Detecting correlation between server resources for system management

Stefania Tosi[a,*], Sara Casolari[a], Michele Colajanni[a]

[a] *University of Modena and Reggio Emilia*
Department of Information Engineering

Abstract

Efficient system management requires a continuous knowledge about the state of system and application resources that are typically represented through time series obtained by monitors. Capacity planning studies, forecasting, state aggregation, anomaly and event detection would be facilitated by evidence of data correlations.

Unfortunately, the high variability characterizing most time series related to system resources affects the accuracy and robustness of existing correlation solutions. This paper proposes an innovative approach that is especially tailored to detect linear and non-linear correlation between time series that are intrinsically characterized by high variability.

We compare the proposed solution and existing algorithms in terms of accuracy and robustness for several synthetic and real settings characterized by low and high variability, linear and non-linear correlation. The results show that our proposal guarantees analogous performance for low variable time series, and improves state of the art in finding correlations in highly variable domains that are of interest for the application context.

Keywords: Correlation model, data analysis, high variability

*Corresponding author
Email addresses: stefania.tosi@unimore.it (Stefania Tosi), sara.casolari@unimore.it (Sara Casolari), michele.colajanni@unimore.it (Michele Colajanni)
[1]Corresponding author contact details: Via Vignolese 905, 41125 Modena, Italy
Tel.: +39 0592056273, Fax: +39 0592056129

1. Introduction

Correlation models are applied to various scientific fields as bases for several statistical analyses, such as forecasting [1], state aggregation [2], anomaly [3, 4] and event detection [5]. For computer systems management, capturing the correlation between time series monitored in a system allows us to discover groups of resources with similar behavior, to reflect changing relationships and, as a global effect, to manage the system in a more efficient way.

There are several correlation models showing robust and accurate results when applied to low variable domains. However, they fail in finding correlation between time series collected from system monitors, where relationships between data are often hidden by high variability. In these contexts, the most popular models, such as the Pearson coefficient [6], the Spearman rank [8], the Kendall rank [9], and the Local Correlation index [5], are unable to capture correlations even when they exist. The approach of filtering highly variable time series and then applying correlation models does not solve the problem and opens other issues.

We propose a new correlation model that is able to capture both linear and non-linear dependences even among time series characterized by high variability. The accuracy and robustness of the proposed solution is achieved through an original approach that separates trend patterns from perturbation patterns, and evaluates correlation by computing the similarity of trend patterns because they reveal the way in which a time series may depend on another one. From this perspective, the idea we pursue is based on the following steps:

1. we extract from time series the information about their trend patterns by removing the perturbations that mask the presence of possible relationships between data;
2. we measure how close the extracted trend patterns of two time series are, thus capturing the presence of linear dependency and of non-linear dependency.

The proposed method represents a step ahead of the literature because the Pearson correlation moment does not cover non-linear dependency [5], the Spearman and Kendall ranks are conditioned by the data distributions [6]. The LoCo model would be able to capture some non-linear relationships, but it infers that the main patterns of a time series are only trend patterns without considering the impact of perturbations. However, system and software resources are characterized by several perturbations due to I/O operations, synchronizations, context switching. We extend the correlation analysis and its applications to domains that are characterized by perturbations and highly variable time series [10, 11]. Time series related to system resources may have linear and non-linear relationships. For example, in Internet services, network traffic volume and service times change in accordance with the volume of user requests [12]. These relationships are typically masked by perturbations intrinsically related to the nature of the applications, but when correlations exist our model is able to identify them.

We illustrate the proposed method applied to real and synthetic time series. We discuss its quantitative and qualitative interpretation, compare it against existing solutions and demonstrate its accuracy and robustness. Moreover, we evaluate that the proposed correlation model is robust and accurate even as a support to different applications, such as tracking analysis, anomaly detection, and prediction analysis.

The remainder of this paper is organized as follows. Section 2 defines the problem of correlation when time series deriving from monitored system resources are characterized by high variability. Section 3 presents the proposed correlation model. Section 4 evaluates the accuracy and robustness of the model for several time series characterized by various statistical properties, and compares its results against those achieved by existing correlation models. Section 5 assesses the complexity of the proposal and of state-of-the-art alternatives. Section 6 analyzes the results of the proposed model when applied to tracking analysis, to anomaly detection, and to forecasting problems. Section 7 compares our contribution with respect to the state of the art. Section 8 concludes

the paper with some final remarks.

2. Problem definition

We consider data obtained from system and software monitors through the periodic sampling of resource measures. As evidenced in many works [10, 13, 11], these measures are extremely variable even at different time scales. We transform these samples in time series. We denote the two considered time series as $\mathbf{x} \equiv [x_1, \ldots, x_n]$ and $\mathbf{y} \equiv [y_1, \ldots, y_n]$, where each one is a vector collecting a time-ordered discrete sequence of data points that can be sampled once.

Existing correlation models do not work on time series exhibiting a high degree of variability, hence we are interested to propose a new *correlation index*, typically denoted as ρ, measuring the similarity between $\mathbf{x}$ and $\mathbf{y}$, as in [14]. The absolute value of the correlation index ranges between 0 and 1. When $\rho = 0$, there is no relationship between the two time series, while $\rho = 1$ indicates a complete correlation between $\mathbf{x}$ and $\mathbf{y}$. The literature offers several guidelines for the best interpretation of the value of the correlation index [14, 6, 7], but all criteria depend on the context and purposes of the analysis. Providing rules for interpreting the meaning of correlation indexes is out of the scope of this paper. In our context, we decide to adopt the common interpretation indicating a *strong correlation* when $\rho > 0.5$, and a *weak correlation* for $\rho \leq 0.5$ (e.g., [14]). Different choices for the threshold may lead to different results but do not impact the main conclusions of this paper.

Many models for capturing correlation are robust and accurate (e.g., [6, 5, 8, 9]) except when they are applied to time series characterized by high variability that is typical of system resource metrics. In accordance with [15], high variability is a phenomenon by which a set of observations takes values that vary over orders of magnitude, with most observations taking values around the time series trend (i.e., *trend pattern*), and some observations departing from it with appreciable frequency, even taking extremely large values with non-

negligible probability (i.e., *perturbation pattern*). Trend patterns represent the tendency of a time series that may be related to the other time series, while perturbation patterns consist of random observations hiding trends. A high standard deviation is the most typical trademark of a highly variable time series. This characteristic implies a trend pattern that is hard to identify because it is masked by perturbations. In this paper, we use standard deviation as a measure of data variability.

The ability of a correlation model in detecting dependency among correlated time series is measured in terms of *accuracy*, while the ability in guaranteeing a stable correlation index when conditions do not change is measured in terms of *robustness*. Accuracy measures how close the correlation index is to the effective level of correlation between two time series, while robustness evaluates the variability of the correlation index among different evaluations carried on under unchanged conditions.

In the case of highly variable time series, the most popular correlation models are affected by two main problems:

1. low accuracy, since they are unable to detect linear and non-linear dependences even among correlated time series;
2. low robustness, since they do not guarantee a stable evaluation of the correlation index, even when the relationships between the time series do not change.

Let us give an example of the above problems. We refer to datasets coming from the resource monitoring of a multi-tier system, whose architecture is illustrated in Figure 1. The four application servers are deployed through the Tomcat servlet container and are connected to two MySQL database servers. The Web switch node, running a modified version of the Apache Web server as in [10], assigns the same amount of requests to the two Apache-based HTTP servers, that run identical applications on identical hardware.

