

DOI:10.13232/j.cnki.jnju.2023.04.002

基于分位数因子模型的高维时间序列因果关系分析

梁慧玲^{1,2}, 刘 慧^{1,2*}, 刘力维^{1,2}, 赵 佳³, 阮怀军³

(1. 山东财经大学计算机科学与技术学院, 济南, 250014; 2. 山东省数字媒体技术重点实验室, 山东财经大学, 济南, 250014; 3. 山东省农业科学院信息技术研究所, 济南, 250000)

摘 要:从观察数据中发现变量之间的因果关系是许多科学研究领域的关键问题,传统 Granger 因果模型受到维度灾难的影响,难以准确地高维时间序列中发现因果关系. 提出一种基于分位数因子模型的 Granger 因果分析新方法 QFM-CGC 用于高维时间序列因果关系的判定. 首先, QFM-CGC 采用赤池信息量准则进行模型选择,避免人为干预设置滞后阶数的操作;然后,对向量自回归(Vector Autoregressive, VAR)模型中的条件变量建立分位数因子模型进行降维,减少 VAR 模型中的待估计系数,对降维后的 VAR 模型重新进行条件 Granger 因果分析;最后,使用蒙特卡洛模拟评估不同方法识别底层系统与观测时间序列的连通性结构的能力. 在不同维度变量的线性仿真系统和两组现实数据集上与基准方法和经典方法进行了比较,实验结果验证了该方法的有效性.

关键词:高维时间序列,分位数因子模型,条件 Granger 因果分析,数据挖掘

中图分类号:TP391

文献标志码:A

Causal relationship analysis of high-dimensional time series based on quantile factor model

Liang Huiling^{1,2}, Liu Hui^{1,2*}, Liu Liwei^{1,2}, Zhao Jia³, Ruan Huaijun³

(1. College of Computer Science and Technology, Shandong University of Finance and Economics, Ji'nan, 250014, China; 2. Key Laboratory of Digital Media Technology of Shandong Province, Shandong University of Finance and Economics, Ji'nan, 250014, China;

3. Institute of Information Technology, Shandong Academy of Agricultural Sciences, Ji'nan, 250000, China)

Abstract: Finding the causal relationship between variables from observed data is a key issue in many scientific research fields. Because the traditional Granger causality model is affected by the curse of dimension, it is difficult to accurately find causality in high-dimensional time series. In this paper, we propose a new Granger causality analysis method based on quantile factor model, QFM-CGC algorithm, which is used to find causality relationship in high-dimensional time series. Firstly, QFM-CGC uses Akaike information criterion to select models, which avoids setting the lag order by human intervention. Then, the quantile factor model is established to reduce the dimensionality of the conditional variables in a vector autoregressive (VAR) model, thus reducing the number of coefficients that need to be estimated. The reduced-dimensional VAR model is used for a conditional Granger causality analysis. Finally, Monte Carlo simulation is applied to evaluate the performance of different methods to identify the connectivity structure between the underlying system and the observation time series. Experiments compare the proposed method with benchmark and classical methods on a linear simulation system with variables in different dimensions and two sets of real data, confirming its effectiveness.

基金项目: 国家自然科学基金(62072274), 山东省科技成果转化项目(2021LYXZ021), 山东省泰山学者特聘专家计划(tstp20221137)

收稿日期: 2023-06-13

* 通讯联系人, E-mail: liuh_lh@sdufe.edu.cn

Key words: high-dimensional time series, quantile factor model, conditional Granger causality analysis, data mining

时间序列是指属于同一统计指标的数值按其时间发生的先后顺序排列而形成的一组随机变量,可以分成一元时间序列和多元时间序列,其中多元时间序列是将多个一元时间序列组合形成的时间序列. 1969年,英国计量经济学家 Granger^[1]首次提出一种经济学上的统计学假设检验方法,通常称为 Granger 因果关系检验,广泛应用于神经科学^[2-3]、计量经济学^[4-5]等其他研究领域^[6]. 根据 Granger 最初提出的概念,如果 Y 的预测模型中包括 X 时, Y 的预测得到了改善,则变量 X 会 Granger 导致变量 Y .

多元时间序列在进行传统 Granger 因果关系分析时忽略了其他变量的存在,因此双变量因果关系测度在估计真实因果关系时不准确^[7]. 针对这些问题,1982年 Geweke^[8]提出条件 Granger 因果关系分析(Conditional Granger Causality Analyse, CGCA). 多元时间序列的因果关系分析依赖于从一个观测变量到另一个观测变量的直接因果关系的估计,并考虑其他观察变量的存在. 相关影响由条件 Granger 因果关系指数(Conditional Granger Causality Index, CGCI)量化,由线性向量自回归(Vector Autoregressive, VAR)模型推导.

随着信息时代的发展,数据的维度不断增加,对于数据的研究也已转变到高维空间^[9-10],然而,目前大部分因果分析研究仍然集中于二元或多元时间序列,对于高维和超高维的时间序列因果分析缺少有效的处理方法. 通常低维主时间序列能提供动态特征的良好表示,易于解释和可视化. 因此,对于高维和超高维的时间序列因果分析需要使用降维方法来限制 VAR 模型.

