TRANSACTIONS OF THE
INSTITUTE OF
MEASUREMENT & CONTROL

Transactions of the Institute of
Measurement and Control
1–15
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/01423312231203030
journals.sagepub.com/home/tim
Sage

Anomaly detection and prediction evaluation for discrete nonlinear dynamical systems

Jan Michael Spoor[1], Jens Weber[2] and Jivka Ovtcharova[1]

Abstract

Anomalies in dynamical systems mostly occur as deviations between measurement and prediction. Current anomaly detection methods in multivariate time series often require prior clustering, training data, or cannot distinguish local and global anomalies. Furthermore, no generalized metric exists to evaluate and compare different prediction functions regarding their amount of anomalous behavior. We propose a novel methodology to detect local and global anomalies in time series data of dynamical systems. For this purpose, a theoretical density distribution is derived assuming that only noise conceals the time series. If the theoretical and the empirical density distribution yield significantly different entropies, an anomaly is assumed. For a local anomaly detection, the Mahalanobis distance using the theoretical noise distribution's covariance is applied to evaluate sequences of predictions and measurements. In addition, the Wasserstein metric enables a comparison of predictions using the distance between the noise and empirical distribution as a measure for selecting the best prediction function. The proposed method performs well on nonlinear time series such as logistic growth and enables a useful selection of a prediction model for satellite orbits. Thus, the proposed method improves anomaly detection in time series and model selection for nonlinear systems.

Keywords

Anomaly detection, dynamical systems, information entropy, Mahalanobis distance, time series, Wasserstein metric

Introduction

When controlling dynamical engineering systems, predictions are used to form expectations of future system behavior and enable a more controlled environment. These predictions are limited by the mathematical model computing the prediction process. During the operation and observation of a system, deviations between a measured actual state and a planned state might trigger a corresponding reaction or a targeted need for action by engineers. The defined plan value can be determined via a simulation using a prediction process and is used as a target value. If the actual state deviates measurably, relevantly, and significantly from a desired plan state, the system will indicate either a warning or a fault. Thus, the risk of a shutdown exists, especially in the case of a long-term inability to correct a deviation. Simulations are hereby a powerful tool to predict and classify malfunction states in advance, to avert possible malfunctions during regular operations, or to start countermeasures in advance. Nevertheless, during the operation and observation of a system, states might still occur, which were not predicted in advance and can represent malfunctions. These would generally be recognized as an anomaly in the data, defined as a substantial deviation from the norm (Mehrotra et al., 2017).

From a system planner's point of view, the anomaly detection process, as well as the evaluation of the precision of a prediction, is therefore a comparison of the expected system state (a planned value) with the actual system state (an actual value). In this definition, anomalies are not the result of noisy data, expected malfunctions, or errors using the prediction models but rather novel deviations not explainable by the underlying prediction process (Spoor et al., 2022). Therefore, anomalies in the measured data are the limitations of these prediction models, if measurement and prediction differ in a substantial manner from the normal data model (Mehrotra et al., 2017). Hereby, global anomalies are referred to as systematic differences between measurements and prediction within the whole time series and local anomalies are distinguishable time frames and spots containing anomalous values. Local anomalies can be further divided into point outlier, contextual anomalies, and collective anomalies (Lindemann et al., 2021). Future challenges in anomaly detection for time series in Internet of Things application are given by Cook et al. (2020) as the development of unsupervised methods, real-time processing, and the generalization of methods.

[1]Institut für Informationsmanagement im Ingenieurwesen, Karlsruhe Institute of Technology, Germany
[2]Faculty of Technology - Mechatronics Trinational, Baden-Wuerttemberg Cooperative State University Lörrach, Germany

Corresponding author:
Jan Michael Spoor, Institut für Informationsmanagement im Ingenieurwesen, Karlsruhe Institute of Technology, Kriegsstraße 77, 76133 Karlsruhe, Germany.
Email: jan.spoor@kit.edu

We propose a novel methodology for the detection and evaluation of global and local anomalies in systems with a discrete measurement and prediction process for multivariate time series data. The idea is to compare the measured covariance matrix and a theoretically derived covariance matrix for which is assumed that only noise conceals the measurements. The comparison is conducted by an entropy measure for finding global anomalies and the application of the Wasserstein metric is used as a measure to compare the amount of anomalous behavior of different predictions. In addition, a local anomaly detection is conducted by applying the Mahalanobis distance using the theoretical noise covariance matrix. This methodology improves the state-of-the-art of anomaly detection of multivariate time series by enabling a local as well as global anomaly detection without prior clustering, training data, or the use of a correct time series as baseline. Furthermore, this methodology provides a novel measure for selecting the best prediction function to improve model selection for nonlinear systems.

This contribution starts with an overview of the current literature for anomaly detection in time series data. Subsequently, the theoretical derivations for the methodology and the setup of the theoretical noise covariance matrix are discussed. Based on the derived methodology, a simulation study using logistic growth as an example of a nonlinear time series is conducted to prove the capabilities of our proposed method for a local and global anomaly detection. In addition, a use case is discussed to compare the amount of anomalous factors in predictors of satellite orbits using the Wasserstein metric. Thereafter, our proposed method is discussed and a conclusion is given.

Literature review

Multiple papers discuss anomaly detection in multivariate and univariate time series. In the case of a local anomaly detection in multivariate time series, Blázquez-García et al. (2021) distinguish model-based approaches (either by prediction or estimation), methods using histograms (for point outliers), and dissimilarity-based approaches. For a global anomaly detection, Blázquez-García et al. (2021) name dissimilarity-based approaches and dimensionality reduction as techniques.

Since information entropy is used as a metric to estimate system complexity (Pincus, 1991), local outliers are detected or anomaly affected areas identified within univariate time series by using the Shannon entropy (He et al., 2021). With this approach, no global anomalies can be detected. However, the Shannon entropy (Germán-Salló, 2018) or the permutation entropy (Bandt and Pompe, 2002) are proven to be, in principle, useful measures to detect anomalies within time series.

Wang et al. (2011) use the correlation of a suspected anomaly affected signal and a known correct signal without anomalies so that global anomalies in the suspected signal are detected. This approach requires an identified second correct system and no local anomaly detection is conducted. Similar to the correlation of two signals, autocorrelation in anomaly detection is widely used (Izakian and Pedrycz, 2013). However, these methods lack the possibility for a global anomaly detection and are applied for univariate time series.

Li et al. (2021) use clustering of multivariate time series and they analyze the data points with a distance measure so that local outliers are detected. The clustering is conducted using a Gaussian Mixture Model solved through the usage of an EM-algorithm, which is enabled by the Mahalanobis distance. In other methods and applications, the Mahalanobis distance provides good results, but a prior clustering is necessary (Sperandio Nascimento et al., 2015) or the data sources are contextually clustered beforehand (Titouna et al., 2019). In the case of clustering, the covariance matrix of a time series is estimated using the priorly set up clusters. In some approaches, the covariance matrix of nonlinear systems is approximated using simulations and evaluated using the Mahalanobis distance (Burr et al., 1994).

Concluding, machine learning is another approach. An advantage of machine learning is that no model assumptions of the analyzed time series are necessary. However, supervised approaches from machine learning, for example, a Support Vector Machine as implemented by Rodriguez et al. (2010), require labeled data sets. Approaches using unsupervised neural network architectures, for example, an Autoencoder as implemented by Audibert et al. (2020), require a prior training phase and the assumption of a training data set without or with only very small amount of anomalies. In recent years, architectures based on long short-term memory (LSTM) are developed but also require extensive training data and sometimes labeling (Lindemann et al., 2021). LSTMs are also used for creating predictions which are then evaluated using a local average with adaptive parameters to detect local anomalies (Tan et al., 2020).

Theoretical derivation of methodology

Applied measurement and prediction model

Following the proposed system description by Spoor et al. (2022), a system state is given by the multivariate description x_i of J real features. This system has a measurement process g, which transforms the real system state into the measured system state $\hat{x}_i$ with D measurable features. This state is affected by noise ϵ_i so that only state $\hat{x}_i^*$ is measured. In addition, for each real operation f transforming the state x_i into state x_{i+1}, a prediction $\hat{f}$ exists, which transforms a measured system state $\hat{x}_i^*$ into a predicted system state $\hat{x}_{i+1}$. The measurement and prediction model can be linear as well as nonlinear. Thus, the system description is applicable for most dynamical systems

$$\begin{aligned} \hat{x}_i^* &= g(x_i) + \epsilon_i \\ \hat{x}_{i+1} &= \hat{f}\left(\hat{x}_i^*\right) \end{aligned} \tag{1}$$

We assume, for most applications, white noise under a normal distribution and an expected mean of zero

$$\epsilon \sim \mathcal{N}\left(0, \sigma^2\right),\ Cov(\epsilon, \epsilon') = 0 \tag{2}$$

In the case of a variance depending on the feature, the variance in the following derivations can simply be adjusted to

$\sigma(x_i)^2$. All following equations can be adjusted for colored noise by applying the changed assumptions. The overall methodology does not change for colored noise.

When modeling a system, a precise knowledge of function f is targeted. If function $\hat{f}$ can predict most system states precisely, the system runs as expected and becomes controllable. When measuring the efficiency of the function $\hat{f}$, the delta between the expected and measured system state becomes an important metric

$$\begin{aligned} \Delta_{i+1} &= \hat{x}^*_{i+1} - \hat{x}_{i+1} \\ &= g(f(x_i)) + \epsilon_{i+1} - \hat{f}(g(x_i) + \epsilon_i) \\ \Leftrightarrow \hat{x}^*_{i+1} &= \hat{f}\left(\hat{x}^*_i\right) + \Delta_{i+1} \end{aligned} \tag{3}$$

Δ_{i+1} includes three linked information (Spoor et al., 2022).

1) Noise and measurement inaccuracy ($\epsilon_{i+1}, \epsilon_i$)
2) Ignorance of the real features of the system (g)
3) Ignorance of the effects of the real operations (f,$\hat{f}$)

To create a precise prediction of future system states, the goal of an engineer is to select $\hat{f}$ so that $\Delta_{i+1} \rightarrow 0$. Deviations between prediction and measurement are the result of the following reasons:

1) Noise in measurements results in distorted predictions of the system
2) Ignorance of the real J features of the system
3) Ignorance of the effects of the real operation f regarding:

   (a) Observable features
   (b) Unobservable features

4) Complexity of the model and limitation of the model due to computational power. Therefore, not all effects on the observable features are precisely modeled

Reason 4 is an additional explanation to reason 1–3 since even if reason 1–3 could be solved, limitations due to computational power still apply and decrease the precision of the predictions and result in noteworthy discrepancies of the expected state and the measured state. Therefore, reason 4 is more of a technical expansion of reason 1–3.

