Asymptotic Random Distortion Testing for Anomaly Detection

Dominique Pastor, Guillaume Ansel

1st IFSA Winter Conference on Automation, Robotics & Communications for Industry 4.0 (ARCI' 2021), 3-5 February 2021, Chamonix-Mont-Blanc, France

Oral Topic: Process Control and Monitoring

Asymptotic Random Distortion Testing for Anomaly Detection

Dominique Pastor, Guillaume Ansel
IMT Atlantique, Lab-STICC, UMR CNRS 6285, F-29238, France
Tel.: +33 229 00 11 11
E-mail: dominique.pastor@imt-atlantique.fr, guillaume.ansel@imt-atlantique.fr

Summary: In connection with cybersecurity issues in ICS, we consider the problem of detecting yet unknown attacks by presenting a theoretical framework for the detection of anomalies when the observations have unknown distributions. We illustrate the relevance of this framework with experimental results.

Keywords: Anomaly detection, industrial control systems (ICS), cybersecurity, industry 4.0, statistical hypothesis testing, random distortion testing.

1. Introduction

Vulnerability of systems in Industry 4.0 and Smart Factories increases as the number of new threats grows with the number of connected devices, especially in ICSs (Industrial Control Systems). It is crucial to devise methods capable of reliably detecting novel types of attacks. Anomaly detection [1, 2] addresses this issue by considering possibly novel attacks as anomalies with respect to nominal system behaviors. However, the diversity of processes and the various variable users' habits and behaviors entail many deviations with respect to nominal system behaviors, even in the absence of attacks. Because of this variability, anomaly detection may yield too many false alarms. It is thus desirable to cast the anomaly detection problem into a theoretical framework to deal with deviations around a nominal model, with guaranteed performance and even optimality.

Although statistical hypothesis testing provides a "*statistically justifiable solution for anomaly detection*" even "*in an unsupervised setting without any need for labeled training data*" [1] , they however assume that observations obey specific distributions, which is questionable in practice because of the aforementioned variations around nominal models, even in absence of attacks.

The Random Distortion Testing (RDT) [3] aims to overcome the aforementioned limitations of statistical methods and anomaly detection. This framework incorporates the existence of unknown deviations and deals with fully unknown probability distributions for the observation. The observation is assumed to result from some signal with unknown distribution observed in additive and independent Gaussian noise, and the RDT approach is optimal with guaranteed performance to decide whether the noisy signal drifts by too much from a certain deterministic model. Since RDT assumes a perfectly known noise distribution, we present below theoretical and experimental results to upgrade the original RDT framework to the White Gaussian Noise (WGN) case with estimated standard deviation.

Notation and terminology. $M(\Omega, \mathbb{R}^d)$ denotes the set of all d-dimensional real random vectors defined on the probability space $(\Omega, \Sigma, \mathbb{P})$. $Q_{d/2}$ is the generalized Marcum function. $\forall(\theta_0, \rho) \in \mathbb{R}^d \times (0, \infty)$, $S_\rho = \{y \in \mathbb{R}^d : \|y - \theta_0\|_2 = \rho\}$, $B_\rho = \{y \in \mathbb{R}^d : \|y - \theta_0\|_2 \leq \rho\}$ and $B_\rho^c = \{y \in \mathbb{R}^d : \|y - \theta_0\|_2 > \rho\}$. For all $t > 0$, we define the test $T_t : \mathbb{R}^d \to \{0, 1\}$ by:

$$\forall y \in \mathbb{R}^d, T_t(y) = \begin{cases} 1 & \text{if } \|y - \theta_0\|_2 > t \\ 0 & \text{otherwise} \end{cases}$$

2. The RDT appoach in the WGN case

The RDT problem can be stated as follows [3]:

$$\begin{cases} \textbf{Data model: } \exists (Y, \Theta) \in M(\Omega, \mathbb{R}^d)^2, \exists X \sim N(0, \sigma^2 \boldsymbol{I}_d), \\ \quad \begin{cases} \Theta \text{ and } X \text{ are independent,} \\ Y = \Theta + X, \\ \forall y \in \mathbb{R}^d, \exists \omega \in \Omega, y = Y(\omega) \end{cases} \\ \textbf{Testing problem: } \text{Given } y = Y(\omega), \text{ test:} \\ \quad H_0 : \Theta(\omega) \in B_\tau \quad \text{vs.} \quad H_1 : \Theta(\omega) \in B_\tau^C \\ \text{with } \tau > 0 \text{ and } \theta_0 \in \mathbb{R}^d \end{cases} \quad (1)$$

The existence of optimal tests for the RDT problem is established via the notions of *conditional size* and *conditional power* defined as follows.

