Preprints of the 21st IFAC World Congress (Virtual)
Berlin, Germany, July 12-17, 2020

# Anomaly Detection of Markov Processes with Evolution Equation and Moments ⋆

**Rafal Wisniewski** * **Manuela L. Bujorianu** **

* *Section of Automation & Control, Department of Electronic Systems, Aalborg University, Fredrik Bajers Vej 7C, 9220 Aalborg, Denmark (e-mail: raf@es.aau.dk).*
** *Maritime Safety Research Center, Department of Naval Architecture, Ocean & Marine Engineering, University of Strathclydes, G4 0LZ Glasgow, Scotland, UK (e-mail: luminita.bujorianu@strath.ac.uk)*

**Abstract:** Our departure point is the evolution equation of a Markov process. It describes the changes in the transition probability as time passes. We compare the transition probability for a priori model with the actual transition probability of the observed process to detect a mismatch between the expected and the measured data. To translate this idea into an algorithm, we characterise the involved measures by their moments. Specifically, a linear dynamic system is put forward that describes the evolution of moments. As the last result, we define a moment divergence as the means of computing the distance between two sequences of moments. We see the work as a step towards merging model-driven and data-driven concepts in control engineering. To elucidate the concepts introduced, we have incorporated several simple examples.

## 1. INTRODUCTION

Fault detection has a rich history (see Hwang et al. [2010], and the references therein). Model-based methods make use of a model (often deterministic) to generate a residual that reflects the deviation between estimated and measured signal. In this work, we strive to merge the model-based with the data-driven fault detection. We limit ourselves to detecting the discrepancy between the distributions of observations and the expected distribution from the model, and call this problem anomaly detection. This work is motivated by the study of anomaly detection in Pauwels and Lasserre [2016], where the measurements were compared with anticipated moments. To this end, the authors used the moment matrix known from the generalised moment method Lasserre [2001] used among others for the polynomial optimisation. The method of Pauwels and Lasserre [2016] applies to random variables and classifies as a data-driven method. This paper aims to extend it by including prior knowledge of the model of the Markov process to-be-examined.

The idea of this paper is to approximate the moments from the data at time $t_0$ and propagate them with the help of the evolution equation to compute the future moments at time $t$. The distance between the propagated moments and the moments computed from the data are first established. Subsequently, they used to determine if there is a discrepancy between the observation and the estimation. This distance indicates whether anomalies such as faults or cyber-attacks took place on the system.

The evolution equation as the means for computing the moments has been used before, for example, in the study of hitting probability in Cho and Stockbridge [2002].

⋆ This work was supported in part by Poul Due Jensen Foundation.

We try to keep the exposition on a level that does not require excessive notions from the probability theory. We use many examples to illustrate introduced notions. However, we also add several remarks which have a more technical character.

The paper is organised as follows. We introduce preliminary concepts and notations in Section 2. The primary object of the study, the evolution equation, is introduced in Section 3. The evolution equation is linear but infinite dimensional. Subsequently, it is approximated by a finite dimensional linear differential equation in Section 4. To this end, we use moments (in monomial basis). In Section 5, we introduce a moment divergence, which corresponds to the distance between two sequences of moments. A variant of the moment divergence will be used in Section 6 to detect anomalies between the moments computed from the observed data and the moments computed from the evolution equation. We will illustrate the method in an example of a one dimensional diffusion process in Section 6.

## 2. NOTATION AND PRELIMINARY CONCEPTS

This section aims to put forward the notation and the preliminary concepts used throughout the paper. In our effort to make the paper accessible by a broader audience in the control society, we will keep the formal definitions to the minimum. However, when referring to a subset of a Euclidean space $\mathcal{X}$, we will mean a Borel measurable set (i.e., it belongs to the Borel sigma algebra on $\mathcal{X}$ denoted by $\mathcal{B}(\mathcal{X})$). Along the same lines, functions $\mathcal{X} \to \mathbb{R}$ will be assumed measurable with respect to $\mathcal{B}(\mathcal{X})$ and $\mathcal{B}(\mathbb{R})$. We assume that the continuous-time processes of concern are Markov. We regard a situation when a Markov process $(X_t)$ at time 0 has an initial distribution $\mu_0$. For a set $A$, the probability $\mathbb{P}[X_0 \in A]$ that $X_0 \in A$ is equal to $\mu_0(A)$. Specifically, for Dirac measure $\delta_x$, $\mu_0 = \delta_x$ amounts to

$X_0 = x$ ($x \in \mathcal{X}$). For a set $A$, the function $I_A(x) = 1$ if $x \in A$, and 0 otherwise, is called the indicator function of $A$. In this work, the set of natural numbers $\mathbb{N}$ is with 0. For $\alpha \in \mathbb{N}^n$, $|\alpha| = \alpha_1 + \ldots + \alpha_n$, and $\binom{n}{\alpha} = \frac{n!}{\alpha_1! \ldots \alpha_n!}$. We refer to $\alpha$ as a multi-index. We use multi-indexes to index the entries of matrices. We use lexicographic order, i.e., $\alpha < \beta$ ($\alpha, \beta \in \mathbb{N}^n$) if and only if $|\alpha| < |\beta|$ else $\alpha_i < \beta_i$ for the first $i$, where $\alpha_i$ and $\beta_i$ differ. For a matrix $M$, $\mathrm{tr}(M)$ stands for its trace.

