Inferring Stochastic Differential Equation from Social Science Time-series

Jin Hong Kuan

October 15, 2024

1 Abstract

There are numerous challenges in analyzing the rich longitudinal datasets that social scientists have gathered, due to the prevalence of noise, missing data, and presence of artifacts pertaining to how the data was collected. One way to address both problems at once is to exploit the recent major advances in inferring non-parametric stochastic differential equations (SDE) by using Gaussian process regression. To this end, we propose adapting these new SDE methodologies to suit the unique needs of social science research, including SDE inference for incomplete datasets, SDE-based imputation, perturbation detection, irreversibility measure, and time-series fitting. We also present the results from applying our proposed methodologies on the Seshat dataset, and discuss the novel insights obtained. Through demonstrating the utility of these methods, we hope to provide a preliminary step towards the adoption of stochastic time-series analyses in a broader range of disciplines.

2 Introduction

With the advent of new data collection and processing methods in recent years, the amount of data available for analyses in the social sciences such as psychology, sociology and economics has grown rapidly. Various statistical methods are employed to extract insights from these large datasets, including but not limited to principal component analysis, factor analysis, and cluster analysis [1].

However, one common challenge posed by these social science datasets is the prevalence of missing values. For instance, an archaelogical dataset may be fragmentary due to difficulty in obtaining a complete sample of artifacts, or a sociological dataset may be skewed due to the answering preferences of surveyed participants. Instead of omitting incomplete data entries, it is worth utilizing methods that are robust against missing data to obtain higher quality insights. There is rich literature in this domain, and the commonly used methods fall generally within the three family of approaches: *imputations*, *weighting*, and *direct analysis of missing data* such as through the use of machine learning or expectation-maximization (EM) algorithms [2].

In this paper, we examine time-series or longitudinal datasets with missing values. This class of data is of particular interest to us because most interesting phenomena in social science unfolds over time. Examples include the Seshat databank which tracks the evolution of human societies over the span of multiple millennia [3], the World Value Survey which explores the shifts in people's opinions and beliefs over time [4]. Through fitting a dynamic process model to time-series, we can learn about the temporal structure of the process governing the system.

We model the underlying dynamic governing a system as a stochastic differential equation (SDE), This is a natural choice because the systems studied in the social sciences is best represented as a noisy generative process. We draw on a recent work that enables the learning of first order Markov processes through Bayesian inference [5]. The robustness of this Bayesian approach against missing data, combined with its potential for high fidelity inference with low number of samples made it an appealing choice for our use case. We further augmented upon this method to address the aforementioned characteristics of social science datasets.

The paper is divided into four sections. In Section 3, we describe the intricacies of the proposed methodology. In Section 4, we explore the application of our method on both real and generated datasets, and explore their implications. We conclude with Section 5 and lay out future research directions.

3 Methods

Stochastic differential equations are conventionally used to model systems in physics and in mathematical finance. It draws upon early works in the study of Brownian motion (i.e. the random movements of a particle suspended in heat bath), and aims to model systems that are driven both by deterministic and random components.

We propose the use of multivariate time-homogeneous SDEs for the modeling of social science systems. The core assumption of this model is that the underlying dynamics is first-order Markovian, i.e. the future state of the system depends only on the present state. We assert, and aim to demonstrate empirically that this is an appropriate assumption to make in many sociological and economical domains, provided that the state vector is rich enough to describe the main driving forces of the system. Expressed mathematically, the model is of the form:

$$d\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t)dt + \sigma(\mathbf{x}_t)dW_t \tag{1}$$

where $\mathbf{x}_t \in \mathbb{R}^D$ is the state vector of our dynamical system containing D variables, $\mathbf{f}(\mathbf{x}_t)$ is the drift component representing deterministic state evolution, while $\sigma(\mathbf{x}_t)$ is the diffusion component representing the coefficient of the multivariate Brownian motion dW_t. At each time step dt, the system takes a small step in the direction of $\mathbf{f}(\mathbf{x}_t)$, plus a Gaussian jump with the coefficient of $\sigma(\mathbf{x}_t)$.

