Knowledge Discovery from Heterogeneous Dynamic Systems using Change-Point Correlations*

Tsuyoshi Idé† Keisuke Inoue†

October 4, 2004

Abstract

Most of the stream mining techniques presented so far have primary paid attention to discovering association rules by direct comparison between time-series data sets. However, their utility is very limited for heterogeneous systems, where time series of various types (discrete, continuous, oscillatory, noisy, etc.) act dynamically in a strongly correlated manner. In this paper, we introduce a new nonlinear transformation, singular spectrum transformation (SST), to address the problem of knowledge discovery of causal relationships from a set of time series. SST is a transformation that transforms a time series into the probability density function that represents a chance to observe some particular change. For an automobile data set, we demonstrate that SST enables us to discover a hidden and useful dependency between variables.

Keywords: time-series, change-point detection, singular-spectrum analysis, hidden dependency

1 Introduction

The frontiers of data mining research are being extended to include knowledge discovery from nontraditional data types such as statically [7] and dynamically [6] structured data. However, little attention has been paid to heterogeneous dynamic systems, where time series of various types (discrete, continuous, oscillatory, noisy, etc.) act dynamically in a strongly correlated manner.

Generally, in strongly correlated dynamic systems, the behavior of the whole system can often be extremely complicated even if the mechanism of correlation between each pair of variables is relatively simple. Therefore knowledge discovery in such systems can be far more difficult than expected. For instance, in an automobile, the individual states of the variables such as engine RPM (revolutions per minutes), engaged gear, fuel flow rate, throttle position (TP) sensor, and air intake oxygen density have almost an infinite number of combinations depending on the environment around the car and human actions. Therefore, it is generally impossible to find a rule like "if variables $x_1, x_2, ...$ have a certain combination of values, then the system would be faulty."

In this paper, we address the issue of discovering causal dependencies hidden deep within the heterogeneous time-series data. We assume that we are not provided with detailed prior knowledge of dependencies. In addition, we assume that each variable exhibits sudden and steep changes in a heterogeneous manner so that traditional approaches that attempt to separate trend and noise components are difficult to use.

Note that this problem setting is different from traditional stream mining. An implicit assumption of Das et al. [2], which is known as a seminal work in this field, was that subsequences of individual variables can be clustered into one of a small numbers of patterns. Except for parts of the data which may exhibit relatively simple behaviors, the utility of their approach is very limited in heterogeneous dynamic systems. Also, Keogh-Lin-Truppel [9] recently pointed out that it is theoretically questionable whether or not one can fit an arbitrarily chosen subsequence into one of the patterns.

In this paper, we tackle this issue by introducing a new nonlinear transformation, singular spectrum transformation (SST) for a set of time series data. SST is a transformation that converts an original time series into a new time series based on change-point scores. The resultant time series can be interpreted as the probability distribution that some change occurs. Since change points in a mechanical system are expected to be caused by a well-defined mechanism, if the score simultaneously has a high value at some time for two different variables, then a causal relationship is likely between them.

The essence of our idea is illustrated in Fig. 1, where two artificially generated heterogeneous data sets (see Subsection 3.2 for details) and their SSTs are shown. While it is difficult to infer any dependency between the two original variables, SST clearly reveals a hidden dependency between them in terms of synchronization

*To appear in Proceedings of 2005 SIAM International Conference on Data Mining (SDM 05), April 21-23, 2005.

†IBM Research, Tokyo Research Laboratory. E-mail: {goodidea, inouek}@jp.ibm.com.