In Figure 2(a), we report 1050 data of the CPU utilization of the two Web servers sampled every 5 minutes. The data are highly variable, although there

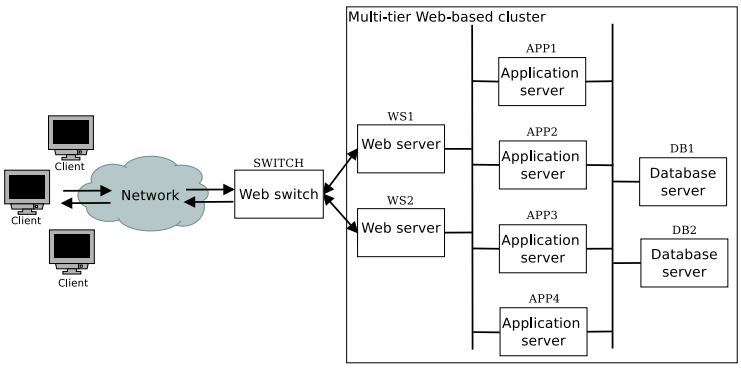
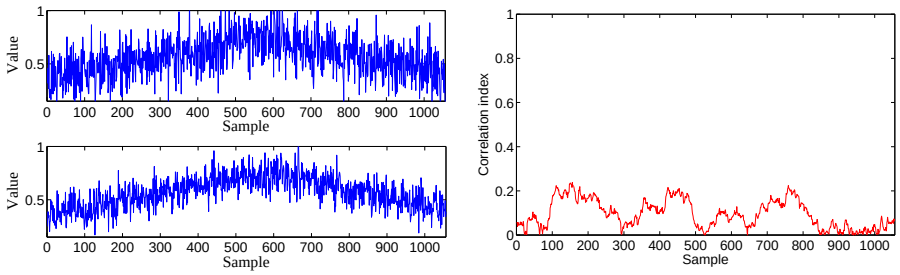


Figure 1: Architecture of a multi-tier Web cluster.

is visual evidence of a high correlation between the two time series. At system level, this correlation is confirmed by the fact that the two hosts receive an analogous amount of requests assigned by the front-end load balancer dispatcher. We evaluate the correlation index between the two time series through the Pearson model [6]. (The Pearson model is used as an example, but other existing models do not change the results). The results of the correlation index are reported in Figure 2(b). Despite the clear relationship between the datasets referring to the two hosts, Figure 2(b) shows that the Pearson model is affected by the two anticipated problems: its results are characterized by low accuracy because the Pearson correlation index remains lower than 0.25 during the entire interval of observation; its results are characterized by low robustness because the correlation index presents marked oscillations even when the correlation between the times series does not change.

A typical approach addressing issues related to highly variable time series is to refer to some rectification algorithms (e.g., [16]). Unfortunately, we will see that filtering data and then applying an existing correlation model does not work and, even worse, opens other issues by moving the problem to the dimension of finding the "right" filter and its parameters. This paper cannot discuss all details of the extensively investigated and hard-to-solve problems related to filters. Interested readers can refer to the huge literature on this field (e.g.,



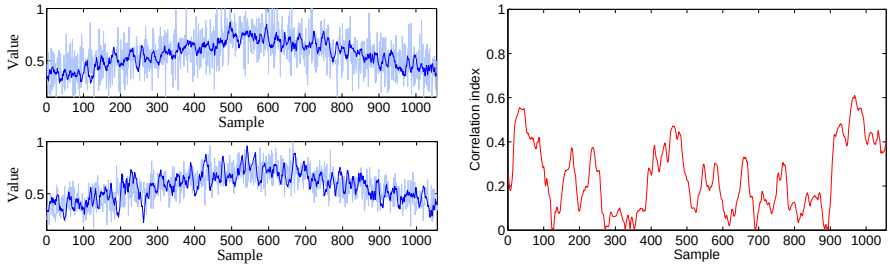
(a) CPU utilization samples **(b) Pearson correlation index**

Figure 2: Result of correlation models applied to highly variable data.

[16, 10, 17]). Basically, we have two alternatives that are equally useless: to apply a weak or a strong filter. A weak filter removes small variability, hence underlying relationships among time series remain undetectable by existing correlation models; a strong filter may remove important information about trend patterns and thus prevent the possibility of finding correlations.

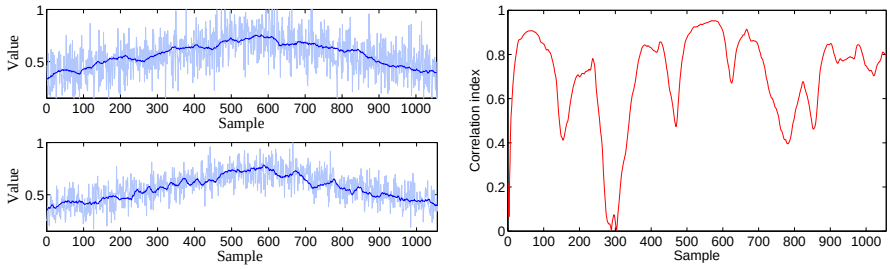
As example, let us consider again the time series represented in Figure 2(a). We choose a popular filter such as the Exponential Weighted Moving Average (EWMA) [16] as a basis, and apply different parameters in order to obtain a weak and a strong filter. The filtered time series are shown in Figure 3(a) and Figure 4(a), respectively. In Figure 3(b), we report the results of the Pearson model applied to the weakly filtered data. As expected, the results are improved with respect to the not filtered case, but the conclusion remains unchanged: the two time series are considered uncorrelated because the correlation index remains lower than 0.5 for almost the entire interval of observation. By increasing the filter strength, most perturbations are discarded, as shown in Figure 4(b). The problem is that a strong filter cancels also information about trend patterns characterizing the time series. The final result is that the correlation index becomes even less robust as it continuously passes from evaluating the time series as correlated, then uncorrelated, correlated again, and so on.

All these results evidence that high variability represents a limit of existing correlation models even when they are integrated with filtering techniques. De-



(a) Weak filtered CPU utilization **(b) Pearson correlation index**

Figure 3: Weak filtering applied before correlation analysis.



(a) Strong filtered CPU utilization **(b) Pearson correlation index**

Figure 4: Strong filtering applied before correlation analysis.

tecting correlations among highly variable data as those obtained by resource monitoring requires some novel approach. The model proposed in the following section has several benefits: it does not require any assumption about statistical properties, any pre-analysis of time series characteristics, and any integration with pre-filtering techniques; it is able to adapt its parameters to data characteristics and to capture linear and non-linear dependences.

3. Correlation model for highly variable time series

In this section, we present a novel correlation model, namely *CoHiVa* (*Co*rrelation for *Hi*ghly *Va*riable data), that is able to evaluate similarity between time series that are characterized by high variability. This model may be viewed as an improvement of the LoCo algorithm [5] that proposes the evaluation of correlation through the analysis of pattern similarity. CoHiVa extends

this idea to the correlation analysis in highly variable domains. Unlike the Pearson model [6] that works well only if time series are linked by a linear relationship, CoHiVa does not assume any data dependency and it is appropriate to discover both linear and non-linear dependencies. Moreover, CoHiVa does not assume any data distribution, as required by the Spearman and Kendall models [8, 9].

The CoHiVa algorithm is based on the following four main steps:

1. we extract from $\mathbf{x}$ and $\mathbf{y}$ the trend patterns and the perturbation patterns;
2. we remove errors contaminating the time series;
3. we select the trend patterns by discarding the perturbation patterns containing highly variable information;
4. we compute the CoHiVa correlation index between the two time series by evaluating the similarity between their trend patterns.

Each step is detailed in the following subsections.

3.1. Pattern extraction

The first goal is to identify the main patterns that are present in the time series $\mathbf{x} \equiv [x_1, \ldots, x_n]$, where patterns correspond to trends (i.e., periodic and seasonal components) and perturbations. To this end, we apply the Singular Value Decomposition (SVD) [5] to the auto-covariance matrix of the time series. Among the spectral decomposition algorithms, SVD is considered as the baseline technique for separating existing patterns without any assumption about the statistical characteristics of the data [18, 19]. In practice, we estimate the full auto-covariance matrix of the time series $\mathbf{x}$ that is defined as:

$$\mathbf{\Phi}(x) = \mathbf{x} \otimes \mathbf{x}, \tag{1}$$

where $\mathbf{\Phi}(x)$ is the auto-covariance matrix of $\mathbf{x}$.

Then, we compute the SVD of the auto-covariance matrix $\mathbf{\Phi}(x)$ as follows:

$$\mathbf{\Phi}(x) = \mathbf{U}(x)\mathbf{\Sigma}(x)\mathbf{V}(x)^T, \tag{2}$$

where $\mathbf{U}(x)$, $\mathbf{\Sigma}(x)$ and $\mathbf{V}(x) \in \mathbb{R}^{n \times n}$.