Our analysis builds upon a recent advent in dynamical systems modeling that enables efficient learning of SDE parameters from data of arbitrary sparseness [**?**]. In brief, the methodology involves inferring the functions $\mathbf{f}(\cdot)$ and $\sigma(\cdot)$ that maximize the expected likelihood of the observed data. Two innovations were introduced to make this learning process tractable. Firstly, drifts and diffusions are themselves modeled as Gaussian processes, thus allowing for the mapping and extrapolation of $\mathbf{f}(\cdot)$ and $\sigma(\cdot)$ from a fixed number of *inducing vectors*. Secondly, the log expected likelihood of observed data, or the log evidence, is computed through Monte Carlo sampling of possible trajectories using the Euler-Maruyama method.

After obtaining the drift and diffusion functions, we are at liberty to utilize them to extract social science insights. Here we elaborate two possible use cases for SDE.

3.1 Perturbation Detection

A dynamical model allows us to quantify the degree to which an observed trajectory or time-series deviates from the prediction of the model. Since the model itself is inferred from an aggregate of time-series, intuitively it captures the uniform trend that is common to all members of the data set. Provided that the model itself is a good fit, a deviation thereby implies an exogenous perturbation that differs from the norm, either in the form of external disturbance to the system or a rare event that is not sufficiently represented in the data set.

For instance, we would like to know whether the jump from $x(t)$ to $x(t+1)$ along a time-series is surprising relative to the model. Through Monte Carlo sampling given $x(t)$ and the inferred SDE, we can generate a distribution of possible values of $x(t+1)$. From this collection of points, we can apply Kernel Density Estimation (KDE) to estimate the probability density function of $x(t+1)$, which in simple terms indicates where the model expects the end point to land on. Since there exists no canonical way of quantifying surprise in this context, we propose a simple metric for perturbation that utilizes the probability density function as a benchmark.

For meaningful comparison across different events within a dataset, this metric must be bounded and normalized to $[0, 1]$. Additionally, this metric is also designed to have the following properties:

1. If $x(t+1)$ lands on the spot with the highest estimated probability density, then perturbation is 0.

2. If $x(t+1)$ lands on the spot with the lowest estimated probability density, then perturbation is 1.

3. The higher the estimated probability density of $x(t+1)$, the lower its perturbation value.

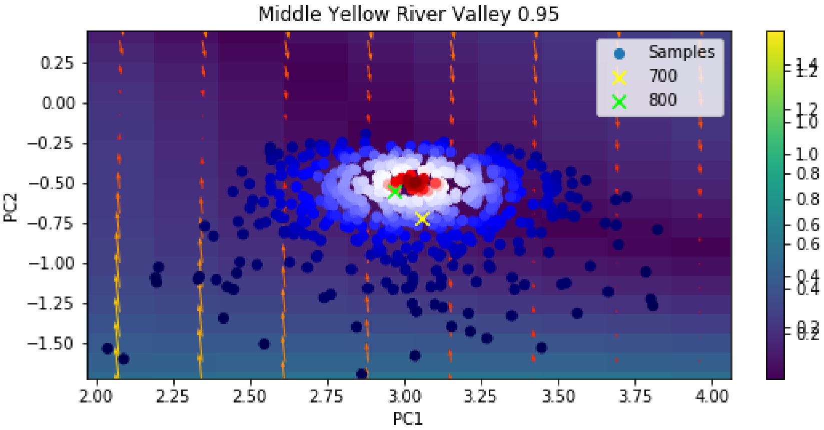

Figure 1: Monte Carlo sampling given a starting point (yellow marker) and underlying SDE. Sample points (dots) are colored according to their estimated probability density, with blue being less probable than actual observed end point (green marker) and red being more probable.

One metric satisfying the above qualities involves finding the ratio in area between regions with lower estimated probability density and regions with higher estimated probability density, compared to $x(t+1)$. Computing the exact amount is intractable, hence we propose an approximation to the quantity by generating a different set of Monte Carlo samples, and ranking their individual estimated probability densities with respect to $x(t+1)$. The exact procedure is as follows:-

1. Draw N_b sample paths starting from $x(t)$ to find endpoints $\{B_i : i = 1, \ldots, N_b\}$ using the Euler-Maruyama (EM) method. These points constitute the bases set.

2. With B as the Gaussian bases, apply kernel density estimation to estimate the probability distribution of projected end points $x(t+1)'$ given $x(t)$ under the stochastic model.

