

基于随机微分方程的结构学习¹

时间序列因果学习的连续建模方法

王一丁

2024 年 7 月 8 日

¹Benjie Wang, Joel Jennings, and Wenbo Gong. “Neural structure learning with stochastic differential equations”. In: ICLR. 2024

目录

- 
- A detailed pencil sketch of the Tsinghua University main gate. The gate is a large, curved structure with the university's name in Chinese characters '清华大学' and English 'TSINGHUA UNIVERSITY' inscribed on it. There are trees and a building in the background.
- 1 绪论
 - 2 预备知识
 - 3 SCOTCH
 - 4 理论分析
 - 5 实验分析
 - 6 总结与展望

问题引入及分析

时间序列数据在现实世界中无处不在，通常由在不同时间间隔记录的一系列数据点组成。理解与时间过程相关的变量之间的底层因果结构对于许多现实世界的应用至关重要。

现有离散时间结构学习的不足之处

- 当真正的底层过程连续，可能推断出不正确的因果关系。
- 难以处理不规则的采样间隔。

本文引入了一种新的框架，采用随机微分方程（SDEs）作为连续的隐过程学习时间序列中的因果结构。

文章贡献

- 与之前使用常微分方程的方法相比，该模型能够从多模态²和非高斯分布的时序数据中准确地学习潜在的因果关系。
- 证明了当 SDEs 直接用于观测过程的建模时，所得到的 SDEs 在全局 Lipschitz 和对角扩散假设下是结构可识别的。
- 对合成的有真实背景的数据集进行了广泛的实验，表明 SCOTCH 可以改进现有的结构学习方法，包括当处理不规则抽样数据³时。

²该模型同时学习多个时间序列共同的因果关系，可能对多模态学习有帮助，但是实验部分并未使用多模态数据进行证实。

³在实验部分，作者随机丢弃时间序列中的数据来模拟不规则抽样。

目录

- 
- A faint, stylized background illustration of the Tsinghua University gate. The gate is a large, curved structure with the university's name in Chinese characters '清华大学' and English 'TSINGHUA UNIVERSITY' inscribed on it. There are trees and a building visible in the background.
- 1 绪论
 - 2 预备知识
 - 3 SCOTCH
 - 4 理论分析
 - 5 实验分析
 - 6 总结与展望

贝叶斯结构学习

目的：从观测数据中学习表示结点之间的有向关系图（因果图）。

给定时间序列数据 $\{\mathbf{X}_{t_i}\}_{i=1}^I$ ，包含 I 个时间点。假设有 N 个独立同分布的数据，给定图 $\mathbf{G} \in \{0, 1\}^{D \times D}$ ，定义图与数据的联合分布：

$$p(\mathbf{G}, \mathbf{X}^{(1)}, \dots, \mathbf{X}^{(N)}) = p(\mathbf{G}) \prod_{n=1}^N p(\mathbf{X}^{(n)} | \mathbf{G}) \quad (1)$$

其中 $p(\mathbf{G})$ 是图的先验分布， $p(\mathbf{X}^{(n)} | \mathbf{G})$ 是似然项。

目的是学习后验分布 $p(\mathbf{G} | \mathbf{X}^{(1)}, \dots, \mathbf{X}^{(N)})$ ：

$$\begin{aligned} p(\mathbf{G} | \mathbf{X}^{(1)}, \dots, \mathbf{X}^{(N)}) &= \frac{p(\mathbf{G}, \mathbf{X}^{(1)}, \dots, \mathbf{X}^{(N)})}{p(\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(N)})} \\ &= \frac{p(\mathbf{G}) \prod_{n=1}^N p(\mathbf{X}^{(n)} | \mathbf{G})}{\sum_{\mathbf{G}} p(\mathbf{G}) \prod_{n=1}^N p(\mathbf{X}^{(n)} | \mathbf{G})} \end{aligned} \quad (2)$$

但是由于图 \mathbf{G} 所在的空间维度很高，难以准确计算上述后验分布，可以使用变分推断与采样来学习。

结构方程模型 (SEM)