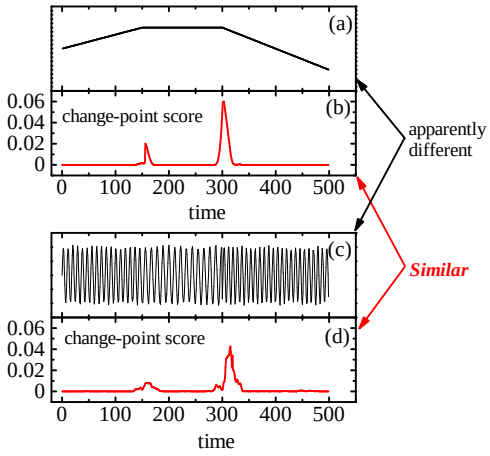

Figure 1: Example of SST in a heterogeneous system. Original time-series in (a) and (c) are transformed into change-point scores in (b) and (d), so that a hidden similarity is revealed.

of their change points. Note that the results in Figs. 1 (b) and (d) were obtained using a common algorithm and a common parameter set. Therefore, we see that, by performing SST, the problem of data mining in heterogeneous systems can be reduced to mining in homogeneous systems without using any detailed knowledge on the behavior of data. To the best of the authors' knowledge, this is the first work that uses change-point correlation in the context of knowledge discovery from dynamic systems with strongly-correlated and heterogeneous natures.

2 Change-point detection

2.1 Extraction of past patterns. Consider a time series $\mathcal{T} = \{x(1), x(2), ..., x(t), ..\}$ and its consecutive subsequence with length w as $\{x(t-w), ...x(t-2), x(t-1)\}$. We define a column vector corresponding to this subsequence as

$$\boldsymbol{s}(t-1) = (x(t-w), ..., x(t-1))^T,$$

where the superscript T represents transpose. We construct a matrix, which is often called a Hankel matrix, using column vectors of this kind as

$$H(t) = [\boldsymbol{s}(t-n), ..., \boldsymbol{s}(t-2), \boldsymbol{s}(t-1)].$$

We call this $w \times n$ matrix a trajectory matrix at t, following Moskvina-Zhigljavsky [11]. By definition, the trajectory matrix is defined over $w+n-1$ elements from $x(t-1)$ to $x(t-w-n+1)$. We denote $w+n-1$ as W. We illustrate the setting in Fig. 2.

The trajectory matrix $H(t)$ can be viewed as a record that contains various change patterns within the

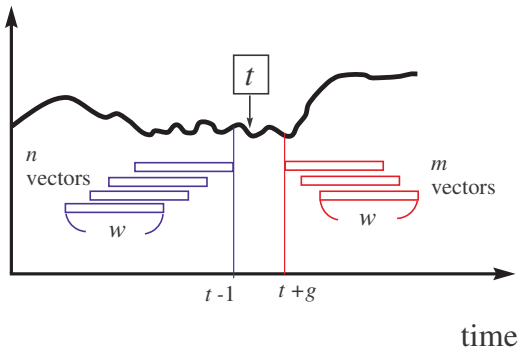

Figure 2: Summary of parameters used in SST. From m and n subsequences at both sides of a time point (t), representative patterns are calculated.

range of the past W points under the length constraint w. Now let us extract the representative patterns from $H(t)$. We write a representative pattern as $\boldsymbol{u}$. It is natural to suppose that this is expressed as a linear combination of $\boldsymbol{s}(t_j)$s:

$$\boldsymbol{u} = c\sum_{i=1}^{n} v_i \boldsymbol{s}(t-i),$$

where c is a normalization constant to satisfy $\boldsymbol{u}^T\boldsymbol{u} = 1$. If we define an n-dimensional vector as $\boldsymbol{v} = (v_1, ..., v_n)^T$, this equation is simply expressed as $\boldsymbol{u} = cH(t)\boldsymbol{v}$. We determine the representative by majority voting among the observed patterns. In particular, we want the direction that produces the strongest constructive interference between $\boldsymbol{s}$s. Mathematically, this direction will be found as

$$\boldsymbol{v}(t) \equiv \arg\max_{\tilde{\boldsymbol{v}}} ||H(t)\tilde{\boldsymbol{v}}||^2, \tag{2.1}$$

where we impose a constraint of $\boldsymbol{v}^T\boldsymbol{v} = 1$. Introducing a Lagrange multiplier λ, this equation is reduced to

$$\frac{\partial}{\partial\tilde{\boldsymbol{v}}}\left[\tilde{\boldsymbol{v}}^T H(t)^T H(t)\tilde{\boldsymbol{v}} - \lambda\tilde{\boldsymbol{v}}^T\tilde{\boldsymbol{v}}\right] = 0.$$