These items describe the reasons for unexpected states despite extensive simulations and knowledge of the system. This raises the question of how these influences can be incorporated into a model. Most notably the Kalman filter enables a correction of predictions, which improves the predictions and state estimations without bias. However, the Kalman filter offers no applicable metric in measuring and comparing the performance of two models and evaluating how inaccurate a model is from the real states. Therefore, it becomes an important task not only to correct the prediction but also to spot the anomalous behavior of a prediction, that is, in which cases the prediction is more inaccurate than in other cases and to define a measure to evaluate the precision or anomalous behavior of the prediction. This evaluation is set up by viewing the occurring deviations from the prediction as a distribution and comparing this distribution with sample distributions only affected by noise and not the ignorance of the real features and operations.

Derivation of theoretical noise covariance matrix

If the knowledge of the underlying operations f and transformation of observation g is ignored, the resulting delta after each transition of states can be described as a statistical process of time following an unknown distribution. For good approximations of the function f, when applying equation (3), Δ_{i+1} should become zero. For two immediately following states i and $i+1$, this process relates to

$$\begin{pmatrix} \Delta_{i+1} \\ \Delta_i \end{pmatrix} \sim \Psi(0, \Sigma_{\Delta_{i+1}, \Delta_i}) \tag{4}$$

Furthermore, we know the distribution Ψ must be influenced by a normal distribution of white noise with an unknown covariance matrix

$$\{\Delta_i\}_{i \in T} \sim \mathcal{N}(0, \Sigma_{Noise}) \tag{5}$$

The distribution is also influenced by an unknown distribution of the ignorance of function f and observation transformation g

$$\{\Delta_i\}_{i \in T} \sim \Phi\left(0, \Sigma_{Ignorance}\right) \tag{6}$$

For further analysis, Δ_{i+1} is written as follows

$$\Delta_{i+1} = \hat{x}^*_{i+1} - \hat{x}_{i+1} = \hat{x}_{i+1} + \epsilon_{i+1} - \hat{f}(\hat{x}_i + \epsilon_i) \tag{7}$$

If the prediction function $\hat{f}$ is perfect for $\hat{x}_{i+1} = \hat{f}(\hat{x}_i)$, then Δ_{i+1} only corresponds to the white noise, which is a combination of ϵ_i and ϵ_{i+1}. Assuming function $\hat{f}$ is a smooth function and infinitely differentiable and the growth is limited by $\hat{f}'' \leqslant \hat{f}'$, a Taylor series for the term $\hat{f}(\hat{x}_i + \epsilon_i)$ is applied (Spoor et al., 2022) as follows

$$\begin{aligned} \hat{f}(\hat{x}_i + \epsilon_i) &= \sum_{k=0}^{\infty} \frac{\epsilon_i^k}{k!} \hat{f}^{(k)}(\hat{x}_i) \\ &= \hat{f}(\hat{x}_i) + \hat{f}'(\hat{x}_i) * \epsilon_i + \hat{f}''(\hat{x}_i) * \frac{\epsilon_i^2}{2} + \cdots \end{aligned} \tag{8}$$

Since ϵ_i is noise, $\| \epsilon_i \| \ll \| \hat{x}_i \|$ is assumed in the case of good measurement equipment. The growth of function f is assumed to be limited by $\| \sum_{k=2}^{\infty} \frac{\epsilon_i^k}{k!} \hat{f}^{(k)}(\hat{x}_i) \| \ll \| \hat{f}'(\hat{x}_i)\, \epsilon_i \|$ (Spoor et al., 2022) and is given as

$$\begin{aligned} \hat{f}(\hat{x}_i + \epsilon_i) &= \hat{f}(\hat{x}_i) + \hat{f}'(\hat{x}_i) * \epsilon_i + \mathcal{O}\left(\hat{f}^{(2)}(\hat{x}_i + \epsilon_i)\right) \\ &\approx \hat{f}(\hat{x}_i) + \hat{f}'(\hat{x}_i) * \epsilon_i \end{aligned} \tag{9}$$

Therefore, Δ_{i+1} simplifies to

$$\begin{aligned} \Delta_{i+1} &= g(f(x_i)) + \epsilon_{i+1} - \hat{f}(g(x_i) + \epsilon_i) \\ &\approx \left(\hat{x}_{i+1} - \hat{f}(\hat{x}_i)\right) + \left(\epsilon_{i+1} - \hat{f}'(\hat{x}_i) * \epsilon_i\right) \end{aligned} \tag{10}$$

It should be noted that $\hat{f}'(\hat{x}_i)$ is the total differential over all features D

$$\hat{f}(\hat{x}_i + \epsilon_i) \approx \hat{f}(\hat{x}_i) + \sum_{d=1}^{D} \frac{\partial}{\partial x_{i,d}} \hat{f}(\hat{x}_i) * \epsilon_{i,d} \tag{11}$$

The term $\hat{x}_{i+1} - \hat{f}(\hat{x}_i) = (g(f(x_i) - \hat{f}(g(x_i))) = \lambda_{i+1}$ describes the ignorance of the operations and observation transformation while the measurement error from noise is described by the term $(\epsilon_{i+1} - \hat{f}'(\hat{x}_i) * \epsilon_i) = \tau_{i+1}$. From this results

$$\begin{aligned} \Delta_{i+1} &= \lambda_{i+1} + \tau_{i+1} \\ \{\lambda_i\}_{i \in T} &\sim \Phi(0, \Sigma_{Ignorance}) \\ \{\tau_i\}_{i \in T} &\sim \mathcal{N}(0, \Sigma_{Noise}) \end{aligned} \tag{12}$$

Both time series are uncorrelated but not independent. Since the time series of $\hat{x}_i$ and ϵ_i are uncorrelated, $\mathbb{E}[\hat{f}(\hat{x}_i)\epsilon_i] = \mathbb{E}[\hat{f}(\hat{x}_i)]\mathbb{E}[\epsilon_i]$, and $\mathbb{E}[\hat{x}_{i+1}\epsilon_i] = \mathbb{E}[\hat{x}_{i+1}]\mathbb{E}[\epsilon_i]$ applies

$$\begin{aligned} Cov(\lambda_{i+1}, \tau_{i+1}) = \; & Cov(\hat{x}_{i+1} - \hat{f}(\hat{x}_i), \epsilon_{i+1} - \hat{f}'(\hat{x}_i)\epsilon_i) \\ &= Cov(\hat{x}_{i+1}, \epsilon_{i+1}) - Cov(\hat{x}_{i+1}, \hat{f}'(\hat{x}_i)\epsilon_i) \\ &- Cov(\hat{f}(\hat{x}_i), \epsilon_{i+1}) + Cov(\hat{f}(\hat{x}_i), \hat{f}'(\hat{x}_i)\epsilon_i) \\ &= - Cov(\hat{x}_{i+1}, \hat{f}'(\hat{x}_i)\epsilon_i) + Cov(\hat{f}(\hat{x}_i), \hat{f}'(\hat{x}_i)\epsilon_i) \\ &= - \mathbb{E}[\hat{x}_{i+1}\hat{f}'(\hat{x}_i)\epsilon_i] + \mathbb{E}[\hat{x}_{i+1}]\mathbb{E}[\hat{f}'(\hat{x}_i)\epsilon_i] \\ &+ \mathbb{E}[\hat{f}(\hat{x}_i)\hat{f}'(\hat{x}_i)\epsilon_i] - \mathbb{E}[\hat{f}(\hat{x}_i)]\mathbb{E}[\hat{f}'(\hat{x}_i)\epsilon_i] \\ &= - \mathbb{E}[\hat{x}_{i+1}\hat{f}'(\hat{x}_i)]\mathbb{E}[\epsilon_i] + \mathbb{E}[\hat{x}_{i+1}]\mathbb{E}[\hat{f}'(\hat{x}_i)]\mathbb{E}[\epsilon_i] \\ &+ \mathbb{E}[\hat{f}(\hat{x}_i)\hat{f}'(\hat{x}_i)]\mathbb{E}[\epsilon_i] - \mathbb{E}[\hat{f}(\hat{x}_i)]\mathbb{E}[\hat{f}'(\hat{x}_i)]\mathbb{E}[\epsilon_i] \\ &= 0 \end{aligned} \tag{13}$$

If we want to calculate the noise term τ_i of the time series $\{\Delta_i\}_{i \in T}$, we have to calculate Σ_{Noise}. The covariance Σ_{Noise} describes the distribution of the time series in the case of a perfect prediction function $\hat{f}$ since in this case $\Delta_{i+1} = \tau_{i+1}$. The variance of ϵ_i is given as $Var(\epsilon_i) = \sigma^2$ as assumed in equation (2). The variance of the measurement noise of τ_{i+1} for a specific feature k is analyzed using e_k as the identity vector of k

$$\begin{aligned} Var(\tau_{i+1,k}) &= Var\left(\epsilon_{i+1,k} - \sum_{d=1}^{D} \frac{\partial}{\partial x_{i,d}} \hat{f}(\hat{x}_i) * \epsilon_{i,d} * e_k\right) \\ &= Var(\epsilon_{i+1,k}) + Var\left(\sum_{d=1}^{D} \frac{\partial}{\partial x_{i,d}} \hat{f}(\hat{x}_i) * \epsilon_{i,d} * e_k\right) \\ &- 2 * Cov\left(\epsilon_{i+1,k}, \sum_{d=1}^{D} \frac{\partial}{\partial x_{i,d}} \hat{f}(\hat{x}_i) * \epsilon_{i,d} * e_k\right) \end{aligned} \tag{14}$$

From $Cov(\epsilon_{i+1,k}, \epsilon_{i,d}) = 0 \; \forall k, d$ follows

$$Var(\tau_{i+1,k}) = \sigma_k^2 + \sum_{d=1}^{D} \left(\frac{\partial}{\partial x_{i,d}} \hat{f}(\hat{x}_i) e_k\right)^2 \sigma_d^2 \tag{15}$$

The covariance for two different features k and l of the same measurement $i + 1$ is as follows

$$\begin{aligned} Cov(\tau_{i+1,k}, \tau_{i+1,l}) = Cov\Bigg(& \epsilon_{i+1,\,k} - \sum_{d=1}^{D} \frac{\partial}{\partial\, x_{i,d}} \hat{f}(\hat{x}_i)\; \epsilon_{i,d}\; e_k\;, \\ & \epsilon_{i+1,\,l} - \sum_{d=1}^{D} \frac{\partial}{\partial\, x_{id}} \hat{f}(\hat{x}_i)\; \epsilon_{i,d}\; e_l \Bigg) \\ = \; & Cov(\epsilon_{i+1,\,k},\; \epsilon_{i+1,\,l}) \; - \; Cov \\ & \left(\sum_{d=1}^{D} \frac{\partial}{\partial\, x_{i,d}} \hat{f}(\hat{x}_i)\; \epsilon_{i,d}\; e_k,\; \epsilon_{i+1,\,l}\right) \\ & - Cov\left(\epsilon_{i+1,\,k},\; \sum_{d=1}^{D} \frac{\partial}{\partial\, x_{i,d}} \hat{f}(\hat{x}_i)\; \epsilon_{i,d}\; e_l\right) \\ & + Cov\Bigg(\sum_{d=1}^{D} \frac{\partial}{\partial\, x_{i,d}} \hat{f}(\hat{x}_i)\; \epsilon_{i,d}\; e_k, \\ & \sum_{d=1}^{D} \frac{\partial}{\partial\, x_{i,d}} \hat{f}(\hat{x}_i)\; \epsilon_{i,d}\; e_l \Bigg) \\ = \; & \sum_{d=1}^{D} \sum_{d'=1}^{D} \frac{\partial}{\partial\, x_{i,d}} \hat{f}(\hat{x}_i)\; e_k \; \frac{\partial}{\partial\, x_{i,d'}} \hat{f}(\hat{x}_i)\; e_l \\ & Cov(\epsilon_{i,d},\; \epsilon_{i,d'}) \end{aligned} \tag{16}$$