主成分分析(Principal Component Analysis, PCA)是目前最主要的降维方法之一^[11],它将原始的高维数据投影到一个较低维的子空间上,使原始高维数据可以由一组低维变量表示^[12]. 但是,PCA 在降维时没有捕获隐藏的因子,值得注意的是相关因子可能会改变时间序列的分布特征(矩或分位数),而不是其均值. Chen et al^[13]提出分位数因子模型(Quantile Factor Models, QFM)

及其估计程序,简称分位数因子分析(Quantile Factor Analysis, QFA). 与其他因子模型不同, QFM 还捕获移动可观测分布的其他相关部分的未观测因子. QFM 的一个重要优点是它能同时提取决定 QFM 因素结构的所有均值和额外(非均值)因子,而 PCA 只能提取平均因子,所以 QFA 克服了 PCA 没有捕捉隐藏因子的能力的问题. 为此,Chen et al^[13]通过蒙特卡洛模拟^[14]说明存在异常值时,使用 QFA 有明显优势^[13].

由于传统 Granger 因果分析方法无法准确区分直接因果关系和间接因果关系,且高维时间序列在 VAR 模型中容易受到维度灾难的影响,难以在高维时间序列中准确地发现因果关系,本文将分位数因子模型与条件 Granger 因果关系分析相结合,提出 QFM-CGC 算法来处理高维数据的因果分析. QFM-CGC 算法将降维技术运用在降低 VAR 模型条件项的计算中,减少 VAR 模型中待估计系数,再对降维后的 VAR 模型重新进行条件 Granger 因果分析,避免传统方法受到的维度灾难的影响. 在线性仿真系统和现实宏观经济数据集上与四种基准方法和经典方法进行对比实验,结果表明,在两个不同维度的仿真实验中,本文提出的方法识别正确因果关系的概率平均提高 6% 和 3.46%. 另外,在现实宏观经济数据的实验中发现,从均方根误差、平均绝对百分误差和对称平均绝对百分比误差三个指标来看,本文提出的方法的因果关系序列的预测效果相较于最优的对比方法,分别降低了 10.47%, 5.18% 和 2.09%. 最后,本文方法的北京空气质量指标(Air Quality Index, AQI)及气象时间序列的预测效果相较于最优的对比方法,三个误差指标分别降低了 11.87%, 14.09% 和 9.79%.

1 相关理论

1.1 条件 Granger 因果关系指数 为了解决传统的 Granger 因果模型在多变量系统中容易生成虚假的因果关系的问题, Geweke^[8]引入条件变量,提出了条件 Granger 因果分析方法.

假设 $X_t = \{X_{1,t}, X_{2,t}, \dots, X_{K,t}\} (t=1, \dots, N)$ 是长度为 N 的 K 维平稳时间序列. 从驱动变量 X_i 到响应变量 X_j 的 CGCI 的定义涉及 X_j 的两个 VAR 模型, 也称动态回归模型^[15]. 第一个模型是无限模型^[16] (U-模型), 表示为:

$$X_{j,t} = \sum_{k=1}^K (a_{jk,1} X_{k,t-1} + \dots + a_{jk,p} X_{k,t-p}) + u_{j,t} \quad (1)$$

其中, p 是模型的阶数, $a_{jk,l} (k=1, \dots, K, l=1, \dots, p)$ 是 U-模型的系数.

第二个模型是从 U-模型排除 X_i 的滞后导出的受限模型 (R-模型), 表示为:

$$X_{j,t} = \sum_{k=1, k \neq i}^K (b_{jk,1} X_{k,t-1} + \dots + b_{jk,p} X_{k,t-p}) + e_{j,t} \quad (2)$$

其中, $b_{jk,l} (k=1, \dots, K \text{ 但 } k \neq i \text{ 且 } l=1, 2, \dots, p)$ 是 R-模型的系数. $u_{j,t}$ 和 $e_{j,t}$ 是均值为零, 方差分别为 σ_U^2 和 σ_R^2 的白噪声. 用普通最小二乘法拟合 U-模型和 R-模型, 得到残差 $\hat{\sigma}_U^2$ 和 $\hat{\sigma}_R^2$ 的估计值. 条件 Granger 因果关系由 CGCI 量化, 定义为 R-模型和 U-模型的误差方差之比的对数^[17], 如式 (3) 所示:

$$CGCI_{X_i \rightarrow X_j} = \ln \frac{\hat{\sigma}_R^2}{\hat{\sigma}_U^2} \quad (3)$$

显然, 当 X_i 不改进 X_j 的预测时, 即 U-模型和 R-模型给出的拟合误差方差大致相同, CGCI 处于零水平; 当 X_i 改进 X_j 预测时, CGCI 获得更大的正值, 表明 X_i Granger 导致 X_j .