给定时间序列数据 $\{\mathbf{X}^{(n)}\}_{n=1}^N$ 和图 $\mathbf{G} \in \{0, 1\}^{D \times D}$ ，可以使用 SEM 描述变量间的结构关系：

$$\mathbf{X}_{t,d} = f_{t,d}(Pa_G^d(< t), \epsilon_{t,d}) \quad (3)$$

$Pa_G^d(< t)$ 表示父节点在之前时刻的数据， $\epsilon_{t,d}$ 表示相互独立的噪声。

该模型需要**离散的时间步长**⁴，这些时间步长通常被假定为遵循规则的采样间隔。现有模型大多遵循此框架。

⁴这里的离散主要说的是数据生成的因果结构是离散的，这导致了如果数据的因果频率与采样频率不一致的话，将很难发现数据中的因果关系。将底层因果结构建模为连续系统的话将没有这种问题，虽然最终经过采样还是会落到离散点，但是底层的连续性保证了采样点足够多便可以逼近真实过程（有点像采样定理）。

伊藤扩散过程 (Itô diffusion)

假设 \mathbf{X}_t 是非时变⁵ (time-homogenous) 的伊藤过程, 则

$$d\mathbf{X}_t = f(\mathbf{X}_t)dt + g(\mathbf{X}_t)d\mathbf{W}_t \quad (4)$$

其中 $f: \mathbb{R}^D \rightarrow \mathbb{R}^D, g: \mathbb{R}^D \rightarrow \mathbb{R}^{D \times D}$ 分别是非时变的漂移项和扩散项。
 \mathbf{W}_t 是测度 P 下的布朗运动。

若全局 Lipschitz 条件满足, 则 (3) 式有唯一强解⁶:

$$\mathbf{X}_t = \lim_{n \rightarrow \infty} \mathbf{Y}_t^{(n+1)} = \lim_{n \rightarrow \infty} \left(\mathbf{X}_0 + \int_0^t f(\mathbf{Y}_s^{(n)})ds + \int_0^t g(\mathbf{Y}_s^{(n)})dB_s \right) \quad (5)$$

其中 $\mathbf{Y}_t^{(0)} = \mathbf{X}_0, \mathbf{Y}_t^{(k)} = \mathbf{Y}_t^{(k)}(\omega)$

但是对于大多数拥有非线性漂移和扩散函数的伊藤扩散过程, 解析解 \mathbf{X}_t 是难以得到的, 因此需要使用**离散化**的方式解决。

⁵非时变是指 $f(\mathbf{X}_t), g(\mathbf{X}_t)$ 不随时间 t 变化。作者在文中提到, 采用非时变有助于证明结构可识别性, 并在展望部分提出可能需要引入新的理论和方法来应对“动态因果”。

⁶Bernt Øksendal and Bernt Øksendal. Stochastic differential equations. Springer, 2003 (Thm 5.2.1) 🔍🔍🔍

固定步长 Δ ，使用 Euler 离散化可得轨迹：

$$\mathbf{X}_{t+1}^{\Delta} = \mathbf{X}_t^{\Delta} + f(\mathbf{X}_t^{\Delta})\Delta + g(\mathbf{X}_t^{\Delta})\eta_t \quad (6)$$

在图 \mathbf{G} 中，定义结点 i 不是结点 j 在时刻 t 的因，若对于任意 k ， $\frac{\partial f_j(\mathbf{X}_t^{\Delta})}{\partial X_{t,i}^{\Delta}} = 0$ ，且 $\frac{\partial g_{j,k}(\mathbf{X}_t^{\Delta})}{\partial X_{t,i}^{\Delta}} = 0$ 。

如果假设 g 输出对角矩阵，则上述 Euler 离散化可以导出时序 SEM⁷，被称为 Euler SEM。

⁷ 此时漂移项部分 $g(\mathbf{X}_t^{\Delta})\eta_t$ 和 t 时刻数据 \mathbf{X}_t^{Δ} 的维度一致，因此可看作是 SEM 中噪声 ϵ_t 的变体。