3. Repeat Step 1 to generate a comparison set of N_c samples, yielding $\{C_i : i = 1, \ldots, N_s\}$. Evaluate their respective probability density using the estimator, i.e. $P(C_i \mid KDE(B))$.

4. Evaluate $P(x(t) \mid KDE(B))$ and rank it against those of the N_c samples in Step 3.

Increasing the value of N_b raises the accuracy of the probability distribution of $x(t+1)$ given the model, whereas increasing the value of N_c improves the precision of the metric. We discuss application of this method on the Seshat dataset in the Results section.

3.2 Irreversibility Measure

Modeling a dynamical system with SDEs also allows us to bring in ideas from stochastic thermodynamics for deeper analysis. Here we explore **local expected irreversibility**, an established concept in thermodynamics used to measure the local entropy flow around a point. In the context of SDE, it is defined as the log of the relative probability of going from x_0 to x_t along trajectory $\vec{x}$, compared to the probability of moving in reverse ($\tilde{\vec{x}}$). This is relevant to our inquiry, since higher irreversibility entails more deterministic flow and a stronger presence of a trend, whereas lower irreversibility represents more diffuse movements.

$$\ln\left[\frac{P(\vec{x}|x_0)}{P(\tilde{\vec{x}}|x_t)}\right] \tag{2}$$

In considering an infinitesimally small step for our irreversibility calculation, we assume that the drift and diffusion values remain constant, and thus the probabilities of eq. (2) can take the form of multivariate Gaussian distributions:

$$\begin{aligned}\ln\left[\frac{P(\vec{x}|x_0)}{P(\tilde{\vec{x}}|x_t)}\right] &= \lim_{\delta t\to 0}\left(\ln\left[\frac{e^{-[(x_{\delta t}-\mu(x_0)\delta t)^2\big/(2\sigma^2\delta t)]}}{e^{-[(-x_{\delta t}-\mu(x_{\delta t})\delta t)^2\big/(2\sigma^2\delta t)]}}\right]\Big/\delta t\right) &(3)\\ &= \frac{2\mu_0^2}{\sigma^2}+\mu'(0) &(4)\end{aligned}$$

where μ is drift, μ' is the differential of drift, and σ is diffusion.

3.3 Imputations using SDE

Another advantage to building a dynamical model for a dataset is that it allows us to make better informed imputations for missing data in the dataset. Our solution involves combining information about observed data points, denoted by D, with the underlying SDE to reach a most probable compromise between the two. As shorthand, we denote the latter by F, representing the full fields defining a heterogeneous SDE. (Yildiz parameterizes F in terms of inducing vectors, etc.)

Specifically, we are interested in $P(d_i^j(t)\,|\,D)$ for some missing value $d_i^j(t)$. $d_i^j(t)$ refers to the j-th component of the i-th timeseries at time t, to account for instances in which only *some* components of the data point is missing. Then we can write

$$P(d_i^j(t))\,|\,D) = \int dF P(F\,|\,D)P(d_i^j(t)\,|\,D,F) \tag{5}$$

Now Yildiz doesn't give us $P(F\,|\,D)$, but some sort of noisy approximation of $F^*(D) := \arg\max_F P(F\,|\,D)$. We are approximating the RHS of eq. (5) with

$$P(d_i^j(t)\,|\,D,F^*(D)) \tag{6}$$

Note that in the limit that $P(F \mid D)$ becomes a delta function centered on $F^*(D)$, this approximation becomes perfectly accurate.

For simplicity, here we consider only the search for the maximum *a posteriori* value $d_i^{j*}(t)$, as defined in eq. (6). The same intuition laid out below can be incorporated with the probability distribution to map out the distribution of probable values.

$$d_i^{j*}(t) = \underset{d_i^j(t)}{\arg\max}\, P(d_i^j(t) \mid D, F^*(D)) \tag{7}$$

To that end, we provide a method for systematically constructing the most probable path $d_i^*(\tau)$ for $t_l \leq \tau \leq t_r$ from which the value of $d_i^{j*}(t)$ can be obtained. Here, t_l is the closest complete data point to the left of t, and t_r is the closest complete data point to the right of t. Since SDE satisfies the Markov property, we can safely assume that $d_i(t)$ is independent of $d_i(\tau)$ for $\tau \geq t_r$ or $\tau \leq t_l$.

