Causal Discovery from Continuous Time Series by Compression

Luofeng Zhou
Wuhan University
Wuhan, China
luofengzhou@whu.edu.cn

Zixuan Cao
Wuhan University
Wuhan, China
zixuancao@whu.edu.cn

Bo Han*
Wuhan University
Wuhan, China
bhan@whu.edu.cn

ABSTRACT

Causal discovery from temporal variables across multiple data sources provides numerous opportunities for scientific discovery and business prediction. The widely used causal discovery method for time series is Granger causality test with a vector autoregressive (VAR) model. However, one single regression model is unable to characterize local segment diversity well over whole temporal sequence. In this paper, we propose an efficient model DISC to discover causal relationship between two continuous time series from the perspective of compression upon the framework of Granger causality. In DISC, compression replaces a VAR regression model to explain the predictability. It characterizes the autocorrelation in time series through a window approach in $O(n)$ time and a weighted approach in $O(n^2)$ time. Both approaches support all kinds of combination of past local segments for prediction, resulting in more accurate causal discovery. Experimental results on both synthetic and real-word data show that the proposed model outperforms Granger causality test in accuracy of inferring causal relations.

CCS CONCEPTS

• **Information systems → Data stream mining**; • **Applied computing → Law, social and behavioral sciences**.

KEYWORDS

Causal Discovery, Continuous Time Series, Compression, Sequential Normalized Maximum Likelihood

ACM Reference Format:
Luofeng Zhou, Zixuan Cao, and Bo Han. 2019. Causal Discovery from Continuous Time Series by Compression. In ,. ACM, New York, NY, USA, 11 pages. https://doi.org/Preprint

1 INTRODUCTION

Causal discovery between time series from different data sources provides numerous opportunities for scientific discovery and business prediction. The widely used causal discovery method for time series is Granger causality test [9]. It asserts that a temporal variable x^t Granger-causes another time series y^t if its past value helps to predict the future value of y^t beyond what could have been done with the past value of y^t only. Within this framework, predictability traditionally is measured in terms of variance of the residues or error terms in a vector autoregressive (VAR) model [4]. The regression model is constructed requiring a predefined lag parameter. The lag is generally given according to prior knowledge or standard information criterion such as BIC (Bayesian Information Criterion). One chosen lag assumes data generating process over whole temporal sequence is stable. In practice, however, only a single process is inadequate to characterize local segment diversity well during the true data generating process. In addition, for ensuring a valid causal relation, a VAR regression model generally works on stationary temporal sequences. Some sequences, such as periodic series, are non-stationary and the potential causality cannot be effectively revealed by the regression model. Therefore, a single lag and assumption of stationarity in regression limit the effective applications of Granger causality test.

Without assumptions about a given lag, [2] proposed a causal inference method CUTE for event sequences by defining causality in terms of compression within Granger causality framework. However, CUTE is only designed for binary variables. Binary variables cannot inadequately express information of highly diversified continuous variables. Additionally, CUTE calculates encoded length in a time point using maximum likelihood over the past sequence in the order from the beginning to the most recent values. The encoded lengths are summed up to form causality statistics. The computation way leads to a self-intensification effect that the earlier information becomes more significant. It is not suitable for many time series where recent temporal points play more important roles than the earlier ones.

Compression involves rich prediction components in information theory. They support partially matching different scales of similar past segments according to a given local context. Compression mechanism provides a more flexible way to fit each local subsequence and can work more accurately as a predictor. In this way, compression offers more powerful capability than the traditional VAR regression model with a fixed lag window. Therefore, combining compression into Granger causality causality framework offers new insights

[*] Corresponding Author.

Preprint, Preprint

ACM ISBN Preprint. . . $Preprint
https://doi.org/Preprint

for measuring predictability in causal discovery. If the involvement of the past values of x^t significantly increases the compression of y^t and not vice versa, it is natural to infer a causal relationship from x^t to y^t since past values of x^t contains unique information in predicting y^t.

Inspired by the above idea, we propose a causal discovery model, DISC (causal **DIS**covery by **C**ompression), for continuous time series through compression. Considering compression works on redundant patterns with finite states, we discretize a continuous variable to multiple states. In this way, compression is feasible while the loss of information in continuous time series is limited. Since the distribution of empirical data sequence is usually unclear, we apply a strategy of sequential normalized maximum likelihood (sNML) to give an alternative to reach the minimum of additional bits (***regret***) to bridge the gap between true distribution and the assumed one. Consequently, prediction can be applied optimally without an assumed data distribution [11, 16]. In order to reflecting the value of neighborhood and mitigate self-intensification effect, we design a window approach to combine segments in a window flexibly for estimation or a weighted approach to give each past value a weight according to their distance to current time point. The two approaches could also be mixed at the same time. The window size here is different from a lag in Granger causality test. For Granger causality test, once a lag is given, all past values within the lagged window should always be taken into consideration for predicting the future. While in DISC model, segments in a window are combined in a selective way according to sNML strategies where the combinations differ from time to time optimally. Thereby, the flexible combination will improve the prediction and result in more accurate causal discovery.

This paper is organized as follows. In Section 2 we give three motivation examples to illustrate several practical challenges in causal inference by Granger causality test and CUTE model. In Section 3, we put forward our DISC model. In Section 4, we validate our encoding method and then test cause-effect pairs on synthetic datasets and three real-world datasets. Section 5 describes the related works. Section 6 draws the conclusion of this paper.

2 MOTIVATION EXAMPLES

During the process of exploring causal relation from time series, we find some typical patterns of sequences where Granger causality test and CUTE model fail.These common patterns in real-world data limit their applications. Here, we illustrate the following three patterns as the motivation examples of our research.

A long sequence occasionally embedded with shifts from another sequence. Figure 1 (a) describes a pattern with occasion noises, which is observable due to signal disturbance in electronics, regulation rule changes in economics, etc. Series 1 is a natural time series. Series 2 has a totally different distribution from series 1, except two noisy points which happen to look like a shift from series 1 in a very short time period. In this situation, both Granger causality test and CUTE model infer a 1 → 2 causal relation which is incorrect through an overall horizon.

Cyclic time series. Figure 1 (b) describes a pattern of cyclic cause-effect pairs. The cause sequence is cyclic. The effect sequence is a right shift of the cause sequence. In this case, CUTE correctly infer their causality. However, Granger causality test is infeasible due to non-stationarity. This pattern is common in many applications.

Continuous time series with different size of amplitude. Figure 1 (c) describes a continuous cause-effect pair with different size of amplitude. This pattern is commonly observable in financial markets, meteorology etc. After detrending, Granger causality test infers correctly from 1 → 2. CUTE infers no causal relation since it cannot distinguish different scopes of directional changes with only binary coding.

3 DISC FRAMEWORK

3.1 Overview

To resolve the challenges described above, we propose a DISC model by using compression to replace VAR regression model within Granger causal framework [9]. Compression is more flexible than a VAR regression model for prediction, resulting in an enhanced causal discovery tool on continuous sequences.

The principal lines of DISC model are as below.

Firstly, we discretize the continuous time series into multiple states sequences according to prior knowledge or the features of data distribution.

Next, we need to provide a robust prediction strategy without knowing the true data distribution. This strategy also takes into account autocorrelation features in time series.