From $Cov(\epsilon_{i,d}, \epsilon_{i,d'}) = 0 \; \forall d, d' : d \neq d'$ follows

$$Cov(\tau_{i+1,k}, \tau_{i+1,l}) = \sum_{d=1}^{D} \left(\frac{\partial}{\partial x_{i,d}} \hat{f}(\hat{x}_i) e_k\right) \left(\frac{\partial}{\partial x_{i,d}} \hat{f}(\hat{x}_i) e_l\right) \sigma_d^2 \tag{17}$$

For the series $\{\tau_i\}_{i \in T}$ between a state i and $i + 1$ of features k and l, the covariance is as follows

$$\begin{aligned} Cov(\tau_{i+1,k}, \tau_{i,l}) = Cov\Bigg(& \epsilon_{i+1,\,k} - \sum_{d=1}^{D} \frac{\partial}{\partial\, x_{i,d}} \hat{f}(\hat{x}_i)\; \epsilon_{i,d}\; e_k\;,\; \epsilon_{i,\,l} \\ & - \sum_{d=1}^{D} \frac{\partial}{\partial\, x_{i-1,d}} \hat{f}(\hat{x}_{i-1})\; \epsilon_{i-1,d}\; e_l \Bigg) \\ = \; & Cov(\epsilon_{i+1,\,k}, \epsilon_{i,\,l}) \; - \; Cov \\ & \left(\sum_{d=1}^{D} \frac{\partial}{\partial\, x_{i,d}} \hat{f}(\hat{x}_i)\; \epsilon_{i,d}\; e_k,\; \epsilon_{i,\,l}\right) \\ & - \; Cov\left(\epsilon_{i+1,\,k},\; \sum_{d=1}^{D} \frac{\partial}{\partial\, x_{i-1,d}} \hat{f}(\hat{x}_{i-1})\; \epsilon_{i-1,d}\; e_l\right) \\ & + \; Cov\Bigg(\sum_{d=1}^{D} \frac{\partial}{\partial\, x_{i,d}} \hat{f}(\hat{x}_i)\; \epsilon_{i,d}\; e_k, \\ & \sum_{d=1}^{D} \frac{\partial}{\partial\, x_{i-1,d}} \hat{f}(\hat{x}_{i-1})\; \epsilon_{i-1,d}\; e_l \Bigg) \end{aligned} \tag{18}$$

Since in the case of white noise $Cov(\epsilon_{i+1,k}, \epsilon_{i,l}) = Cov(\epsilon_{i+1,k}, \epsilon_{i-1,l}) = Cov(\epsilon_{i,k}, \epsilon_{i-1,l}) = 0 \; \forall k, l$ applies, the covariance is as follows

$$\begin{aligned} Cov(\tau_{i+1,k}, \tau_{i,l}) &= -Cov\left(\sum_{d=1}^{D} \frac{\partial}{\partial x_{i,d}} \hat{f}(\hat{x}_i)\epsilon_{i,d} e_k, \epsilon_{i,l}\right) \\ &= -\sum_{d=1}^{D} \frac{\partial}{\partial x_{i,d}} \hat{f}(\hat{x}_i) e_k Cov(\epsilon_{i,d}, \epsilon_{i,l}) \end{aligned} \tag{19}$$

From $Cov(\epsilon_{i,d}, \epsilon_{i,l}) = 0 \; \forall l, d : l \neq d$ follows

$$Cov(\tau_{i+1,k}, \tau_{i,l}) = -\frac{\partial}{\partial x_{i,l}} \hat{f}(\hat{x}_i) e_k \sigma_l^2 \tag{20}$$

If the state i is known through a measurement $\hat{x}_i^*$, it is possible for a prediction to set $\hat{x}_i = \hat{x}_i^*$. This term is used to create a prediction of state $i+1$ using $\hat{f}(\hat{x}_i)$. Therefore, the covariance matrix Σ_{Noise} becomes computable and we describe the time series of τ_i as follows

$$\{\tau_i\}_{i \in T} \sim \mathcal{N}(0, \Sigma_{\tau_i}) \tag{21}$$

The covariance matrix of the time series $\Sigma_{\tau_{i+1}} \in \mathbb{R}^{2D} \times \mathbb{R}^{2D}$ consists of the sub-matrices $\Sigma_{\tau_{i,i}}, \Sigma_{\tau_{i+1,i+1}}, \Sigma_{\tau_{i+1,i}}, \Sigma_{\tau_{i,i+1}} \in \mathbb{R}^D \times \mathbb{R}^D$, that is

$$\Sigma_{\tau_{i+1}} = \begin{pmatrix} \Sigma_{\tau_{i+1,i+1}} & \Sigma_{\tau_{i+1,i}} \\ \Sigma_{\tau_{i,i+1}} & \Sigma_{\tau_{i,i}} \end{pmatrix} \tag{22}$$

Using equations (15), (17), and (20), the diagonal sub-covariance matrices are constructed as

$$\Sigma_{\tau_{i+1,i+1}} = \begin{pmatrix} \sigma_1^2 + \sum\limits_{d=1}^{D} \left(\frac{\partial}{\partial x_{i,d}} \hat{f}(\hat{x}_i) e_1\right)^2 \sigma_d^2 & \dots & \sum\limits_{d=1}^{D} \left(\frac{\partial}{\partial x_{i,d}} \hat{f}(\hat{x}_i) e_1\right)\left(\frac{\partial}{\partial x_{i,d}} \hat{f}(\hat{x}_i) e_D\right)\sigma_d^2 \\ \dots & \dots & \dots \\ \sum_{d=1}^{D} \left(\frac{\partial}{\partial x_{i,d}} \hat{f}(\hat{x}_i) e_1\right)\left(\frac{\partial}{\partial x_{i,d}} \hat{f}(\hat{x}_i) e_D\right)\sigma_d^2 & \dots & \sigma_D^2 + \sum\limits_{d=1}^{D} \left(\frac{\partial}{\partial x_{i,d}} \hat{f}(\hat{x}_i) e_D\right)^2 \sigma_d^2 \end{pmatrix} \tag{23}$$

The matrix $\Sigma_{\tau_{i,i}}$ is computed analogously to $\Sigma_{\tau_{i+1,i+1}}$ as

$$\Sigma_{\tau_{i+1,i}} = \Sigma_{\tau_{i,i+1}}{}^T = \begin{pmatrix} -\frac{\partial}{\partial x_{i,1}} \hat{f}(\hat{x}_i) e_1 \sigma_1^2 & \dots & -\frac{\partial}{\partial x_{i,D}} \hat{f}(\hat{x}_i) e_1 \sigma_D^2 \\ \dots & \dots & \dots \\ -\frac{\partial}{\partial x_{i,1}} \hat{f}(\hat{x}_i) e_D \sigma_1^2 & \dots & -\frac{\partial}{\partial x_{i,D}} \hat{f}(\hat{x}_i) e_D \sigma_D^2 \end{pmatrix} \tag{24}$$

It should be noted that for linear systems the covariance matrix $\Sigma_{\tau_{i+1}}(\hat{x}_i, \hat{x}_{i-1})$ is static, while for nonlinear systems the covariance becomes dynamic. Therefore, the covariance of noise is able to describe dynamical systems without limitations regarding linearity, while also creating valid solutions for linear cases. In the case of colored instead of white noise, the derivation of equations (15), (17), and (20) must be adjusted for the corresponding correlated terms and the adjusted equations are then used to construct the different sub-covariance matrices in equations (23) and (24). Thus, the model is compatible with the assumption of colored noise but requires a more extensive calculation.

This theoretical covariance matrix can be used as a test measure if the measured Δ_{i+1} only depends on noise or if the ignorance of the functions f and g results in differences of the empirical distribution of the process Δ_{i+1}. An estimation of the parameter σ can be conducted with observations of sensors under a halting operation $\hat{x}_{i+1} = \hat{f}_{halt}(\hat{x}_i) = (t_i + \Delta t, \hat{x}_{i,1}, \dots, \hat{x}_{i,D-1})$ since the resulting time series should primarily be influenced by white noise of the measurement of each observed feature.

Global anomaly detection via entropy of the density distribution

For an anomaly detection, the existence of the ignorance must be tested. It is deduced that an anomaly is present within the system when the empirical covariance $\hat{\Sigma}_i$ significantly differs from the pure noise covariance Σ_{τ_i} because in absence of an error term λ_i, $\Sigma_{\lambda_i} \rightarrow 0$ also applies. Therefore, without noise, the relation $\hat{\Sigma}_i = \Sigma_{\tau_i}$ applies. As a possible test for anomalies, the comparison of the empirical covariance and theoretical computed noise covariance matrices given by equation (22), via equations (23) and (24), using Box's M-Test (Box, 1949) or similar tests (cf. Marques and Coelho, 2018) can be applied. The tested hypothesis is as follows

$$H_0 : \hat{\Sigma}_i = \Sigma_{\tau_i} \tag{25}$$

If this hypothesis is rejected, an unknown covariance matrix with a density distribution $f_\lambda(\lambda) \neq 0$ exists, which describes the influence on the real operation due to the ignorance of f and g. If the distribution of λ does not follow a normal distribution, Box's M-test does not apply since it assumes a normal distribution for the compared covariances' underlying distributions (Manly and Navarro Alberto, 2017). The same limitation would occur if the empirical covariance matrix is compared to the theoretical covariance matrix using a matrix norm or metric since the empirical covariance matrix Σ_{λ_i} of a non-statistical and non-normal distributed process does not accurately reflect the real distribution. Therefore, a more general comparison is necessary.