The intuition for our method goes as follows: we know that the path $d_i^*(\tau)$ is a continuous curve. Start by approximating it with two straight lines, one extending from $d_i^*(t_l)$ to $d_i^*(\frac{t_l+t_r}{2})$, and the other from $d_i^*(\frac{t_l+t_r}{2})$ to $d_i^*(t_r)$. The choice of $\frac{t_l+t_r}{2}$ is arbitrary, for any $t_l < \tau < t_r$ will do.

This approximation allows us to reduce the problem to finding $d_i^*(\frac{t_l+t_r}{2})$. We take advantage of the Markov property of SDE to express $d_i(\frac{t_l+t_r}{2})$ as a joint probability:

$$d_i^*(\frac{t_l+t_r}{2}) = \underset{d_i(\frac{t_l+t_r}{2})}{\arg\max}\, P(d_i(\frac{t_l+t_r}{2})|d_i(t_l), F^*(D))P(d_i(t_r)|\frac{t_l+t_r}{2}, F^*(D)) \tag{8}$$

Which can be obtained through stochastic gradient descent with respect to $d_i^*(\frac{t_l+t_r}{2})$.

Aside (to be put into the text): To properly calculate $P(F \mid D)$ one would have to go beyond Yildiz, and introduce a "birth location / time" pdf, $p_{BL}(x,t)$, over $\mathbb{R}^N \times \mathbb{R}$. We would then replace the part of the Yildiz code that starts the Euler algorithm from the first points in the m time-series with a code that starts the Euler algorithm from a location-time given by sampling $p_{BL}(x,t)$. (Going further still, one should consider a stochastic process model in which the times at which data-point are missing and the fields that are missing in other data points are random variables, possibly coupled to F — we are assuming that there is no such coupling.)

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{9}$$

4 Results

Our proposed methods are empirically tested on the Seshat data set. This data set tracks the evolution of human societies in thirty regions around the globe and contains diverse variables covering economical, sociological and geographical domains.

We exploited previous analysis on this data set in our investigation of it. First, the original investigation applied principal component analysis (PCA) to Seshat. This revealed that the principal component along the 9-dimensional state space accounted for 77% of the variance, suggesting that key aspects of human societal organizations co-evolve in predictable ways [3].

Subsequent research then found that the second principal component, despite accounting only for 6% of total variance, has patterns which have significant implications regarding societal developments [6]. Through sliding window analysis of the first and second principal components, polities are shown to go through multiple phases of development. Initially, there exists minimal changes in information-processing (IP) capabilities such as writing and infrastructure, while scale variables such as population and territorial sizes experience significant changes and growth ('Scale Threshold'). Then, once these societies reach a certain threshold in scale, the developmental trajectory changes dramatically, with scale variables remaining stagnant while the (successful) societies experience rapid development of information-processing capabilities ('Information Threshold'). Once a certain maturity of information-processing capabilities is reached, the developmental path reverts back to having minimal changes in information-processing capabilities and exhibits greater variance in scale variables. The point in state space around which the switch between Scale and Information Theshold occurs is known as the first hinge point, and the point that marks the end of the Information Threshold is known as the second hinge point.

We follow up on this study with further statistical analyses for rigor, as well as dynamical analyses as described above.

4.1 Further Statistical Analysis

Since the sliding window analysis in [6] only considers the first and second principal components, we wanted to conduct further analysis to verify the conclusion about slopes corresponding with each phase. To that end, we separated the data points into three distinct subsets (hereon referred to as Phases 1, 2 and 3) according to their PC1 values, and performed Gaussian fit independently on each. The result is shown in fig. 2. As expected, a significant variance is observed for scale variables, and minimal variance for IP variables in Phases 1 and 3, and conversely for Phase 2. Interestingly, the variable *infrastructure*, categorized as IP variable in [6], also had significant variance along Phase 2, suggesting that a more nuanced distinction between Scale and Information Thresholds may be warranted.