From this, we immediately see that $\boldsymbol{v}$ is the normalized solution of an eigenvalue equation

$$H(t)^T H(t)\boldsymbol{v} = \lambda\boldsymbol{v}.$$

Also, $\boldsymbol{u}$ is the normalized solution of the eigenvalue equation of $H(t)H(t)^T$, i.e.

$$H(t)H(t)^T\boldsymbol{u} = \lambda\boldsymbol{u}. \tag{2.2}$$

These results show that the representative pattern $\boldsymbol{u}$ and its coefficient vector $\boldsymbol{v}$ are the left and right singular

vectors of $H(t)$, respectively. The singular value is equal to $\sqrt{\lambda}$.

Let us denote the singular values and the left singular vectors as $\{(\sigma_1, \boldsymbol{u}_1), (\sigma_2, \boldsymbol{u}_2), ..., (\sigma_l, \boldsymbol{u}_l)\}$ in descending order of the singular values. The parameter l represents the number of representative patterns under consideration. The greater the singular value is, the more dominant the corresponding pattern is. If a singular value (≥ 0) is small, then the corresponding pattern can be considered to be a noise component.

As described above, the method to find the dominant components using singular value decomposition (SVD) on the Hankel matrices is called singular spectrum analysis. [1]

2.2 Extraction of the current pattern. On the future side of the trajectory matrix, we again take a column vector with length w as

$$\boldsymbol{r}(t+g) = (x(t+g), ..., x(t+g+w-1))^T .$$

This is the same as $\boldsymbol{s}(t+g+w-1)$, but we introduce this new notation to represent a symmetry between both sides of t. We again define a Hankel matrix, which we will call a test matrix at t, using m $\boldsymbol{r}$s

$$G(t) = [\boldsymbol{r}(t+g), \boldsymbol{r}(t+g+1), ..., \boldsymbol{r}(t+g+m-1)] .$$

As in Eq. (2.2), the present representative pattern is given by the solution of

$$G(t)G(t)^T \boldsymbol{u} = \mu \boldsymbol{u}. \tag{2.3}$$

We call the normalized largest eigenvector the test vector, and represent it as $\boldsymbol{\beta}(t)$.

2.3 Change-point score. We have obtained the past representative patterns $\{\boldsymbol{u}_i | i = 1, .., l\}$ and the present representative pattern as $\boldsymbol{\beta}(t)$. Let us define an anomaly metric using these patterns. If $\boldsymbol{\beta}(t)$ is sufficiently similar to some of the frequent patterns, it should be on the hyperplane spanned by $\{\boldsymbol{u}_i | i = 1, .., l\}$. Otherwise, $\boldsymbol{\beta}(t)$ would be directed outside of the hyperplane.

To quantitatively evaluate how far $\boldsymbol{\beta}(t)$ is from the hyperplane, let us define a matrix U_l as

$$U_l = [\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_l] .$$

Using this matrix, the normalized projection of $\boldsymbol{\beta}(t)$ onto the hyperplane is given by

$$\boldsymbol{\alpha}(t) \equiv \frac{U_l^T \boldsymbol{\beta}(t)}{\left|\left| U_l^T \boldsymbol{\beta}(t) \right|\right|}.$$

[1] While it is called "spectrum analysis," we should emphasize that it has nothing to do with the classical Fourier analysis.

Now we can define the change-point score as

$$z(t) \equiv 1 - \boldsymbol{\alpha}(t)^T \boldsymbol{\beta}(t). \tag{2.4}$$

By definition, this quantity is limited to the range from zero to 1. It is small when there is little change compared to the past patterns and large when the present pattern is quite different from the past patterns.