If an analysis of a time series with enough data points ($\geqslant D$) is conducted, the empirical covariance $\hat{\Sigma}_i$ of the time series $\{\Delta_i\}_{i \in T}$ is an estimator for the covariance matrix Σ_i. In general, the Kullback–Leibler divergence (KL-divergence) can be used to compare distributions using covariance matrices. However, since the function $\hat{f}(\hat{x}_i)$, which is necessary to compute the covariance matrix Σ_{τ_i}, is assumed to be nonlinear, a comparison using the KL-divergence cannot be conducted because the metric requires a static covariance matrix. In addition, the KL-divergence makes the assumption of a normal distribution of the time series. Therefore, if the assumption $\{\lambda_i\}_{i \in T} \sim \mathcal{N}(0, \Sigma_{Ignorance})$ is rejected, an anomaly detection without underlying assumptions regarding the distributions $\{\lambda_i\}_{i \in T}$ and $\{\Delta_i\}_{i \in T}$ is required. In addition, the dynamic characteristic of the covariance matrix Σ_{τ_i} must be considered.

Thus, the distribution is analyzed using the Shannon entropy without assumptions regarding the distribution. The entropy of the distribution is given by

$$\mathbb{H}(f_\Delta) = -\sum_k f_{\Delta,k} \log(f_{\Delta,k}) \tag{26}$$

The cross-entropy between the two density distributions of $f_\Delta(\Delta)$ and $f_\tau(\tau)$ is defined as follows

$$\mathbb{H}(f_\Delta, f_\tau) = -\sum_k f_{\Delta,k} \log(f_{\tau,k}) \tag{27}$$

The density distributions are not measured directly, but the density distributions can be approximated by using a histogram. If the measured values of pairs of Δ_{i+1} and Δ_i are ordered within a histogram using $K^{2D} = K \times \cdots \times K$ bins, the amount of values within a bin is countable as $h_{\Delta,k}$. The method can be compared to HBOS (Goldstein and Dengel, 2012), where the difference of two histograms is analyzed regarding bins with high differences. The entropy calculation changes to

$$\hat{\mathbb{H}}(f_\Delta) = -\sum_{k \in K^{2D}} h_{\Delta,k} \log(h_{\Delta,k}) \tag{28}$$

$$\hat{\mathbb{H}}(f_\Delta, f_\tau) = -\sum_{k \in K^{2D}} h_{\Delta,k} \log(h_{\tau,k}) \tag{29}$$

In conclusion, the KL-divergence is adjusted to

$$\hat{D}_{KL}(f_\Delta, f_\tau) = \sum_{k \in K^{2D}} h_{\Delta,k} \log\left(\frac{h_{\Delta,k}}{h_{\tau,k}}\right) \tag{30}$$

Since the KL-divergence is not symmetric, both directions should be calculated and added together for analyzing the comparison. If both distributions are identical, $h_{\Delta,k} \approx h_{\tau,k} \Rightarrow \log(h_{\Delta,k}/h_k) \approx 0 \Rightarrow \hat{D}_{KL}(f_\Delta, f_\tau) \approx 0$ applies.

As a test value, the comparison of entropies using $\frac{\hat{\mathbb{H}}(f_\Delta)}{\hat{\mathbb{H}}(f_\tau)} \approx 1$ or $\hat{D}_{KL}(f_\Delta, f_\tau) + \hat{D}_{KL}(f_\tau, f_\Delta) \approx 0$ is applied in the hypothesis test given by equation (25). This test evaluates whether the term λ_i is not zero and results in a significant difference due to the ignorance of operations and therefore unknown system behavior should be assumed.

A sample distribution $\{\tau_i\}_{i \in T}$ is used for building pure noise histograms as comparison. Random values are picked from a multivariate normal distribution using the theoretical covariance matrix Σ_{τ_i} or the values are simulated using the function $\hat{f}(\hat{x}_i)$ together with a white noise term.

Evaluation and comparison of predictions via Wasserstein metric

It is not only important whether a prediction differs from the measurements so that an unknown system behavior is assumed, but it is also important to measure how strong the anomalous behavior and difference between the prediction and the measurement is. Thus, a metric is necessary to measure how inaccurate a prediction is. For time series and nonlinear systems, the influence of parameters and the Wasserstein metric as evaluation criterion of such systems is studied by Muskulus (2010). The metric is based on the comparison of both distributions $\{\tau_i\}_{i \in T}$ and $\{\Delta_i\}_{i \in T}$ by their histograms. The distance between two histogram bins is given by the Manhattan distance C of the two corresponding bins, which is the L_1 distance. The position of the bin k is a vector of the bin position (a, b). For the distance between two bins follows.

$$C(k, k') = |a - a'| + |b - b'| \tag{31}$$

The Wasserstein metric can be visualized as the optimal transport flow between the two observed distributions. One distribution acts as α_k and the other distribution acts as $\beta_{k'}$ demand. The distributions are normalized so that $\sum_k \alpha_k = \sum_{k'} \beta_{k'} = 1$ and all values of α_k and $\beta_{k'}$ are positive values. This results in two measures for the discretized distributions where δ_x denotes the Dirac delta distribution as follows

$$\begin{aligned} \nu &= \sum_k \alpha_k \delta_{h_{\Delta,k}} \\ \upsilon &= \sum_{k'} \beta_{k'} \delta_{h_{\tau,k'}} \end{aligned} \tag{32}$$

Thus, the histogram bins act as sources of entries flowing toward sinks. Therefore, the amount of values of all bins of the first distributions are the sources and the values of all bins of the second distribution are the sinks. This results in source and sink conditions

$$\begin{aligned} \sum_{k'} q_{k,k'} &= \alpha_k \\ \sum_k q_{k,k'} &= \beta_{k'} \end{aligned} \tag{33}$$

The first-order Wasserstein distance becomes as follows

$$W_1(\nu, \upsilon) = \min \sum_{k,k'} q_{k,k'}\ C(k, k') \tag{34}$$

The value $W_1(\nu, \upsilon)$ can be computed within an acceptable time, if the bins are limited. Alternatively, a two-dimensional (2D) sliced Wasserstein metric can be used as described by Bonneel et al. (2015). The Wasserstein metric can then be used as a measure of distance between the two distributions and enables a comparison between two predictions $\hat{f}_1$ and $\hat{f}_2$ on which prediction is more suitable to describe the system and has less anomalous properties. By using the Wasserstein metric, it is possible to analyze the difference between the empirical and theoretical distribution without computing the empirical covariance matrix. This is important since the empirical covariance matrix does not reflect the non-normal distributed density distribution of the histograms.

Local anomaly detection via Mahalanobis distance

If only single data points when measuring Δ_{i+1} derivate from the distribution, a local outlier detection is necessary. The theoretical covariance matrix of noise can still be used and is adapted for each delta. The Mahalanobis distance applies as follows

$$D(\Delta_{i+1}) = \sqrt{(\Delta_{i+1}, \Delta_i)^T \Sigma_{\tau_{i+1}}(\hat{x}_i, \hat{x}_{i-1})^{-1} (\Delta_{i+1}, \Delta_i)} \tag{35}$$

This distance metric is applicable to all states i and all measured triplets of $\hat{x}_{i+1}, \hat{x}_i, \hat{x}_{i-1}$.

Since the Mahalanobis distance follows the chi-square distribution (Fauconnier and Haesbroeck, 2009), the chi-square distribution with 2D degrees of freedom is applied to test the measured Δ_{i+1} for outliers for a chosen significance level α as

$$(\Delta_{i+1}, \Delta_i)^T \Sigma_{\tau_{i+1}}(\hat{x}_i, \hat{x}_{i-1})^{-1}(\Delta_{i+1}, \Delta_i) > \chi^2_{2D, 1-\alpha} \qquad (36)$$

Since the covariance matrix is known beforehand and can be computed in advance to the measurement using the function $\hat{f}(\hat{x}_i)$, only this concise test has to be conducted for a valid and useful outlier detection. This enables the method to compute and evaluate outliers in a real-time detection in time series with prior known prediction functions.

When counting the amount of detected outliers, the detected amount is compared to the expected amount of false-positive detected outliers. Using the significance level α and the properties of the chi-square distribution, a probability, if the detected amount is within the expected amount of false-positive outliers, is calculated. Therefore, a global anomaly score is computed by the given probability that the counted outliers are statistically significant for belonging to a chi-square distribution.

Proposed algorithm

As algorithm for a functional global anomaly detection, the pseudo code of Algorithm 1 is proposed. The assumption is that if $\hat{\mathbb{H}}(f_\Delta) \neq \hat{\mathbb{H}}(f_\tau)$, the distribution of measured values does not follow the pure noise distribution.

As algorithm for a local outlier detection, Algorithm 2 is proposed. This algorithm can be applied to a time series in real time since only the function $\hat{f}$ is required and no further prior knowledge about the time series is necessary.

In general, the parameters for the algorithms are comparably easy to estimate. Regarding the amount of simulations of the pure noise distributions, a sufficient sample size S should be selected so that the mean of the noise distribution is meaningful. The size of the histogram bins should be selected that the bins are large enough that each includes some data points. If the bins are too small, the entropy computation might not work and might not result in meaningful values since it assumes at least one data point per bin. The significance level α of the test should reflect the amount of knowledge about the system. If the system follows a strict physical differential equation and is modeled comprehensively, a more strict significance level might be necessary. For the estimation of the white noise $\hat{\sigma}^2$ of the sensors, a measurement during system standstill can be conducted and evaluated. In this case, it is assumed that $\hat{\sigma}^2 = Var(\{\Delta_i\}_{i \in N})$. This is applicable since it is assumed that the ignorance of f and g only conceals the measurement when operations of the systems are conducted.

Algorithm 1 Unsupervised histogram entropy global anomaly detection

```
Input:
    $N$ measurements of $\hat{x}_i^*$
    Prediction function $\hat{f}(\hat{x}_i)$
Parameter:
    Noise estimation $\hat{\sigma}^2$
    Set of histogram bins $K$
    Amount of simulations $S$
    Significance level $\alpha$
Output:
    Boolean value $A$ for anomaly existence
for $i = 1, i++$ do
    while $i \leqslant N - 1$ do
        Compute $\Delta_{i+1} \Leftarrow \hat{x}_{i+1}^* - \hat{f}(\hat{x}_i)$
        Compute $\Sigma_{\tau_{i+1}}(\hat{x}_i^*, \hat{x}_{i-1}^*)$
        Draw S random variables $(\tau_{i+1}, \tau_i)^T \sim \mathcal{N}(0, \Sigma_{\tau_{i+1}}(\hat{x}_i^*, \hat{x}_{i-1}^*))$
    end while
end for
while $k \in K$ do
    while $s \in S$ do
        $h_{\tau,k,s} \Leftarrow$ amount of $(\tau_{i+1}, \tau_i)^T$ in bin $k$ of simulation $s$
    end while
    $h_{\Delta,k} \Leftarrow$ amount of $(\Delta_{i+1}, \Delta_i)^T$ in bin $k$
end while
Compute $\hat{\mathbb{H}}(f_\Delta) \Leftarrow -\sum_{k \in K^{2D}} h_{\Delta,k} \log(h_{\Delta,k})$ for all $h_{\Delta,k} \neq 0$
Compute $\overline{\hat{\mathbb{H}}(f_\tau)} \Leftarrow -\frac{1}{S}\sum_{s=1}^{S} \sum_{k \in K^{2D}} h_{\tau,k,s} \log(h_{\Delta,k,s})$ for all
    $h_{\tau,k,s} \neq 0$
Compute $\hat{\sigma}_{\mathbb{H}} \Leftarrow \sqrt{Var(\{\hat{\mathbb{H}}(f_\tau)_s\}_{s \in S})}$
if $\overline{\hat{\mathbb{H}}(f_\tau)} - t_{(1-\alpha)} * \hat{\sigma}_{\mathbb{H}} \leqslant \hat{\mathbb{H}}(f_\Delta) \leqslant \overline{\hat{\mathbb{H}}(f_\tau)} + t_{(1-\alpha)} * \hat{\sigma}_{\mathbb{H}}$ then
    $A \Leftarrow$ False
else
    $A \Leftarrow$ True
end if
```

