Cointegration analysis method for fault detection based on sensor data

R V Faizullin* and S Hering
Izhevsk State Technical University, 7 Studencheskaya street, Izhevsk, Russia, 426069
*rf85@mail.ru

Abstract. Sensors are a popular source of information about the operation of complex dynamic technical systems. Considering data from sensors as a multidimensional time series is also used to describe cyber-physical systems. The article proposes a method for detecting system malfunctions based on the method of analyzing cointegration dependencies. It is determined that in the data for analysis it is possible to reveal cointegration dependences as facts of interdependence of data from different sensors. Calculations are given on the example of a system with 52 parameters. Out of 1,326 data pairs, 75 are cointegrated. The conducted analysis shows that the proposed method enables one to clearly illustrate situations with changes in behavior. By tracking deviations of observations from the calculated equation, it is possible to identify system errors.

1. Introduction
The Fourth Industrial Revolution is already in progress. In the face of all these elemental changes, a vast majority of industries will have to reorganize. This radical reorganization is referred to in this text as ‘Industry 4.0’.

Wildemann points out that the basic idea of Industry 4.0 consists of two elements: the global cross-system networking of people, plants and products as well as the independent and decentralized organization and control of these production units in real time [1]. According to him, the technological-economic paradigm of Industry 4.0 is based on the merging of the physical world with the virtual world into a cyber-physical system (CPS) [1]. In addition to the Internet of Things, current trends in information and communication technology, such as Big Data, Cloud Services and major technical advances in sensor technology, form the foundation for expanding cyber-physical systems in line with Industry 4.0.

The IoT connects decentralized physical objects to the Internet. In a factory, a CPS is created by interlinking Smart Products, intelligent workpieces, machines and transport units with each other. As already mentioned, those objects synchronize and collaborate automatically and interact with humans via interfaces to form a real-time industrial network. This enables intelligent automation.

Modern technical systems are complex and dynamic, therefore, to describe the operation of such a system, it is necessary to track a large number of parameters that change over time. That is, complex dynamic systems can be described in the form of a multidimensional time series, where some series will describe data on technological processes, and others on management action. Controlling process conditions is critical to proper operation.
Data is rightly seen as the most important raw material. To use and manage the ever-growing volumes of data, innovative concepts of Data Mining are required. This also helps generating new business models. A popular application example is ‘predictive maintenance’: the machine tells the user when it has to be maintained, reducing its down time.

The modern stage of development enables you to receive and use more and more information about technology. Modern sensors do not simply measure classic switching signals, they also analyse their environment. From this analysis, output signals are generated to control actuators, which leads to predetermined mechanical actions. Smart Sensors not only enable ‘predictive maintenance’. They entail a much more radical automation, including machine learning, which provides for added value in process design. This opens up a huge potential for planning security, quality standards, flexibility and productivity.

Fault diagnosis in continuous dynamic systems can be challenging, since the variables in these systems are typically characterized by autocorrelation, as well as time variant parameters, such as mean vectors, covariance matrices, and higher order statistics, which are not handled well by methods designed for steady state systems. In dynamic systems, steady state approaches are extended to deal with these problems, essentially through feature extraction designed to capture the process dynamics from the time series. [2]

Modern CPS are typical examples of complex dynamic engineering systems. Such systems are now equipped with sensors and control mechanisms. Such systems generate large amounts of data that need to be analyzed and if the data do not meet expectations, then this may be a sign of a possible breakdown, accident, or the need for human intervention into management.

The Mining Data Correlation from Multi-faceted Sensor Data is one of the sources of information for making decisions about a complex dynamic object in the concept of the Internet of Things. The Internet of Things uses sensors to analyze data. It is important to consider a multidimensional time series (a data set in the form of a time series from each sensor and control mechanism), because it enables one to consider the relationships between the data. Analyzing these data derived from sensors is an essential task and can find the useful latent information besides the data itself [3]. Researchers use correlation analysis [3, 4, 5, 6], analysis for the presence of cointegration [7, 8, 9, 10], use principal component methods [11, 12, 13], neural networks [14, 15, 16, 17] to analyze CPS data and predict breakdowns. However, even the most complex forecasting systems do not give 100% results. Methods for identifying the fact that the time series has changed (a structural shift has occurred), for example, by the Perron-Bay method [18], are based on the fact that a sufficient number of observations must be obtained after the shift for the estimates to be reliable.

2. Methodology
Let us consider cointegration as a property of time series, for this it is also necessary to mention the concepts of stationarity/nonstationarity of time series.