4.2 Dynamical Analysis

Having established PC1-PC2 as satisfactory proxies for the evolution of the system, we fed the longitudinal values of PC1 and PC2 into the Yildiz SDE

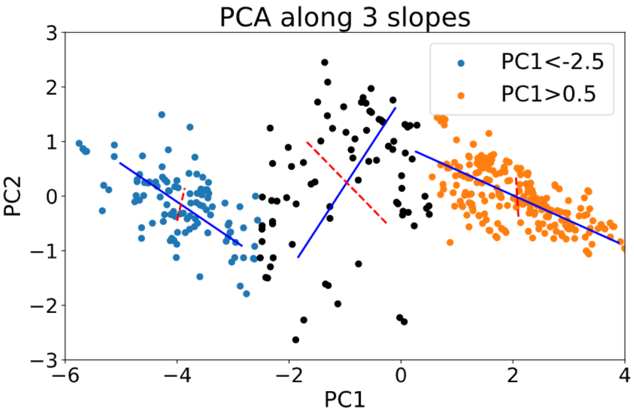

Figure 2: Gaussian fit on each subset of data points. Blue and red lines represent and major and minor axes of the Gaussian curves respectively. Note that both data points and curve axes are projected onto PC1-PC2 space for visualization.

inference library. Since multiple hyperparameters (e.g. number of inducing vectors, learning rate) can be modified to increase the fit of the model, we cross-validated our results and chose the set of hyperparameters that maximize the cross-validation score.

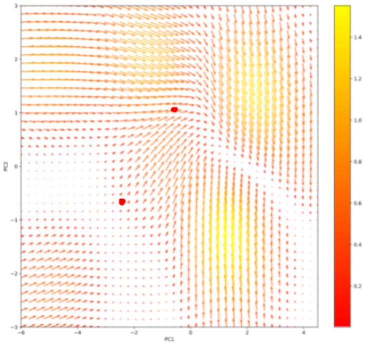

Figure 3: Flow field generated via SDE inference based on Yildiz et al. The red dots mark the two hinge points reported in Shin et al.

The inferred drift and diffusion functions are visualized as flow fields (shown in fig. 3). Through visual inspection, we can reasonably conclude that the second hinge point (top-right) indeed does exist, as characterized by the sharp change in direction of vectors in the flow field. However, it was found that there was no dynamical explanation behind the first hinge point, as the the magnitude of drift term around that region is relatively low.

This was a surprising result, and indicated a clear disparity with the ear-

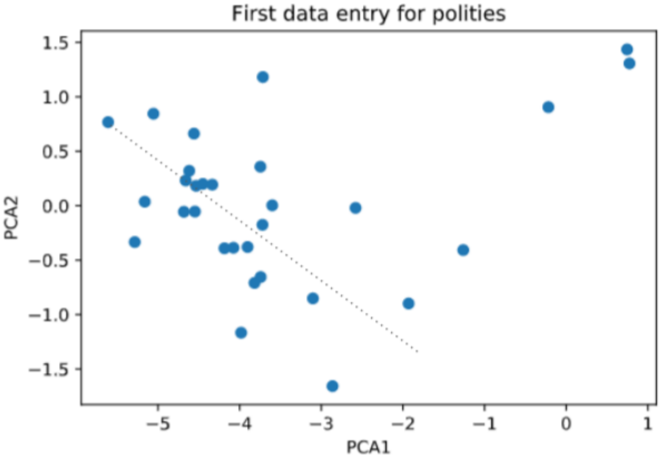

Figure 4: The first time-series entry of each polity reported in the Seshat dataset, in PC1-PC2 space

lier statistical analyses. Why was the shift from Scale Threshold to Information Threshold not captured in the dynamical model? We hypothesize that the Scale Threshold was due to birth processes rather than dynamical processes. When we graphed the first data points for each trajectory, it was found that the distribution of the data points formed a slope that corresponds with the Scale Threshold (fig. 4). One possible social science interpretation is that there exists a much greater variance in the scale characteristics of early polities, such as population or territorial size, compared to their information-processing capabilities at the stage at which they became traceable for archaeologists. It was the presence of information-processing technologies such as writing and infrastructures that enabled different polities to be rediscovered and added to the Seshat dataset, hence the early records of said polities will reasonably be more similar in those aspects than in categories that are not directly tied to information-processing.