Algorithm 2 Unsupervised distance-based local anomaly detection

```
Input:
    Measurements of $\hat{x}_{i+1}^*, \hat{x}_i^*, \hat{x}_{i-1}^*$
    Prediction function $\hat{f}(\hat{x}_i)$
Parameter:
    Noise estimation $\hat{\sigma}^2$
    Significance level $\alpha$
Output:
    Array $L$ containing outlier data points
Compute $\Delta_{i+1} \Leftarrow \hat{x}_{i+1}^* - \hat{f}(\hat{x}_i)$
Compute $\Sigma_{\tau_{i+1}}(\hat{x}_i^*, \hat{x}_{i-1}^*)$
Compute $D(\Delta_{i+1})$
if $D(\Delta_{i+1}) > \chi^2_{2D, 1-\alpha}$ then
    $L \Leftarrow \hat{x}_{i+1}^*$
end if
```

Simulation study: Anomaly detection in logistic growth

Applied global and local anomaly detection

For an analysis with synthetic data, we assume a system with a system state z_i with $J = 4$ real unknown features and with $D = 3$ observable features. For simplicity, we assume the observed features are direct measures of the real features and one real feature is completely unknown. One of the observed features is the linear increasing time of the system. The measurement is also concealed by white noise with a standard deviation of $\sigma = 0.01$. For the operation of the system, only one real operation f is assumed. This operation transforms state i in state $i+1$ within one time unit. The feature x_1 and x_2 are logistic growths with $r_1 = 3$ and $r_2 = 3.5$. The feature x_2 is obscured additively by feature y scaled with the signal strength s. The feature y is a time-dependent sine wave. The real-time series is as follows

$$z_{i+1} = f(z_i) = \begin{pmatrix} t_i + 1 \\ 3 * x_{i,1} * (1 - x_{i,1}) \\ 3.5 * x_{i,2} * (1 - x_{i,2}) + s * y_i \\ \sin(0.7 * t_i) \end{pmatrix} \tag{37}$$

Since only the observed features are known, the prediction function $\hat{f}(\hat{z}_i)$ is as follows

$$\hat{z}_{i+1} = \hat{f}(\hat{z}_i) = \begin{pmatrix} t_i + 1 \\ 3 * x_{i,1} * (1 - x_{i,1}) \\ 3.5 * x_{i,2} * (1 - x_{i,2}) \end{pmatrix} \tag{38}$$

As soon as z_i is observed with the usage of function $\hat{f}$, the next state z_{i+1} is predicted.

The operation f is applied multiple times, and the outcome of the real unknown values z_i, the measured values $\hat{z}_i^*$ and predicted values $\hat{z}_i$ are analyzed for a selected amount of executions. As assumed, the observed features behave within the prediction as logistic growth. The Δ_{i+1} between measurement and expectation are analyzed in Figure 1.

With the knowledge that the additive signal follows a sine wave function, the signal can sometimes be guessed within the Δ_{i+1} of feature x_2. However, during multiple tests, the sine wave is often not clearly visible, even when applying a Fourier transformation. Thus, we use the proposed method to systematically prove that the measurement in feature x_2 is obscured by a signal. With knowledge of function $\hat{f}$, it is possible to compute the theoretical white noise covariance matrix as follows

$$\Sigma_{\tau_i} = \begin{pmatrix} 2 & 0 & 0 & -1 & 0 & 0 \\ 0 & 36x_{i,1}^2 - 36x_{i,1} + 10 & 0 & 0 & 6x_{i,1} - 3 & 0 \\ 0 & 0 & 49x_{i,2}^2 - 49x_{i,2} + 13.25 & 0 & 0 & 7x_{i,2} - 3.5 \\ -1 & 0 & 0 & 2 & 0 & 0 \\ 0 & 6x_{i,1} - 3 & 0 & 0 & 36x_{i,1}^2 - 36x_{i,1} + 10 & 0 \\ 0 & 0 & 7x_{i,2} - 3.5 & 0 & 0 & 49x_{i,2}^2 - 49x_{i,2} + 13.25 \end{pmatrix} \sigma^2 \tag{39}$$

For the purpose of a better visualization of the relation between state i and $i+1$ of both observed features, the empirical density distribution and covariance matrix are split up into two separate distributions since there is no relation between feature x_1 and x_2. If there were a relation, conducting this split would not be recommended and it would cause limitations in the analysis and anomaly detection. These spilt-up density distributions are illustrated in Figure 2(a) top left and bottom right. Since no information is lost by conducting the analysis with the split-up density distributions, but a better visualization is achieved to describe the relation between state i and state $i+1$, we will conduct the further analysis based on this simplification.

As a second step, the white noise must be estimated. Therefore, we apply a halting operation as proposed. Within our time series, it is also applicable to measure the time's standard deviation directly since it is not concealed by a signal and follows a linear relation. The measured standard deviation of the Δ_{i+1} of time is $\hat{\sigma}_t \approx 0.014$. As we see in the theoretical noise covariance matrix, we have to correct the measured value in the Δ_{i+1} of time by $\sqrt{2}$. We then receive a standard deviation of $\hat{\sigma}_t \approx 0.01$, which is exactly as modeled. Using the white noise, it is possible to generate random variables using the theoretical covariance matrix and to create a theoretical density distribution. This density distribution is illustrated in Figure 2(b). For lower sample sizes, only little differences are visible to the empirical density distribution. For a higher amount of samples N, the difference becomes more obvious.

When analyzing the entropy for the distribution of the time measurement, no significant differences occur between the empirical density and the theoretical noise density. This is expected since there is no signal concealing the time measurement. Since the time is linear, we can cross-check the analysis with the measured correlation of time between state i and $i+1$. The measured correlation is $\hat{\rho}_{t_i, t_{i+1}} \approx -0.474$ and therefore close to the expected theoretical value of -0.5.

The entropy of the distributions of the features x_1 and x_2 is evaluated for the empirical density distribution and compared with the mean theoretical density distributions' entropy over $S = 30$ simulations. In this evaluation, $N = 100$ executions of function f, a signal strength of $s = 0.02$, and $K = 9 \times 9$ histogram bins are applied. This results in an empirical entropy of $\hat{\mathbb{H}}(f_\Delta)_{x_1} = -144.7$ for the first feature and $\hat{\mathbb{H}}(f_\Delta)_{x_2} = -71.1$ for the second feature.

An entropy of $\overline{\mathbb{H}(f_\tau)_{x_1}} = -143.8 \pm 10.2$ for the first feature and an entropy of $\overline{\mathbb{H}(f_\tau)_{x_2}} = -102.3 \pm 8.9$ for the second feature is computed for the mean over $S = 30$ theoretical density distributions of white noise. The empirical entropy of the system for feature x_2 is over $z_{0.99} \approx 2.576$ standard deviations different than the theoretical entropy of a white noise distribution. The feature x_1 does not show any significant differences. Therefore, based on the entropy comparison, an anomaly within the time series of feature x_2 is assumed.

Using Algorithm 2, a local anomaly detection is conducted. Since the sine wave signal is only intense compared to the noise in the minimum and maximum values, we expect outliers to occur right after these extreme values. Other data points in the time series might be more compatible with the noise assumptions. The detected outliers are marked in Figure 3 using an $\alpha = 0.01$. Overall, nine outliers are detected within the data points of the time series of feature x_2. Since $\alpha = 0.01$, it is assumed that out of 100 measurements, only one is false positive. The possibility that nine false positives are detected is $p \approx 0\%$. Thus, by using the local outlier detection, it is reasoned that a global anomaly is present in the time series.

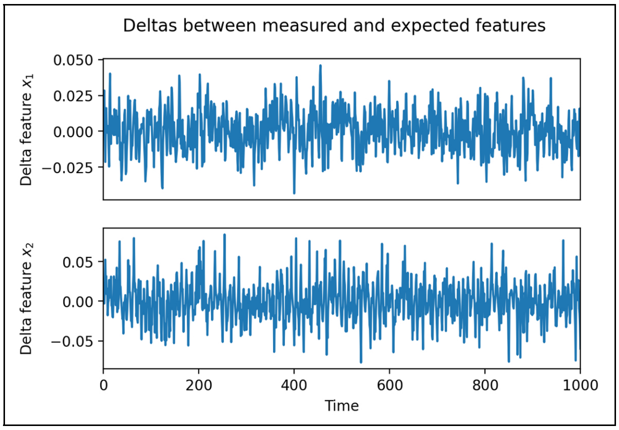


Figure 1. Delta between measured and expected values of feature x_1 and x_2 in a sample run with $N = 1000$ executions of function f and a signal strength of $s = 0.02$.

It is visible that some marked outliers occur directly after the high and low points of the concealing signal feature y. This is coherent since in these areas the signal $s * y_i$ is higher in relation to the noise. Therefore, the signal is able to influence the measures of Δ_{i+1} in the time series more. Outliers are also detected for comparably low values of Δ_{i+1}. This has two reasons: the dynamic covariance yields a lower variance for this area of the time series or, since state i and $i + 1$ are compared, the occurring difference between Δ_i of state i and Δ_{i+1} of state $i + 1$ is considered anomalous by the density distribution of state i and $i + 1$.

Using a signal intensity $s = 0.02$ and $N = 100$ measurements, it is possible to successfully detect multiple anomalies within the time series. This proves the applicability of the proposed algorithms for linear and nonlinear time series.