Given the predictions, we will investigate if a sequence of x significantly helps the compression of another sequence. If the improvement is higher than a threshold, we will derive their causal relationship.

The details of the important steps above are explained in Section 3.2.

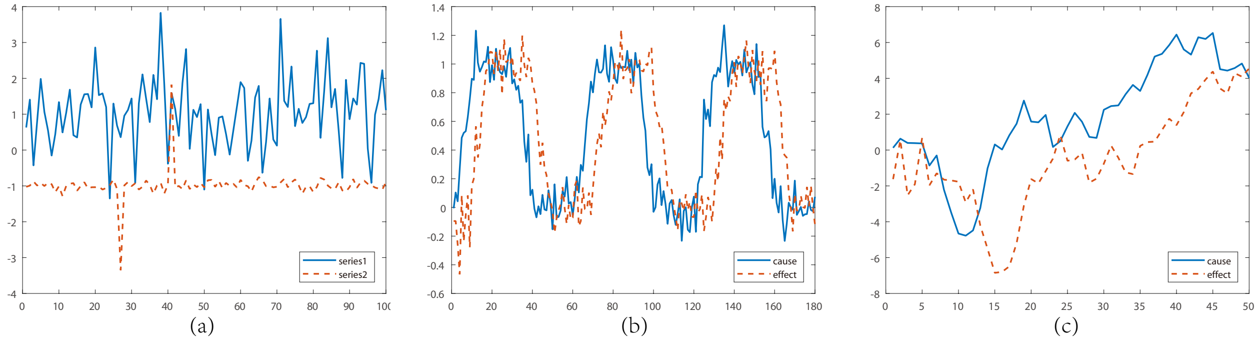
3.2 Preliminaries

3.2.1 Granger Causality. We follow the idea of [9] to define statistical causality as following:

Definition 1. (Granger Causality) *A time series x^t* ***Granger-causes*** *another series y^t if the prediction function holds $P(y_{t+1}, \theta | x^t, y^t) > P(y_{t+1}, \theta | y^t)$.*

This definition concisely reveals the basic ideas of Granger causality: (1) cause precedes the effect in time; (2) cause has unique information for predicting the future values of effect.

3.2.2 Robust Prediction by Normalized Maximum Likelihood. The causal inference in Granger's framework relies on prediction ability improvement by involving additional information. For modeling the prediction function P, the maximum likelihood estimation (MLE) for parameters is computed by solving a normal maximization problem with an assumed data distribution. However, if the assumption of data distribution is not consistent with the real case, the ML strategy



(a) (b) (c)

Figure 1: Three motivation examples. (a) A long sequence occasionally embedded with shifts from another sequence; (b) Cyclic time series; (c) Continuous time series with different size of amplitude.

will result in arbitrarily bad outcomes. Thereby, a robust prediction function should be constructed by minimizing the difference between the true data distribution and the assumed ones, or say, deriving the minimax of ***regret***. In this paper, we use sequential normalized maximum likelihood (sNML) strategy due to its easy calculation [16].

In sNML strategy, the prediction function of x_t under x^{t-1} is:

$$P_{sNML}(x_t\ |x^{t-1}) = \frac{\mathcal{L}(\hat{\theta}_{x^{t-1},x_t}; x^{t-1}, x_t)}{\sum_x \mathcal{L}(\hat{\theta}_{x^{t-1},x}; x^{t-1}, x)}, \qquad (1)$$

where $\hat{\theta}_{x^{t-1},x_t}$ denotes the MLE based on x^{t-1}, x_t and $\hat{\theta}_{x^{t-1},x}$ denotes the MLE based on x^{t-1}, x.

3.2.3 Discount Function. By considering autocorrelation features in time series, we propose two approaches. One approach, called window approach, concerns only the subsequences within a small window. The window length is set through application domain theories or other prior knowledge. The other approach, called weighted approach, applies a discount factor to give more weights on neighborhood subsequences. The discount factor is also exogenously given. The two approaches above can be combined together for characterizing autocorrelation.

Note that this window length or discount factor here is different from the lag operator in Granger causality test to some extent. In DISC model, only the selected subsequences within the window will be used for prediction. These selected subsequences are not necessarily successive. While in Granger test, all consecutive observations within a lag window are taken into account.

In brief, we define discount function as below.

Definition 2. (Discount Function) *The discount function $\delta(d)$ is a multiplier of sequences where d denotes the distance from the multiplied bit to the latest one. The discounted sequence $\delta(d) \cdot x^t$ is described as the weight of x^i is specified as $\delta(t-1)$. When the window approach is selected, $\delta(i) = 0$ for i greater than the exogenously given window length. When the weighted approach is selected, $\delta(d)$ is positive and monotonically decreasing w.r.t. d.*

From the above definition, only neighborhood subsequence is considered by a window approach. Weighted approach gives a decreasing discount function with distance of each timing point. When all weights are set to a constant, the discount function can describe non-discount situation.

3.2.4 Compression. From the perspective of compression, the difference between ideal self-compressed length for y^t and conditional-compressed length on x^t for y^t is recognized as the information contained in x^t which increases the predictability of y^t.

Definition 3. (Ideal Self-Compressed Length) *The ideal self encoded length of x_t is given by: $len^{\delta}(x_t) = -logP_{\hat{\theta}}(x_t|\delta(d) \cdot x^{t-1})$*[1] *where $-logP_{\hat{\theta}}(x_t|\delta(d) \cdot x^{t-1})$ denotes a prediction of x_t under the information of $\delta(d) \cdot x^{t-1}$ and the estimated parameter $\hat{\theta}$. Thus the t-period total self-compressed length of x^t is given by: $len^{\delta}(x^t) = \sum_{t=1}^{T} -logP_{\hat{\theta}}(x_t|\delta(d) \cdot x^{t-1})$.*

Definition 4. (Ideal Conditional-Compressed Length) *The ideal conditional-compressed length of x_t conditional on the information of y^t is given by: $len^{\delta}(x_t|\delta(d) \cdot y^{t-1}) = -logP_{\hat{\theta}}(x_t|\delta(d) \cdot x^{t-1}, \delta(d) \cdot y^{t-1})$. The t-period total conditional-compressed length of x^t conditional on y^t is given by: $len^{\delta}(x^t|\delta(d) \cdot y^{t-1}) = \sum_{t=1}^{T} -logP_{\hat{\theta}}(x_t\ |\delta(d) \cdot x^{t-1}, \delta(d) \cdot y^{t-1})$.*

Specifically, in this paper, the prediction function is given by sNML strategy. $len^{\delta}(y_t)$ shows the predictability by using the past realizations of itself and $len^{\delta}(y_t|\delta(d) \cdot x^{t-1})$ involves the past realizations of another series. Hence, their difference between them measures the extra predictability of y_t uniquely contributed by the realizations of $\delta(d) \cdot x^{t-1}$.

3.2.5 Causal Inference by Compression. After computing self- and conditional-compressed length, x^t contains more unique and useful information for y^t than vice versa, a causal relationship is inferred from $x^t \rightarrow y^t$. So we define the causal dependence as below.