Definition 1. *For $\rho > 0$ and $T : \mathbb{R}^d \to \{0, 1\}$, we set:*

Conditional power:

$$\forall \Theta \in M(\Omega, \mathbb{R}^d), \mathbb{P}\big[T(\Theta + X) = 1 \mid \Theta \in S_\rho\big]$$

Conditional size:

$$\alpha_T = \sup_{\Theta \in M(\Omega, \mathbb{R}^d) : P[\Theta \in B_\tau] \neq 0} \mathbb{P}[T(\Theta + X) = 1 \mid \Theta \in B_\tau]$$

Definition 2. *Given $\Theta \in M(\Omega, \mathbb{R}^d)$ and $\rho \geq 0$, a test T is said to have constant conditional power function (CCPf) given $\Theta \in S_\rho$ if for every $\theta \in S_\rho$:*

$$\mathbb{P}\big[T(\Theta + X) = 1 \mid \Theta \in S_\rho\big] = \mathbb{P}[T(\theta + X) = 1]$$

We can then exhibit optimal tests as follows.

Theorem 1. *Given $\gamma \in (0,1)$, if $\lambda_\gamma(\tau)$ is such that $Q_{d/2}\left(\tau, \lambda_\gamma(\tau)\right) = 1 - \gamma$, $T_{\lambda_\gamma(\tau)}$ is optimal in that:*

i. $\alpha_{T_{\lambda_\gamma(\tau)}} = \gamma$;

ii. *$\forall \Theta \in M(\Omega, \mathbb{R}^d)$ and for $\mathbb{P}(\|\Theta - \theta_0\|_2)^{-1}$-almost every $\rho > \tau$, $T_{\lambda_\gamma(\tau)}$ has CCPf given $\Theta \in S_\rho$ and for all T with $\alpha_T \leq \gamma$ and CCPf given $\Theta \in S_\rho$:*

$$\mathbb{P}\left[T_{\lambda_\gamma(\tau)}(\Theta+X)=1 \mid \Theta\in S_\rho\right] \geq \mathbb{P}\left[T(\Theta+X)=1 \mid \Theta\in S_\rho\right]$$

3. Asymptotic RDT

For any $t>0$, let $\tilde{T}_t$ be the function defined by:

$$\begin{aligned}\tilde{T}_t:\mathbb{R}^d\times\mathbb{R}^d\times(0.\infty) &\rightarrow \{0,1\}\\ (y,\theta,\sigma) &\mapsto \begin{cases}1 & \text{if } \|y-\theta\|_2>\sigma t\\ 0 & \text{otherwise}\end{cases}\end{aligned}$$

We herafter consider consistent estimators $\hat{\theta}_n$ and $\hat{\sigma}_n$ of θ_0 and σ_0 respectively. Set $Z=(Y,\theta_0,\sigma_0)$ and $\forall n\in\mathbb{N}, Z_n=(Y,\hat{\theta}_n,\hat{\sigma}_n)$. $\forall\rho\geq 0$ and $\forall n\in\mathbb{N}$, define $\Pi_n(\rho,\cdot)$ and $\Pi(\rho,\cdot)$ by setting for all Borel set $A\subset\mathbb{R}^{2d+1}$:

$$\begin{aligned}\Pi_n(\rho,A) &= \mathbb{P}[Z_n\in A \mid \|\Theta-\theta_0\|_2=\rho]\\ \Pi(\rho,A) &= \mathbb{P}[Z\in A \mid \|\Theta-\theta_0\|_2=\rho]\end{aligned}$$

Theorem 2 (Asymptotic level)**.** *If* $S=\{\Xi\in M(\Omega,\mathbb{R}^d):\forall n\in\mathbb{N},\Xi$ *and* $(\hat{\theta}_N,\hat{\sigma}_N)$ *are independent*$\}$, *then:*

$$\limsup_n \sup_{\Xi\in S:\mathbb{P}[\Xi\in B_\tau]\neq 0}\mathbb{P}\left[\tilde{T}_{\lambda_\gamma(\tau)}(\Xi+X,\hat{\theta}_N,\hat{\sigma}_N)=1 \mid \Xi\in B_\tau\right]\leq\gamma$$

Define the critical region of $\tilde{T}:\mathbb{R}^{2d+1}\rightarrow\{0,1\}$ as: $K_{\tilde{T}}=\{(y,\theta,\sigma)\in\mathbb{R}^d\times\mathbb{R}^d\times(0,\infty):\tilde{T}(y,\theta,\sigma)=1\}$

Theorem 3. *If* $\tilde{T}:\mathbb{R}^{2d+1}\rightarrow\{0,1\}$ *is such that* $\tilde{T}(\cdot,\theta_0,\sigma_0)$ *has asymptotic level* γ *and constant conditional power function given* $\|\Theta-\theta_0\|_2=\rho$ *for* $\mathbb{P}\|\Theta-\theta_0\|_2^{-1}$*-almost every* ρ *and if the critical region* $K_{\tilde{T}}$ *is a* $\mathbb{P}Z^{-1}$*-continuity set, then:*

$$\limsup_n\left(\Pi_n\left(\rho,K_{\tilde{T}_{\lambda_\gamma(\tau)}}\right)-\Pi_n(\rho,K_{\tilde{T}})\right)\geq 0 \qquad (2)$$