Sensitivity analysis of global anomaly detection via entropy

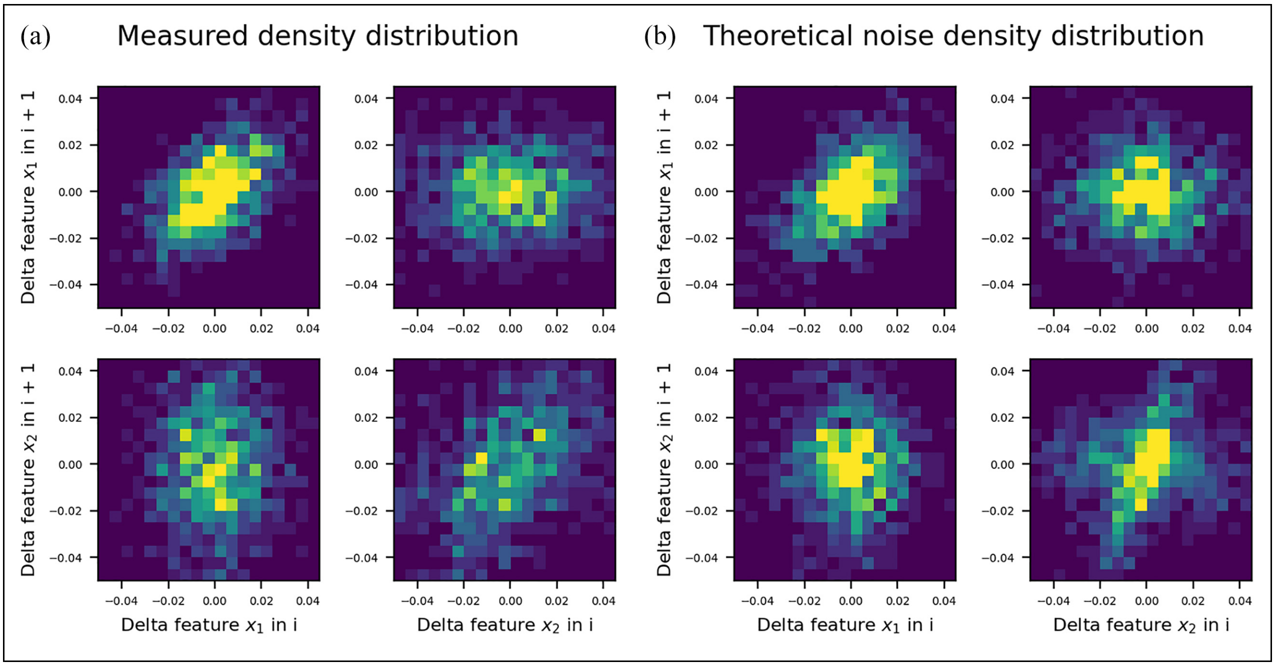
If the signal strength s is varied, a comparison of the entropy of all cross-sections can be conducted. The signal-to-noise ratio is defined as follows

$$S/N = \frac{\max\limits_{i \in N}(s * y_i)}{\hat{\sigma}_x} \tag{40}$$

Therefore, the significance of the anomaly detection is validated by the evaluation of the entropy using varying S/N ratios in Figure 4.

Figure 4 shows that the empirical entropy of feature x_2 starts to differ significantly from the theoretical entropy of a pure noise scenario at $S/N \approx 0.75$. The entropy of feature x_1 and the time measurement are unchanged since no unknown influences conceal these measurements. This analysis shows precisely in which feature the unknown influence is detected and therefore helps to identify the relevant features for an anomaly cause analysis. This enables an easier problem identification and correction of the prediction function $\hat{f}(x_i)$. The algorithm is also capable of detecting unknown influences with a signal strength lower than the noise in some cases, as well as identifying the related feature.

For different amounts of executions of the time series N and varying signal-to-noise ratios S/N, different sensitivities of the anomaly detection are measured using a constant $\alpha = 0.01$. The results are listed in Table 1. In the cases of


(a) Measured density distribution

(b) Theoretical noise density distribution

Figure 2. Theoretical noise density distribution and empirical density distribution of state i and $i + 1$ for the exemplary time series with $N = 1000$ executions of function f and a signal strength of $s = 0.02$ using $K = 19 \times 19$ bins. (a) Empirical density distribution and (b) noise density distribution.

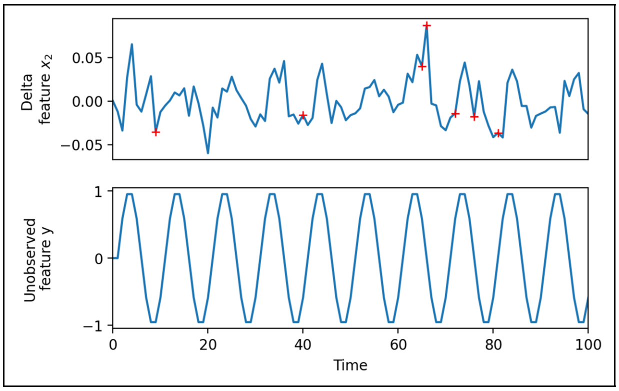


Figure 3. Identified outliers of feature x_2 in a sample run with $N = 100$ executions of function f and a signal strength of $s = 0.02$.

Table 1. Sensitivity analysis of the proposed algorithm using varying sample sizes N and signal-to-noise ratios S/N for a constant $\alpha = 0.01$.

	Sample size N					
S/N	30	50	100	300	500	1000
0.2	20%	20%	16%	16%	26%	12%
0.4	20%	22%	18%	18%	30%	28%
0.6	24%	24%	24%	24%	28%	62%
0.8	22%	28%	36%	36%	52%	84%
1.0	18%	22%	32%	68%	82%	98%
1.2	20%	32%	48%	86%	98%	100%
1.4	30%	30%	60%	96%	100%	100%
1.6	34%	48%	92%	100%	100%	100%
1.8	42%	60%	92%	100%	100%	100%
2.0	64%	78%	92%	100%	100%	100%

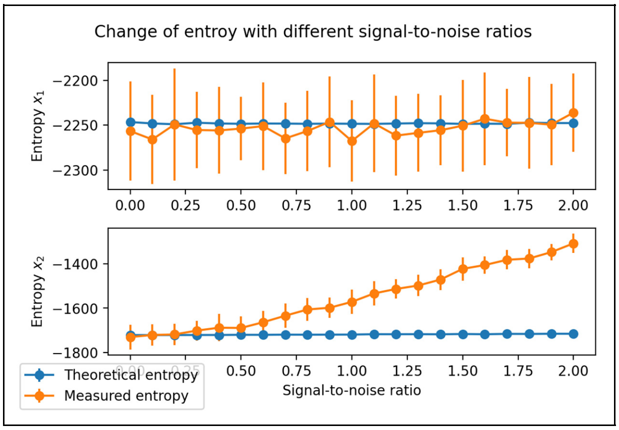


Figure 4. Mean theoretical and measured entropy for feature x_1 and x_2 for varying S/N ratios with S = 100 samples and N = 1000 executions of function f.

small signal-to-noise ratios, even a Fourier transformation often fails to visually separate the underlying sine wave. The analysis shows that the proposed algorithm is capable of detecting global anomalies in the case of small signal-to-noise ratios. Therefore, the model is recommended in practice in order to find small anomalous signals in a large sample size or large signals in a small sample size.

Use case: Evaluation of numeric predictors for satellite orbits

Orbital mechanics

Since satellites follow an easy to predict path using physical models, that is, newton mechanics, they also have a prediction and a measurement process, which is necessary for implementing our proposed method. Furthermore, satellites follow an elliptic path in orbit and are therefore not a linear system. In order to demonstrate the proposed method, satellite data already researched by Puente et al. (2021) and provided by the International Data Analysis Olympiad (IDAO, 2020) are used. The data are given for the period of January 2014 for 600 satellites. The previous research makes the data set a good choice for benchmarking and comparison.

First, the physical models need to be set up. The main information in the data set are the coordinates of the satellites along the x, y, and z-axis. Since the data are analyzed by Puente et al. (2021) and also provided in cartesian coordinates, we do not transform them into the more commonly used polar coordinates. Besides the coordinates, the velocity along these coordinates is given.

Each satellite has a specific radius $r(t) = (x(t), y(t), z(t))^T$ from Earth (the origin of the coordinate system) at each time. The velocity along the radius is given as $v(t) = (v_x(t), v_y(t), v_z(t))^T$. The gravitational constant $G = 6.674 \times 10^{-20}$ $\mathrm{km}^3/(\mathrm{kg} * \mathrm{s}^2)$ and the mass of earth $M = 5.972 \times 10^{24}$ kg are treated as parameters. The mass of the satellite and its gravitational force are neglected. Also, Earth is assumed to be a point mass. The first-order differential equations are given as follows

$$\dot{r}(t) = v(t) \tag{41}$$

$$\dot{v}(t) = -GM\frac{r(t)}{\|r(t)\|^3} \tag{42}$$

A common solver for these differential equations is the Euler method or the Runge–Kutta method of order 4 (RK4). The RK4 and Euler method use first-order differential equations. As an additional solver, a LSODA method, a variant with automatic method selection of the Livermore Solver for Ordinary Differential Equations (LSODE), as implemented by Hindmarsh (1983) is used as a very precise predictor of the orbits.

Derivation of applied predictions

The Euler method is the historic way to calculate orbits and is the simplest of the family of Runge–Kutta methods, but it therefore has a high error-proneness for computing the orbits. The prediction of the velocity for step $i + 1$ using step i is given by

$$v_{i+1} = v_i - GM\frac{r_i}{\|r_i\|^3} * h \tag{43}$$

The prediction of the radius uses the prediction of the velocity

$$\begin{aligned} r_{i+1} &= r_i + v_{i+1} * h \\ &= r_i + \left(v_i - GM \frac{r_i}{\|r_i\|^3} * h \right) * h \\ &= r_i + v_i * h - GM \frac{r_i}{\|r\|_i^3} * h^2 \end{aligned} \tag{44}$$

The deviations are calculated so that the full theoretical covariance matrix is constructed. As an example and to keep the covariance matrix smaller, only the x-coordinate is checked for anomalies while the error in time is neglected. Thus, the theoretical covariance matrix $\Sigma_{\tau_{x_i}}$ of the Euler method for approximating the orbits is as follows

$$\Sigma_{\tau_{x_i}} = \begin{pmatrix} \sigma_x^2 + A & -B \\ -B & \sigma_x^2 + A \end{pmatrix} \tag{45}$$

with

$$\begin{aligned} A = &\left(1 + \frac{GM\left(2x_i^2 - y_i^2 - z_i^2\right)h^2}{\sqrt{x_i^2 + y_i^2 + z_i^2}^5} \right)^2 \sigma_x^2 \\ &+ \left(\frac{3GMx_iy_ih^2}{\sqrt{x_i^2 + y_i^2 + z_i^2}^5} \right)^2 \sigma_y^2 + \left(\frac{3GMx_iz_ih^2}{\sqrt{x_i^2 + y_i^2 + z_i^2}^5} \right)^2 \sigma_z^2 + h^2\sigma_{v_x}^2 \end{aligned} \tag{46}$$

$$B = \left(1 + \frac{GM\left(2x_i^2 - y_i^2 - z_i^2\right)h^2}{\sqrt{x_i^2 + y_i^2 + z_i^2}^5} \right) \sigma_x^2 \tag{47}$$

The theoretical covariance matrix of the velocity is given as follows

$$\Sigma_{\tau_{v_{x_i}}} = \begin{pmatrix} \sigma_{v_x}^2 + C & -\sigma_{v_x}^2 \\ -\sigma_{v_x}^2 & \sigma_{v_x}^2 + C \end{pmatrix} \tag{48}$$

with

$$\begin{aligned} C = &\left(\frac{GM(2x_i^2 - y_i^2 - z_i^2)h}{\sqrt{x_i^2 + y_i^2 + z_i^2}^5} \right)^2 \sigma_x^2 \\ &+ \left(\frac{GM(2y_i^2 - x_i^2 - z_i^2)h}{\sqrt{x_i^2 + y_i^2 + z_i^2}^5} \right)^2 \sigma_y^2 \\ &+ \left(\frac{GM(2z_i^2 - x_i^2 - y_i^2)h}{\sqrt{x_i^2 + y_i^2 + z_i^2}^5} \right)^2 \sigma_z^2 + \sigma_{v_x}^2 \end{aligned} \tag{49}$$

The theoretical covariance matrix is used to compute the theoretical noise density distributions, which is compared with the empirical density distribution in order to spot anomalous behavior in the x-coordinate.