Definition 5. (Causal Dependence based on Compression) *The **causal dependence** from y_t to x_t and vice versa*

[1] All $logs$ in this paper refer to log_2.

are given by

$$\Delta^{\delta}_{y^t \to x^t} = len^{\delta}(x^t) - len^{\delta}(x^t|\delta(d) \cdot y^{t-1}),$$
$$\Delta^{\delta}_{x^t \to y^t} = len^{\delta}(y^t) - len^{\delta}(y^t|\delta(d) \cdot x^{t-1}),$$

Following the framework of Granger causality, we form our belief of causal direction as following:

- If $\Delta^{\delta}_{y^t \to x^t} > \Delta^{\delta}_{x^t \to y^t}$, we infer $x^t \to y^t$.
- If $\Delta^{\delta}_{y^t \to x^t} < \Delta^{\delta}_{x^t \to y^t}$, we infer $y^t \to x^t$.
- If $\Delta^{\delta}_{y^t \to x^t} = \Delta^{\delta}_{x^t \to y^t}$, we are undecided.

The problem is how to determine the statistic threshold to infer causality. Its essence is to determine the probability of wrongly refusing of null hypothesis which is said to be non-causal relationship. [18] gives the answer for the problem.

In the context of our research, the null hypothesis is that there is no causal relationship from a given direction. When given the significant level α, the decision rule is specified as:

- If $\Delta^{\delta}_{y^t \to x^t} - \Delta^{\delta}_{x^t \to y^t} > -log\alpha$, we infer $x^t \to y^t$.
- If $\Delta^{\delta}_{x^t \to y^t} - \Delta^{\delta}_{y^t \to x^t} > -log\alpha$, we infer $y^t \to x^t$.
- If $\left|\Delta^{\delta}_{y^t \to x^t} - \Delta^{\delta}_{x^t \to y^t}\right| < -log\alpha$, we are undecided.

3.3 DISC Model

Following the principal lines discussed above, we state the whole algorithm of DISC model in **Algorithm 1**.

In **Algorithm 1**, i and j denotes the value of next bit of two series respectively. For any given $t_0 \in [0, t]$ and the combinations of series x^{t_0-1} and y^{t_0-1}, we denote the series with smallest discounted sum as $z_0^{t_0-1}$ and the biggest one as z_2^{t-1}. Their corresponding discounted sum is recorded as $\Sigma_0^{\delta} z^{t_0-1}$ and $\Sigma_2^{\delta} z^{t_0-1}$ respectively. The imbalance between 2 and 0 for z_0^{t0-1}, I_0, denotes the difference between the number of 2 in it and that of 0. And I_2, the imbalance between 2 and 0 for $z_2^{t_0-1}$, is calculated through the same procedure. The boolean indicator B shows whether there exists 0-1 pair or 1-2 pair at a same digit for the two series in the past.

There are two main issues in DISC implementation. Firstly, we need to assume distribution functions of each series for prediction. Since sNML strategy is robust for any distributions, we choose a commonly used binomial distribution for sNML implementation. Their computing steps are described in Section 3.3.1. Secondly, we consider the realization of window approach and weighted approach. DISC model allows any combination of subsequences in a window or any weight combination for prediction. It will result in expensive computation cost. In Section 3.3.2, we put forward two efficient algorithms to maximize the conditional sNML strategy.

3.3.1 sNML Implementation. For the sNML strategy for multi-state sequences, we consider binomial distributions for implementation $P_{\theta}(X = k) = C_n^k \theta^k (1-\theta)^{n-k}, k = 0, 1, ..., n$. The computation involves a discount function $\delta(d)$. We deduce **Theorem 1** and **Theorem 2** to compute self- and conditional-compressed sNML strategy.

Algorithm 1 DISC model

```
Input:
    Two time series, $x^t$, $y^t$;
    The discount function, $\delta(d)$ ;
Output:
    The inferred causal relationship between $x^t$ and $y^t$
Determine which approach is selected, denoted as $app$, through $\delta(d)$;
Initialize intermediate variables: $\Sigma_0^{\delta} z^0 = \Sigma_2^{\delta} z^0 = I_0 = I_2 = B = 0$;
Initialize intermediate series: $z_0^0 = z_2^0 = \{\}$;
Set the loop indicator k=0;
while $k < t$ do
    $k+ = 1$;
    $i = x^k$, $j = y^k$;
    Calculate self-compressed sNML strategies;
    Calculate conditional-compressed sNML strategies according to app;
    $z_0^k = z_0^{k-1}, min\{x_k, y_k\}$; $z_2^k = z_2^{k-1}, max\{x_k, y_k\}$;
    Update intermediate variables;
end while
Calculate t-period self- and conditional-compressed length for $x^t$ and $y^t$ respectively and then calculate the causal dependences in Definition 5;
return The inferred causality through decision rule;
```

Theorem 1. *For binomial distribution $P_{\theta}(X = k) = C_n^k \theta^k (1-\theta)^{n-k}, k = 0, 1, ..., n$ and the discount function $\delta(d)$, the self-compressed sNML strategy is the function of $\sum_{j=1}^{t-1} \delta(t-j) x_j := \Sigma^{\delta} x^{t-1}$ but irrelevant to the specific constituent or permutation. To be more specific, denote $\Sigma\delta_{t-1} := \sum_{j=1}^{t-1} \delta(j)$, then the self-compressed sNML strategy for binomial distribution with $P_{\theta}(X = k) = C_n^x \theta^x (1-\theta)^{n-k}, k = 0, 1, ..., n$ is specified as:*

$$P_{sNML}(x_t|\delta(d)x^{t-1}) = \frac{f^{x_t}}{\sum_{i=0}^{n} f^i}, \tag{2}$$

where for any i = 0, 1,. .., n:

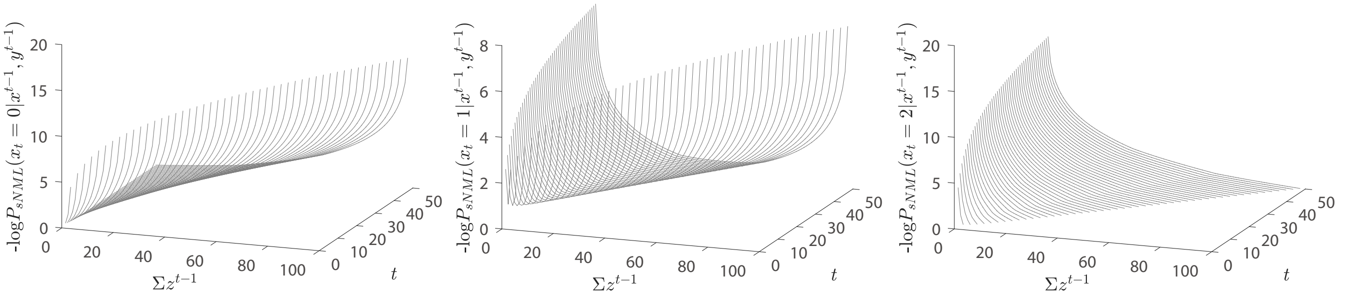
$$f^i = C_n^i (i + \Sigma^{\delta} x^{t-1})^{(i+\Sigma^{\delta} x^{t-1})} \cdot (n(\Sigma\delta_{t-1} + 1) - i - \Sigma^{\delta} x^{t-1})^{(n(\Sigma\delta_{t-1}+1)-i-\Sigma^{\delta} x^{t-1})}.$$

Figure 2 plots the negative logarithm of prediction function given by conditional-compressed sNML strategy for x^t conditional on y^t, which gives intuition showing that the maximaum of prediction function is attained when the difference between mean of the past values and the value of next bit is minimized. A formal theorem which specifies the maximum of conditional-compressed sNML strategy is described and detailed proof is affixed in appendix.

