Time-Series Causality With Missing Data

Bo Yuan Chang Vision and Image Processing Group, University of Waterloo, ON, Canada
Mohamed A. Naiel Vision and Image Processing Group, University of Waterloo, ON, Canada
Steven Wardell ATS Automation, Cambridge, ON, Canada
Stan Kleinikkink ATS Automation, Cambridge, ON, Canada
John S. Zelek Vision and Image Processing Group, University of Waterloo, ON, Canada
Email: {by2chang, mohamed.naiel, jzelek}@uwaterloo.ca, {swardell, skleinikkink}@atsautomation.com

Abstract

Over the past years, researchers have proposed various methods to discover causal relationships among time-series data as well as algorithms to fill in missing entries in time-series data. Little to no work has been done in combining the two strategies for the purpose of learning causal relationships using unevenly sampled multivariate time-series data. In this paper, we examine how the causal parameters learnt from unevenly sampled data (with missing entries) deviates from the parameters learnt using the evenly sampled data (without missing entries). However, to obtain the causal relationship from a given time-series requires evenly sampled data, which suggests filling the missing data values before obtaining the causal parameters. Therefore, the proposed method is based on applying a Gaussian Process Regression (GPR) model for missing data recovery, followed by several pairwise Granger causality equations in Vector Autoregssive form to fit the recovered data and obtain the causal parameters. Experimental results show that the causal parameters generated by using GPR data filling offers much lower RMSE than the dummy model (fill with last seen entry) under all missing values percentage, suggesting that GPR data filling can better preserve the causal relationships when compared with dummy data filling, thus should be considered when dealing with unevenly sampled time-series causality learning.

1 Introduction

Modelling time-series data is an important problem in the field of causal discovery. Due to the complicated nature of time-series data (i.e. seasonal, trend, stochastic term, interventions etc.) it is difficult enough to work with. To make the problem even more challenging missing data are often ubiquitous in many real world data [1]. Evenly sampled time-series data is essential for causal discovery. But it is often difficult to obtain this regularly samples data in the many industry sectors due to varies reasons (i.e. hardware limitation, cost of maintenance etc.). Thus, methods to fill in missing values are required before preceding with causal learning.

Gaussian Processes (GP) is a very powerful non-parametric algorithm that can be applied to solve both complicated regression and classification problems [2]. Generally speaking the GP algorithm is mainly applied in the area of supervised learning [2] while there are also some work done in areas like un-supervised learning [3] and reinforcement learning [4]. The application for the paper will be in the area of filling missing values which is under supervised learning umbrella.

There are mainly two types of approaches for time series data filling [5]. The first category is the *parametric approach*, which is to simply consider a linear (or high degree polynomial function) and find the line of best fit of the training dataset. This approach is simple but at the same time, the degree of order must be defined in advance. In real world datasets it is often very challenging to find a single line of best fit to represent the entire dataset, suggesting that the first approach is not as practical as one would hoped when dealing with complicated real world datasets. In contrast to the first category, the *non-parametric approach* gives a prior probability to every possible function where higher probabilities are given to functions that we have higher confidence in (based on the training dataset). Gaussian process can be used to generalize the Gaussian probability distribution, and allows us to compute and select functions from an uncountable infinite set of possible functions.

In this paper, we perform missing data recovery using Gaussian Process Regression technique for filling missing values in time-series data to obtain pairwise Granger Causality parameters. In addition, we compare the quality of filling the missed data by comparing the Granger causality parameters estimated using original time-series data against its GPR filled version where the RMSE values under each filling percentage is calculated. The same procedure is repeated using dummy filling and the two sets of RMSE values are compared for evaluation of the performance.

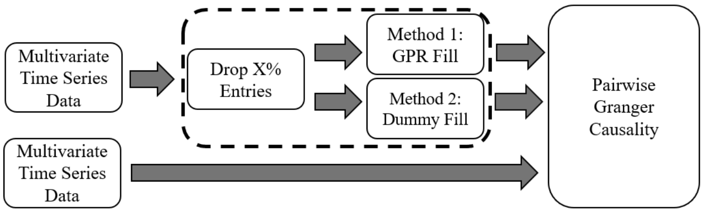
Multivariate Time Series Data
Drop X% Entries
Method 1: GPR Fill
Method 2: Dummy Fill
Multivariate Time Series Data
Pairwise Granger Causality

Fig. 1: The proposed pipeline for Granger causality from irregularly sampled data, where GPR and Dummy methods refer to the Gaussian process regression method and filling data with the last seen value, respectively.