Many processes can be described as autoregressive models. Indeed, time series of any parameters are often inertial processes of some kind, that is, each value depends on the previous one.

\[ y_t = \phi y_{t-1} + \varepsilon_t \]  

Where random component \( \varepsilon_t \) has zero mathematical expectation and stable variance \( \varepsilon_t \sim N(0, \sigma^2) \). Then, depending on the value of coefficient \( \phi \), three types of time series are distinguished [19, 20]. These are stationary time series (for \( \phi <0 \)); nonstationary time series (for \( \phi >0 \)); random walk (for \( \phi =0 \)).

We can say that the stationary process does not depend on time, but in the non-stationary process there are some dependences on time. The great popularity of studies of time series in general and stationarity of time series in particular came from econometrics, because many economic processes are nonstationary, and most methods impose the requirement of stationarity. In technology, time series are also often found, which demonstrate trends, fairly regular cycles or other manifestations of non-stationary behavior. If any of the indicators is growing steadily over time, then it can be brought to a
stationary form by simply removing the trend line. If this transformation results in a stationary series, it is said to be difference-stationary [20]. This is not the only way to make the series stationary. It is possible to go from the original series to the first-order differences \( \Delta \gamma_t = \gamma_t - \gamma_{t-1} \). An example of a difference-stationary process is the random walk series, defined by Equation (1). This series is an \( \phi = 1 \) integrated series of order 1, denoted I(1) [20]. For such a series, Equation (1) yields

\[
\Delta \gamma_t = \gamma_t - \gamma_{t-1} = \epsilon_t \tag{2}
\]

If, for reduction to a stationary form, it is necessary to take second-order differences, then the process is considered I(2) and so on. In most cases, the first-order differences are sufficient, the second-order differences are used much less frequently, and the differences above 2 are almost never.

Having defined our understanding of stationary time series and reduction of non-stationary time series to a stationary form, let us proceed to the description of cointegration.

If a linear combination of some set I(1) of time series \( Y_t = (y_{1t}, y_{2t}, \ldots, y_{nt})^T \) has the properties of a stationary series, then they are called cointegrated.

\[
\beta^T Y_t = \beta_1 y_{1t} + \beta_2 y_{2t} + \cdots + \beta_n y_{nt} \sim I(0) \tag{3}
\]

If there is at least one such vector \( \beta \), then the dataset is considered cointegrated. It is clear that if the original series are non-stationary, then for their linear combination to be stationary, it must include at least 2 time series. The use of a large number of time series in the cointegration relationship is complicated by interdependencies between non-stationary time series, which causes the effect of multicollinearity.

It is clear that any cointegration vector is not unique, because (4)

\[
k \cdot \beta^T Y_t = (\beta^*)^T Y_t \sim I
\]

Usually, the cointegration relation leads to the expression of dependence of one time series on others, then the cointegration vector can be written as (5)

\[
\beta = (1, -\beta_2, \ldots, -\beta_n)^T
\]

Thus, the cointegration ratio can be represented as

\[
\beta^T Y_t = y_{1t} - \beta_2 y_{2t} - \cdots - \beta_n y_{nt} \sim I(0) \tag{6}
\]

or

\[
y_{1t} = \beta_2 y_{2t} + \beta_3 y_{3t} + \cdots + \beta_n y_{nt} + u_t \tag{7}
\]

where \( u_t = \beta^T Y_t \sim I(0) \)

It turns out that we assume the presence of a certain long-term equilibrium equation (8)

\[
y_{1t} = \beta_2 y_{2t} + \beta_3 y_{3t} + \cdots + \beta_n y_{nt} \tag{8}
\]

If for the presence of cointegration it is necessary that a combination of time series forms a stationary process, then in the presence of several time series, there can be several such combinations, and they speak of a set of cointegration vectors (9):

\[
B^T Y_t = \begin{pmatrix}
\beta_1^T Y_t \\
\vdots \\
\beta_f^T Y_t
\end{pmatrix}
= \begin{pmatrix}
u_{1t} \\
\vdots \\
u_{rt}
\end{pmatrix} \sim I \tag{9}
\]

One of Granger’s findings is that in the case of cointegration of I(1) series, the time series dependence model can be represented in the form of an error-correction model (ECM) [21]. That is, there is a certain functional dependence (10)

\[
\Delta y_{1t} = f(y_{1t-1} - \beta_2 y_{2t-1} + \beta_3 y_{3t-1} + \cdots + \beta_n y_{nt-1}) = f(u_{t-1}) \tag{10}
\]
On the one hand, we know that random component $u_t$ has zero mathematical expectation, and on the other hand, that the next change in the time series depends on the value of the random component at the previous step. Thus, if the processes are cointegrated, then they tend to the values that are laid down by the long-term equilibrium (8). Using this, one can monitor the value of random component $u_t$, and if its analysis shows that at some point this value began to behave neither like a stationary process, then one can suspect the occurrence of some changes in the processes, that is, note a violation of long-term equilibrium, which can mean some kind of abnormal situation in technology, breakdown, accident, the need to replace equipment, and so on.