目录

- 
- A faint, artistic sketch of the Tsinghua University gate serves as the background. The gate is a large, curved structure with the university's name in Chinese characters '清华大学' and English 'TSINGHUA UNIVERSITY' inscribed on it. Trees and foliage are sketched around the gate, and the overall style is a light, hand-drawn illustration.
- 1 绪论
 - 2 预备知识
 - 3 SCOTCH
 - 4 理论分析
 - 5 实验分析
 - 6 总结与展望

模型引入

考虑一个动力系统，既拥有内在的随机性，也被外界的噪声干扰。例如，在医疗保健中，患者的病情将随机发展，而不是确定性。此外，患者状况的测量也会受到设备精度的影响，其中噪声与内在随机性无关。

因此引入潜在随机微分方程模型：

$$\begin{aligned} d\mathbf{Z}_t &= f_{\theta}(\mathbf{Z}_t)dt + g_{\theta}(\mathbf{Z}_t)d\mathbf{W}_t \quad (\text{隐过程}) \\ \mathbf{X}_t &= \mathbf{Z}_t + \epsilon_t \quad (\text{带噪观测过程}) \end{aligned} \tag{7}$$

其中 \mathbf{Z}_t 是系统内部状态的潜在变量， \mathbf{X}_t 是具有相同维度的观测数据， \mathbf{W}_t 是 Wiener 过程（物理意义为布朗运动）。

模型假设

假设 1: 全局 Lipschitz

上式中的漂移和扩散函数满足全局 Lipschitz 约束。也就是说

$$|\mathbf{f}_\theta(\mathbf{x}) - \mathbf{f}_\theta(\mathbf{y})| + |\mathbf{g}_\theta(\mathbf{x}) - \mathbf{g}_\theta(\mathbf{y})| \leq C|\mathbf{x} - \mathbf{y}| \quad (8)$$

假设 1 是大多数 SDE 文献要求的标准假设, 以确保存在强解。

假设 2: 对角扩散

扩散函数输出一个非零对角线矩阵。也就是说, 它可以简化为向量值函数 $\mathbf{g}_\theta(\mathbf{X}_t): \mathbb{R}^D \rightarrow \mathbb{R}^D$

假设 2 帮助证明了结构可识别性。

过程先验

由于潜过程在看到任何观测值之前就诱导了潜轨迹的分布，因此也将其称为先验过程。本文使用类似掩码的方法，将图 G 与隐空间表征做内积来学习因果。并将神经网络嵌入在漂移函数 $f_{\theta}(\cdot, G)$ 与扩散函数 $g_{\theta}(\cdot, G)$ 的表示中：

$$f_{\theta,d}(\mathbf{Z}_t, G) = \zeta \left(\sum_{i=1}^D G_{i,d} l(\mathbf{Z}_{t,i}, \mathbf{e}_i), \mathbf{e}_d \right) \quad (9)$$

其中 ζ, l 是神经网络， \mathbf{e}_i 是共享的可训练的结点嵌入。

综上，过程先验可表示为：

$$d\mathbf{Z}_t = f_{\theta}(\mathbf{Z}_t, G)dt + g_{\theta}(\mathbf{Z}_t, G)d\mathbf{W}_t \quad (\text{先验过程}) \quad (10)$$

该先验过程中的参数可随着训练过程迭代。

图先验

$$p(\mathbf{G}) \propto \exp(-\lambda_s \|\mathbf{G}\|_F^2) \quad (11)$$

其中 λ_s 是稀疏系数。

似然

$$p(\{\mathbf{X}_{t_i}\}_{i=1}^I | \{\mathbf{Z}_{t_i}\}_{i=1}^I, \mathbf{G}) = \prod_{i=1}^I \prod_{d=1}^D p_{\epsilon_d}(X_{t_i,d} - Z_{t_i,d}) \quad (12)$$

该似然表示了噪声的分布。

变分推断

假设 $\{\mathbf{X}^{(n)}\}_{n=1}^N$ 是系统中的观测数据，目标是学习数据中的图结构 $p(\mathbf{G}|\{\mathbf{X}^{(n)}\}_{n=1}^N)$ ⁸，这是无法直接学习的，因此采用变分推断。