The columns $\mathbf{v}_i$ of $\mathbf{V}(x) \equiv [\mathbf{v}_1, \ldots, \mathbf{v}_n]$ are the right singular vectors of $\mathbf{\Phi}(x)$. Similarly, the columns $\mathbf{u}_i$ of $\mathbf{U}(x) \equiv [\mathbf{u}_1, \ldots, \mathbf{u}_n]$ are the left singular vectors of $\mathbf{\Phi}(x)$. Finally, $\mathbf{\Sigma}(x) \equiv \text{diag}[s_1, \ldots, s_n]$ is a diagonal matrix with positive values s_i, called the singular values of $\mathbf{\Phi}(x)$.

3.2. Removing errors

The singular vectors corresponding to singular values often contain some approximation errors [21] that usually contaminate the measured variables. The contribution of these errors must be discarded by eliminating the singular vectors corresponding to the smallest singular values [22]. By retaining just the principal vectors corresponding to the highest k singular values ($k < n$) we can reconstruct a *k-dimensional approximation* of the correlation matrix:

$$\bar{\mathbf{\Phi}}(x) \equiv \bar{\mathbf{U}}(x)\bar{\mathbf{\Sigma}}(x)\bar{\mathbf{V}}(x)^T, \tag{3}$$

where $\bar{\mathbf{U}}(x) \equiv [\bar{\mathbf{u}}_1, \ldots, \bar{\mathbf{u}}_k]$, $\bar{\mathbf{V}}(x) \equiv [\bar{\mathbf{v}}_1, \ldots, \bar{\mathbf{v}}_k]$ and $\bar{\mathbf{\Sigma}}(x) \equiv \text{diag}\ [\bar{s}_1, \ldots, \bar{s}_k]$.

Literature on SVD gives little importance to the problem of dynamically selecting the appropriate number of principal vectors that capture the patterns (e.g., [5, 18]). A common approach is to choose a fixed number of principal vectors independently of data characteristics, but this choice is unsuitable to time-varying contexts where the statistical properties of data continuously change in time. Hence, we choose a threshold-based method that takes into account the characteristics of considered data [23]. We select the principal vectors contributing to 90% of variation, and discard singular vectors contributing for less than 10%. This number of principal vectors capturing the main patterns of the time series is variable and dependent to the amount of variation in data. The variable number of principal vectors capturing the main patterns of the time series is denoted by k.

3.3. Selection of trend patterns

We now analyze the main patterns of $\mathbf{x}$ in order to understand the information they carry. The goal is to retain just trend patterns but, in the context of

interest for this paper, we can expect that the main patterns may include also some perturbation patterns [24]. As these last patterns prevent the identification of correlation between time series, we have to discard them. The idea is to build a new matrix that is based on a subspace of the $\hat{k} \leq k$ principal vectors that capture trend patterns.

To remove perturbation patterns from $\bar{\mathbf{U}}(x)$, we compute the Hurst exponent H of the k principal vectors by means of the R/S analysis of Hurst [25]. The Hurst exponent measures whether the data have pure random variability or are characterized even by some underlying trends [26]. For each principal vector $\bar{\mathbf{u}}_i \in \bar{\mathbf{U}}(x)$ of length n, we first define its cumulative deviate series:

$$Z_t = \sum_{j=1}^{t} (\bar{u}_j - m), \tag{4}$$

where $t = 1, 2, \ldots, n$, $\bar{u}_j$ is the $j-$th element of the $i-$th principal vector and m is the mean of $\bar{\mathbf{u}}_i$.

To define the rescaled range $\frac{R(n)}{S(n)}$ of Hurst, we compute the range as:

$$R(n) = \max\left(Z_1, Z_2, \ldots, Z_n\right) - \min\left(Z_1, Z_2, \ldots, Z_n\right), \tag{5}$$

and the standard deviation as:

$$S(n) = \sqrt{\frac{1}{n} \sum_{j=1}^{n} (\bar{u}_j - m)^2}. \tag{6}$$

The *Hurst exponent* H is defined in terms of the asymptotic behavior of the rescaled range as a function of the time span of a time series as follows [25]:

$$\mathrm{E}\left[\frac{R(n)}{S(n)}\right] = Cn^H \quad \text{as } n \to \infty, \tag{7}$$

where $E[\frac{R(n)}{S(n)}]$ is the expected value of the rescaled range, n is the number of observations in a time series, and C is a constant.

If the estimated Hurst exponent H_i of a principal vector $\bar{\mathbf{u}}_i \in \bar{\mathbf{U}}(x)$ is close to 0.5, then we can conclude that $\bar{\mathbf{u}}_i$ contains perturbations. On the other hand, if the Hurst exponent remains far from 0.5, then we can assume that $\bar{\mathbf{u}}_i$ is a trend principal vector.

Our model builds up a new $\hat{\mathbf{U}}(x)$ matrix containing only the $\hat{k}$ trend principal vectors $\hat{\mathbf{u}}_j$, $j = 1, \ldots, \hat{k}$ as follows:

$\forall \bar{\mathbf{u}}_i \in \bar{\mathbf{U}}(x)$, $i = 1, \ldots, k$:

$$\hat{\mathbf{u}}_j \equiv \bar{\mathbf{u}}_i \quad \text{if} \quad H_i < 0.5 - \frac{\delta}{2} \quad \text{or} \quad H_i > 0.5 + \frac{\delta}{2}, \tag{8}$$

where δ is a two-sided 95% confidence interval empirically computed depending on the number of samples as in [26].

This separation approach allows us to remove perturbation patterns in the time series and to focus only on trend patterns. By focusing on the $\hat{k} \ll n$ trend patterns, we are able to construct a new approximation of the $\mathbf{\Phi}(x)$ matrix that we name *trend approximation*. Given $\hat{\mathbf{U}}(x) \equiv [\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_{\hat{k}}]$, the corresponding singular values and the right singular vectors form the matrices $\hat{\mathbf{\Sigma}}(x) \equiv \text{diag}\,[\hat{s}_1, \ldots, \hat{s}_{\hat{k}}]$ and $\hat{\mathbf{V}}(x) \equiv [\hat{\mathbf{v}}_1, \ldots, \hat{\mathbf{v}}_{\hat{k}}]$, respectively. Through these matrices, the trend approximation of the correlation matrix using only trend patterns is given by:

$$\hat{\mathbf{\Phi}}(x) \equiv \hat{\mathbf{U}}(x)\hat{\mathbf{\Sigma}}(x)\hat{\mathbf{V}}(x)^T. \tag{9}$$

The matrix $\hat{\mathbf{\Phi}}(x)$ approximates the trend behavior of $\mathbf{\Phi}(x)$ by removing both error information and perturbation patterns that affect the identification of correlations. This approach retains trend information, as well as seasonal and oscillatory behavior in the time series.

3.4. Computation of the CoHiVa index

After the extraction of the main trends, we can evaluate whether they are correlated or not by computing how close their trend patterns are. As example, we compute the correlation index of the time series $\mathbf{x}$ and $\mathbf{y}$ by measuring their trend similarity. When the matrices $\hat{\boldsymbol{U}}(x)$ and $\hat{\boldsymbol{U}}(y)$ are similar, the time series $\mathbf{x}$ and $\mathbf{y}$ follow similar (linear or non-linear) trends, and we can consider that the original time series are correlated. In geometric terms, if two time series are correlated, then the trend principal vectors of each time series should lie within the subspace spanned by the trend principal vectors of the other time series.

For this reason, we project the trend principal vectors of the time series $\mathbf{x}$ into the trend principal vectors of $\mathbf{y}$, as following:

$$\rho(\mathbf{x}, \mathbf{y}) = \frac{1}{\hat{k}(x) + \hat{k}(y)} (\|\hat{\mathbf{U}}(x)^T \hat{\mathbf{U}}(y)\| + \|\hat{\mathbf{U}}(y)^T \hat{\mathbf{U}}(x)\|), \tag{10}$$

where $\hat{k}(x)$ and $\hat{k}(y)$ are the numbers of trend principal vectors of $\mathbf{\Phi}(x)$ and $\mathbf{\Phi}(y)$, respectively, while $\hat{\mathbf{U}}(x)$ and $\hat{\mathbf{U}}(y)$ are the trend principal vectors matrices collecting them.