We also tested our proposed methods for perturbation detection on the Seshat dataset. In fig. 5, we plotted out a timeline of perturbation values for the Middle Yellow River Valley polity. To give context, we also added labels for significant historical events that may influence the developmental pathway of this polity.

Based on the visualization, it would appear that with respect to our stochastic model, the polity experienced more unexpected developments in certain centuries compared to others. For instance, there is a notable dip in model fit in the century that covers the unification of China by the Qin dynasty from 300 B.C.E to 200 B.C.E., suggesting that the political and social landscape shifted in significant ways that are unpredicted. This is contrasted with the century from 700 C.E. to 800 C.E., in which the relatively peaceful century that followed

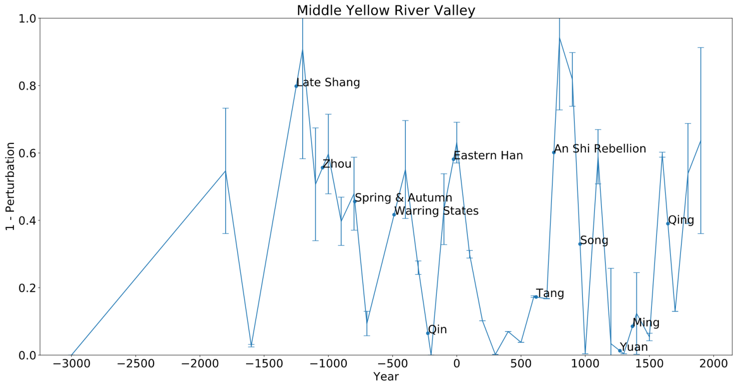

Figure 5: Timeline of perturbation values for the Middle Yellow River Valley polity (China), with increasing values in the y-axis indicating better fit to model prediction.

the establishment of the Tang dynasty saw the polity grow in a manner that corroborates with our predictions.

Additionally, we were also able to generate an irreversibility graph from the drift and diffusion flow fields (fig. 6). It was found that there is high local irreversibility around the second hinge point, which may imply that once a polity achieves a certain level of maturity in information-processing capabilities, it enters a new phase of developments distinct from its previous developmental history. An example will be that of Latium, which after moving past the second hinge point in the state space (corresponding roughly to the era Roman expansion in the Italian peninsula), does not revert back to pre-threshold states despite experiencing general decline in the first millennium C.E. This finding thus indicates the importance of information-processing capabilities in bounding the growth trajectory of a polity.

5 Discussion

We presented the application of our methodologies on the Seshat dataset, and made preliminary social science inferences about the results. The comparison of showed the utility of SDE inference in disentangling birth processes from dynamical processes, while the proof-of-concept applications of novel methodologies such as perturbation detection, irreversibility measure and SDE-based imputations are discussed.

For SDE inference to be rigorously adopted in the analysis of social science datasets, it is imperative that the confidence interval of the inferred values be determinable. This unfortunately is a missing feature from the Bayesian-based library that we used for our analyses. As this will involve obtaining the probability distribution of the inferred values, which is an intractable task, we plan to

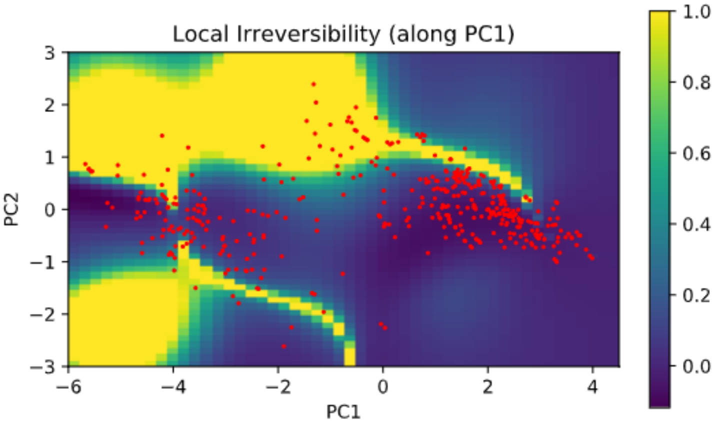

Figure 6: Local irreversibility (local entropy flow) along PC1. Higher value represents higher irreversibility.

take advantage of the variational inference library provided by the probabilistic programming framework Pyro to approximate and bound the confidence intervals.