## 3. EVOLUTION EQUATION

At the outset, we introduce the main object of study, the evolution equation. It describes how the transition probability (entrance law) evolves in time. For a set $A$, the transition probability $\mu_t(A)$ tells what is the probability that $X_t$, with the initial distribution $\mu_0$, is in $A$. In particular, it is given by

$$\mu_t(A) \equiv \mathbb{E} I_A(X_t) = \mathbb{P}[X_t \in A], \tag{1}$$

where $\mathbb{E}$ is the expected value with respect to the probability $\mathbb{P}$, and $I_A$ is the indicator function of the subset $A \subset \mathcal{X}$. If the initial probability is $\mu_0 = \delta_x$, we will denote this specific occupation measure by $\mu_t^x$.

Having defined the measure $\mu_t$, we can integrate a measurable bounded function $h$ with respect to this measure. We call this integral the action of the measure $\mu_t$ on the function $h$

$$\mu_t h \equiv \int_{\mathcal{X}} h(x) \mu_t(dx).$$

To define the evolution equation, we will use the notion of infinitesimal operator. Its generalisation, an extended generator, will be discussed in the following after the Example 1. Following Arnold [1974], the (weak) infinitesimal generator $\mathcal{L}$ of a process $(X_t)$ is defined by

$$\mathcal{L}h(x) = \lim_{t \searrow 0} \frac{\mu_t^x h - h(x)}{t},$$

where the limit is point-wise for each $x$. We gather the functions for which the limit above is defined in the set $\mathcal{D}(\mathcal{L})$; this is the domain of $\mathcal{L}$.

*Remark 1.* We introduce the so-called extended generator Davis [1993] of the considered Markov process $(X_t)$, which is a generalisation of the infinitesimal generator. It plays an important role when dealing with more general processes than diffusions, for example: switched diffusion processes, piecewise-deterministic Markov processes, jump diffusion processes, or stochastic hybrid systems. Let $\mathcal{D}(\mathcal{L})$ be the set of measurable functions $h : (\mathcal{Y}, \mathcal{B}(\mathcal{Y})) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ having the property that there is a measurable function $g : (\mathcal{X}, \mathcal{B}(\mathcal{X})) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that the function $t \mapsto g(X_t)$ is almost surely (a.s.) integrable for each $y \in Y$, and the process $C_t^h$ given by

$$C_t^h \equiv h(X_t) - h(X_0) - \int_0^t g(X_s) ds \tag{2}$$

is a local martingale. We write $\mathcal{L}h = g$ and call $(\mathcal{D}(\mathcal{L}), \mathcal{L})$, or even $\mathcal{L}$, an extended generator. Notice that the extended generator $\mathcal{L}$ is possibly multi-valued. Nonetheless, if $g_1$ and $g_2$ are two values of the extended generator corresponding to $h$, then $g_1(x) \neq g_2(x)$ only on a subset $A$ where $\int_0^\infty I_A(X_t) dt = 0$ a.s. for $y \in \mathcal{Y}$, i.e., the process $(X_t)$ spends no time in $A$.

We are ready to define the evolution equation. For any function $h$ in the domain $\mathcal{D}(\mathcal{L})$ of the infinitesimal (or extended) generator $\mathcal{L}$, the following equation holds

$$\int_{\mathcal{X}} h(x) \mu_t(dx) - \int_{\mathcal{X}} h(x) \mu_0(dx) = \int_0^t \int_{\mathcal{X}} \mathcal{L}h(x) \mu_\tau(dx) d\tau. \tag{3}$$

*Remark 2.* The relation (3) follows from (2), hence

$$\mathbb{E}[h(X_t)] = \mathbb{E}[h(X_0)] + \mathbb{E}\left[\int_0^t \mathcal{L}h(X_s)\right] ds. \tag{4}$$

Taking the limit of $t$ going to 0 in relation (3), we formulate the following differential equation

$$\frac{d}{dt} \int_{\mathcal{X}} h(x) \mu_t(dx) = \int_{\mathcal{X}} \mathcal{L}h(x) \mu_t(dx). \tag{5}$$

Note that (5) is a differential form on the space of probability distributions, which is infinite dimensional. From now on, we suppose that $\mathcal{X}$ is a bounded subset of $\mathbb{R}^n$ and use monomials $x^\alpha$, where $\alpha \in \mathbb{Z}^n$, as the test functions $h$, $h(x) = x^\alpha$. Alternatively, the test function might be chosen arbitrary basis functions spanning a space dense in the domain of the generator $\mathcal{L}$.