2 Proposed Method

Figure 1 illustrates the proposed pipeline in order to study the performance of causal discovery with irregularly sampled data. Given a multivariate time-series data, the proposed method randomly drops X% of the original data entry and the missing values is then filled using either (a) Gaussian Process Regression (Section 2.1) or (b) Dummy model approach in which the data is filled with last seen entry. Next, the two recovered datasets are then used to obtain the parameters of the pairwise Granger Causality (Section 2.2). Finally, the root mean square error (RMSE) for each filling technique is calculated with respect to the causal parameters obtained from the original dataset.

2.1 Gaussian Process Regression for Data Filling

Although the GP requires an entire training set to perform prediction and lose efficiency with higher dimensions [6], it offers probabilistic predictions and allow the incorporation of different kernels which leads to flexibility in implementation.

In [2], it was stated that GP process can be interpreted with two views: weight-space view and function-space view. A quick discussion regarding GP's hyper-parameters as well as GP sampling function is also included. However, for more details about GPR the reader is referred to [2].

Weight-Space View: The equation from the Bayesian analysis of the standard linear regression model can be written as [2]:

$$f(x_t) = {x_t}^T w \tag{1}$$

$$y = f(x_t) + \epsilon_i \tag{2}$$

where w is the weight vector, ϵ_i is the noise term and it follows a normal distribution in such $\epsilon_t \sim \mathcal{N}(0, {\sigma_\epsilon}^2)$ with 0 mean and ${\sigma_\epsilon}^2$ as the variance from the Bayes' rule. The posterior distribution can be obtained as [7]:

$$p(w|y_t, X_t) \; \alpha \; p(y_t|X_t, w)p(w) \tag{3}$$

$$p(w|y_t, X_t) = \mathcal{N}({\sigma_\epsilon}^{-2} A_t^{-1} X_t y_t, A_t^{-1}) \tag{4}$$

where $A_t = \sum^{-1} + {\sigma_\epsilon}^{-2} X_t X_t^T$, $\sum$ is the covariance matrix and we want to predict y_t for a new input point x_t with the information obtained prior to instant t. Please refer to the original paper for detailed derivation [8].

The form shown above is often referred to as the weight space view of regression [7]. In order to predict the y^* at new point x^*, we can average over all the possible parameter values that are provided by the function f, predicting $f(x^*) = y^* + \epsilon_*$. Again without going into the actual derivation, the predictive distribution with respect to the Gaussian posterior can be written as [7]:

$$p(f(x^*)|x^*, X_t, y_t) = \int p(f(x^*)|x^*, w)p(w|X, y)dw \tag{5}$$

After performing integration equation (5) can be expressed as [7]:

$$p(f(x^*)|x^*, X_t, y_t) = \mathcal{N}(\sigma_\epsilon{}^{-2} x^{*T} A^{-1} X, y, X^{*T} A^{-1} x^*) \tag{6}$$

Weights are first generated from this posterior distribution and the final predictions are generated using the weight generated previously. The term can be generalized from 1-dimensional spacing to higher dimensional space [2]. The model now becomes :

$$f(x) = \phi(x)^T w \tag{7}$$

where $\phi(x) = (1, x, x^2, x^3, ..., x^n)$. The predictive distribution then becomes [2]:

$$p(f(x^*)|x^*, X, y) \sim \mathcal{N}(\sigma_n{}^{-2} \phi(x^*)^T A^{-1} \phi_y, \phi(x^*)^T A^{-1} \phi(x^*)) \tag{8}$$

Function-Space View: Another way to understand the GP algorithm is to focus directly on its distribution over functions [2]. As stated previously, GP algorithm defines a distribution over several functions: if we pick any two (or more) points inside a function, our observations at the selected points follows a joint multivariate Gaussian distribution [9]. In [2], the Gaussian process is defined as a collection of random variables, any finite number of which have a joint Gaussian distribution. Similar to the assumption made in linear regression, we can write the Gaussian Process regression equation as:

$$y = f(x) + \epsilon \tag{9}$$

where the noise term $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$, reflects the randomness or uncertainty of our observation. Based on the definition provided, we can specify a Gaussian Process by its mean function and covariance matrix function, thus Gaussian process can be expressed as follow :

$$f(x) \sim GP(m(x), k(x, x')) \tag{10}$$

where $m(x)$ is the mean function and the $k(x, x')$ is the covariance function (also known as the kernel function) for the randomly selected two points x and x'. Equations 2.1 can be expressed as following :

$$m(x) = \mathbb{E}[f(x)] \tag{11}$$

$$k(x, x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x')] \tag{12}$$

To simplify calculation equation the prior mean function is often set to $m(x) = 0$. This can reduce heavy posterior computations via covariance function [7]. The covariance function, k, is more commonly referred to as the kernel of the GP algorithm [10]. There are many kernel functions available and the choice of which kernel function to use is based on the prior knowledge of the data (i.e. information such as will variable b be effected when variable a is larger,if so to what degree etc.). The choice of the kernel function is also based on factors such as the smoothness and the cycle patterns of the observed values.