4. Performance evaluation

We evaluate the performance of the proposed correlation model, and we compare it against the results of several state-of-the-art alternatives. As terms of comparison, we consider the following correlation models: the Pearson product moment (Pearson) [6], the Spearman rank (Spearman) [8], the Kendall rank (Kendall) [9], and the Local Correlation index (LoCo) [5]. Moreover, we consider the performance of a model that is integrated with a pre-filtering technique, namely Pearson with filtering.

To evaluate the accuracy and robustness of all models, we initially refer to synthetic time series that allow us to have full control over their actual degree of correlation. Then, in the following section we will consider real time series.

In Section **??**, we first evaluate the impact of linear and non-linear correlations on the models performance. Then, in Section 4.2 we evaluate how the performance of the correlation models changes in case of time series affected by different levels of variability.

4.1. ***Linear and non-linear correlations***

Here, we consider three types of time series: correlated with linear dependence, correlated with non-linear dependence, not correlated. The time series of each scenario take values in the range $[0, 1]$. In order to evaluate the ability of the correlation models to capture different types of dependency for different levels of variability, we introduce perturbations from $N(0, \sigma)$, where

$\sigma \in \{0.01, 0.05, 0.1, \ldots, 0.5\}$ is the standard deviation that quantifies the intensity of perturbations added to data [27, 28].

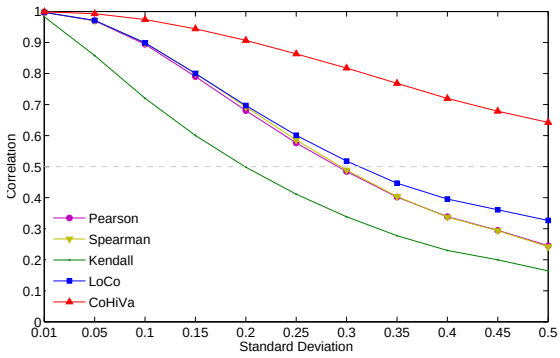
The performance of the correlation models is evaluated in terms of accuracy (Section 4.1.1) and robustness (Section 4.1.2) over 1000 independent generations of data for each σ value in each scenario.

4.1.1. Accuracy

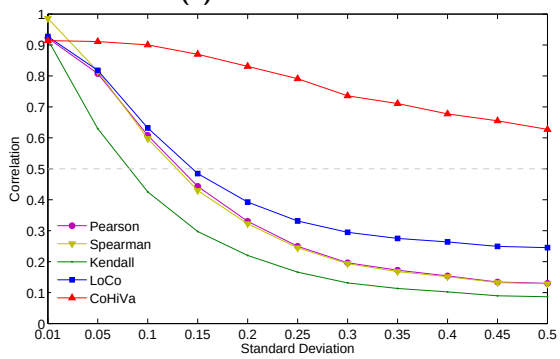
We define the *accuracy* of a correlation model as its ability to capture correlation when data present some linear or non-linear relationships, and in categorizing as not correlated time series having no dependence. For example, an accurate model should obtain a correlation index close to 1 in both linear and non-linear scenarios, and an index close to 0 in the uncorrelated scenario.

The first set of experiments evaluates the accuracy of the correlation models when time series are characterized by different intensities of perturbations. In Figure 5, we report the mean correlation index of all considered models computed over 1000 generations of correlated data with different σ values. We remind the reader that we consider a strong correlation when $\rho > 0.5$, and a weak correlation for $\rho \leq 0.5$ [14]. The results of Figure 5(a) refer to the linear scenario. As expected, we see a decrease of all correlation indexes for increasing values of σ, but the impact of perturbations is different for the considered models. When the dispersion is low ($\sigma \leq 0.2$), all models are able to capture the strong correlation among data. When the dispersion increases ($\sigma > 0.2$), the Kendall model is the first to lose its ability to detect data correlation. In higher variable contexts ($\sigma > 0.3$), only the CoHiVa model captures the strong data correlation for each variability level, because its index is always higher than 0.65.

The accuracy of all the models deteriorates when we pass to a scenario where the correlation between data is non-linear. A comparison between Figure 5(a) and Figure 5(b) gives a first idea about the overall results. Only the CoHiVa model is able to detect a strong correlation for any σ when the relationship between data is non-linear. On the other hand, all existing models are affected



(a) Linear scenario



(b) Non-linear scenario

Figure 5: Analysis of accuracy of correlation models without data filtering.

by a low accuracy for increasing values of σ. (They estimate a weak correlation even when data are perturbed by very low levels of dispersion, such as $\sigma = 0.15$.) It is interesting to observe that the Spearman rank, which is expressly oriented to capture non-linear dependencies [8], exhibits the best accuracy when the dispersion is very low (that is, $\sigma < 0.05$), but it loses its capacity as soon as the time series are characterized by higher perturbations.

To address issues related to high variability, the state-of-the-art models may increase their accuracy by working on a filtered representation of the original time series. We anticipated in Section 2 that this approach does not work, but for the sake of an exhaustive comparison we compare the performance of CoHiVa against a Pearson model combined with a pre-filtering technique. We

have to specify that the choice of the best filtering model and of its parameters is a serious issue by itself, and is out of the scope of this paper. We integrate the Pearson correlation model with an EWMA filter that we have experimentally evaluated as giving good results. We do not claim that we are applying an optimal filter with optimal parameter setting, even because the definition of optimum is improper in this context.

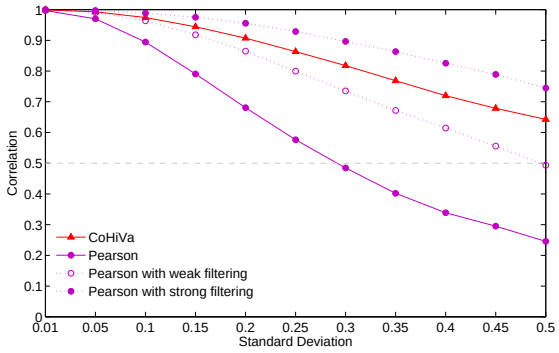
Figure 6 shows the results obtained by applying the Pearson model to data filtered through a weak and a strong filter. If we compare the results of Pearson without filter to the results of Pearson with any kind of filtering, we can appreciate that the filter in fact improves accuracy: the correlation value is higher for every σ and for linear and non-linear scenarios. In the linear scenario shown in Figure 6(a), the Pearson model with filtering is able to detect a correlation index higher than 0.5 for any σ. On the other hand, if we consider the non-linear scenario shown in Figure 6(b), there is an evident decrease of the correlation index. Both weak and strong filtering are useless because they estimate a $\rho \leq 0.5$ when $\sigma > 0.2$ and $\sigma > 0.3$, respectively.

These results demonstrate that filters do not guarantee accurate results, without considering the further problems related to the choice of the filter and its parameters in a highly variable context.

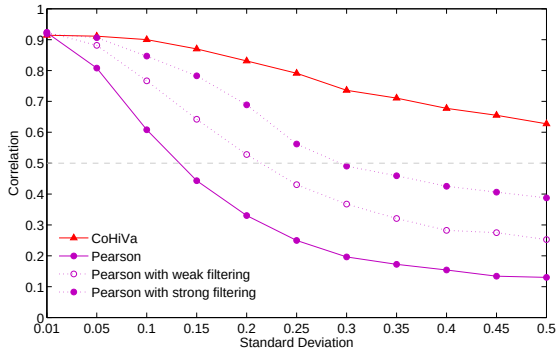
To complete the accuracy evaluation of the models, we report some results obtained in the third scenario characterized by time series with no dependence. Despite the level of variability, we see in Figure 7 that all the models are accurate and detect a weak correlation between time series having no dependence. This result shows that CoHiVa joins the high performance of capturing linear and non-linear correlations to the ability of detecting the absence of correlation.

4.1.2. Robustness

The accuracy of a correlation model must be combined with information about its *robustness*, that assesses the reliability of correlation model results across different evaluations. We quantify the robustness in terms of the *coefficient of variation* (CoV) of the correlation indexes computed over the different



(a) Linear scenario



(b) Non-linear scenario

Figure 6: Analysis of accuracy of correlation models with data filtering.

