy hybrid model (AAEKF) that is capable of identifying governing equations from incomplete measurement data. Furthermore, we added ways to make the update process using available system observations more robust by co-estimating the process respectively measurement noise covariance matrices and by introducing sparsity criteria similar to the original SINDy model by Brunton et al. (2016). We then applied our proposed model onto a reduced model of influenza virus infection dynamics, whereby only few system states are observable respectively measurable. From our experiments, we can conduct that the Kalman filter SINDy hybrid yields promising approximation results. This work opens up several promising avenues for future work, as both the topic of Kalman filters and SINDy approaches are vast.

Data Availability: All implementations and experiments are available upon request as MATLAB code at a GitLab repository (<https://gitlab.fh-ooe.at/fe/sparserf>) under the MIT license that is authored by the University of Applied Sciences Upper Austria.