*Example 1.* Consider the following stochastic differential equation on $\mathbb{R}^n$

$$dX_t = f(X_t) dt + \sigma(X_t) dB_t, \tag{6}$$

where $(B_t)$ is the Brownian motion with values in a Euclidean space $\mathbb{R}^l$. The infinitesimal generator $\mathcal{L}$ is given as follows: For any differentiable function $h : \mathbb{R}^n \to \mathbb{R}$

$$\mathcal{L}h = \frac{\partial h}{\partial x} f + \frac{1}{2} \mathrm{tr}\left(\sigma \sigma^T D^2 h\right),$$

where tr() stands for the trace, $\frac{\partial h}{\partial x} f = \sum \frac{\partial h}{\partial x_i} f_i$ and $D^2 h = [\frac{\partial^2 h}{\partial x_i \partial x_j}]$ is the Hessian of $h(t, \cdot)$.

*Example 2.* A switched diffusion process (SDP) consists of a family of diffusion processes and a switched mechanism, which allows spontaneous change among them. An SDP is a two component process $(q_t, X_t)$ with values in $\mathcal{Z} \equiv Q \times \mathcal{X}$ ($Q$ is a finite set of discrete modes) that satisfies stochastic differential equation (7) and the stochastic integral (8)

$$dX_t = f(q_t, X_t) dt + \sigma(q_t, X_t) dB_t, \tag{7}$$

where the probability of switch from the mode $i$ to $j$ is

$$\mathbb{P}[q_{t+\delta} = j |\ q_t = i,\ X_s,\ q_s, s \leq t] = \lambda_{ij}(X_t)\delta + o(\delta) \tag{8}$$

for $i \neq j$, and for intensity functions $\lambda_{ij}$.

By Baran et al. [2013], the generator is characterised as follows. For any function $h : \mathcal{Z} \to \mathbb{R}$ with $h(i, \cdot) \in C^2(\mathbb{R}^n)$, $i \in Q$, the generator $\mathcal{L}$ is defined by

$$\mathcal{L}h(i,x) = \frac{\partial h(i,x)}{\partial x} f(i,x) + \frac{1}{2} \mathrm{tr}(\sigma(i,x) \sigma^{\mathrm{T}}(i,x) \mathrm{D}^2 h(i,x)) + \sum_{j \in Q, j \neq i} \lambda_{ij}(x)(h(j,x) - h(i,x)). \tag{9}$$

## 4. FINITE DIMENSIONAL APPROXIMATION

In this section, we will develop finite dimensional approximation of (5). To this end, we will assume that polynomials are dense in the domain of the considered generator $\mathcal{L}$, and represent $h = \sum_\alpha a_\alpha x^\alpha$ in monomial basis. Motivated by Examples 1 and 2, we also suppose that $\mathcal{L}$ is a variant of a differential operator. Since a differential operator is a linear one that satisfies the Leibniz rule, $\mathcal{L}$ acting on $x^\alpha$ is of the form

$$\mathcal{L}x^\alpha = \sum_\beta l(\alpha,\beta)x^\beta. \tag{10}$$

In Example 3, we will show the specific form of (10) for a diffusion process, and in Example 4 for an SDP.

We shall denote

$$m_\alpha(t) = \int_{\mathcal{Y}} x^\alpha \mu_t(dx).$$

Throughout the work, we suppose that $m_\alpha$ are finite.

From (5), we have the following differential moment equation

$$\frac{d}{dt}m_\alpha(t) = \sum_\beta l(\alpha,\beta)m_\beta(t).$$

We define matrix $L = [l(\alpha,\beta)]$ with entries $l(\alpha,\beta)$, then the time evolution of $m(t) = [m_\alpha(t)]$ is governed by

$$\dot{m}(t) = Lm(t). \tag{11}$$

As a result, the probability distribution evolution equation is transformed into a moment evolution equation.