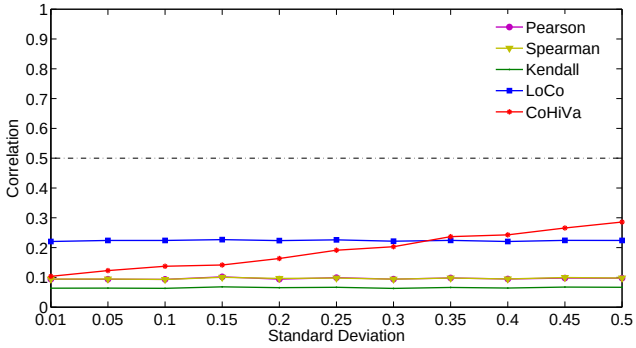


Figure 7: Analysis of accuracy in a not correlated scenario.

data generations. The coefficient of variation is defined as the ratio of the standard deviation to the mean of the correlation index over all the experiments. A

lower CoV denotes a better robustness of the correlation model.

We evaluate the robustness of the results obtained in Section **??**. Table 1 reports the CoV of each considered correlation model computed over 1000 generations of time series in a linear scenario. The columns refer to the increasing values of perturbations intensity σ, while the rows report the correlation models. The CoV of all correlation models increases when σ increases. Compared to existing models, the CoHiVa model is able to keep the lowest CoV for any σ value. Thanks to a CoV always lower than 0.15, the proposed correlation model guarantees a high robustness in capturing linear correlations also among highly variable data.

	σ					
	0.01	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**
Pearson	0.0232	0.0299	0.0914	0.1992	0.3437	0.4817
Spearman	0.0227	0.0304	0.0905	0.2036	0.3496	0.4874
Kendall	0.0371	0.0486	0.1098	0.2170	0.3606	0.4936
LoCo	0.0220	0.0284	0.0835	0.1653	0.2452	0.2735
Pearson with weak filtering	0.0001	0.0069	0.0324	0.0785	0.1206	0.1787
Pearson with strong filtering	0.0001	0.0123	0.0497	0.0917	0.1318	0.1736
CoHiVa	0.0073	0.0086	0.0217	0.0498	0.0888	0.1343

Table 1: Coefficient of Variation in the linear scenario.

	σ					
	0.01	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**
Pearson	0.0274	0.1296	0.3704	0.5843	0.6838	0.7052
Spearman	0.0266	0.1457	0.3894	0.5919	0.6892	0.7104
Kendall	0.0391	0.1632	0.4014	0.5997	0.6960	0.7194
LoCo	0.0258	0.1136	0.2550	0.2923	0.3073	0.2924
Pearson with weak filtering	0.0070	0.0835	0.2026	0.2215	0.2224	0.2236
Pearson with strong filtering	0.0083	0.0944	0.2120	0.2468	0.2434	0.2473
CoHiVa	0.0380	0.0172	0.1467	0.1614	0.1711	0.1926

Table 2: Coefficient of Variation in the non-linear scenario.

As expected, the non-linear context worsens the robustness of all the models. This main conclusion is confirmed by the CoV values reported in Table 2. These results demonstrate that only the CoHiVa model is able to guarantee a CoV

lower than 0.2 for any perturbation intensity. On the other hand, state-of-the-art models show poor results even for medium-low values of σ ($\sigma \leq 0.2$). With the exception of the LoCo and the Pearson model integrated with filters, all the other correlation models are totally unreliable in highly variable contexts because they reach CoV values around 0.7. These results confirm that they cannot be used to capture non-linear relationships among highly variable time series.

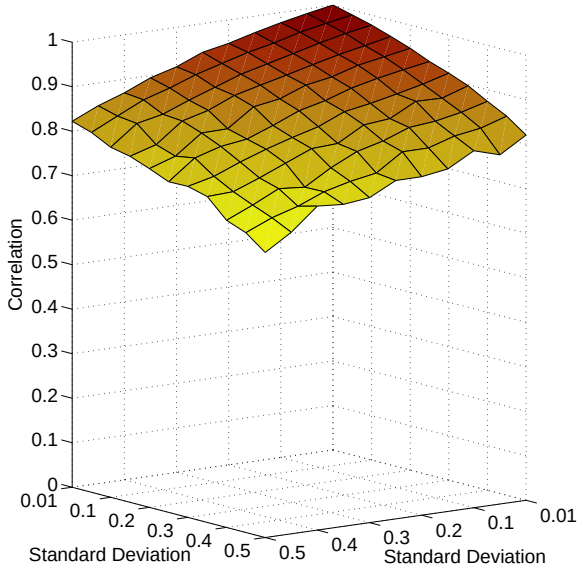
Our analyses confirm that the most popular correlation models without filtering are affected by low accuracy and robustness when data exhibit high variabilities and/or non-linear dependency. Filtering the time series and then applying a correlation model is not sufficient to solve the problems. The use of the proposed CoHiVa model allows to guarantee good performance for any considered type of correlation and variability in the time series.

4.2. ***Different variability levels***

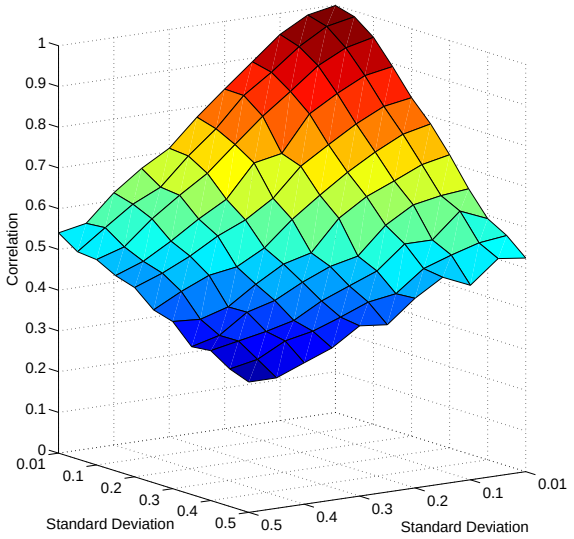
So far, we considered a variability level σ and simulated two time series both affected by that same level of variability. Now, we consider the case in which the variability level that affects the two time series is different. The goal is to evaluate how the difference in time series variability levels impacts on the models performance. Again, we evaluate the models performance in terms of both accuracy and robustness over 1000 generations of linear correlated data taking values in the range $[0, 1]$. We use this set to test the models over all possible combinations of time series with variability levels in $\{0.01, 0.05, 0.1, \ldots, 0.5\}$.

Figure 8 shows the accuracy results of the CoHiVa and LoCo correlation models over all the combinations of variabilities in the time series. (We selected LoCo as term of comparison, since it has shown to be our best competitor).

As expected, both indexes have better performance when working on sets where one or both time series have low variability. As the variability of one time series increases, the model performance decreases. Besides that, the increasing of variability has very different impact over the performance of the two models. In Figure 8(a), we see that the performance of CoHiVa slightly depend on the



(a) CoHiVa



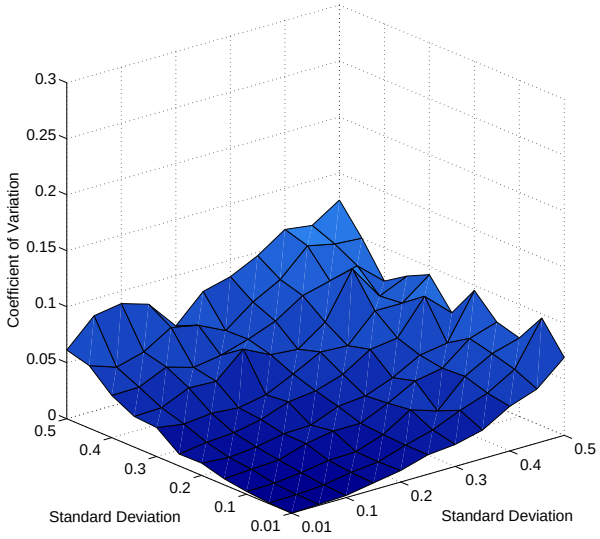
(b) LoCo

Figure 8: Analysis of accuracy at different variability levels.