Theorem 2. *The maximum of conditional-compressed sNML strategy $P_{sNML}(x_t = i|\delta(d)x^{t-1}, \delta(d)y^{t-1})$ is attained by the series which satisfies:*

$$\Sigma_j^{\delta} z^{t-1} = \min_{z_i \in \{x_i, y_i\}} \sum_{i=1}^{t-1} \delta(t-i) \cdot (z_i - j) + \Sigma^{\delta} \cdot j. \tag{3}$$

Proposition 1. *The conditional-compressed sNML strategy for binomial distribution: $P_{\theta}(X = k) = C_n^k \theta^k (1-\theta)^{n-k}$,*



(a) (b) (c)

Figure 2: The conditional-compressed sNML strategy for x^t conditional on y^t when the current x_t equals to 0 (a), 1 (b) and 2 (c). The x-axis is the sum of z_i where $z_i \in \{x_i, y_i\}$ and the y-axis is the past length.

$k = 0, 1, ..., n$ *is characterized as:*

$$P_{sNML}(x_t = i|\delta(d)x^{t-1}, \delta(d)y^{t-1}) = \frac{f_i^i}{\sum_j f_i^j}, \qquad (4)$$

where for any i = 0, 1,. .., n:

$$f_j^i = C_n^i(i+\Sigma_j^\delta z^{t-1})^{(i+\Sigma_j^\delta z^{t-1})} \cdot (n(\Sigma\delta_{t-1}+1) - i - \Sigma_j^\delta z^{t-1})^{(n(\Sigma\delta_{t-1}+1)-i-\Sigma_j^\delta z^{t-1})},$$

$$\Sigma_j^\delta z^{t-1} = \min_{z_i \in \{x_i, y_i\}} \sum_{i=1}^{t-1} \delta(t-i) \cdot (z_i - j) + \Sigma^\delta \cdot j.$$

Since in practice the function f^i is always too big to be calculated forthright, a *log-sum-exp* is applied to solve the overflow error.

Note that the theorems of this paper are the more general form of that for CUTE model.

However, when calculating the conditional-compressed s-NML strategy, things become a little different. In binary series, only one time of scanning is needed to calculate the optimal conditional-compressed sNML strategy through the calculation of XOR. When it comes to multi-state data, the corresponding calculate is totally a different story from binary sequences in which an XOR operator can easily solve this problem. However, if we use enumerative method, the time complexity is exponential. So the maximum likelihood of the intermediate situation should be well identified.

3.3.2 *Realizing Conditional-Compressed sNML strategy.*

From now on we assume the parameter of binomial distribution n to be 2 and consider ternary data.

Window Approach and Non-Discount Situation. Let us begin from the simplest case: window approach situation. When the window length is set to be the length of the whole sequence, it becomes the non-discount situation. So an algorithm is designed for both cases. The core idea of this algorithm is to make full use of outcomes from previous steps to simplify the time complexity, described in the next section in detail.

First, we focus on the simplest case where the discount function remains a constant. **Algorithm 2** gives specific actions to determine whether the optimal sum is attainable for any given $t_0 \in [0, t]$. If the feasible interval does not include the unconstrained optimal solution, then the optimal solution within the feasible interval is one of the corner points. However, even if when the unconstrained optimal solution falls into the feasible interval, it may not always be feasible too when the difference of lower-bound and the optimal solution is odd but all the feasible incremental chances give only an even increase. The algorithm is telling exactly the same story in a formal and logical way.

Algorithm 2 Conditional-Compressed sNML Prediction for Window Approach and Non-Discount Situation

```
Input:
    The next bit of the series, $x_{t_0} = i$;
    The past discounted sum interval, $[\Sigma_0^\delta z^{t_0-1}, \Sigma_2^\delta z^{t_0-1}]$ ;
    Corresponding boundary sequences $z_0^{t_0-1}$ and $z_2^{t_0-1}$;
    Imbalance between 2 and 0 for two series, $I_0$, $I_2$;
    Boolean indicator, $B$;
Output:
    $P_{sNML}(x_{t_0} = i|\delta(d)x^{t_0-1}, \delta(d)y^{t_0-1})$;
if i=0 or i=2 then
    $\Sigma^\delta z^{t_0-1} = \Sigma_i^\delta z^{t_0-1}$;
else if $I_0 \cdot I_2 > 0$ then
    $\Sigma^\delta z^{t_0-1} = \arg\min_{\Sigma_i} |I_i|$;
else if $I_0$ is even or $B$ =1 then
    $\Sigma^\delta z^{t_0-1} = \Sigma\delta_{t_0-1}$;
else
    $\Sigma^\delta z^{t_0-1} = \Sigma\delta_{t_0-1} + 1$;
end if
Calculate $P_{sNML}(x_{t_0} = i|\cdot)$ as a function of $\Sigma^\delta z^{t_0-1}$
    through Theorem 1;
return $P_{sNML}(x_{t_0} = i|\delta(d)x^{t_0-1}, \delta(d)y^{t_0-1})$;
```

When it comes to the case where window approach is applied to DISC model, it can be seen as a certain step in non-discount situation. If the length of window is chosen as l, then each step can be regarded as the l^{th} step in the case described above.

Weighted Approach. For continuous discount method, since different subsequence has different weights, it is hard to utilize the previous outcome of calculation directly. Intuitively, for different kind of data, the discount functions vary a lot. When the next bit is 0 or 2, the sNML strategy is easily applied through **Theorem 1**. For a given discount function $\delta(d)$, since the problem becomes a *knapsack problem*, a practical greedy algorithm is applied to get the suboptimal sNML strategy for the next bit equals to 1 with lower time complexity. **Algorithm 3** describes the detailed procedure for any given $t_0 \in [0, t]$.

Algorithm 3 Conditional-Compressed sNML Prediction for Weighted Approach

```
Input:
    The next bit of the series, $x_{t_0} = i$;
    The past discounted sum interval, $[\Sigma_0^\delta z^{t_0-1}, \Sigma_2^\delta z^{t_0-1}]$ ;
    Corresponding boundary sequences $z_0^{t_0-1}$ and $z_2^{t_0-1}$;
    The difference sequence for two sequences, $d^{t_0-1}$ ;
Output:
    $\tilde{P}_{sNML}(x_{t_0} = i|\delta(d)x^{t_0-1}, \delta(d)y^{t_0-1})$;
if $i = 0$ or $i = 2$ then
    $\Sigma^\delta z^{t_0-1} = \Sigma_i^\delta z^{t_0-1}$;
else if $I_0 \cdot I_2 > 0$ then
    $\Sigma^\delta z^{t_0-1} = \arg\min_{\Sigma_i} |I_i|$;
else
    Solve the knapsack problem with $\delta(d)z_0^{t_0-1}$ according to $d^{t_0-1}$ and the constrain $\Sigma\delta_{t_0-1} - \Sigma_0^\delta z^{t_0-1}$, $\Sigma_-$;
    Solve the knapsack problem with $\delta(d)z_1^{t_0-1}$ according to $d^{t_0-1}$and the constrain $\Sigma_2^\delta z^{t_0-1} - \Sigma\delta_{t_0-1}$, $\Sigma_+$;
    $\Sigma^\delta z^{t_0-1} = \arg\min_{\Sigma_j} |\Sigma\delta_{t_0-1} - \Sigma_j^\delta z^{t_0-1}|$, $j = +, -$;
end if
Calculate $\tilde{P}(x_{t_0} = i|\cdot)$ as a function of $\Sigma^\delta z^{t_0-1}$ through Theorem 1 and update input variables;
return $\tilde{P}_{sNML}(x_{t_0} = i|\delta(d)x^{t_0-1}, \delta(d)y^{t_0-1})$;
```

3.3.3 Weighted Window Approach. The window approach and weighted approach can be combined together to form a weighted window approach. This combination makes intuitive sense since the discount function in weighted approach for too early information has little impact to the prediction function. This approach can be applied through simple combination of **Algorithm 1** and **Algorithm 2**.