There is a need to test time series for stationarity and cointegration.

As an example of using the method, we will use the ready-made Tennessee Eastman Process (TEP) data set, which is actively used by researchers to solve similar problems [2, 9, 11]. The dataset contains the data referenced in Rieth et al. The TEP was created by the Eastman Chemical Company to provide a realistic industrial process for evaluating process control and monitoring methods. The test process is based on a simulation of an actual industrial process where the components, kinetics, and operating conditions have been modified for proprietary reasons. The data in the training and testing sets included all the manipulated and measured variables for a total of $m = 52$ observation variables [22].

The process diagram is shown in the Figure.

![Figure 1. Diagram of Tennessee Eastman Process.](image)

Let’s consider a certain simplification for finding cointegration vectors. Considering the fact that if a linear combination of a pair of non-stationary time series forms a stationary time series, then such pairs can again form many other cointegration relations. Even looking only at pairs, we are faced with an increase in the number of possible options. For the above example with 52 variables, we get the following number of possible combinations (11)

$$C_n^k = \frac{n!}{k!(n-k)!}$$

That is, we have to check $C_{52}^2 = 1326$ pairs for cointegration. Having identified cointegrated pairs, we can follow them, and if cointegration has ‘disappeared’, that is, at some new time interval we can no longer talk about the presence of a cointegration ratio, then something has changed in the process itself. In practice, this means either a change in technology (which the operator knows about), or a breakdown/accident/failure, due to equipment errors, changes in some parameters of the resources used.
In the latter case, such information (that the process has changed) can be used to attract attention in general, which may ultimately lead to the need for equipment repair or maintenance or readjustment, etc. Changes in the pair will be estimated as the calculation of the root-mean-square deviation of value $u_t$.

3. Results

Obviously, the operator is not able to manually track a large number of parameters, and even more so all pairs. In our example, we would have to track 1326 parameter pairs.

We checked the dataset for cointegration in 1326 pairs using Python. To do this, we used the coint function.

Time series analysis shows that 75 pairs are cointegrated. Let’s look at some of them in Figures 2-9.

In Figure 1 (left), we see a graph of the values of two variables in normal operation (xmeas_33 and xmeas_40), the relationship between which is difficult to visually determine, but the analysis shows that a pair of variables is cointegrated. In Figures 2, 4, 6 and 8 (on the left), we see quite similar pictures of the initial values during normal operation, however, there are changes after an accident in the system (FaultNumber = 18) and the graphs have become completely different (right parts of the figures).

In Figures 3, 5, 7 and 9, we see diagrams of the range of root-mean-square deviations (RMSD) of the random component $u_t$ for the corresponding data pairs, for all variants of errors (1-20). Figures 2-9 and calculations were made in the R language (library tidyverse).

**Figure 2.** The graph of changes in variables xmeas_33 and xmv_5 during normal operation (left) and after error No. 18

**Figure 3.** Box plots of RMSD of random component $u_t$ for variables xmeas_33 and xmv_5 for different FaultNumbers

**Figure 4.** The graph of changes in variables xmeas_8 and xmv_2 during normal operation (left) and after error No. 18

**Figure 5.** Box plots of RMSD of random component $u_t$ for variables xmeas_8 and xmv_2 for different FaultNumbers
The box plots of root-mean-square deviations of random component $u_t$ for variables $x_{meas\_29}$ and $x_{meas\_41}$ for different FaultNumbers (Figure 5) show that for FaultNumbers 1, 7, 8, 12, 18 even the minimum values of the root-mean-square errors are higher than the maximum values during normal operation. Other pairs of data show similar results.

4. Conclusion
The analysis shows that the proposed method enables one to clearly illustrate situations with changes in behavior. As an example of using the method, we used the ready-made Tennessee Eastman Process (TEP) data set. Different pairs of data may have the ability to identify different errors. All errors cause a change in the behavior of one or several pairs of data, thus tracking the behavior of the value of random component $u_t$ enables identifying cases of deviation of the process from long-term equilibrium (in terms of cointegration), that is, cases of failure from the normal system operation.

The results obtained are clear and objective and can be used by process operators or by a source for automatic process control.

Acknowledgements
The work was carried out with the financial support by Kalashnikov Izhevsk State Technical University within the framework of the grant no. FRV/20-55-25.