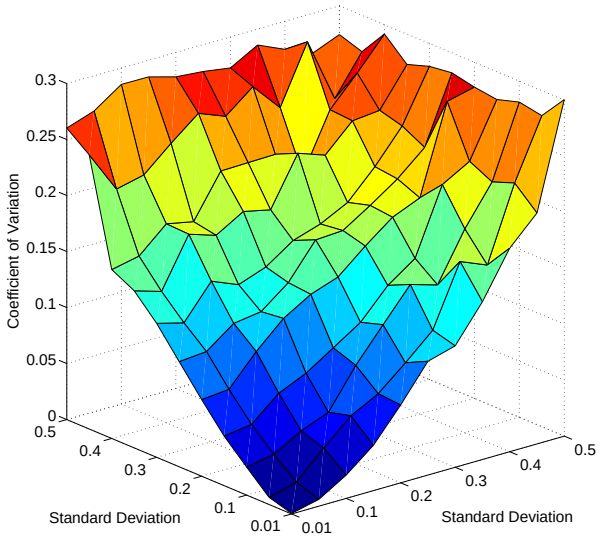
different variability levels for the two time series. Despite the fact that a time series has low ($\sigma = 0.01$) or high ($\sigma = 0.5$) variability, the correlation level captured by CoHiVa always maintains high, with an index that never goes below 0.64. On the contrary, to combine different variability levels has evident impact on LoCo performance in Figure 8(b). In particular, the increase in variability of one of the two time series strongly decreases the LoCo correlation index. For example, if one time series has $\sigma = 0.1$, LoCo gives very different results in case that the other time series is low variable ($\sigma = 0.01$) or highly variable ($\sigma = 0.5$). In the first case, the LoCo index sees a high correlation between the two time series with $\rho = 0.94$; in the latter case, ρ drops to 0.47 and the two time series are seen as not correlated. This drastic difference (≈ 0.5) in LoCo accuracy results due to the change of variability in only one time series are due to the inability of LoCo to adapt the choice of number of principal components to the characteristics of the time series. Through CoHiVa and its adaptive selection of trend patterns, accuracy is maintained high despite the level of variability of the time series.

To complete the evaluation, Figure 9 shows the robustness results of the CoHiVa and LoCo correlation models over all the combinations of variabilities in the time series. Again, changes in the variability levels of the two time series have small effects on the robustness of CoHiVa, as we can see in Figure 9(a). CoV values rise above 0.1 only in case that both time series are affected by high variability levels (i.e., $\sigma \geq 0.4$) and they never exceed 0.12. By comparing Figure 9(b) to Figure 9(a), we see an evident increase in both values and instability of LoCo CoVs with respect to CoHiVa ones. This means that LoCo robustness, as well as its accuracy, is strongly dependent to the variability of the two time series. For example, if one time series has $\sigma = 0.01$, the LoCo robustness can span from a CoV value of 0 to a CoV value of 0.3 on the basis of the variability level of the other time series.

From this analyses we see that both different types of correlation (i.e., linear and non-linear) and different levels of variabilities strongly affect the accuracy and robustness of existing correlation models. The main result is that the pro-



(a) CoHiVa



(b) LoCo

Figure 9: Analysis of robustness at different variability levels.

posed CoHiVa model is able to guarantee good performance for any considered scenario.

5. Complexity

This section estimates the computational complexity and memory space requirements of the considered models: Pearson, Spearman and Kendall, LoCo, CoHiVa, and the possibility of using a filter model.

The majority of these models are characterized by a linear computational complexity as a function of the components n of the time series window. The Spearman and Kendall ranks, and the Pearson index computing correlation between two time series of n data points requires $\mathcal{O}(n)$ operations. Both them require $\mathcal{O}(n)$ space for storing the time series data.

The LoCo index requires $\mathcal{O}(n^2k)$ operations to compute the k largest eigenvectors of the auto-covariance matrix of a time series of length n. LoCo authors [5] set the parameter k to a small value ($k = 4$) in all their experiments, hence the complexity remains quadratic. For the same reason, the memory requirements can be considered linear, because LoCo needs $\mathcal{O}(nk)$ space for storing k eigenvectors and $\mathcal{O}(n)$ space for storing the time series values, for a total of $\mathcal{O}(nk + n) = \mathcal{O}(nk)$ space.

The CoHiVa model has an approach different from LoCo, because CoHiVa does not set the number of patterns, but it selects k and $\hat{k}$ according to the characteristics of the time series. For this reason, we can distinguish two phases in CoHiVa: startup and operation. The startup phase, during which we evaluate k and $\hat{k}$, requires the SVD decomposition of the auto-covariance matrix that has a complexity in the order of $\mathcal{O}(n^3)$. After this startup phase, we incrementally update the CoHiVa index in a streaming setting. By employing the subspace tracking algorithms as in [29], the update of the trend approximation matrix requires $\mathcal{O}(n\hat{k})$ operations. The space requirement is in the order of $\mathcal{O}(n\hat{k})$. In all our experiments on real datasets, we observed that typically $\hat{k} \ll n$. Hence, the cost of the operation phase of CoHiVa tends to be linear as a function of

the length of the time series in terms of time and space. These values are comparable to the complexity of existing correlation solutions.

As example, Table 3 reports the average runtime for the different techniques to compute correlation between linear correlated time series over one of the experiments. The experimental setting used as example is $\sigma = 0.3$ and $n = 100$.

	Pearson	**Spearman**	**Kendall**	**LoCo**	**CoHiVa**	
					startup	**operation**
Time (s)	0.01	0.02	0.04	0.37	0.51	0.06

Table 3: Average runtime over an example experiment.

For highly variable time series, we can also consider the possibility of the combined use of filtering data and applying a correlation algorithm. In this case, the overall computational complexity depends on the filtering technique. Some filtering techniques, such as EWMA, are characterized by a linear complexity. Other more efficient filters, such as Fast Fourier Transform [30], require $\mathcal{O}(nlog(n))$ computations. As a consequence, the total cost can span from linear to quadratic and even more. In terms of space requirements, pre-filtering data and then computing correlation requires at least $\mathcal{O}(n^2)$ space for storing the original and the filtered time series.

6. System management applications

We evaluate the performance of the CoHiVa model for three types of system management applications that can benefit from an accurate and robust evaluation of correlation: tracking analysis (Section 6.1), anomaly detection (Section 6.2), and prediction studies (Section 6.3).

6.1. Tracking analysis

For system management, it is useful to evaluate how the correlation between system resource measures evolves. To this purpose, correlation is evaluated on a *sliding window* containing the most recent samples of time series. Several correlation models can be adapted to tracking correlation analysis. We compare

the performance of four correlation models: Pearson, LoCo, Pearson integrated with a filter, and CoHiVa. The right size for the sliding window and the progression rule that eliminates older samples and incorporates new ones depend on the application context. For the sake of a fair comparison we choose the same window size and the same progress rule for all considered models.

As a summary of a larger set of experiments, we consider the time series referring to the CPU utilization of the two Web servers in the architecture described in Section 2. The results are reported in the four graphs of Figure 10. The CoHiVa model (Figure 10(a)) maintains a stable index around 0.85 during the entire interval of observation. As the two time series are correlated but characterized by a high variability that tends to mask their correlation, the CoHiVa results confirm that this model is accurate and robust even for tracking analysis. On the other hand, the indexes of Pearson (Figure 10(b)) and of LoCo (Figure 10(c)) are affected by low accuracy and low robustness. They oscillate continuously and tend to conclude that the two time series are weakly correlated for most of the time, while the opposite is true. The integration of the Pearson model with a filter (Figure 10(d)) improves the accuracy of the Pearson results but not their robustness, since the variability of the correlation index remains too high.

6.2. Anomaly detection

There are several strategies for anomaly detection [4], but in this paper we are interested to those founded on tracking correlation analysis (e.g., [3]) that are based on an intuitive idea. If a correlation between two time series exists and, at a certain point, this relationship disappears, we can assume that an anomaly occurs in the observed system. Similar conclusions can be achieved when two unrelated time series become correlated.