变分近似： $q_\phi(\mathbf{G}) \approx p(\mathbf{G}|\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(N)})$ 。

变分下界计算：

$$\log p(\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(N)}) \geq \mathbb{E}_{q_\phi(\mathbf{G})} \left[\sum_{n=1}^N \log p_\theta(\mathbf{X}^{(n)}|\mathbf{G}) \right] - D_{\text{KL}}(q_\phi(\mathbf{G})\|p(\mathbf{G})) \quad (13)$$

由于存在隐空间 $\mathbf{Z}^{(n)}$ ，故 $p_\theta(\mathbf{X}^{(n)}|\mathbf{G})$ 无法直接计算，因此首先使用变分框架估计后验分布 $p(\mathbf{Z}^{(n)}|\mathbf{G}, \mathbf{X}^{(n)})$ ，变分后验 $q_\psi(\tilde{\mathbf{Z}}^{(n)}|\mathbf{G}, \mathbf{X}^{(n)})$ 被定义为如下系统的解：

$$\begin{aligned} \tilde{\mathbf{Z}}_{t,0}^{(n)} &\sim \mathcal{N}(\mu_\psi(\mathbf{G}, \mathbf{X}^{(n)}), \Sigma_\psi(\mathbf{G}, \mathbf{X}^{(n)})) \quad (\text{后验初始状态}) \\ d\tilde{\mathbf{Z}}_t^{(n)} &= h_\psi(\tilde{\mathbf{Z}}_t^{(n)}, t; \mathbf{G}, \mathbf{X}^{(n)})dt + g_\psi(\tilde{\mathbf{Z}}_t^{(n)})d\mathbf{W}_t \quad (\text{后验过程}) \end{aligned} \quad (14)$$

⁸ 文章假设所有数据来自同一个因果系统生成，fMRI 分析可能不太适用，因为每个个体可能存在差异。

$$\tilde{\mathbf{Z}}_{t,0}^{(n)} \sim \mathcal{N}(\mu_\psi(\mathbf{G}, \mathbf{X}^{(n)}), \Sigma_\psi(\mathbf{G}, \mathbf{X}^{(n)})) \quad (\text{后验初始状态})$$

$$d\tilde{\mathbf{Z}}_t^{(n)} = h_\psi(\tilde{\mathbf{Z}}_t^{(n)}, t; \mathbf{G}, \mathbf{X}^{(n)})dt + g_\psi(\tilde{\mathbf{Z}}_t^{(n)})d\mathbf{W}_t \quad (\text{后验过程})$$

初始状态由神经网络学习到的均值和方差表示。后验扩散函数与先验相同（由于扩散函数类似噪声，与数据无关）。

后验漂移函数与先验不同的两点理由

- 后验漂移函数取决于时间，即使因果是非时变的。不同时间下即使是相同的数据，可能有不同的运动趋势。
- 不受限于先验掩码结构，从而可以表示更加复杂的系统。

接着，根据之前文献⁹的理论， $\log p(\mathbf{X}^{(n)}|\mathbf{G})$ 有下界：

$$\log p(\mathbf{X}^{(n)}|\mathbf{G}) \geq \mathbb{E}_{q_\psi} \left[\sum_{i=1}^I \log p(\mathbf{X}_{t_i}^{(n)}|\tilde{\mathbf{Z}}_{t_i}^{(n)}, \mathbf{G}) - \int_0^T \|\mathbf{u}^{(n)}(\tilde{\mathbf{Z}}_t^{(n)})\|^2 dt \right] \quad (15)$$

其中：

$$\mathbf{u}^{(n)}(\tilde{\mathbf{Z}}_t^{(n)}) = g_G(\tilde{\mathbf{Z}}_t^{(n)})^{-1} (h_\psi(\tilde{\mathbf{Z}}_t^{(n)}, t; \mathbf{G}, \mathbf{X}^{(n)}) - f_G(\tilde{\mathbf{Z}}_t^{(n)})) \quad (16)$$