References:
[1] Wildemann H 2018 Produktivität durch Industrie 4.0 (München: TCW Transfer-Centrum GmbH)
[2] Aldrich C 2019 Process Fault Diagnosis for Continuous Dynamic Systems Over Multivariate Time Series Time Series Analysis – Data, Methods, and Applications. URL:
https://www.intechopen.com/books/time-series-analysis-data-methods-and-applications/process-fault-diagnosis-for-continuous-dynamic-systems-over-multivariate-time-series

[3] Dong C, Xiuquan Q, Gelernter J, Xiaofeng L & Luoming M 2011 Mining data correlation from multi-faceted sensor data in the Internet of Things China Communications 8 1 132–138

[4] Momtazpour M, Zhang J, Rahman S, Sharma R & Ramakrishnan N 2015 Analyzing invariants in cyber-physical systems using latent factor regression Proc. of the 21th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (Sydney: NSW) pp 2009–18

[5] Lokhov A Y, Lemons N, McAndrew T C, Hagberg A & Backhaus S 2016 Detection of cyber-physical faults and intrusions from physical correlations The 16th Int. Conf. on Data Mining Workshops (Barcelona) pp 303–10

[6] Lavrova D & Pechenkin A 2015 Applying correlation and regression analysis to detect security incidents in the internet of things Int. J. of Communication Networks and Information Security 7 3 131

[7] Souza Goncalves E & Rosa P F F 2017 Sensor fusion with cointegration analysis for IMU in a simulated fixed-wing UAV The Int. Conf. on Military Technologies (Brno: Jihočeský kraj) pp 493–9

[8] Wang G, Liu C, Cui Y & Feng X 2014 Tool wear monitoring based on cointegration modelling of multisensory information Int. J. of Computer Integrated Manufacturing 27 5 479–87

[9] Li G, Qin J S & Yüan T 2014 Nonstationarity and cointegration tests for fault detection of dynamic processes IFAC Proc. 47 3 10616–21

[10] Dao P B, Staszewski W J, Barszcz T & Uhl T 2018 Condition monitoring and fault detection in wind turbines based on cointegration analysis of SCADA data Renewable Energy 116 107–22

[11] Chiang L H, Russell E L & Braatz R D 2000 Fault diagnosis in chemical processes using Fisher discriminant analysis, discriminant partial least squares, and principal component analysis Chemometrics and Intelligent Laboratory Systems 50 2 243–52

[12] Dunia R, Qin S J, Edgar T F & McAvoy T J 1996 Identification of faulty sensors using principal component analysis AIChE J. 42 10 2797–812

[13] Yu T, Wang X & Shami A 2017 Recursive principal component analysis-based data outlier detection and sensor data aggregation in IoT systems IEEE Internet of Things J. 4 6 2207–16

[14] Zhang D, Lindholm G & Ratnaweera H 2018 Use long short-term memory to enhance Internet of Things for combined sewer overflow monitoring J. of Hydrology 556 409–18

[15] Kotenko I, Saenko I, Skorik F & Bushuev S 2015 Neural network approach to forecast the state of the internet of things elements The XVIII Int. Conf. on soft computing and measurements (St. Petersburg) pp 133–5

[16] Mahdavinejad M S, Rezvan M, Barekatain M, Adibi P, Barnaghi P & Sheth A P 2018 Machine learning for Internet of Things data analysis: A survey Digital Communications and Networks 4 3 161–75

[17] Goh J, Adepu S, Tan M & Lee Z S 2017 Anomaly detection in cyber physical systems using recurrent neural networks The 18th Int. Symp. on High Assurance Systems Engineering (Singapore) pp 140–5

[18] Bai J & Perron P 1998 Estimating and Testing Linear Models with Multiple Structural Changes Econometrica 66 1 47

[19] Tsay R S 2005 Analysis of Financial Time Series (Wiley, New York) 576

[20] Zivot E & Wang J 2006 Modeling Financial Time Series with S-PLUS (New York: Springer) 1016

[21] Granger C W 2001 Essays in econometrics: collected papers of Clive WJ Granger (Vol. 32) (Cambridge: Cambridge University Press) 523
[22] Rieth C A, Amsel B D, Tran R & Cook M B 2017 Additional Tennessee Eastman Process Simulation Data for Anomaly Detection Evaluation (Cambridge: Harvard Dataverse)
[23] Hamilton J D 1994 Time Series Analysis (Princeton: Princeton Univ. Pr.) 816
[24] Johansen S 1988 Statistical Analysis of Cointegration Vectors J. of Economic Dynamics and Control 12 231–54
[25] Engle R F & Granger C W 1987 Co-integration and Error Correction: Representation, Estimation, and Testing Econometrica 55 251–76
[26] Dickey D A & Fuller W A 1979 Distribution of the Estimators for Autoregressive Time Series With a Unit Root J. of the American Statistical Association 366 74 427–31