*Example 3.* Consider a simple diffusion on the real line given by the following stochastic differential equation

$$dX_t = aX_t dt + bdB_t, \tag{12}$$

for some nonzero reals $a$, $b$. Specifically, for $j \geq 2$

$$\mathcal{L}x^j = jax^j + j(j-1)\frac{1}{2}b^2x^{j-2}.$$

The equation for the evolution of moments are

$$\dot{m}_j = jam_j + j(j-1)\frac{1}{2}b^2m_{k-2} \quad \text{for } j \geq 2,$$

and $\dot{m}_1 = am_1$. Taking the vector $m = (m_0, m_1, m_2)$, we compute

$$\dot{m}(t) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & a & 0 \\ b^2 & 0 & 2a \end{bmatrix} m(t). \tag{13}$$

As a result, $m_0(t) = 1$, $m_1(t) = m_1(0)e^{at}$, and $m_2(t) = (m_2(0) + \frac{b^2}{2a})e^{2at} - \frac{b^2}{2a}$.

*Example 4.* Consider an SDP

$$dX_t = a_iX_t dt + b_i dB_t,\ i \in \{1,2\}$$

with $\lambda_{12}(x) = \lambda_1$, and $\lambda_{21}(x) = \lambda_2$.

The generator is of the form

$$\mathcal{L}(h_1,h_2) \tag{14}$$
$$= \begin{bmatrix} \dfrac{\partial h_1(x)}{\partial x}a_1x + \dfrac{1}{2}b_1^2\dfrac{\partial^2 h_1}{\partial x^2} + \lambda_1(h_2(x) - h_1(x)) \\ \dfrac{\partial h_2(x)}{\partial x}a_2x + \dfrac{1}{2}b_2^2\dfrac{\partial^2 h_2}{\partial x^2} + \lambda_1(h_1(x) - h_2(x)) \end{bmatrix}.$$

The moments are defined by the family of pairs

$$\begin{bmatrix} m_\alpha \\ m_\beta \end{bmatrix} = \begin{bmatrix} \int_{\mathcal{X}} x^\alpha \mu_t(1,dx) \\ \int_{\mathcal{X}} x^\beta \mu_t(2,dx) \end{bmatrix},$$

where $\mu_t(\{i\},A) = \mathbb{P}[(q_t,X_t) \in \{i\} \times A]$.

To illustrate an instance of moments, we compute the evolution of the pair $(m_2,m_1)$ of moments applying $h(x) = (x^2,x)$ in (14),

$$\frac{d}{dt}\begin{bmatrix} m_2 \\ m_1 \end{bmatrix} = \begin{bmatrix} 2am_2 + b_1^2 + \lambda_1(m_1 - m_2) \\ am_1 + \lambda_2(m_2 - m_1) \end{bmatrix}.$$

### *4.1 Observations*

We will model the observations as a stationary Markov process $(Y_t)$ on a bounded subset $\mathcal{Y}$ of the Euclidean space $\mathbb{R}^l$. We define its conditional distribution $P_Y(dy,x)$, where $P_Y(A,x)$ is the probability that $Y_t \in A$ provided that $X_t = x$.

We let $P_Y$ act on a polynomial $f \in \mathbb{R}[Y]$

$$P_Y(f,x) \equiv \int_{\mathcal{Y}} f(y)P_Y(dy,x)$$

and

$$P_Y(f,t) \equiv \int_{\mathcal{X}}\int_{\mathcal{Y}} f(y)P_Y(dy,x)\mu_t(dx).$$

Specifically $\mathbb{P}[Y_t \in A] = P_Y(I_A,t)$.

Suppose that there exists a probability density function $c(x,y)$ of $P_Y(dy,x)$ with respect to the Lebesgue measure $dy$, i.e.,

$$P_Y(dy,x) = c(x,y)dy.$$

As a consequence

$$P_Y(f,t) = \int_{\mathcal{X}}\int_{\mathcal{Y}} f(y)c(x,y)dy\mu_t(dx). \tag{15}$$

We approximate the distribution $c$ by a polynomial $\overline{c} \in \mathbb{R}[X,Y]$ on $\mathcal{X} \times \mathcal{Y}$.

Firstly, consider the integral $\int_{\mathcal{X}} \overline{c}(x,y)\mu_t(dx)$ with $\overline{c}(x,y) = \sum_{(\alpha,\beta)} \overline{c}_{(\alpha,\beta)}x^\alpha y^\beta$

$$\int_{\mathcal{X}} \overline{c}(x,y)\mu_t(dx) = \sum_{(\alpha,\beta)} \overline{c}_{(\alpha,\beta)}m_\alpha(t)y^\beta.$$