因此总体证据下界为：

$$\mathbb{E}_{q_\phi} \left[\sum_{n=1}^N \mathbb{E}_{q_\psi} \left[\sum_{i=1}^I \log p(\mathbf{X}_{t_i}^{(n)}|\tilde{\mathbf{Z}}_{t_i}^{(n)}, \mathbf{G}) - \int_0^T \|\mathbf{u}^{(n)}(\tilde{\mathbf{Z}}_t^{(n)})\|^2 dt \right] \right] - D_{\text{KL}}(q_\phi(\mathbf{G})\|p(\mathbf{G})) \quad (17)$$

实际计算中使用 Monte-Carlo 方法近似上式中的期望。

⁹Xuechen Li et al. “Scalable gradients for stochastic differential equations”. In: [International Conference on Artificial Intelligence and Statistics](#). PMLR. 2020, pp. 3870–3882.

$$\mathbb{E}_{q_\phi} \left[\sum_{n=1}^N \mathbb{E}_{q_\psi} \left[\sum_{i=1}^I \log p(\mathbf{X}_{t_i}^{(n)} | \tilde{\mathbf{Z}}_{t_i}^{(n)}, \mathbf{G}) - \int_0^T \|\mathbf{u}^{(n)}(\tilde{\mathbf{Z}}_t^{(n)})\|^2 dt \right] \right] - D_{\text{KL}}(q_\phi(\mathbf{G}) \| p(\mathbf{G}))$$

Algorithm 1 SCOTCH training

Input: i.i.d time series $\{\mathbf{X}^{(n)}\}_{n=1}^N$; drift functions \mathbf{f}_G , \mathbf{h}_ψ , diffusion function \mathbf{g}_G , SDE solver Solver, initial condition $\tilde{\mathbf{Z}}_0^{(n)}$, training iterations L

for $l = 1, \dots, L$ **do**

 Sample time series mini-batch $\{\mathbf{X}^{(n)}\}_{n=1}^S$ with batch size S .

for $n = 1, \dots, S$ **do**

 Draw graph $\mathbf{G} \sim q_\phi(\mathbf{G})$

 Draw initial latent state $\tilde{\mathbf{Z}}_0^{(n)} \sim \mathcal{N}(\boldsymbol{\mu}_\psi(\mathbf{G}, \mathbf{X}^{(n)}), \boldsymbol{\Sigma}_\psi(\mathbf{G}, \mathbf{X}^{(n)}))$

 Solve (sample from) the posterior process $(\tilde{\mathbf{Z}}^{(n)}, L) = \text{Solver}((\tilde{\mathbf{Z}}_0^{(n)}, 0), \mathbf{f}_G, \mathbf{h}_\psi, \mathbf{g}_G)$

end for

 Maximize ELBO eq. (15) w.r.t. ϕ, ψ, θ

end for

图: SCOTCH 算法

目录

- 
- A faint, stylized background illustration of the Tsinghua University gate. The gate is a large, curved structure with the university's name in Chinese characters '清华大学' and English 'TSINGHUA UNIVERSITY' inscribed on it. There are trees and a building visible in the background.
- 1 绪论
 - 2 预备知识
 - 3 SCOTCH
 - 4 理论分析
 - 5 实验分析
 - 6 总结与展望

结构可识别性定理

假设观察过程是作为 Itô 扩散给出的：

$$d\mathbf{X}_t = f_G(\mathbf{X}_t)dt + g_G(\mathbf{X}_t)d\mathbf{W}_t \quad (18)$$

则模型在结构上可识别的充分条件由下述定理提供：

定理 1：观测过程的结构可识别性

给定上述方程，设有另一过程 $\bar{\mathbf{X}}_t, \mathbf{G} \neq \bar{\mathbf{G}}, \bar{f}_{\bar{\mathbf{G}}}, \bar{g}_{\bar{\mathbf{G}}}, \bar{\mathbf{W}}_t$ ，则在共同的初始值与假设 1, 2 下，解 $\mathbf{X}_t, \bar{\mathbf{X}}_t$ 有不同的分布。