3 Singular spectrum transformation

3.1 Definition. As discussed, the change-point score can be defined at arbitrary t by calculating representative patterns for both the trajectory and test matrices. This can be viewed also as a transformation from an original time-series $\mathcal{T}$ to a new time-series $\mathcal{T}_c$, i.e.

$$\mathcal{T} \rightarrow \mathcal{T}_c(w, l, g, m, n).$$

We define this transformation as the singular spectrum transformation (SST). As expressed in the parenthesis, there are five major parameters in SST. This transformation defines a nonlinear transformation in that it does not satisfy the principle of superposition. Hereafter, the integrated area of the transformed time-series is assumed to be normalized to one. Under this condition, the transformed time-series is interpreted as the probability density that some change occurs at time t.

The occurrence of a change-point should be independent of any apparent variety such as discrete, continuous, noisy, oscillatory, etc. Thus, one may think of the new time-series $\mathcal{T}_c(w, l, g, m, n)$ as the signs of causality hidden behind the apparent variation of the original time series: If the similarity between a pair of variables is high for a set of SST series, then some dependency between them is strongly suggested. SST can be a powerful tool to discover hidden dependencies among variables.

3.2 Example. An example of SST is shown in Fig. 1. The time series (a) was generated using three linear functions with slopes of 1/300, 0, and $-1/200$. The other time-series (c) was generated using a sine function $x(t) = \sin(2\pi t/\lambda)$, for $\lambda = \sqrt{80}$, $\sqrt{120}$, and $\sqrt{70}$. In (c), we also added random fluctuations to the amplitude and the periods of up to $\pm 7.5\%$ and $\pm 0.5\%$, respectively, to simulate fluctuations in realistic observations. For both data sets, the change points are located at $t = 150$ and 300. The results of SST in Figs. 1 (b) and (d) was calculated with $w = -g = m = n = 20$ and $l = 3$. In spite of the apparent differences in the original data, we see that SST strikingly reveals the similarities without any ad hoc tuning for individual time series. It is evident that existing methods such as differentiation [5] and wavelet-based approaches [8] fail to detect the change points if a common parameter set is used for both sets.

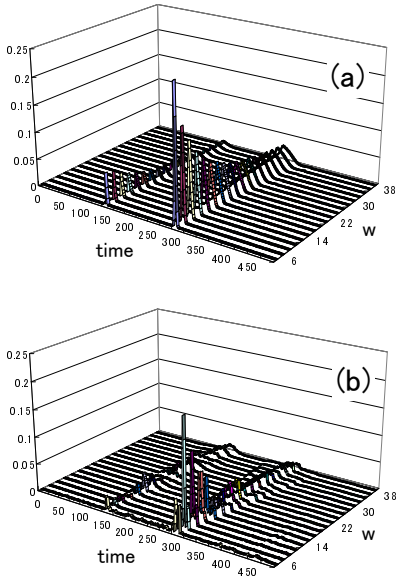

Figure 3: The dependence of SST on w for (a) the linear function and for (b) the oscillatory function shown in Fig. 1 (a) and (c), respectively.

The dependence on w is of particular interest in SST. We calculated SST as a function of w under $w = -g = m = n$ and $l = 3$. The results are shown in Fig. 3. It is surprising that the essential features remain unchanged over a very wide range of w, $6 \lesssim w \lesssim 40$, while the widths of the major features become broader as w increases. This robustness is quite suitable for heterogeneous systems.

4 Experiment

4.1 Data set. The goal of this experiment is to identify the pair of variables that correlates the most in terms of causality, without using any prior knowledge of the variables. The data set used in this section was generated by a specialized simulator for the power train control module of a vehicle, and was taken for one minute with a sampling interval of 0.1 sec. It includes fuel flow rate (x_1), engaged gear (x_2), vehicle speed (x_3), engine RPM (x_4), and manifold absolute pressure (x_5). Figure 4 (a) shows all five of these time-series. For visibility, the signals from x_1 to x_4 are shifted vertically in the figure.