A visual exemplification of the transition from a normal to an anomalous behavior is presented in Figure 11 and Figure 12, respectively. Let us consider the two servers in the architecture receiving the same number of requests through a load balancer. Figure 11(a) shows the network packet rate during two days

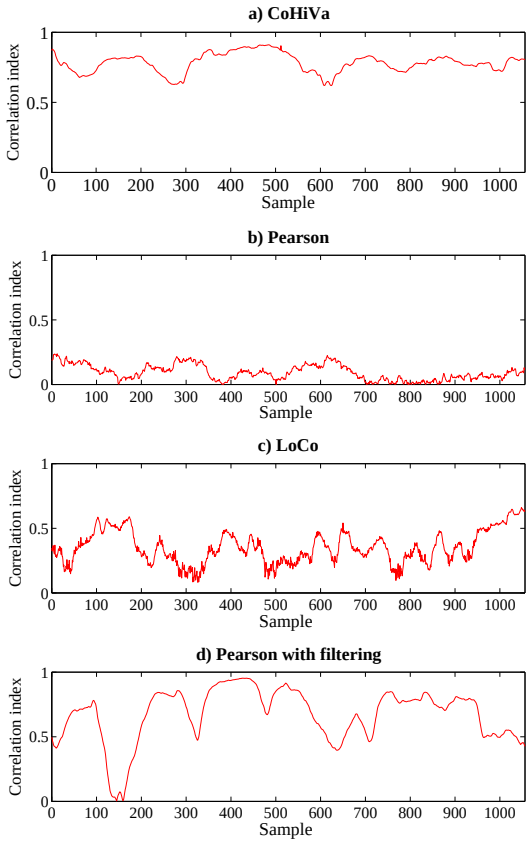
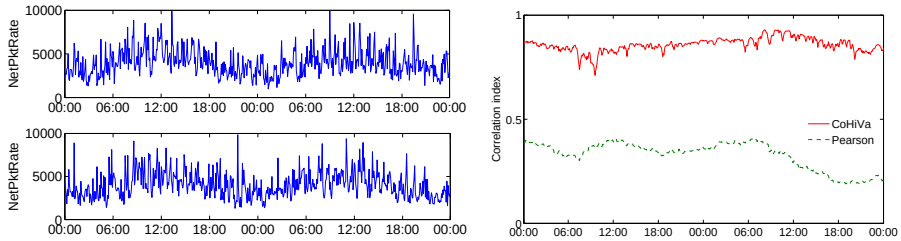


Figure 10: Performance of four correlation models: CoHiVa, Pearson, LoCo, and Pearson integrated with a filter.

of normal behavior, and Figure 11(b) displays the results of the CoHiVa model (continuous line) and the Pearson model (dotted line) in tracking the correlation between the two time series. This figure reveals that the CoHiVa index is able to capture the correlation between the two metrics despite the high variability perturbing the data. On the other hand, the Pearson index remains lower than 0.5, thus concluding that there is no correlation between the two servers.

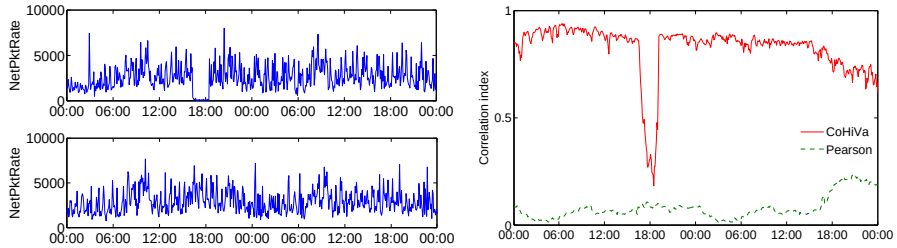
During the following two days reported in Figure 12(a), a system problem in the load balancer causes one of the two serves to not receive requests. The problem was temporary and normal service resumed after two days. In correspondence with this event, the CoHiVa correlation index (continuous line in



(a) Network Packet Rate of two servers **(b) Correlation indexes**

Figure 11: Normal behavior.

Figure 12(b)) shows a progressive drop to the extent that the two time series are evaluated as no longer correlated. On the other hand, the Pearson index (dotted line in Figure 12(b)) remains lower than 0.5 during the entire period of sampling, thus resulting useless as a basis for tracking anomalies.



(a) Network Packet Rate of two servers **(b) Correlation indexes**

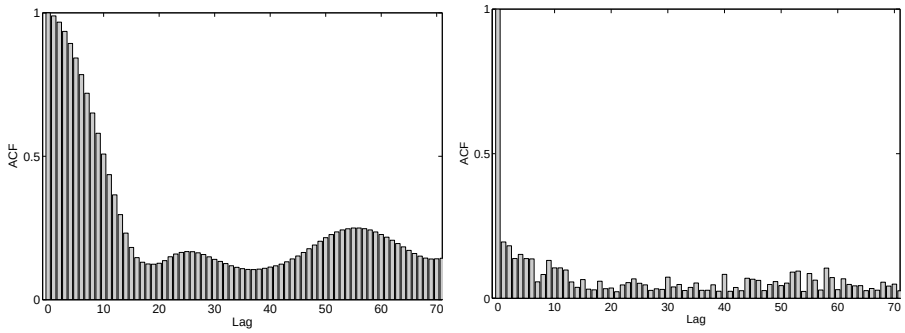
Figure 12: Anomalous behavior in one server.

6.3. Time series prediction

Time series prediction is often adopted in system management contexts (e.g., [31]). In order to establish whether a time series is predictable or not, we measure its autocorrelation, that is, the correlation among values of a time series at different lags in time. This information determines the presence of a statistical dependency among the values of a time series. Prediction models are considered applicable if the decay in the autocorrelation function (ACF) of the time series is exponential [1]. The lag at which the autocorrelation function becomes negligible determines how many past values it is convenient to include

for the estimation of future values [28].

Let us consider as an example the time series at the bottom of Figure 11(a) to illustrate the importance of a robust evaluation of the autocorrelation even for highly variable data. To determine if the considered time series is predictable, we compute its ACF through the CoHiVa and Pearson correlation indexes. Figure 13(a) shows the ACF obtained by computing the CoHiVa correlation index at different lags ranging in the interval $[1, 72]$. The exponential decay of the curve means that a relationship between the samples exists and therefore, by following CoHiVa, we can conclude that the time series can be predicted. On the other hand, by computing the ACF on the basis of the Pearson index, we obtain the quickly decaying curve of Figure 13(b) suggesting that the time series cannot be predicted.



(a) CoHiVa index **(b) Pearson index**

Figure 13: Autocorrelation functions.

To validate that the considered time series is actually predictable, we refer to an autoregressive model (AR) that is a weighted linear combination of p past values. These values are weighted by p linear coefficients that are the first p values of the ACF function evaluated on time series. The p order of the AR model is defined by a statistical test based on the partial auto-correlation function that is described in [28].

We set AR parameters according to the ACF results obtained through CoHiVa and on the basis of this evaluation we conclude that the AR(12) is the best

autoregressive model for the considered time series. In Figure 14, we show the results of applying an AR(12) model for predicting 6 hours of network packet rate on the basis of past data.

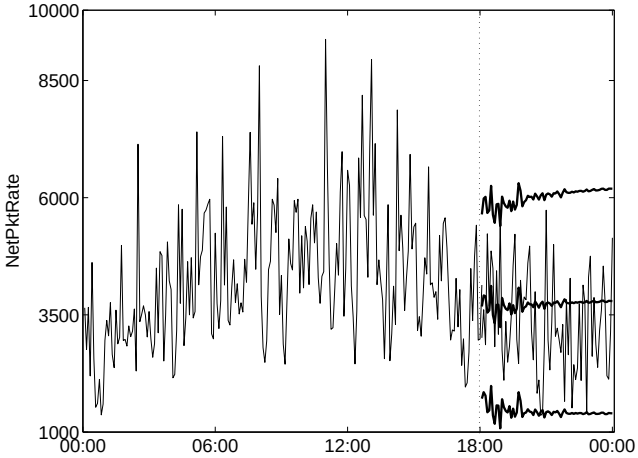


Figure 14: Predicted time series.

The line before 18:00 represents past samples. After that hour, the line represents the future values that we aim to predict. The three bold lines represent the predicted values including their confidence interval. We can appreciate that the actual values are always contained in the prediction interval, hence the time series is predictable as anticipated by the ACF evaluation through CoHiVa.