A more precise method is the Runge–Kutta method of order 4. Therefore, the RK4 is used in comparison to the Euler method. The difference between the theoretical noise density distributions and the measured empirical density distribution of the Euler method is assumed to be greater than in the case of the RK4 method, marking the RK4 as a more viable prediction method. For a defined time step h, the RK4 coefficients for predicting r_{i+1} are given as follows

$$\begin{aligned} V_{1_r}(r_i, v_i) &= v_i \\ V_{2_r}(r_i, v_i) &= v_i + \tfrac{h}{2} V_{1_v} \\ V_{3_r}(r_i, v_i) &= v_i + \tfrac{h}{2} V_{2_v} \\ V_{4_r}(r_i, v_i) &= v_i + hV_{3_v} \end{aligned} \tag{50}$$

$$\begin{aligned} V_{1_v}(r_i, v_i) &= -GM \frac{r_i}{\|r\|_i^3} \\ V_{2_v}(r_i, v_i) &= -GM \frac{r_i + \frac{h}{2}V_{1_r}}{\left\|r_i + \frac{h}{2}V_{1_r}\right\|^3} \\ V_{3_v}(r_i, v_i) &= -GM \frac{r_i + \frac{h}{2}V_{2_r}}{\left\|r_i + \frac{h}{2}V_{2_r}\right\|^3} \\ V_{4_v}(r_i, v_i) &= -GM \frac{r_i + hV_{3_r}}{\|r_i + hV_{3_r}\|^3} \end{aligned} \tag{51}$$

This results in predictions of the state $i + 1$ depending on only r_i and v_i as follows

$$r_{i+1} = r_i + \frac{h}{6} \left(V_{1_r}(r_i, v_i) + 2V_{2_r}(r_i, v_i) + 2V_{3_r}(r_i, v_i) + V_{4_r}(r_i, v_i)\right) \tag{52}$$

$$v_{i+1} = v_i + \frac{h}{6} \left(V_{1_v}(r_i, v_i) + 2V_{2_v}(r_i, v_i) + 2V_{3_v}(r_i, v_i) + V_{4_v}(r_i, v_i)\right) \tag{53}$$

Either the theoretical covariance matrix of these predictions is computed or the noise is simulated S-times by adding a random normal-distributed $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ to r_i and v_i and calculating the resulting predictions as a comparison base. In both cases, an exemplary density distribution is computed and compared with the empirical density distribution. As an alternative method for computing more complex numeric predictions, the deviation of the prediction function for feature j can be locally estimated using an infinitesimal change Δq_{ij} of feature j as follows

$$\frac{\partial}{\partial q_{ij}} \hat{f}(q_i) \approx \frac{\hat{f}\left(q_i + e_j * \Delta q_{ij}\right) - \hat{f}(q_i)}{\Delta q_{ij}} \tag{54}$$

By evaluating the prediction using the numeric solution at an infinitesimal change, the resulting values are used to construct the theoretical covariance matrix. This estimation is used for computing the theoretical covariance matrix of the LSODA predictions.

For real-time applications using Algorithm 2, a full calculation or estimation of the covariance matrix is necessary, while for those using Algorithm 1, an amount of sample runs under noise is sufficient.

Evaluation and comparison of prediction methods for satellite orbits

The satellite orbits are assumed to be quite anomalous since the simple two-body problem as presented in equations (41) and (42) does not include other astronomical objects, that is, the moon and the sun, as well as man-made satellites and

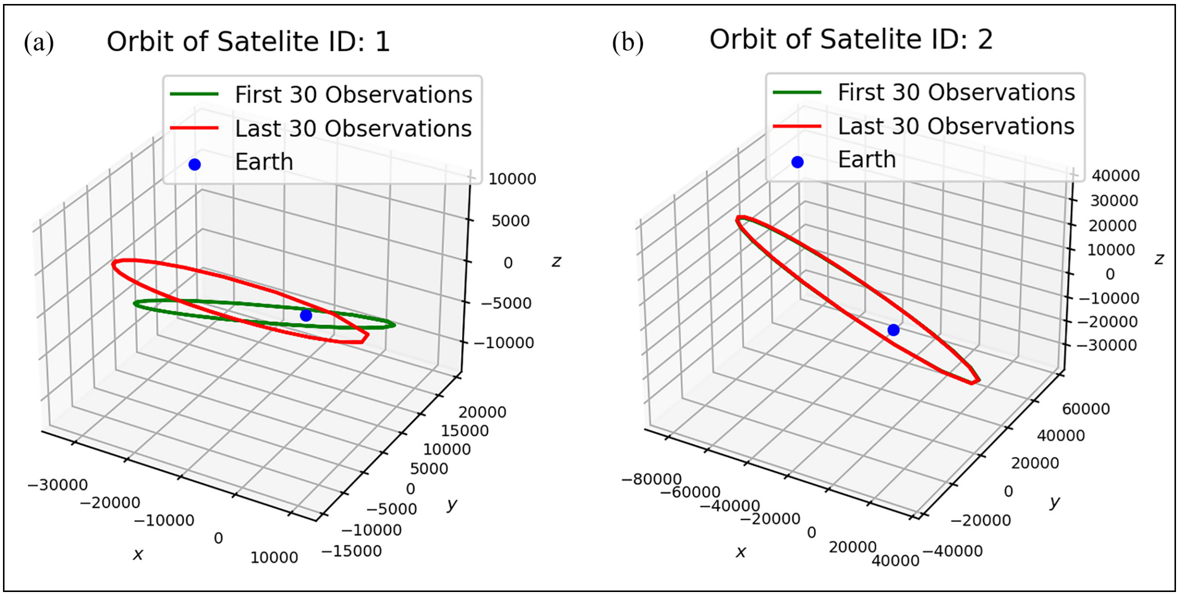


Figure 5. Orbits of satellites with ID 1 and 2 using cartesian coordinates in kilometers of the first and last 30 observations. Earth is at point (0, 0, 0). (a) Satellite orbit of satellite ID 1 and (b) satellite orbit of satellite ID 2.

other objects within Earth's orbit. Furthermore, it assumes that Earth is a point mass and neglects any relativistic effects. Since it is expected that the proposed method will find anomalies quite easily and that these anomalies can even be spotted by a visual comparison of the delta values without further analysis, the evaluation is rather a comparison of the precision of the Euler method prediction, the RK4 prediction, and the prediction using LSODA. This is achieved by applying the Wasserstein metric between the empirical and theoretical (or by equation (54) estimated) density distributions for each prediction and comparing the resulting distances. It is assumed that the Euler method performs worst and the LSODA prediction best. Also, the more stable the orbit is, the better the predictors are.

Two satellite orbits with ID 1 and ID 2 are analyzed in detail. The orbit of satellite 1 is quite unstable and is subject to strong other effects besides Earth's gravity and satellite 2 is stable in its orbit around Earth. A visualization of the first 30 and last 30 orbits after the measurements in January is given in Figure 5(a) and (b). It is easily visible that the orbits of satellite 1 are very different after the time frame, while the orbits of satellite 2 are still overlapping.

Equation (45) is used for the calculation of the noise covariance matrix in the case of the Euler method. Equation (54) is applied for the estimation of the noise covariance matrix of the RK4 method and LSODA. For a better visualization, only the prediction of the x-coordinate is discussed. However, an analysis of the other coordinates is also applicable and produces the same results and derivations. The variance is estimated for the position coordinates as $\sigma_x \approx 0.3$ km and for the velocity as $\sigma_{v_x} \approx 5 \times 10^{-5}$ km/s. The estimation of the variance takes the precision of the given data as well as the mean derivation of the predictions into account.

Even for the stable satellite orbit 2, the deviations between measurement and predictions are important and visible without further analysis within the data only by observation of the empirical density plots. The difference between predictions and measurements of RK4 and LSODA are within the same magnitude as the noise of the theoretical covariance matrix. The difference between predictions and measurements of the Euler method are, as assumed, multiple times the magnitude of the noise. The evaluation is plotted using the density distribution histograms. The histograms for satellite ID 2 are given exemplary for the RK4 in Figure 6. The empirical density distribution again highlights the necessity of using histogram bins and the Wasserstein metric since the covariance matrix would not fully encompass the complexity of the distribution.

By applying the proposed anomaly detection, all prediction methods would be classified as anomalous. Since it is not relevant in this case whether an anomaly is present but rather which prediction method is a better predictor, the Wasserstein metric is applied using the implementation by Flamary et al. (2021) with the sliced 2D Wasserstein metric by Bonneel et al. (2015) to determine which prediction is the most precise. The results are given in Table 2.

The results of the metric are as expected, with the exception that the Euler predictor performs more precisely in the unstable orbit of satellite 1 than in the stable orbit. This might be the result of the worse performance of the Euler method in orbits with high eccentricity since satellite 2 has a less round shape with a higher eccentricity. For the LSODA and RK4, the predictor performs better for the stable orbit. In addition, the analysis suggests that the LSODA performs better than the RK4 method, while the Euler method performs significantly worse than the other methods. This result is no surprise

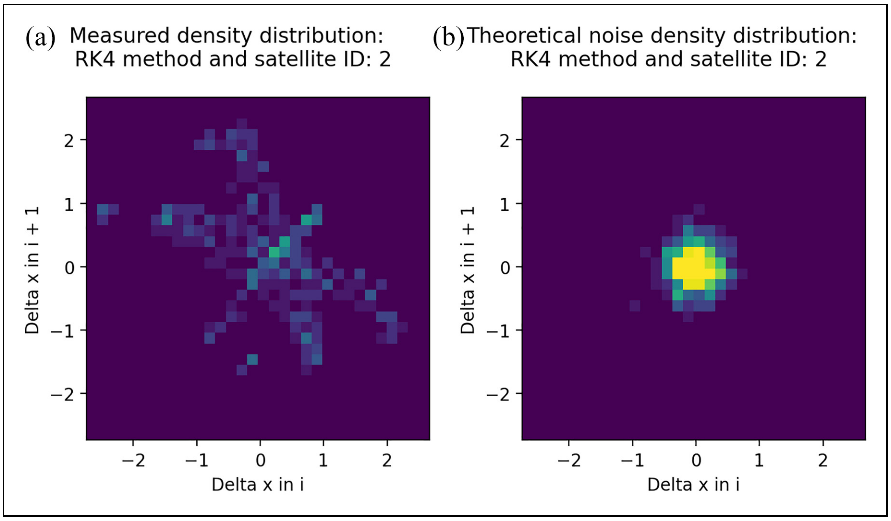


Figure 6. Theoretical noise density distribution and empirical density distribution of the *x*-coordinate between measurement *i* and *i* + 1 of satellite ID 2 for the RK4 method. (a) Empirical density distribution and (b) noise density distribution.