Hyper-Parameter-Based Kernels: The hyper-parameters refer to the pre-defined constant terms inside the kernel functions. Since there are many possible kernel functions, there will be different hyper-parameters for each kernel function. A simple example will be given in this report to communicate the idea of hyper-parameters but please bare in mind that this is only one type of kernel function alongside with its hyper-parameters. Readers are encouraged to explore more kernel functions if interested. A very popular kernel function is the radial basis function kernel, or RBF kernel in short [11]. The kernel function can be expressed as the follow:

$$k(x, x') = \sigma_f^2 \exp(-\frac{\|x - x'\|^2}{2\lambda^2}) \tag{13}$$

where $\|\cdot\|$ denotes the euclidean distance. Term x and x' are the two points passed into the kernel function. There are two hyper-parameters inside the radial basis kernel function: λ and σ_f^2. The term λ refers to the length scale while the term σ_f^2 is the data variance of the kernel function. These two hyper-parameters can be increased or decreased to better fit the working dataset. Usually this is an iterative process for user should test out values before finding the most optimal hyper-parameter values for the working dataset. The GP can then be used to draw prior functions once the mean function and the kernel functions are selected.

Sampling From GP: Let X^* be a matrix that contains all the new input points where $x_i^*, i = 1, 2, ..., n$. The kernel function in (13) are constructed for all the pairs between the input points. The expression can be displayed in a matrix form as follow [7]:

$$K(X^*, X^*) = \begin{pmatrix} k(x_1^*, x_1^*) & k(x_1^*, x_2^*) & \cdots & k(x_1^*, x_n^*) \\ k(x_2^*, x_1^*) & k(x_2^*, x_2^*) & \cdots & k(x_2^*, x_n^*) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_n^*, x_1^*) & k(x_n^*, x_2^*) & \cdots & k(x_n^*, x_n^*) \end{pmatrix} \tag{14}$$

Where $k(x_1^*, x_2^*)$ is the kernel function constructed using point x_1^* and x_2^* selected from X^* which contains all input values.
To simplify the equation obtained from (10), the mean function $m(x)$ is set to 0 and (14) is substituted. The following term for a normal distribution is obtained as [7]:

$$f(x^*) \sim \mathcal{N}(0, K(X^*, X^*)) \tag{15}$$

Where the notation $f(x^*)$ represents the samples from the defined function. Our observed values defined in previous section is $D_t = \{(x_i, y_i)|i = 1, 2, ..., n\}$ and we would like to draw new entry X^*'s predictions from function $f(x^*)$ using the posterior distribution. Let x_t (value at instant t) be the value drawn from X^*. Then the matrix form of the distribution can be expressed as follows [7]:

$$\begin{bmatrix} y_t \\ f(x^*) \end{bmatrix} = \mathcal{N}(0, \begin{bmatrix} K(x_t, x_t) + \sigma_\epsilon^2 I & K(x_t, x^*) \\ K(x^*, x_t) & K(x^*, x^*) \end{bmatrix}) \tag{16}$$

where σ is the noise level term and I is the identify matrix. By implementing the Gaussian Identities Theorem for conditional distribution $p(f(x^*)|X_t, y_t, X^*)$ provided in [2], we can rewrite equations (16) and (8) as the following expression:

$$f(x^*)|X^*, x_t, y_t \sim \mathcal{N}\big(m_t(x), k_t(x, x')\big) \tag{17}$$

where the mean function and the kernel function in equations (11) and (12), respectively, can now be expressed as follow [7, 12]:

$$m_t(x) = K(X^*, x_t)[K(x_t, x_t) + \sigma_\epsilon^2 I]^{-1} y_t \tag{18}$$

$$k_t(x, x') = K(X^*, X^*) - K(X^*, x_t)K(X^*, X^*)^{-1}K(x_t, X^*) \tag{19}$$

The sample functions $f(x^*)$ can now be sampled using (18) and (19) stated above.