Besides justifying the discussed use cases, the confidence of an SDE fit will also reveal important inherent qualities about the dataset itself. Specifically, it can be used to verify or refute the assumptions that the generative process behind a dataset can be adequately described as first-order Markovian. Here we list a few candidates for testing the model: Compustat for the dynamics of firms growth, PSID for the career dynamics of workers, Microsoft Academic for researchers' career/activities and/or evolution of research disciplines, IPEDS for the evolution of higher education systems, Congressional speech data for the change of political debates in US congress

Despite the numerous technical roadblocks in its application, we believe that stochastic differential equations is a promising candidate for future uses in modeling the underlying process behind many social science processes.

6 Appendix

6.1 Inference with missing data

The analysis methods we proposed above are implementation-agnostic. However, since we primarily applied the Yildiz inference method for our analysis, we made some modifications to the algorithm to allow it to process partially missing data. Such modification broadens the use case of this algorithm to include many areas of the social sciences, where data points may be incomplete due to challenges in data collection.

The method involves learning a discrete number of inducing vectors and locations to approximate the drift and diffusion of any point in a stochastic system with Gaussian processes. Specifically, for inducing locations Z and inducing values $\mathbf{u}_f$, $\mathbf{u}_\sigma$, we construct f and σ, which are functions for the drift and diffusion term respectively. Ω is used to represent the Gaussian noise variance of the stochastic dataset, and is itself a learned parameter.

Through stochastic gradient descent, the method seeks to maximize the likelihood of the data given the parameters:

$$\begin{aligned} p(Y \mid f, \sigma, \Omega) &= \prod_{i=1}^{N} \mathbb{E}_{p(x|t_i;f;\sigma)}[N(y_i \mid x, \Omega)] && (10) \\ &\simeq \prod_{i=1}^{N} \frac{1}{N_s} \sum_{s=1}^{N_s} N(y_i \mid x_i^{(s)}, \Omega), x^{(s)} \sim p(x_{0\ldots t}|u_f, u_\sigma, Z) && (11) \end{aligned}$$

Here y_i represents the i-th observed value along the trajectory and $x_i^{(s)}$ represents the corresponding value from the sampled path. We propose to generalize this formulation to include y_i where only some of the components y_i^j, $j \in D_i$ are known:

$$p(Y \mid f, \sigma, \Omega) = \prod_{i=1}^{N} \frac{1}{N_s} \sum_{s=1}^{N_s} \prod_{j \in D_i} N(y_i^j \mid x_i^{(s)j}, \Omega), x^{(s)} \sim p(x_{0\ldots t}|u_f, u_\sigma, Z) \quad (12)$$

References

[1] S. P. Mukherjee, Bikas K. Sinha, and Asis Kumar Chattopadhyay. *Statistical methods in social science research.* Springer, 2018.

[2] Roderick JA Little and Donald B. Rubin. The analysis of social science data with missing values. *Sociological Methods & Research*, 18(2-3):292–326, 1989. Publisher: Sage Publications.

[3] Peter Turchin, Rob Brennan, Thomas E. Currie, Kevin C. Feeney, Pieter Francois, Daniel Hoyer, Joseph G. Manning, Arkadiusz Marciniak, Daniel Austin Mullins, and Alessio Palmisano. Seshat: The global history databank. *Cliodynamics: The Journal of Quantitative History and Cultural Evolution*, 2015.

[4] Sarah Fleche, Conal Smith, and Piritta Sorsa. Exploring determinants of subjective wellbeing in OECD countries: evidence from the World Value Survey. 2012. Publisher: OECD.

[5] Cagatay Yildiz, Markus Heinonen, Jukka Intosalmi, Henrik Mannerstrom, and Harri Lahdesmaki. Learning stochastic differential equations with gaussian processes without gradient matching. In *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2018.

[6] Jaeweon Shin, Michael Holton Price, David H. Wolpert, Hajime Shimao, Brendan Tracey, and Timothy A. Kohler. Scale and information-processing thresholds in Holocene social evolution. *Nature Communications*, 11(1):2394, May 2020. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Archaeology;Cultural evolution;Social anthropology Subject_term_id: archaeology;cultural-evolution;social-anthropology.