We denote the moments of the observation process $Y_t$ by $w_\beta(t) \equiv P_Y(y^\beta,t)$. From (5), the moment $w_\beta(t)$ is

$$w_\beta(t) = \sum_\alpha \overline{c}_{(\alpha,\beta)}m_\alpha(t).$$

Hence, for $w(t) \equiv [w_\beta(t)]$, and $C \equiv [\overline{c}_{\alpha,\beta}]$

$$w(t) = Cm(t). \tag{16}$$

In conclusion, the measurements of the observation process are governed by the linear system

$$\begin{aligned} \dot{m}(t) &= Lm(t) \\ w(t) &= Cm(t). \end{aligned} \tag{17}$$

*Example 5.* Consider the diffusion equation in Example 1 with the measurement governed by a linear map $D : \mathbb{R}^n \to \mathbb{R}^l$, $x \mapsto y = Dx$. Suppose $K \in \mathbb{N}^{l\times n}$. We denote the ith raw of $K$ by $K(i,:)$, i.e., the vector $K(i,:) \equiv [K(i,1),\ldots,K(i,l)]$.

We use the notation

$$\binom{n}{K} \equiv \binom{n}{K(1,:)} \cdots \binom{n}{K(l,:)},$$
$$D^K = \prod_{(i,j)\in\{1,\ldots,l\}\times\{1,\ldots,n\}} D(i,j)^{K(i,j)}.$$

Then

$$w^\beta = \sum_{\substack{|K(1,:)| = \beta_1 \\ \cdots \\ |K(l,:)| = \beta_l}} \binom{n}{K} D^K m(t)^{K(1,:)+\ldots+K(l,:)}.$$

*Remark 3.* Using the standard linear control theory, we can estimate $m(t)$ provided that the pair $(L, C)$ is observable. For example, defining standard observer dynamics

$$\frac{d}{dt}\overline{m}(t) = L\overline{m}(t) + K(w(t) - C\overline{m}(t)), \tag{18}$$

we strive to optimize

$$\int_0^\infty (m(t) - \overline{m}(t))^T Q(m(t) - \overline{m}(t)) + (w(t) - C\overline{m}(t))^T R(w(t) - C\overline{m}(t))dt.$$

subject to $K$ satisfying the relation in (18).

The solution is given by $K = PC^T R^{-1}$, where $P$ is the solution of the following Riccati equation

$$-PA^T - AP + PC^T R^{-1} CP - Q = 0.$$

*Remark 4.* Not each sequence $\overline{m}$ corresponds to a moment vector of a measure. In fact, by Riesz-Haviland theorem, for a closed $\mathcal{X}$, there exists a finite Borel measure $\mu$ on $\mathcal{X}$ such that $\int_{\mathcal{X}} x^\alpha \mu(dx) = \overline{m}_\alpha$ if and only if $\sum_\alpha p_\alpha \overline{m}_\alpha$ for all non-negative real polynomials $\sum_\alpha p_\alpha x^\alpha$. An equivalent formulation in terms of moment- and localizing-matrices in given in Theorem 3.8 in Lasserre [2010].

## 5. MOMENT DIVERGENCE

In this section, we will establish the means of computing the distance between two sequences of moments. For this reason, we introduce the concept of a moment matrix.

### *5.1 Moment Matrix*

Let $m \equiv (m_\alpha)$ be the moments of a measure $\mu$ on a space $\mathcal{X} \equiv \mathbb{R}^n$. Let $\mathbb{N}^n_d \equiv \{\gamma \in \mathbb{N}^n | \ |\alpha| \leq d\}$. The moment matrix $M_{\mathcal{X},d}(m)$ of order $d$ is defined by

$$M_{\mathcal{X},d}(m)(\alpha,\beta) = m_{\alpha+\beta} \ \forall \alpha, \beta \in \mathbb{N}^n_d.$$

It is $s(k,d)\times s(k,d)$ real matrix with $s(k,d) = \binom{k+d}{k}$. From the definition, the moments matrices are symmetric.

*Example 6.* For the measure $\mu$ defined on $\mathbb{R}^2$, the moment matrix $M_{\mathcal{X},2}(m)$ is

$$\begin{bmatrix} m_{00} & m_{01} & m_{10} & m_{02} & m_{11} & m_{20} \\ m_{01} & m_{02} & m_{11} & m_{03} & m_{12} & m_{21} \\ m_{10} & m_{11} & m_{20} & m_{13} & m_{21} & m_{30} \\ m_{02} & m_{03} & m_{12} & m_{04} & m_{13} & m_{22} \\ m_{11} & m_{12} & m_{21} & m_{13} & m_{22} & m_{31} \\ m_{20} & m_{21} & m_{30} & m_{22} & m_{31} & m_{40} \end{bmatrix}.$$

*Example 7.* We continue with Example 3. The moment matrix $M_{\mathcal{X},2}(m(t))$ is

$$\begin{bmatrix} m_0(t) & m_1(t) \\ m_1(t) & m_2(t) \end{bmatrix} = \begin{bmatrix} 1 & m_1(0)e^{at} \\ m_1(0)e^{at} & \left(m_2(0) + \frac{b^2}{2a}\right)e^{2at} - \frac{b^2}{2a} \end{bmatrix}.$$

Important for the next section is that the moment matrix is positive semi-definite Lasserre [2010].