7. Related work

The task of capturing correlation between time series has received much attention in literature. Many correlation models have been proposed [6, 8, 9, 5], but all of them suffer of poor performance when dealing with some statistical properties of time series. For example, the Pearson product moment [6] is not effective in capturing non-linear correlations, because it assumes that time series are related through a linear dependency. The Spearman rank [8] and the Kendall rank [9] manage linear and non-linear dependences, but the efficacy of their results depends on data distributions [6]. The correlation index pro-

posed in [5] overcomes these limits by looking for main patterns in data and by computing correlation through an estimation of the pattern similarity. This approach is promising, but it does not work when time series are characterized by high variability. In highly variable domains, considering the main pattern of a time series as the only information for capturing relationships between data is inadequate for two reasons: more than one pattern typically carries on useful information about time series similarity; the main patterns of highly variable time series do not include just trends, but they are likely to include perturbations that mask correlation between time series.

An alternative approach to face the high variability problem is to reduce the amount of perturbation before computing correlation, by applying random matrices [32] or some filtering algorithms [10]. Unfortunately, this approach opens more issues than solutions. Random matrices require the a-priori knowledge of the distribution of the random matrix, as well as the possibility of modeling the perturbation behavior, that is infeasible in most contexts. Filtering algorithms are affected by the well known trade-off between the extent of perturbation removal and the quality of the retained trends [20]. This complicates the choice of the best filtering technique and of its parameters. In highly variable contexts, an inappropriate choice of the parameters either results in reduced perturbation removal compromising the correlation accuracy result, or in excessive smoothing which nullifies the robustness of the models.

The CoHiVa model proposed in this paper solves most problems of existing algorithms. It does not require any assumption about data distribution as Spearman and Kendall's ranks do [8, 9], and its performance does not depend on the type of dependency existing between time series, as it is necessary for the Pearson model [6]. Unlike LoCo [5], CoHiVa does not rely upon a fixed number of main patterns, hence it is able to disclose all relationships between time series. By selecting all trend patterns and discarding patterns related to perturbations, the proposed model is able to capture correlations among highly variable time series without requiring any pre-analysis of time series characteristics. All these features make CoHiVa an accurate and robust solution that can

support a large range of applications for system management.

8. Conclusion

Having an accurate and robust model for capturing correlations between time series is of crucial importance for system management. However, when relationships between time series are hidden by highly variable perturbations, the accuracy and robustness of existing correlation models are limited. We propose a novel model for detecting correlations that is able to capture the presence of dependency also between highly variable time series as those coming from samples of monitored system resources. It is based on the extraction of the main patterns of the time series, the removal of perturbations, and the selection of just the trend patterns. The evaluation carried out on synthetic and real datasets characterized by different levels of variability demonstrate that the proposed model improves the state of the art in terms of accuracy and robustness. Our promising results evidence the possibility of using the proposed model as a support for system management applications in all domains characterized by high variability where existing correlation models do not work.

References

[1] C. Granger, P. Newbold, Forecasting Economic Time Series, Academic Press, 1986.

[2] T. Warren Liao, Clustering of time series data - a survey, Pattern Recognition, 38 (11), pp. 1857–1874, 2005.

[3] C. Kruegel, G. Vigna, Anomaly detection of web-based attacks, in: Proc. of the 10th ACM conference on Computer and communications security, Washington D.C., USA, 2003.

[4] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: A survey, ACM Computing Surveys, 41 (3), pp. 15:1–15:58, 2009.

[5] S. Papadimitriou, J. Sun, P. S. Yu, Local Correlation Tracking in Time Series, IEEE International Conference on Data Mining, Los Alamitos, CA, USA, 2006.

[6] J. Cohen, Applied multiple regression/correlation analysis for the behavioral sciences, L. Erlbaum Associates, 2003.

[7] F. Gorunescu, Data Mining: Concepts, Models and Techniques, Springer, 2011.

[8] C. Spearman, The proof and measurement of association between two things, The American Journal of Psychology, 100 (3-4), pp. 441-471, 1904.

[9] M.G. Kendall, Rank correlation methods, Charles Griffin & Company Ltd., 1962.

[10] M. Andreolini, S. Casolari, M. Colajanni, Models and framework for supporting run-time decisions in web-based systems, ACM Transaction on the Web 2 (3), pp. 17:1–17:43, 2008.

[11] Q. Zhang, A. Riska, W. Sun, E. Smirni, G. Ciardo, Workload-Aware Load Balancing for Clustered Web Servers, IEEE Transaction on Parallel and Distributed Systems, 16 (3), pp. 219-233, 2005.

[12] S. Ghosh, M.S. Squillante, Analysis and control of correlated web server queues, Computer Communications, 27 (18), pp. 1771–1785, 2004.

[13] M. Dahlin, Interpreting Stale Load Information, IEEE Transaction on Parallel and Distributed Systems, 11 (10), pp. 1033–1047, 2000.

[14] A. Buda, A. Jarynowski, Life-time of correlations and its applications, Wydawnictwo Niezalezne, 2010.

[15] W. Willinger, D. Alderson, L. Li, A pragmatic approach to dealing with high-variability in network measurements, in: Proc. of the 4th ACM SIGCOMM conference on Internet measurement, Taormina, Italy, 2004.

[16] D. C. Montgomery, Introduction to Statistical Quality Control, John Wiley and Sons, 2008.

[17] S. G. Mallat, A Theory of Multiresolution Signal Decomposition: The Wavelet Decomposition, IEEE Transaction on Pattern Analysis and Machine Intelligence, 11 (7), pp. 674–693, 1989.

[18] S. Papadimitriou, P.S. Yu, Optimal multi-scale patterns in time series streams, in: Proc. of the 2006 ACM SIGMOD international conference on Management of data, Chicago, USA, 2006.

[19] S. Papadimitriou, S. Jimeng, C. Faloutsos, Streaming Pattern Discovery in Multiple Time-Series, in: Proc. of the 31st International Conference on very large data bases, Trondheim, Norway, 2005.

[20] M. N. Nounou, B. Bakshi, On-Line Multiscale Filtering of Random and Gross Errors without Process Models, American Institute of Chemical Engineers Journal, 45 (5), pp. 1041-1058, 1999.

[21] B.R. Bakshi, Multiscale PCA with application to multivariate statistical process monitoring, AIChE Journal, 44 (7), pp. 1596-1610, 1998.

[22] B. Abrahao, A. Zhang, Characterizing application workloads on cpu utilization in utility computing, Technical Report HPL-2004-157, Hewllet-Packard Labs, 2004.

[23] R. Khattree, D. N. Naik, Multivariate data reduction and discrimination with SAS software, SAS Institute Inc., 2000.

[24] A. Lakhina, K. Papagiannaki, M. Crovella, C. Diot, E. D. Kolaczyk, N. Taft, Structural analysis of network traffic flows, in: Proc. of the Joint International Conference on Measurement and Modeling of Computer Systems, New York, USA, 2004.

[25] H. E. Hurst, Long-term storage capacity of reservoirs, Transaction of the American Society of Civil Engineers, 116, pp. 770-799, 1951.

[26] R. Weron, Estimating long range dependence: finite sample properties and confidence intervals, Physica A, 312 (1-2), pp. 285-299, 2002.

[27] M. Dobber, R. Van det Mei, G. Koole, A prediction method for job runtimes in shared processors: Survey, statistical analysis and new avenues", Performance Evaluation, 64 (7-8), pp. 755-781, 2007.

[28] B.L. Brockwell, R.A. Davis, Time Series: Theory and Methods, Springer-Verlag, 1987.

[29] B. Yang, Projection approximation subspace tracking, IEEE Transaction on Signal Processing, 43 (1), pp. 95-107, 1995.

[30] A.V. Oppenheim, R.W. Schafer, J.R. Buck, Discrete-time signal processing, Prentice Hall, 1999.

[31] G.E. Box, G.M. Jenkins, G.C. Reinsel, Time Series Analysis Forecasting and Control, Prentice Hall, 1994.

[32] H. Kargupta, K. Sivakumar, S. Ghosh, Dependency Detection in MobiMine and Random Matrices, in: Proc. of the 6th European Conference on Principles of Data Mining and Knowledge Discovery, London, UK, 2002.