3.4 Time Complexity Analysis

To put the DISC model into practice, we should calculate self- and conditional-compressed length respectively. To compute self-compressed length $len^\delta(x^t)$, the self-compressed sNML strategy is first imposed. According to **Theorem 1**, we can calculate it in constant time if we know the sum within the given window or the whole previous series, which is realized through temporal variables and these variables are updated for each step. So we compute self-compressed length in $O(n)$ time. So we discuss the time complexity for the calculation of conditional-compressed length for different situations.

3.4.1 Window Approach and Non-Discount Situation. To compute conditional-compressed length $len^\delta(x^t|\delta(d) \cdot y^{t-1})$, the conditional-compressed sNML strategy is given by **Algorithm 2**. If we know all the inputs we can calculate it in constant time. In the same way, we use extra space to store these values and arrays and update them every time the current step ends. The updating also just costs constant time. Therefore, we can compute conditional-compressed length in $O(n)$ time.

In summary, the worst-case time complexity is $O(n)$ for window approach and non-discount situation.

3.4.2 Weighted Approach. In this case, the time complexity of conditional-compressed length depends through **Algorithm 3**. When the next bit is 0 or 2, the complexity reduces to $O(n)$. However, if the next bit is 1, the problem changes into *knapsack problems* or *subset sum problems* whose optimality cannot be achieved in linear time complexity. To be more specific, the time complexity of dynamic programming algorithm for these problems in this paper is $O(n^3)$. Thus we apply greedy algorithm instead to simplify the calculations. In this way, the time complexity reduces to $O(n^2)$. Note that for special discount functions such as exponential discount function, there exist easy algorithms since the outcome of the last application can be reused.

In summary, the worst-case computational complexity is $O(n^2)$ for weighted approach.

3.4.3 Weighted Window Approach. Due to the high time complexity of weighted approach, in practice, it can reduce computational cost by conducting weighted window approach. Because the knapsack problems are computed within just a short window in which the computing time is $O(1)$, for the whole compression process the time complexity reduces to $O(n)$.

Figure 3 illustrates the time complexity of DISC model for the three approaches.

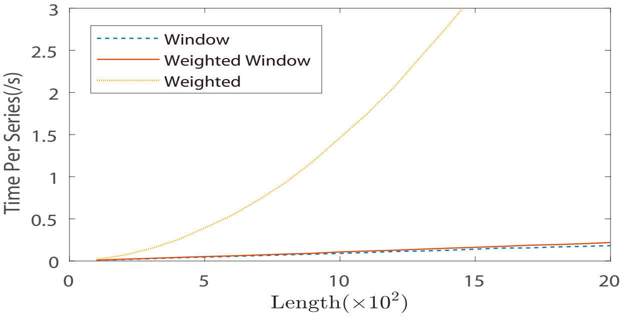


Figure 3: Time Complexity of DISC model

4 EXPERIMENTAL RESULTS

4.1 Experimental settings

Our source code and datasets are public and details are included in appendix. We implemented DISC model in Python and provide the source code for research purposes along with the datasets and synthetic dataset generator. All experiments were executed on a laptop computer with an Intel Core i7 1.8GHz CPU and 16 GB main memory.

We compare DISC model with Granger causality test and CUTE model. For Granger causality test, the lag operator in VAR is selected through prior knowledge for testing accuracy towards synthetic cause-effect and AIC is applied otherwise. For CUTE model, it receives stable results in binary data. However, if it is applied without a window discount to binary data transferred from continuous sequences, CUTE cannot achieve good results because the autocorrelation effect has not been taken into consideration. So for the rest of this paper, we use window approach to modify original CUTE model to become an improved CUTE model.

In the following paper, ***G-test*** denotes the Granger causality test under a chosen VAR model. ***iCUTE*** denotes the improved CUTE model.

In order to control false discovery rate for our repeated experiments in DISC model and improved CUTE model, we follow [2] and apply the most widely used *Benjamini-Hochberg procedure* [1]. Let $H_1, ..., H_m$ be the null hypotheses tested, and their corresponding p-values $p_1, ..., p_m$. First sort these p-values in ascending order. For a significance level of α, we reject the null hypothesis or all H_i, where $i = 1, ..., k$ for the largest k that $p_k \leq k/m \cdot \alpha$.

4.2 Validation on Encoding Method

In our experiment, we choose multi-state discretization to encode continuous data because our basic concern lies on the probably non-informative events. Our DISC model gives a third intermediate state between upward state and downward state. Take CUTE model for example, the encoding method is obviously problematic since it cannot distinguish a drastic increase (e.g. 100%) and a slight increase (e.g. 0.01%). Granger causality test, to another extreme, considers the causality linearly. When the regression function is badly specified, the results may be misleading. Our DISC model falls in between. Although we admit that slight oscillation could be non-informative, the drastic changes tell a rather different story. According to this belief, the encoding process should base on the distribution of the dataset and what kind of impact is concerned informative or non-informative. In synthetic data, since the distributions are Gaussian for both sequences, we choose thresholds -0.68σ and 0.68σ of Gaussian distribution to let the ratio of three states be $1 : 2 : 1$.

In order to ascertain the coding method of our model does not change the causality of the original time series, we do the following simulation experiments. Unless otherwise stated, we choose the window size from $[6, 9]$ and the significance level as 0.05. For each synthetic experiment, we generate 1000 samples. In the case of different lengths, we test the Granger causality before and after the encoding process and the corresponding model (DISC for ternary and iCUTE for binary) with the synthetic data of the most obvious case (linear, noise 0). In addition, although **equation (1)** guarantees the probability with null hypothesis falsely rejected to be no more than any given significance level, the difference from continuous distributions and discrete distributions could also bias the effectiveness. So the robustness towards non-causal series is also tested. Results are shown in Figure 4.

The results show that the Granger causality test is consistent before and after ternary coding, which shows that the coding method does not change the intrinsic causality. Plus, the consistency between Granger causality test and DISC model is also salient. However, binary coding method for iCUTE model is not that promising for non-causal series.