2.2 Granger Causality

The concept of Granger Causality [13] is commonly adapted in the area of cause-effect relationships in time-series analysis. The Granger causality is a concept that is made out of two fundamental principals that can be summarized as follows:

1. Only the input/intervention from the past can Granger cause the outcome in the future. Future input/intervention cannot Granger cause any past values.
2. If having information about variable A can improve the predictability of variable B, then variable A Granger causes variable B.

The most straight forward method to exterminate Granger causality is the vector autoregressive (VAR) model [14]. A pairwise lag-n VAR model can be written as follow:

$$y_t = \alpha_1 y_{t-1} + ... + \alpha_n y_{t-n} + \beta_1 x_{t-1} + ... + \beta_n x_{t-n} + \mu t + C + \epsilon \tag{20}$$

where μ, C and ϵ are the slope, a constant and noise term, respectively, and y_t is the time-series value of dependent variable at time t, x_{t-i} is the time-series of the independent variable x at time $t - i$, and α and β are their corresponding parameter values, respectively.

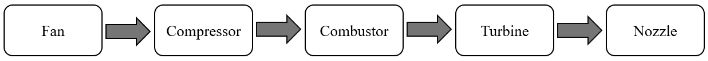

Fig. 2: Simplified diagram for engine components in the PHM08 dataset [15].

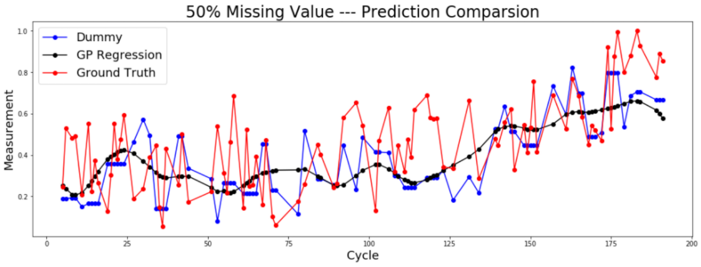

Fig. 3: Comparison between GP Regression prediction, dummy prediction and ground truth value for engine 1 sensor 7 that contains 50% missing values, where the PHM08 dataset is used.

3 Experimental Results

Data Description: In order to validate the performance of recovery of the proposed method, we use a public dataset called the Prognostics and Health Management (PHM08) system dataset [15, 16]. The PHM08 [15] is a turbofan engine degradation simulation dataset created by NASA using the Commercial Modular Aero Propulsion System Simulation Tool (C-MAPSS). The engine is simulated to failure point and the average sensor/operational measurements are recorded for each cycle. Engines inside the training set lasted anywhere from 130 cycles to 362 cycles before failure point.

Although the ground truth data for causality is not available for this PHM08 dataset, it is safe to make the assumption that causality relationships did exist in between these sensor measurements. In real world scenario it is often rare to spot breakdown of a complicated system caused by all intermediate components fail at one instant. It is more common to have breakdown of one component (sensor) which leads to failure of surrounding components and ultimately leads to the malfunctioning of the system. Figure 2 is an illustration of a simplified jet engine diagram. The first 11 engines in the first training set of the PHM08 [15] dataset are selected for this experiment. There are 9 constant sensor readings (with little to no fluctuation) out of the given 24 time-series, thus are neglected for this experiment and the remainder 15 sensor data are used.

Discussion: To evaluate the proposed scheme, as we indicated in Figure 1, the selected data is dropped by 10%, 20%,..., or 80% of its' original entries to simulate an unevenly sparsely sampled time-series data. Gaussian Process Regression[1] is then used to recovery those missing values and finally the recovered multivariate time-series data is feed into the VARs model[2] to calculate the causality parameters. Those parameter values are then compared against the parameter values obtained from the original dataset and the average RMSE, for all the considered engines, values are recorded under each missing value percentage. The same test is repeated using the dummy filling which is to fill in the value with the last seen entry.

Figure 3 shows the comparison between the GP Regression prediction, Dummy model prediction and the ground truth values for engine 1 sensor 7 with 50% missing data. It is clear that GPR is able to follow the changes in the time-series data better than the dummy model. In addition GPR filling is able to provide smoothing effect to reduce noise level. The RMSE values in predicting the causal parameters for the proposed method and the dummy model under different filling percentage are also summarized and plotted in Figure 4. As shown in this figure, the GPR filled data can better preserve the pairwise causal relationships in the original data when compared against the dummy approach.