4.2 Comparison between raw and SST time-series. SST was done with the parameters $w = m = n = -g = 25$ (2.5 sec) and l=2 for the five time series. Since SVD is not invariant with respect to translation of the origin of the column vectors in the matrix, we

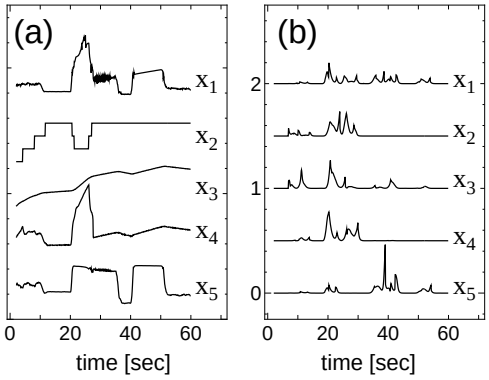

Figure 4: (a) Time-series data from an automobile and (b) the resulting SST series.

standardized the time series so that each of the averages is three times the standard deviation. The result is shown in Fig. 4 (b).

By comparing (a) with (b), we see that each feature in Fig. 4 (b) corresponds to a change in the original data. Interestingly, the SST series of x_2 and x_4 exhibit some similarity in terms of the synchronization of the change-points, in spite of the fact that they seem to behave totally differently in the original series. Similarly, x_1 and x_3 seem to have some in common in Fig. 4 (b) while the original data are quite different. This result demonstrates that SST can make the variables of different types be comparable with each other. In other words, SST converts a heterogeneous system into a different "homogeneous" system.

4.3 Visualization via MDS. To compare the inter-dependencies of the variables, we used the classical solution of multidimensional scaling (MDS) [10]. For the definition of the distance matrices, we took the L_1 and L_2 distances for SST and the raw time series, respectively. Each of the time series was normalized in advance so that $\int dt x(t)^2 = 1$ or $\int dt z(t) = 1$ holds. To remove the unwanted effects of noisy fluctuations of the signals, we performed Gaussian convolution with the standard deviation of 1.5 seconds before computing the distance matrix for SST.

The results of MDS are shown in Fig. 5. Since the definition of the distance metrics are not common in the raw and SST cases, only the relative locations within each plot are meaningful. In Fig. 5 (a), the variables x_1, x_4, and x_5 can be attributed to one cluster. We see that they are actually similar in shape in Fig. 4 (a). Similarly, the variables x_2 and x_3 form the other cluster due to the similarity in their increasing trends in Fig. 4 (a).

On the other hand, the two clusters collapse in Fig. 5 (b). Specifically, the closest pair is x_2 and x_4. This is very interesting because they have totally different trends in the original sequences shown in Fig. 4 (a). This result is due to the synchronizations of the change points in both data sets. In reality, the variables x_2 and x_4 are the engaged gear and the engine RPM, respectively. The close dependency of x_2 and x_4 corresponds to "the value of engine RPM increased after shifting to a lower gear." It is worth noting that we could discover a part of the causal relationships without using any prior knowledge. This result demonstrates that SST can reveal the signs of causality hidden deep inside of the heterogeneous correlated systems.

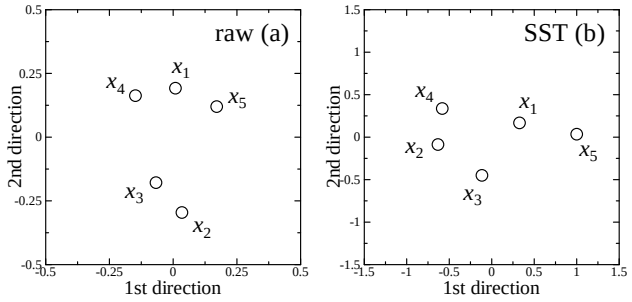

Figure 5: Two dimensional MDS plot for (a) raw data, and (b) SST data.