4.3 Results on Synthetic Data

In general, we generate synthetic simple linear cause-effect pairs through a single standard Gaussian distribution. The cause series, x^t, is just the copy of the random generated standard Gaussian distribution with a certain length and the effect series, y^t, is the generated by $y_t = x_{t+L}$ where L is an exogenous scalar indicating the lag operator. The first L digits of the effect series are randomly chosen through standard Gaussian distribution.

When it comes to nonlinear monotonic cause-effect pairs, we test two kinds of nonlinear and monotonic function $f(x) = tanh(x)$ and $g(x) = e^x$ which is widely used in many subjects. In each case, the nonlinear monotonic function is applied to the effect series only and the causality is tested under three models.

We test 3 models with length 150, 250, 350 and 450. At each length, the noise-free case is tested and noises of 0.1, 0.2 and 0.3 are applied to the time series respectively. Since the outcome of two discount approaches are similar, we always display results of window approach.

The result can be seen in Figure 3. We can see that our model and improved CUTE perform similarly and perform better than Granger causality test in longer time series. But Granger causality test performs better with shorter and noisier time series.

4.4 Results on Real-World Data

We apply three data sources of cause-effect pairs to examine the performance of DISC model.

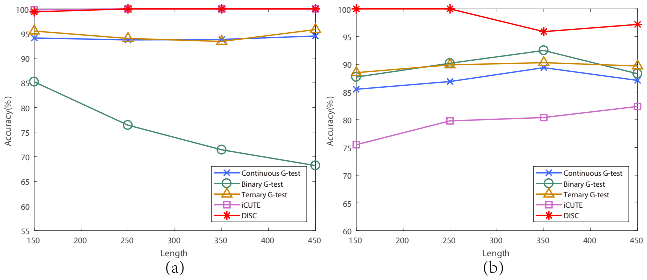


Figure 4: The consistency before and after encoding. (a) cause-effect pairs. (b) non-causal series.

Table 1: Accuracy (%) on Synthetic Cause-Effect Pairs

Type		Simple Linear				Nonlinear Monotonic: tanh				Nonlinear Monotonic: exp			
Model	Noise	t=150	t=250	t=350	t=450	t=150	t=250	t=350	t=450	t=150	t=250	t=350	t=450
G-test	0	95.3	94.3	93.8	95.9	94.6	95.3	96.8	95.4	95.3	95.4	95.1	95.1
iCUTE	0	99.9	100	100	100	99.6	100	100	100	99.9	100	100	100
DISC	0	99.3	100	100	100	99.8	100	100	100	99.6	100	100	100
G-test	0.1	95.4	93.9	94.8	95.2	95.5	95.9	95.9	95.3	93.8	94.1	94.1	94.2
iCUTE	0.1	97.7	99.9	100	100	97.3	100	100	100	98.4	100	100	100
DISC	0.1	98.1	100	100	100	97.7	100	100	100	98.6	99.9	100	100
G-test	0.2	96.2	95.8	95	94.3	96.1	94.3	94.9	96.1	94	94.7	95.3	96
iCUTE	0.2	95	98.5	99.3	100	94.8	98.6	99.7	99.8	95.3	98.5	99.3	100
DISC	0.2	93.8	99.7	99.9	100	93.8	99.3	99.9	100	95	99.6	99.9	100
G-test	0.3	94.4	95.7	94.3	95.3	93.4	95.3	96.2	95.8	94.4	95.8	95.5	95.2
iCUTE	0.3	87.6	93.8	96.8	99.2	89.5	94.8	97.4	98.9	88.3	94.3	97	98.4
DISC	0.3	87.2	97.4	99.4	100	85.9	97.5	99.6	99.6	86.7	97.8	99.3	99.9

River data is applied in [2]. This dataset collects water level data for several monitoring points in the two rivers of Saar and Rhein, and the ground truth is *upstream* → *downstream*.

Meteorolog and Environment data is retreived from Tübingen cause-effect benchmark pairs[2], in which datasets are tagged with the corresponding ground truth. We use 4 cause-effect pairs to test the DISC model. Three pairs are about meteorology and the ground truth is *temperature* → *ozone*. Another pair is about environment and the ground truth is *outdoor* → *indoor*.

Engineering data describes a mechatronic engineering scenario, i.e. the mechanical arm, where the input is a signal in degrees and outputs represent the displacement on both sides of the torsion spring. Noises are caused by micro-motion of user's muscles.We use Simulink in Matlab to simulate and the generator is provided in our source code. The ground truth is *input* → *output1,2*.

Table 2 gives an overview of all the empirical results on real-world data for each model.

Table 2: Empirical Results on Real World Data

Application Field	Samples	G-test	CUTE	iCUTE	DISC
River	9	4/9√	8/9√	8/9√	8/9√
Meteorology	3	2/3√	1/3√	1/3√	√
Environment	1	×	√	√	√
Engineering	2	×	×	√	√

River Data. It can be seen that the Granger causality test incorrectly identifies the causal direction, and our model is completely consistent with the results of the CUTE model. This proves the consistency of our model with the CUTE model.

Meteorology Data. DISC model correctly tells the causality from temperature to ozone. Granger causality test makes only one correct judgment. CUTE and iCUTE tell the correct direction for two cases.

Environment Data. Only Granger causality test indicates wrong causal relationship between indoor and outdoor temperature.

Mechatronic Engineering Data. Mechatronic engineering data are shown in Figure 5. Since the lag between input signal and output signals is obviously less than 100, we choose the window size as 100.

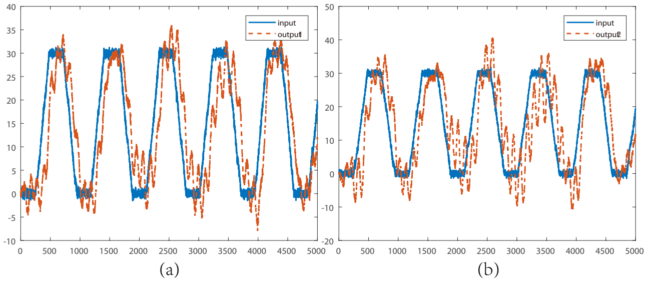


Figure 5: Mechatronic Causality: one noisy input causes the change of two output series: (a) and (b).

Despite of noises, our model still makes correct inferences. But at this time, CUTE makes mistakes and infers $y_2^t \to x^t$. As these sequences are cyclic and thus nonstationary, Granger causality test gives a bidirectional causality judgment between x^t and either y_1^t or y_2^t.

In summary, by discretizing the continuous data, interference of noise on causal inference can be reduced while at the same time the original time series information is well kept. DISC model performs better than other methods in both synthetic data and real-world data.

5 RELATED WORKS

Granger causality test. [22] and [9] introduced the basic notion of statistical-based causality, also known as Granger causality, to study the cause-effect relationships between time series. [20] began to use the lag of one series to predict another time series and [4] extended its idea and defined the Granger

[2] http://webdav.tuebingen.mpg.de/cause-effect/

causality to vector autoregressive (VAR) regression models. Although some recent theoretical works focus on nonlinear and nonparametric test for Granger causality ([12–14, 21]), there are also a lot of causal discovery research using linear specification in VAR forms. For example, the temporally subsampled [6] and aggregated [7] time series and mixed frequency time series [5, 8] are investigated in VAR regression models. Our study falls in the Granger causality framework, but applies compression to replace VAR regression model for improving predictability.