In the sequel, we will explore the moment matrix corresponding to the observation process on a subspace $\mathcal{Y} \equiv \mathbb{R}^l$. For now, we consider a vector $m \in \mathbb{R}^{s(k,2d)}$ of moments up to degree $2d$ of a random variable defined on $\mathcal{X}$ and the vector of moments $w \in \mathbb{R}^{s(l,2d')}$ (up to degree $2d'$) of another random variable defined on $\mathcal{Y}$. We suppose that $w$ and $m$ are related as in (16) by $w = Cm$. We have the following commutative diagram

$$\begin{array}{ccc} \mathbb{R}^{s(k,2d)} & \xrightarrow{M_{\mathcal{X},d}} & \mathbb{R}^{s(k,d)} \times \mathbb{R}^{s(k,d)} \\ \Big\downarrow C & & \vdots\, C' \\ \mathbb{R}^{s(l,2d')} & \xrightarrow[M_{\mathcal{Y},d'}]{} & \mathbb{R}^{s(l,d')} \times \mathbb{R}^{s(l,d')} \end{array}$$

The maps $M_{\mathcal{X},d}$ and $M_{\mathcal{Y},d'}$ are bijections, and the induced linear operator $C'$ takes $M_{\mathcal{X},d}(m)$ to $C'M_{\mathcal{X},d}(m)$ defined by

$$C'M_{\mathcal{X},d}(m) = M_{\mathcal{Y},d'}(Cm).$$

### *5.2 Moment Divergence*

In this section, we establish the means of computing the distance between two measures on a space $\mathcal{Y}$, both with finite moments. Suppose that there are two measures $\mu_A$ with the moment matrix $M_A$ and $\mu_B$ with the moment matrix $M_B$. We use the property of the moment matrix of being positive semidefinite. Symmetric matrices constitute a finite dimensional Hilbert space with scalar product given by

$$\langle M_A, M_B \rangle = \mathrm{tr}(M_A M_B).$$

Since $M_A$ and $M_B$ are the moment matrices, the product $\langle M_A, M_B \rangle$ is non-negative. To show it, we define a vector of monomials up to degree $d$ by

$$v_d(y) \equiv (y^\alpha)_{|\alpha|\leq d} = \left[1, y_n, \ldots, y_1, y_n^2, y_n y_{n-1}, \ldots, y_1^d\right]^T. \tag{19}$$

Notice that the monomial in (19) are ordered in the lexicographic order. Nonetheless, any other order can be applied and consistently used in the numbering the entries of the moment matrix. It follows that

$$\begin{aligned} \langle M_A, M_B \rangle &= \int_{\mathcal{Y}} \langle M_A v_d(y) v_d(y)^T \rangle \mu_B(dy) \\ &= \int_{\mathcal{Y}} v_d(y)^T M_A v_d(y) \mu_B(dy) \geq 0. \end{aligned}$$

We will define two moment divergences: the first for a non-singular moment matrix $M_A$, and the second without this assumption. It will be shown in the next section that the former is particularly useful for anomaly detection.

We define a moment divergence by

$$\pi(M_A, M_B) \equiv \log^2 \frac{\langle M_A, M_B \rangle}{||M_A||^2}. \tag{20}$$

Since $\pi(M_A; M_A) = 0$, $\pi$ is a proximity of the distance of $\mu_B$ from $\mu_A$. The moment divergence $\pi$ can be interpreted as the projection of the moments of measure $\mu_B$ on the moments of $\mu_A$.

We suppose now that $M_A$ is non-singular, subsequently we define another moment divergence $\rho$ by

$$\rho(M_A, M_B) \equiv \log^2 \frac{\langle M_A^{-1}, M_B \rangle}{s(l,d)}, \tag{21}$$

where $l = \dim \mathcal{Y}$, and $d$ is the order of the moment matrices $M_A$ and $M_B$.