4. Experimental results for signal detection in non-asymptotic regimes

Consider the detection problem (3), where $\Delta\in M(\Omega,\mathbb{R}^d)$ is a bounded random interference and assume that we have a consistent estimator of σ. The standard Neyman-Pearson test and the GLRT cannot be used since the distribution of Δ is unknown. We can cast (3) in (1) with $\Theta=\varepsilon\theta_0+\Delta$ and thus use $\tilde{T}_{\lambda_\gamma(\tau)}$ to perform the decision. The false alarm rate (FAR) of $\tilde{T}_{\lambda_\gamma(\tau)}$ tends to γ by upper values as N increases.

$$\begin{cases}\textbf{Observation: } Y=\varepsilon\theta_0+\Delta+X, \text{ where:}\\ \quad\begin{cases}\varepsilon\in\{0,1\} \text{ is unknown}\\ \theta_0\in\mathbb{R}^d \text{ is known, } X\sim N(0,\sigma^2 I_d)\\ \Delta\in M(\Omega,\mathbb{R}^d) \text{ has unknown distribution}\\ \|\Delta\|_2\leq\tau \text{ for a known } \tau\geq 0\\ \Delta \text{ and } X \text{ are independent}\end{cases}\\ \textbf{Testing problem: } \text{given } y=Y(\omega), \text{ test:}\\ \qquad H_0:\varepsilon=0 \quad \text{vs.} \quad H_1:\varepsilon=1\end{cases} \qquad (3)$$

It is then possible to determine τ^* so that $\tilde{T}_{\lambda_\gamma(\tau^*)}$ maintains the FAR below γ for the detection problem (3). This adjustment is achieved as follows. Consider a uniform distribution for Δ on S_τ because this distribution is the least favorable in that it maximizes the FAR. Seek the value $k=1,2,\ldots$ such that the FAR of $\tilde{T}_{\lambda_\gamma(k\tau)}$ drops below γ. For the value of k thus found, fix $\tau_l=(k-1)\tau$ and $\tau_u=k\tau$. Then, by dichotomy, calculate $\tau^*\in(\tau_l,\tau_u)$ such that the FAR of $\tilde{T}_{\lambda_\gamma(\tau^*)}$ approximates at best γ without exceeding this level. In our experiments below, the dichotomy was stopped after 10 steps. The Monte-Carlo simulations were carried out with the following parameters: $d=2$; $\sigma=1$; $\tau=1.77$, which corresponds to a Distortion-to-Noise ratio of 2 dB; $\|\theta_0\|_2=5.62$, which corresponds to an SNR of 12 dB; $\hat{\sigma}$ was the Maximum Likelihood Estimate from N iid standard Gaussian samples, with $N=\{20,50,100\}$. The ROC curves of $\tilde{T}_{\lambda_\gamma(\tau)}$ in Fig. 1 exhibit two important features. First, they are identical, which can be proved mathematically. Second, although $\tilde{T}_{\lambda_\gamma(\tau^*)}$ is not optimal in the sense of Theorem 3, it maintains the FAR below γ without much performance loss in comparison with the optimal $T_{\lambda_\gamma(\tau)}$ that requires a known σ.

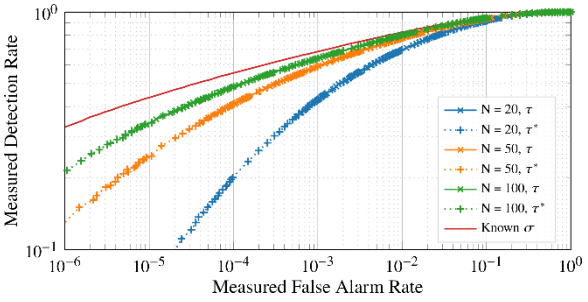

Fig. 1. ROC curves of $\tilde{T}_{\lambda_\gamma(\tau^*)}$ and $\tilde{T}_{\lambda_\gamma(\tau)}$

5. Conclusion and perspectives

In this paper, with respect to current issues in cybersecurity and ICS, we have presented Asymptotic RDT, which extends the initial RDT approach so as to take estimation of the model and the noise variance into account. We have illustrated the relevance of the approach through simulations and presented a way to compensate the effect of the estimation when detection is performed via Asymptotic RDT. The approach is very promising due to its genericity. Applications to change detection are in progress. Future works involve extension to the case of an unknown noise covariance matrix and to noise distributions other than Gaussian.

References

[1] Varun Chandola, Arindam Banerjee, and Vipin Kumar. "Anomaly Detection: A Survey". In: ACM Computing Surveys 41.3 (July 1, 2009), pp. 1–58. issn: 03600300. doi: 10.1145/1541880.1541882.

[2] Marco A.F. Pimentel et al. "A Review of Novelty Detection". In: Signal Processing 99 (June 2014), pp. 215–249. issn: 01651684. doi: 10.1016/j.sigpro.2013.12.026.

[3] Dominique Pastor and Quang-Thang Nguyen. "Random Distortion Testing and Optimality of Thresholding Tests". In: IEEE Transactions on Signal Processing 61.16 (Aug. 2013), pp. 4161–4171. issn: 1053-587X. doi: 10.1109/TSP.2013.2265680.