Compression and Prediction. Data compression which encodes information into fewer bits is closely related to data prediction. [3] draw techniques from data compression to form a theoretical basis for a prediction task. [17] addresses the problem of online prediction for time series by compression-based methods. Recently, [19] summarized that how data compressors can be used for solving prediction problems in time series. These researches reveal compression techniques can work for prediction. In fact, many compressors consist of predictors. [10] show that a gaze predictor can help to improve the sequence compression effects. [15] improve compression rates by using adaptive block differential prediction. Thereby, there exists intimacy between compression and prediction. The compression methods can apply multiple coding or prediction method to increase the compression ratios and improve predictability. They can dynamically capture local segment patterns along sequence and provide more flexible and accurate prediction than a VAR regression model with a fixed lag window. This is the motivation why we apply compression to replace a regression model within Granger causality framework.

Causal discovery based on Compression. [2] put forward a causality test model CUTE through ideal encoded (compressed) length to analyze causality on event sequence. Although his idea is novel, CUTE model is only designed for binary causal inference. To our best knowledge, our proposed DISC model is among the very few works that combine compression into Granger causality framework on causal discovery from continuous time series, which is closer to real-world causal discovery nature.

6 CONCLUSION

In this paper, we propose a DISC model for causal discovery from continuous time series through compression upon Granger causality framework. It takes advantages of all combinations of subsequences for prediction by optimization with a discount function. The prediction dynamically captures local segment patterns along sequence and provide more flexible and accurate prediction than a VAR regression model with a fixed lag window. Next we compute the compression length through sNML predictions and derive the causal relations. We test the DISC model on synthetic and real-world datasets. The experiments show that DISC model outperforms Granger causality test and CUTE in accuracy.

In practice, an effect time series may be driven by multiple cause time series. At some interval, one cause-effect causal relation becomes notable while disappears in other intervals. In the future, we are interested in exploring a model for discovering dynamically changing and mixed causality.

REFERENCES

[1] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)* (1995), 289–300.

[2] Kailash Budhathoki and Jilles Vreeken. 2018. Causal Inference on Event Sequences. In *Proceedings of the 2018 SIAM International Conference on Data Mining*. SIAM, 55–63.

[3] I-Cheng K Chen, John T Coffey, and Trevor N Mudge. 1996. Analysis of branch prediction via data compression. *ACM SIGPLAN Notices* 31, 9 (1996), 128–137.

[4] Jean-Marie Dufour and Eric Renault. 1998. Short run and long run causality in time series: theory. *Econometrica* (1998), 1099–1125.

[5] Eric Ghysels, Jonathan B Hill, and Kaiji Motegi. 2016. Testing for Granger causality with mixed frequency data. *Journal of Econometrics* 192, 1 (2016), 207–230.

[6] Mingming Gong, Kun Zhang, Bernhard Schoelkopf, Dacheng Tao, and Philipp Geiger. 2015. Discovering temporal causal relations from subsampled data. In *International Conference on Machine Learning*. 1898–1906.

[7] Mingming Gong, Kun Zhang, Bernhard Schölkopf, Clark Glymour, and Dacheng Tao. 2017. Causal discovery from temporally aggregated time series. In *Uncertainty in artificial intelligence: proceedings of the... conference. Conference on Uncertainty in Artificial Intelligence*, Vol. 2017. NIH Public Access.

[8] Thomas B Götz, Alain Hecq, and Stephan Smeekes. 2016. Testing for Granger causality in large mixed-frequency VARs. *Journal of Econometrics* 193, 2 (2016), 418–432.

[9] Clive WJ Granger. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society* (1969), 424–438.

[10] Sigrid Klerke, Yoav Goldberg, and Anders Søgaard. 2016. Improving sentence compression by learning to predict gaze. *arXiv preprint arXiv:1604.03357* (2016).

[11] Wojciech Kotlowski and Peter Grunwald. 2012. Sequential normalized maximum likelihood in log-loss prediction. In *Information Theory Workshop (ITW), 2012 IEEE*. IEEE, 547–551.

[12] Daniele Marinazzo, Mario Pellicoro, and Sebastiano Stramaglia. 2006. Nonlinear parametric model for Granger causality of time series. *Physical Review E* 73, 6 (2006), 066216.

[13] Daniele Marinazzo, Mario Pellicoro, and Sebastiano Stramaglia. 2008. Kernel method for nonlinear Granger causality. *Physical Review Letters* 100, 14 (2008), 144103.

[14] Yoshihiko Nishiyama, Kohtaro Hitomi, Yoshinori Kawasaki, and Kiho Jeong. 2011. A consistent nonparametric test for nonlinear causalitySpecification in time series regression. *Journal of Econometrics* 165, 1 (2011), 112–127.

[15] Cristian Perra. 2015. Lossless plenoptic image compression using adaptive block differential prediction. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 1231–1234.

[16] Teemu Roos and Jorma Rissanen. 2008. On sequentially normalized maximum likelihood models. In *Workshop on Information Theoretic Methods in Science & Engineering*.

[17] Boris Ryabko. 2009. Compression-based methods for nonparametric prediction and estimation of some characteristics of time series. *IEEE Transactions on Information Theory* 55, 9 (2009), 4309–4315.

[18] Boris Ryabko and Jaakko Astola. 2005. Application of data compression methods to hypothesis testing for ergodic and stationary processes. In *International Conference on Analysis of Algorithms DMTCS proc. AD*, Vol. 399. 408.

[19] Boris Ryabko, Jaakko Astola, and Mikhail Malyutov. 2016. *Compression-based methods of statistical analysis and prediction of time series*. Springer.

[20] Christopher A Sims. 1972. Money, Income, and Causality. *The American Economic Review* 62, 4 (1972), 540–552.

[21] Abderrahim Taamouti, Taoufik Bouezmarni, and Anouar El Ghouch. 2014. Nonparametric estimation and inference for conditional density based Granger causality measures. *Journal of Econometrics* 180, 2 (2014), 251–264.

[22] Norbert Wiener. 1956. The theory of prediction. *Modern Mathematics for Engineers* (1956).

A REPRODUCIBILITY

Our source code has been uploaded to a GitHub account without any personal information. The experimental datasets are also included.

The hyper-link is: https://github.com/EZnlp3aUSfR8R7Ny/Causal-Discovery-from-Continuous-Time-Series-by-Compression.

See README.md for more details.