*Example 8.* Specifically, the moment divergence $\pi$ between $M \equiv M_{\mathcal{X},2}(m(t))$ and $M' \equiv M_{\mathcal{X},2}(m'(t))$ is

$$\pi(M, M') = \log^2 \frac{1 + 2m_1(t)m_1'(t) + m_2(t)m_2'(t)}{1 + 2m_1^2(t) + m_2^2(t)}.$$

Whereas, $\rho$ is given by

$$\rho(M, M') = \log^2 \left( \frac{m_0(t)m_2'(t) - 2m_1(t)m_1'(t) + m_2(t)m_0'(t)}{2(m_0(t)m_2(t) - m_1^2(t))} \right).$$

## 6. ANOMALY DETECTION

The moment matrix will be used in this section for evaluating probability that the process $(X_t)$ is a subject to change of its characteristics. To this end, we strive to define a function $r : \mathcal{Y} \to \mathbb{R}$ defined on the observation space and to evaluate the probability that the anomaly takes place. We denote the observation process predicted from the model (17) by $\overline{Y}_t$, and its moments by $\overline{w}(t)$. Suppose that the initial distribution is the same for both $Y_0$ and $\overline{Y}_0$. At time $t$, the moments of the two processes are $M_{\mathcal{Y},d}(w(t))$ and $M_{\mathcal{Y},d}(\overline{w}(t))$:

$$M_{\mathcal{Y},d}(w(0)) \xrightarrow{t} M_{\mathcal{Y},d}(w(t))$$
$$M_{\mathcal{Y},d}(w(0)) \xrightarrow{t} M_{\mathcal{Y},d}(\overline{w}(t)).$$

The discrepancy between the modelled and observed moment matrices is captured by the moment divergences $\pi$ and $\rho$. Specifically in the section, we assume that $M_{\mathcal{Y},d}(w(t))$ is non-singular for $t \geq 0$ and focus on "re-scaled" $\rho$

$$\langle M_{\mathcal{Y},d}(\overline{w}(t))^{-1}, M_{\mathcal{Y},d}(w(t) \rangle$$

Specifically, we study a quadratic form $r : \mathbb{R}^{s(l,2d')} \to \mathbb{R}$ defined by

$$r : \ v \mapsto v^T M_{\mathcal{Y},d}(\overline{w}(t))^{-1} v.$$

The quadratic form $r$ is positive definite, since $M_{\mathcal{Y},d}(w(t))$ has been assumed to be a positive definite matrix.

*Proposition 1.* Suppose that $M_{\mathcal{Y},d}(\overline{w}(t))$ is non-singular.

$$\mathbb{P}[r(v_d(Y_t)) \geq \epsilon] \leq \frac{1}{\epsilon} \langle M_{\mathcal{Y},d}(\overline{w}(t))^{-1}, M_{\mathcal{Y},d}(w(t)) \rangle.$$

*Example 9.*

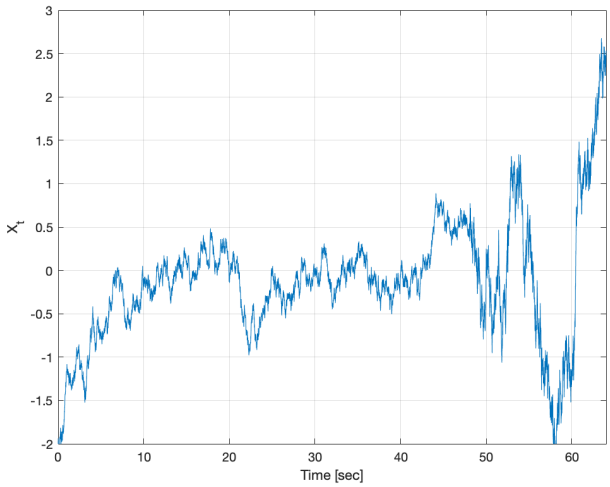


Fig. 1. The diffusion process with the drift $x \mapsto -0.2x$ and diffusion $x \mapsto 0.3$ . At the time instance 48 sec, the diffusion increases from 0.3 to 0.8.

Before we prove the proposition, we will discuss its importance. Indeed, the proposition demonstrates that

$$r_t \equiv r(v_d(Y_t))$$

is a residual applicable for anomaly detection. If there is no anomaly, $M_{\mathcal{Y},d}(w(t)) = M_{\mathcal{Y},d}(\overline{w}(t))$, then

$$\mathbb{P}[r_t \geq \epsilon] \leq \frac{s(l,d)}{\epsilon}.$$

Specifically, we pick a number $p$, the probability that

$$r_t \geq \kappa \equiv \frac{s(l,d)}{p}.$$

Whenever we observe that $r_t \geq \kappa$, we declare that the anomaly takes place.

**Proof.** By Markov inequality, we have

$$\mathbb{P}[r(v_d(Y_t)) \geq \epsilon] \leq \frac{1}{\epsilon} \mathbb{E}[r(v_d(Y_t)]. \tag{22}$$

We compute the expected value of $r_t \equiv r(v_d(X_t))$

$$\mathbb{E}[r_t] = \int_{\mathcal{Y}} r(v_d(y)) \nu_t(dy), \tag{23}$$

where $\nu_t(dy) = \int_{\mathcal{X}} P_Y(dy, x) \mu_t(dx)$ is the distribution of $Y_t$.