5 Related work and our contribution

The problem of change detection has been studied for a long time, and various methods such as CUSUM (cumulated summation) [1], wavelet analysis [8], inflection point search [5], and Gaussian mixtures [13] have been proposed. These existing methods, however, are not applicable to our task without using ad hoc tuning for individual signals. Similarly, time-series correlation methods based on these techniques in a few application domains [5, 4, 12] are inapplicable to our task.

Moskvina-Zhigljavsky [11] used the singular spectrum analysis technique [3] for change detection, based on SVD of the Hankel matrix. Mathematically, SVD can be performed for almost any kind of matrix. Thus, the method can be applicable to various sorts of time series without any ad hoc tuning. Our contribution is to have defined the problem of knowledge discovery from heterogeneous dynamic systems and to have proved that their method is one of the most suitable solutions for this problem. Theoretically, our contribution is to have adopted a dimensionless definition of the score, and to have given an algorithm that is pseudo-invariant with respect to time inversion. In other words, our algorithm is invariant with respect to $t \to -t$ for $l = 1$ and $m = n$.

Acknowledgements

The authors thank W. Nathaniel Mills III for fruitful discussions. T.I. thanks Akihiro Inokuchi for providing valuable information on stream mining.

References

[1] M. Basseville and I. Nikiforov. *Detection of Abrupt Changes*. Prentice Hall, Englewood Cliffs, New Jersey, 1993.

[2] G. Das, K.-I. Lin, H. Mannila, G. Renganathan, and P. Smyth. Rule discovery from time series. In *Proc. the Fourth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1998.

[3] M. Ghil, M. R. Allen, M. D. Dettinger, K. Ide, D. Kondrashov, M. E. Mann, A. W. Robertson, A. Saunders, Y. Tian, F. Varadi, and P. Yiou. Advanced spectral methods for climatic time series. *Reviews of Geophysics*, 40:1–41, 2002.

[4] H. Guo, J. Crossman, Y. Murphey, and M. Coleman. Automotive signal diagnostics using wavelets and machine learning. *IEEE Trans. Vehicular Technology*, 49:1650–1662, 2000.

[5] S. Hirano and S. Tsumoto. Mining similar temporal patterns in long time-series data and its application to medicine. In *Proc. 2002 IEEE International Conference on Data Mining*, pp. 219–226, 2002.

[6] T. Idé and H. Kashima. Eigenspace-based anomaly detection in computer systems. In *Proc. the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 440–449, 2004.

[7] A. Inokuchi, T. Washio, and H. Motoda. Complete mining of frequent patterns from graphs: Mining graph data. *Machine Learning*, 50(3):321–354, 2003.

[8] S. Kadambe and G. Boudreaux-Bartels. Application of the wavelet transform for pitch detection of speech signals. *IEEE Trans. Information Theory*, 38:917–924, 1992.

[9] E. Keogh, J. Lin, and W. Truppel. Clustering of time series subsequences is meaningless: Implications for previous and future research. In *Proc. IEEE International Conference on Data Mining*. IEEE, 2003.

[10] K. Mardia, J. Kent, and J. Bibby. *Multivariate Analysis*. Academic Press, 1980.

[11] V. Moskvina and A. Zhigljavsky. An algorithm based on singular spectrum analysis for change-point detection. *Communications in Statistics—Simulation and Computation*, 32(4):319–352, 2003.

[12] M. Thottan and C. Ji. Anomaly detection in IP networks. *IEEE Trans. Signal Processing*, 51(8):2191–2204, 2003.

[13] K. Yamanishi and J. Takeuchi. A unifying framework for detecting outliers and change points from non-stationary time series data. In *Proc. the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 676–681, 2002.