B PROOF OF CORE THEOREMS

B.1 Proof of Theorem 1

Proof. For any x = 0, 1,. .., n, the likelihood function of x^{t-1}, x is :

$$\mathcal{L}(\theta; x^{t-1}, x) = \prod_{i=0}^{n} [C_n^i \theta^i (1-\theta)^{n-i}]^{t_i} \cdot C_n^x \theta^x (1-\theta)^{n-x}.$$

To maximize this likelihood function, the first order condition is applied and the following equation is drawn as:

$$\frac{x + \sum_{i=0}^{n} i t_i}{\hat{\theta}_{x^{t-1},x}} = \frac{n - x + \sum_{i=0}^{n} (n-i) t_i}{1 - \hat{\theta}_{x^{t-1},x}}.$$

To simplify the above equation,

$$\hat{\theta}_{x^{t-1},x} = \frac{x + \sum_{i=0}^{n} i t_i}{nt},$$

Then the likelihood function in sNML is:

$$\begin{aligned} P_{sNML}(x|x^{t-1}) &= \prod_{i=0}^{n} [C_n^i \hat{\theta}^i (1 - \hat{\theta}^{n-i})]^{t_i} \cdot C_n^x \hat{\theta}^x (1-\hat{\theta})^{n-x} \\ &= \frac{\prod_{i=0}^{n} [C_n^i]^{t_i}}{(nt)^n} C_n^x (x+\Sigma)^{x+\Sigma} (nt - x - \Sigma)^{nt-x-\Sigma} \\ &= \frac{\prod_{i=0}^{n} [C_n^i]^{t_i}}{(nt)^n} f^x. \end{aligned}$$

Now the unconditional sNML strategy can be calculated directly through **equation (1)**:

$$\begin{aligned} P_{sNML}(x_t \,|x^{t-1}) &= \frac{\mathcal{L}(\hat{\theta}_{x^{t-1},x_t}; x_t)}{\sum_x \mathcal{L}(\hat{\theta}_{x^{t-1},x}; x)} \\ &= \frac{f^{x_t}}{\sum_{i=0}^{n} f^i}. \end{aligned}$$

□

B.2 Proof of Theorem 2

First, we prove the following lemma which is simpler but sufficient for situation in this paper. For n greater than 2 and for discount function not always to be a constant, since the proof of **Lemma 1** is from continuous analysis, the sequential maximization problem remains similar procedure of being proved.

Lemma 1. *For ternary series, the maximum of conditional-compressed sNML strategy $P_{sNML}(x_t = 0|x^{t-1}, y^{t-1})$ is attained by the series which preserve the minimal weighted sum Σx^{t-1}; the maximum of conditional-compressed sNML strategy, $P_{sNML}(x_t = 2|x^{t-1}, y^{t-1})$ is attained by the series which preserve the maximal weighted sum; the maximum of conditional-compressed sNML strategy $P_{sNML}(x_t = 1|x^{t-1}, y^{t-1})$ is attained by the series which preserve the weighted sum closest to $t-1$.*

Proof. There are three situations when measuring sNML strategy for $n = 2$ according to the value of the present bit.

- **Situation 1**: $x^t = 0$
- **Situation 2**: $x^t = 1$
- **Situation 3**: $x^t = 2$

First we prove **Situation 1**. Assume $F(\lambda, t) = (t+\lambda)^{(t+\lambda)}(t-\lambda)^{(t-\lambda)}$. For a given t, denote $f(\lambda) = F(\lambda, t)$ and $f(\lambda)$ is an even function. The we take first order derivative as follow:

$$\begin{aligned} g_1(\lambda) =: & \frac{d}{d\lambda} \frac{f(\lambda+1)}{f(\lambda)} \\ &= \frac{f'(\lambda+1)}{f(\lambda)} - \frac{f'(\lambda) \cdot f(\lambda+1)}{(f(\lambda))^2} \\ &= \frac{f(\lambda+1)}{f(\lambda)} (\frac{d}{d\lambda} lnf(\lambda+1) - \frac{d}{d\lambda} lnf(\lambda)) \\ &= \frac{f(\lambda+1)}{f(\lambda)} (ln(t^2 - (\lambda+1)^2) - ln(t^2 - \lambda^2)). \end{aligned}$$

When λ equals to -0.5, this derivative equals to 0. When λ is greater than -0.5, the derivative is negative and vice versa.

For the same reason we have:

$$\frac{d}{d\lambda} \frac{f(\lambda+2)}{f(\lambda)} = \frac{f(\lambda+2)}{f(\lambda)} (ln(t^2 - (\lambda+2)^2) - ln(t^2 - \lambda^2)).$$

When λ equals to -1, this derivative equals to 0. When λ is greater than -1, the derivative is negative and vice versa.

Here the optimization problem is changed into minimizing:

$$G_0(\lambda) = 2 \cdot \frac{f(\lambda+1)}{f(\lambda)} + \frac{f(\lambda+2)}{f(\lambda)}.$$

Through direct derivative we have:

$$(\lambda + \frac{3}{4}) \cdot G_0'(\lambda) < 0 \; for \; |\lambda + \frac{3}{4}| \geq \frac{1}{4}.$$

Then :

$$\begin{aligned} \lambda^* &= arg \min_{\lambda \in N} G_0(\lambda). \\ &= arg \min_{\lambda \in \Lambda} G_0(\lambda), \end{aligned}$$

where $\Lambda = \{-t, -1, 0, t-2\}$.

Because $f(t) > f(t-2) > 0$, we have:

$$\begin{aligned} G_0(-t) - G_0(t-2) = & 2f(t-1) \cdot (\frac{1}{f(t)} - \frac{1}{f(t-2)}) \\ & + (\frac{f(t-2)}{f(t)} - \frac{f(t)}{f(t-2)}) < 0, \end{aligned}$$

And:

$$G_0(-1) - G_(0) = 2 \cdot (\frac{f(0)}{f(1)} - \frac{f(1)}{f(0)}) + (1 - \frac{f(2)}{f(0)}) < 0,$$

$$G_0(0) > G_0(-1) > G_0(-t).$$

Hence $\lambda^* = -t$.

According to duality, in **Situation 3** $\lambda^* = t$. Next we prove **Situation 2**.

$$\begin{aligned} g_2(\lambda) :=& \frac{d}{d\lambda}\frac{f(\lambda-1)}{f(\lambda)} \\ =& \frac{f(\lambda-1)}{f(\lambda)}(ln(t^2-(\lambda-1)^2)-ln(t^2-\lambda^2)). \end{aligned}$$

When λ equals to 0.5, this derivative equals to 0; when λ is greater than 0.5, the derivative is positive and vice versa.

Here our aim is equivalent to minimize:

$$G_1(\lambda) = \frac{f(\lambda+1)}{f(\lambda)} + \frac{f(\lambda-1)}{f(\lambda)}.$$

When $\lambda < -0.5$, $g_1(\lambda)$ is positive while $g_2(\lambda)$ is negative. Thus it is useful to compare the absolute of these two values. Since

$$(\lambda+0.5)\cdot(\frac{f(\lambda+1)}{f(\lambda)}-1) \geq 0,$$

Note that $g_1(\lambda) = g_2(-\lambda)$. Then we have:

$$\lambda \cdot G_1'(\lambda) > 0 \; for \; |\lambda| \geq \frac{1}{2}.$$

Next we consider $g_1'(\lambda)$:

$$\begin{aligned} g_1'(\lambda) =& \frac{f(\lambda+1)}{f(\lambda)}\{[ln(t^2-(\lambda+1)^2)-ln(t^2-\lambda^2)]^2 \\ &+(\frac{-2(\lambda+1)}{t^2-(\lambda+1)^2}+\frac{2\lambda}{t^2-\lambda^2})\} \end{aligned}$$

It is easy to prove that $g_1'(\lambda) < 0$, $for \; \lambda \in [-0.5, 0.5]$. So when characterizing $H(\lambda)$, it can be naturally proven that:

$$\lambda \cdot G_1'(\lambda) \geq 0 \; for \; any \; \lambda.$$

Thus,

$$\lambda^* = arg \min_{\lambda \in N} G_1(\lambda) = 0.$$

□