Let $v \equiv v_d(y)$, $\overline{M} \equiv M_{\mathcal{Y},d}(\overline{w}(t))$, and $M \equiv M_{\mathcal{Y},d}(w(t))$. From (23),

$$\mathbb{E}[r_t] = \int_{\mathcal{Y}} v^T \overline{M}^{-1} v d\nu_t = \mathrm{tr} \int_{\mathcal{Y}} \overline{M}^{-1} v v^T d\nu_t = \langle \overline{M}^{-1}, M \rangle.$$

□

*Remark 5.* Suppose that the moments $m$ in (17) are estimated from the moments of the observation $w$ as in Remark 3. We denote this estimate by $\overline{m}$. By the continuity of the map $M$, and assuming a sufficiently small estimation error $e \equiv m - \overline{m}$, if $M_{\mathcal{Y},d}(Cm) > 0$, then $M_{\mathcal{Y},d}(C\overline{m}) > 0$.

We consider the diffusion process in Example 3. The specific stochastic differential equation is given by

$$dX_t = -0.2X_t + 0.3dB_t.$$

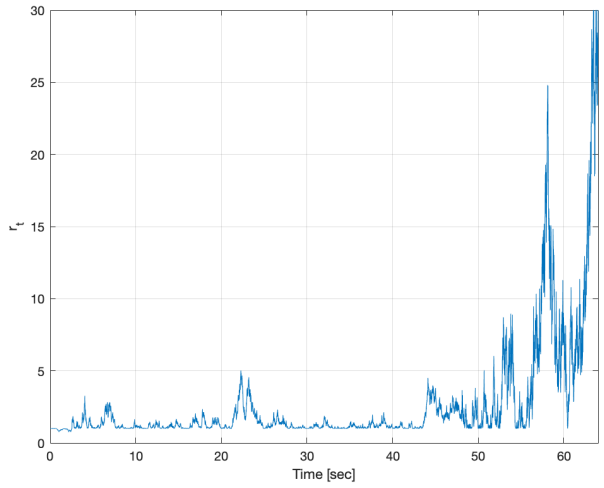


Fig. 2. The residual process $r_t$ indicates if an anomaly takes place. If there is no anomaly, the values of $r_t$ being above 20 have probability below 0.1.

The observation process $Y_t = X_t$.

During the first 10 sec, the moments are computed. Subsequently, the vector $m(t)$ of moments is propagated according to (13). At the time instance 48 sec, the anomaly is generated, which is the increase of the diffusion term from 0.3 to 0.8. A realization of the diffusion process is shown in Figure 1. The process $r_t$ is shown in Figure 2. To illustrate, in the case of no anomaly, the probability that $r_t \geq 20$ is less or equal to 0.1. The values of $r_t$ above this level indicate the occurrence of an anomaly.

## 7. CONCLUSIONS

In the paper, we developed a method for detecting anomalies in Markov processes. To this end, we used the time evolution of moments, and we introduced a moment divergence that allows comparing the expected distribution of the process with the actual observed data. We see the future of this work as verification of the method on a concrete industrial use-case.

## REFERENCES

L. Arnold. *Stochastic differential equations: theory and applications.* Wiley, 1974. ISBN 9780471033592.

Nicholas A. Baran, George Yin, and Chao Zhu. Feynman-Kac formula for switching diffusions: connections of systems of partial differential equations and stochastic differential equations. *Adv. Difference Equ.*, pages 2013:315, 13, 2013. ISSN 1687-1847.

Moon Jung Cho and Richard H. Stockbridge. Linear programming formulation for optimal stopping problems. *SIAM J. Control Optim.*, 40(6):1965–1982, 2002. ISSN 0363-0129. doi: 10.1137/S0363012900377663.

M. H. A. Davis. *Markov models and optimization.* Chapman & Hall, 1993. ISBN 041231410X.

I. Hwang, S. Kim, Y. Kim, and C. E. Seah. A survey of fault detection, isolation, and reconfiguration methods. *IEEE Transactions on Control Systems Technology*, 18(3):636–653, May 2010. doi: 10.1109/TCST.2009.2026285.

Jean B. Lasserre. Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization*, 11(3):796–817, March 2001. ISSN 1052-6234. doi: 10.1137/S1052623400366802.

Jean Bernard Lasserre. *Moments, positive polynomials and their applications*, volume 1 of *Imperial College Press Optimization Series.* Imperial College Press, London, 2010. ISBN 978-1-84816-445-1; 1-84816-445-9.

Edouard Pauwels and Jean B Lasserre. Sorting out typicality with the inverse moment matrix sos polynomial. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 190–198. Curran Associates, Inc., 2016.