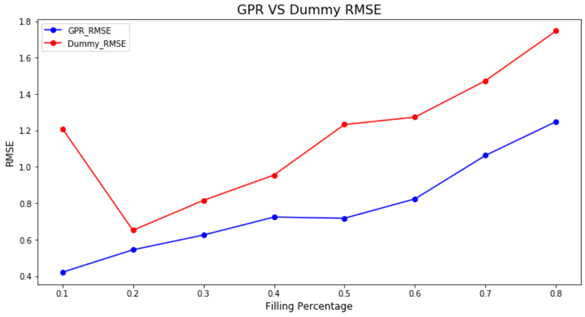

Fig. 4: RMSE Comparison Between GPR Filling and Dummy Filling wrt Pairwise Granger Causality Parameters

4 Conclusion

In this paper, we have studied the ability of Gaussian Process Regression to recover missing time-series data values for the purpose of determining the pairwise Granger causality. The proposed method has been tested by using the PHM08 dataset subjected to different missing value percentages can effect the causal parameter values obtained from pairwise Granger causality. The results show that the Gaussian Process recovered data is better preserved for the pairwise Granger causality relations when compared to those obtained by the dummy filling.

Acknowledgments

The authors would like to thank the Ontario Centres of Excellence (OCE), Canada program, Natural Sciences and Engineering Research Council (NSERC) - Voucher for Innovation and Productivity II (VIP II) - and ATS Automation Tooling Systems Inc., Cambridge, ON Canada, for supporting this research work.

References

[1] I. Pratama, A. Permanasari, I. Ardiyanto, and R. Indrayani, "A review of missing values handling methods on time-series data," 10 2016, pp. 1–6.

[2] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[3] M. Kac and A. J. F. Siegert, "An explicit representation of a stationary gaussian process," *Ann. Math. Statist.*, vol. 18, no. 3, pp. 438–442, 09 1947.

[4] M. P. Deisenroth, "Efficient reinforcement learning using gaussian processes," 2010.

[5] T. M. Mitchell, *Machine Learning*, 1st ed. USA: McGraw-Hill, Inc., 1997.

[6] D. Duvenaud, "Automatic model construction with gaussian processes," Ph.D. dissertation, 11 2014.

[7] E. Schulz, M. Speekenbrink, and A. Krause, "A tutorial on gaussian process regression: Modelling, exploring, and exploiting functions," *Journal of Mathematical Psychology*, vol. 85, pp. 1 – 16, 2018.

[8] C. K. I. Williams, *Prediction with Gaussian Processes: From Linear Regression to Linear Prediction and Beyond*. Dordrecht: Springer Netherlands, 1998, pp. 599–621. [Online]. Available: https://doi.org/10.1007/978-94-011-5014-9_23

[9] J. Bernardo, J. Berger, A. Dawid, A. Smith *et al.*, "Regression and classification using gaussian process priors," *Bayesian statistics*, vol. 6, p. 475, 1998.

[10] F. Jäkel, B. Schölkopf, and F. Wichmann, "A tutorial on kernel methods for categorization," *Journal of Mathematical Psychology*, vol. 51, no. 6, pp. 343–358, Dec. 2007.

[1] GPR function in pymc3 is used `https://github.com/pymc-devs/pymc3`

[2] VARs package in R is used `https://cran.r-project.org/web/packages/vars/vars.pdf`

[11] J. Vert, K. Tsuda, and B. Schölkopf, "A primer on kernel methods," *Kernel Methods in Computational Biology, 35-70 (2004)*, 01 2004.

[12] C. E. Rasmussen and H. Nickisch, "Gaussian processes for machine learning (gpml) toolbox," *Journal of Machine Learning Research*, vol. 11, no. 100, pp. 3011–3015, 2010. [Online]. Available: http://jmlr.org/papers/v11/rasmussen10a.html

[13] C. W. J. Granger, "Investigating causal relations by econometric models and cross-spectral methods," 11 1969.

[14] *Vector Autoregressive Models for Multivariate Time Series*. New York, NY: Springer New York, 2006, pp. 385–429. [Online]. Available: https://doi.org/10.1007/978-0-387-32348-0_11

[15] A. Saxena, K. Goebel, D. Simon, and N. Eklund, "Damage propagation modeling for aircraft engine run-to-failure simulation," *International Conference on Prognostics and Health Management*, 10 2008.

[16] D. Frederick, J. DeCastro, and J. Litt, "User's guide for the commercial modular aero-propulsion system simulation (c-mapss)," *NASA Technical Manuscript*, vol. 2007–215026, 01 2007.