Table 2. Evaluation of the satellite orbit predictions of ID 1 and 2 using the 2D sliced Wasserstein metric by Bonneel et al. (2015).

	Applied prediction method		
Object	Euler method	RK4 method	LSODA
Satellite ID 1	832 ± 25	5.25 ± 0.03	1.003 ± 0.002
Satellite ID 2	1796 ± 71	5.07 ± 0.03	0.711 ± 0.006

since the Euler method is a Runge–Kutta Method of order 1 and therefore lacks the precision of a higher order method. Also, the RK4 is considered a less precise method than the more advanced LSODA predictors, which is reflected by our results. A reason for the worse performance of the RK4 are some very high outliers of the x-coordinate at the vertex points. To summarize, the Wasserstein metric enables a measure to evaluate predictors of an applied model.

Discussion

The use case and simulation study show the capabilities of our proposed methodology. However, some limitations exist. First, the expected function $\hat{f}$ of the operation must be a smooth function or always differentiable. This would not be the case for a sawtooth signal. In non-differentiable regions of the function, problems would arise in determining the theoretical covariance matrix of the measurement noise. However, the method would still be applicable in differential regions.

Second, the measurement noise must not be so large that the true operation is completely obscured. In this case, the model would be insufficient to obtain information about the true operation. The focus in the application would then be to first eliminate the measurement noise or to increase the number of samples.

Third, the runtime scales linearly with the number of samples N, but with smaller S/N ratios the required samples become larger by a factor of $10^{S/R}$. Therefore, a large sample size might be needed for very small signals, which increases the runtime. In general, a larger sample size improves the quality of the analysis.

Fourth, a prediction function $\hat{f}$ is necessary. If there is no model-based prediction function, the model can be applied analogously to any type of prediction function and combined with any forecasting or prediction processes, for example, AR(1) processes. Thereby, prediction processes can also be applied in nonlinear contexts. If the model size is extended from an AR(1) process to several past influences with an AR(q) process, the analyzed pairs $\hat{x}_{i+1}, \hat{x}_i$ increase linearly to the model size M to $\hat{x}_{i+M}, \ldots, \hat{x}_i$. The computation of the covariance matrix is analogous.

As a main difference to other methods, this contribution focuses on the prediction function as the subject of interest for anomaly detection and thus, error correction. Therefore, our proposed method emphasizes the validation of a system model using a measurement and prediction process. This model can be based on physical properties and derived differential equations but also on, for example, autoregressive models. It is not discussed nor are the cases differentiated within our method, whether the cause for differences between prediction and measurement is explained by inaccurate predictions or external factors creating an anomaly.

The procedure is able to precisely detect unexpected influences in the operations of a system and to assign them to the corresponding features and operation. No assumptions have to be made about the underlying distribution, and the necessary parameters are relatively easy to estimate in order to initialize the model. In addition, the approach is unsupervised and does not require any prior analysis of the results or a labeling of data points. However, the methodology assumes modeling and thus knowledge of the normal or expected system state. A further advantage is that the covariance matrix is computed analytically and no estimation with a prior clustering is necessary. This improves the real-time detection of outliers in time series with prior known prediction functions since the Mahalanobis distance is a well-researched and tested measure for outlier detection. In comparison to other models, this enables a global and local anomaly detection and a model identification process.

Conclusion

This work contributes to performing predictive anomaly detection more efficiently since the analysis is conducted without a clustering or other estimations, except a prior knowledge of the prediction function. Moreover, a contribution could be made especially for anomaly detection in nonlinear systems for which many of the conventional methods of prediction formation and anomaly detection have limitations. Furthermore, the systematic evaluation of prediction functions is an important task for practitioners setting up and controlling complex dynamical systems. Therefore, a main contribution of our approach is that it provides a useful measure to compare prediction functions using the Wasserstein metric, enabled by the analytically derived covariance matrix and the distribution of deltas via a histogram.

Through the knowledge of the unexpected states in a system and the affected features, a system engineer is subsequently able to transfer the unexpected states into a prediction formation to perform better simulations. Thus, the proposed anomaly detection and prediction evaluation improve the prediction formation in dynamical and nonlinear systems. Further research regarding possible applications within engineering and a benchmarking of the performance in different use cases compared to other models and algorithms for time series needs to be conducted. In addition, we want to analyze the possibility of using the information about the existence of an outlier or a global anomaly in the time series in order to develop a methodology to systematically improve the prediction function and, therefore, improve the capability of a system engineer to run simulations.

Acknowledgement

The research was prepared within the framework of the doctoral program of the Institut für Informationsmanagement im Ingenieurwesen at the Karlsruhe Institute of Technology.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research is funded by the Mercedes-Benz Group AG.

ORCID iD

Jan Michael Spoor https://orcid.org/0000-0001-8936-5695

References

Audibert J, Michiardi P, Guyard F, et al. (2020) USAD: UnSupervised anomaly detection on multivariate time series. In: *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining, Virtual Event, CA*, 6–10 July, pp. 3395–3404. New York: IEEE.

Bandt C and Pompe B (2002) Permutation entropy: A natural complexity measure for time series. *Physical Review Letters* 88(17): 174102.

Blázquez-García A, Conde A, Mori U, et al. (2021) A review on outlier/anomaly detection in time series data. *ACM Computing Surveys* 54(3): 56.

Bonneel N, Rabin J, Peyré G, et al. (2015) Sliced and radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision* 1(51): 22–45.

Box G (1949) A general distribution theory for a class of likelihood criteria. *Biometrika* 36(3–4): 317–346.

Burr T, Mullen M and Wangen L (1994) Process fault detection and nonlinear time series analysis for anomaly detection in safeguards. In: *International symposium on nuclear material safeguards, Vienna*. Available at: https://www.osti.gov/biblio/10120300

Cook A, Mısırlı G and Fan Z (2020) Anomaly detection for IoT time-series data: A survey. *IEEE Internet of Things Journal* 7(7): 6481–6494.

Fauconnier C and Haesbroeck G (2009) Outliers detection with the minimum covariance determinant estimator in practice. *Statistical Methodology* 6(4): 363–379.

Flamary R, Courty N, Gramfort A, et al. (2021) POT: Python optimal transport. *Journal of Machine Learning Research* 22(78): 1–8.

Germán-Salló Z (2018) Measure of regularity in discrete time signals. *Procedia Manufacturing* 22: 621–625.

Goldstein M and Dengel A (2012) Histogram-based outlier score (HBOS): A fast unsupervised anomaly detection algorithm. *Poster and Demo Track of the 35th German Conference on Artificial Intelligence (KI-2012)* 9: 59–63.

He J, Liu J, Shang P, et al. (2021) Dynamic Shannon entropy (DySEn): A novel method to detect the local anomalies of complex time series. *Nonlinear Dynamics* 104(4): 4007–4022.

Hindmarsh A (1983) ODEPACK, a systematized collection of ODE solvers. *IMACS Transactions on Scientific Computation* 1: 55–64.

International Data Analysis Olympiad (IDAO) (2020) Competition data set. Available at: https://disk.yandex.ru/d/0zYx00gSraxZ3w (accessed 9 March 2022).

Izakian H and Pedrycz W (2013) Anomaly detection in time series data using a fuzzy c-means clustering. In: *2013 joint IFSA world congress and NAFIPS annual meeting (IFSA/NAFIPS)*, Edmonton, Canada, 24–28 June 2013, pp. 1513–1518. New York: IEEE.

Li J, Izakian H, Pedrycz W, et al. (2021) Clustering-based anomaly detection in multivariate time series data. *Applied Soft Computing* 100: 106919.

Lindemann B, Maschler B, Sahlab N, et al. (2021) A survey on anomaly detection for technical systems using LSTM networks. *Computers in Industry* 131: 103498.

Manly B and Navarro Alberto J (2017) Multivariate Statistical Methods: A Primer (4th edn). New York: Chapman & Hall.

Marques F and Coelho C (2018) The simultaneous test of equality and circularity of several covariance matrices. *Journal of Statistical Theory and Practice* 12(4): 861–885.

Mehrotra K, Mohan C and Huang H (2017) Anomaly Detection Principles and Algorithms. Cham: Springer.

Muskulus M (2010) Distance-based analysis of dynamical systems and time series by optimal transport. PhD Thesis, Universiteit Leiden, Leiden.

Pincus S (1991) Approximate entropy as a measure of system complexity. *Proceedings of the National Academy of Sciences* 88(6): 2297–2301.

Puente C, Sáenz-Nuño M, Villa-Monte A, et al. (2021) Satellite orbit prediction using big data and soft computing techniques to avoid space collisions. *Mathematics* 9(17): 2040.

Rodriguez A, Bourne D, Mason M, et al. (2010) Failure detection in assembly: Force signature analysis. In: *2010 IEEE international conference on automation science and engineering*, Toronto, ON, Canada, 21–24 August, pp. 210–215. New York: IEEE.

Sperandio Nascimento E, Tavares O and De Souza A (2015) A cluster-based algorithm for anomaly detection in time series using Mahalanobis distance. In: *International conference on artificial intelligence (ICAI'2015)*, Las Vegas, NV, 27–30 July 2015, pp. 622–628. CSREA Press.

Spoor JM, Weber J and Ovtcharova J (2022) A definition of anomalies, measurements, and predictions in dynamical engineering systems for streamlined novelty detection. In: *2022 8th international conference on control decision and information technologies (CoDIT)*, Istanbul, Turkey, 17–20 May, pp. 675–680. New York: IEEE.

Tan Y, Hu C, Zhang K, et al. (2020) LSTM-based anomaly detection for non-linear dynamical system. *IEEE Access* 8: 103301–103308.

Titouna C, Titouna F and Ari A (2019) Outlier detection algorithm based on Mahalanobis distance for wireless sensor networks. In: *2019 international conference on computer communication and informatics (ICCCI)*, Coimbatore, India, 23–25 January, pp. 1–6. New York: IEEE.

Wang T, Cheng W, Li J, et al. (2011) Anomaly detection for equipment condition via cross-correlation approximate entropy. In: *MSIE*, Harbin, China, 8–11 January, pp. 52–55. New York: IEEE.