结构可识别性定理

下述定理表明，在某些条件下，即使没有直接观察到 SDE 的解，即存在隐过程 \mathbf{Z} ，结构可识别性也可以保持。

定理 2：潜在过程的结构可识别性

考虑由隐过程公式定义的两组分布 p, \bar{p} ，分别对应 $(\mathbf{G}, \mathbf{Z}, \mathbf{X}, \mathbf{f}_{\mathbf{G}}, \mathbf{g}_{\mathbf{G}})$, $(\bar{\mathbf{G}}, \bar{\mathbf{Z}}, \bar{\mathbf{X}}, \bar{\mathbf{f}}_{\bar{\mathbf{G}}}, \bar{\mathbf{g}}_{\bar{\mathbf{G}}})$ ，其中 $\mathbf{G} \neq \bar{\mathbf{G}}$ 。令 t_1, \dots, t_I 是观测的时间序列，则在假设 1, 2 下，有

- 若对任意 i ，有 $t_{i+1} - t_i = \Delta$ ，则欧拉离散化
$$p^\Delta(\mathbf{X}_{t_1}, \dots, \mathbf{X}_{t_I}) \neq \bar{p}^\Delta(\bar{\mathbf{X}}_{t_1}, \dots, \bar{\mathbf{X}}_{t_I})$$
- 如果固定时间范围 $[0, T]$ ，则采样点数趋近无穷时
$$p(\mathbf{X}_{t_1}, \dots, \mathbf{X}_{t_I}) \neq \bar{p}(\bar{\mathbf{X}}_{t_1}, \dots, \bar{\mathbf{X}}_{t_I})$$

在结构可识别性的基础上，可以证明变分公式的一致性。也就是说，在无限数据限制下，可以通过使用足够表达的后验过程和正确指定的模型最大化 ELBO 来恢复真值图。

定理 3: 变分表述的一致性

若假设 1-2 满足潜在公式，则对于固定的观测时间范围 $[0, T]$ ，作为观测值数趋于无穷，当 ELBO 最大化时， $q_\phi(\mathbf{G}) = \delta(\mathbf{G}^*)$ ，其中 \mathbf{G}^* 为真值图。

目录


- 
- A detailed pencil sketch of the Tsinghua University main gate, featuring the university's name in Chinese characters '清华大学' and English 'TSINGHUA UNIVERSITY' on the archway, with trees and a building in the background.
- 1 绪论
 - 2 预备知识
 - 3 SCOTCH
 - 4 理论分析
 - 5 实验分析
 - 6 总结与展望

目录

- 
- The background of the slide features a detailed pencil sketch of the main entrance gate of Tsinghua University. The gate is a large, curved structure with the university's name in Chinese characters '清華大學' and English 'TSINGHUA UNIVERSITY' inscribed on it. Behind the gate, there are several large, leafy trees. The entire illustration is rendered in a light, sketchy style, serving as a subtle background for the text.
- 1 绪论
 - 2 预备知识
 - 3 SCOTCH
 - 4 理论分析
 - 5 实验分析
 - 6 总结与展望

- 本文提供的连续时间因果发现框架解决了离散学习中遇到的采样频率与真实因果频率不一致等问题。
- 本文的结构可识别性证明框架奠定了连续时序因果发现的基础。
- 使用了广泛的可下载的数据集。
- 可以考虑将其拓展至动态因果。

- [1] Benjie Wang, Joel Jennings, and Wenbo Gong. “Neural structure learning with stochastic differential equations”. In: ICLR. 2024.
- [2] Bernt Øksendal and Bernt Øksendal. Stochastic differential equations. Springer, 2003.
- [3] Xuechen Li et al. “Scalable gradients for stochastic differential equations”. In: International Conference on Artificial Intelligence and Statistics. PMLR. 2020, pp. 3870–3882.

A faint, stylized background illustration of the Tsinghua University gate. The gate is a large, curved structure with the university's name in Chinese characters '清华大学' and English 'TSINGHUA UNIVERSITY' visible. There are trees and a path leading to the gate.

谢谢!