1.2 分位数因子模型 分位数因子模型 (QFM) 是代表高维面板数据的一类新的因子模型, 旨在提取大型面板数据集分布均值处的公共因子^[13]. 设 $\{X_{it}\}$ 是一个由 N 个观测单位组成的面板, 且每个变量都有 T 个观测值. 那么, 在 $\tau \in (0, 1)$ 处 X_{it} 有如下的因子结构:

$$Q_{X_{it}}[\tau | f_i(\tau)] = \lambda_i'(\tau) f_i(\tau) \quad (4)$$

s.t. $i=1, 2, \dots, N$ 且 $t=1, 2, \dots, T$

其中, 公共因子 $f_i(\tau)$ 是 $r(\tau) \times 1$ 维不可观测因子向量, $\lambda_i(\tau)$ 是 $r(\tau) \times 1$ 维因子载荷向量且 $r(\tau) \ll N$, $f_i(\tau)$ 和 $\lambda_i(\tau)$ 在每个 τ 上可能不同. 为了简化符号, 下文中抑制 $f_i(\tau)$, $\lambda_i(\tau)$ 和 $r(\tau)$ 对 τ 的依赖性, 分别改写为 f_i , λ_i 和 r , 因此, 式 (4) 中的因子结构可以通过最小化目标函数获得, 如式 (5) 所示:

$$L_{NT}(\theta) = \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \rho_\tau(X_{it} - \lambda_i' f_t) \quad (5)$$

其中, $\theta = (\lambda_1', \dots, \lambda_N', f_1', \dots, f_T')'$ 为待估参数, $\rho_\tau(u) = (\tau - 1\{u \leq 0\})$ 为分位数损失函数. 对因子与因子载荷进行以下规范化:

$$\frac{1}{T} \sum_{t=1}^T f_t f_t' = I_r \quad (6)$$

$$\frac{1}{N} \sum_{i=1}^N \lambda_i \lambda_i' \text{ 为对角元素非增的对角}$$

Chen et al^[13] 给出了当样本矩阵 Y 的维度趋于无穷时因子模型估计量的渐近性质, 提出迭代分位数回归 (Iterative Quantile Regression, IQR) 算法, 可以有效地找到目标函数的平稳点. 令:

$$\Lambda = (\lambda_1, \lambda_2, \dots, \lambda_N)'$$

$$F = (f_1, f_2, \dots, f_T)'$$

并定义以下平均值:

$$M_{i,T}(\lambda, F) = \frac{1}{T} \sum_{t=1}^T \rho_\tau(X_{it} - \lambda' f_t)$$

$$M_{N,t}(\Lambda, f) = \frac{1}{N} \sum_{i=1}^N \rho_\tau(X_{it} - \lambda_i' f)$$

IQR 的迭代过程如下:

(1) 随机选择起始参数 $F^{(0)}$;

(2) 给定 $F^{(t-1)}$, 对 $i=1, \dots, N$, 计算 $\lambda_i^{(t-1)} = \arg \min_{\lambda} M_{i,T}(\lambda, F^{(t-1)})$;

给定 $\Lambda^{(t-1)}$, 对 $t=1, \dots, T$, 计算 $f_t^{(t-1)} = \arg \min_f M_{N,t}(\Lambda^{(t-1)}, f)$;

(3) 对于 $l=1, \dots, L$, 迭代第二步直到 $M_{NT}(\theta^{(L)})$ 接近 $M_{NT}(\theta^{(L-1)})$, 其中 $\theta^{(l)} = (\text{vech}(\Lambda^{(l)})', \text{vech}(F^{(l)})')'$;

(4) 规范化 $\Lambda^{(l)}$ 和 $F^{(l)}$, 使其满足 (3) 的规范化.

通过 Chen et al^[13] 提出的基于秩最小化的方法来确定不可观测因子的数量.

2 基于分位数因子模型的条件 Granger 因果关系分析

基于分位数因子模型的条件 Granger 因果关系分析模型包括以下几个部分.

2.1 平稳性分析 由于 Granger 因果模型的时间序列具有平稳性, 即该时间序列均值和方差没有系统的变化且严格消除了周期性变化, 因此, 在

建模前需要对数据进行平稳性分析. 本文采用 Augmented Dickey-Fuller (ADF)^[18] 检验, 确定序列中是否存在单位根, 帮助判断该序列是否平稳, 如式(7)所示:

$$\Delta X_t = \alpha + \beta t + \delta X_{t-1} + \sum_{i=1}^m \beta_i \Delta X_{t-i} + \varepsilon_t \quad (7)$$

其中, α 是一个常数; β 是趋势项系数; δ 是自回归系数, 描述数据的平稳性; m 是自回归模型的阶数; ε_t 是误差项. 若序列不存在单位根, 表示时间序列是平稳的, 可以直接进行 Granger 因果关系分析; 反之, 时间序列是非平稳的, 需要进行差分, 实现时间序列平稳化后再进行因果关系分析.

2.2 模型选择 VAR 模型中, 如果解释变量的最大滞后阶数 p 太小, 残差可能存在自相关, 导致参数估计不一致. 虽然适当增加滞后阶数 p 可以解决此问题, 但 p 过大会使得待估计参数增多, 严重降低自由度, 最终影响模型参数估计的有效性^[19], 所以 VAR 模型中解释变量的最大滞后阶数 p 的选择很重要. 本文采用 AIC (Akaike Information Criterion)^[20] 来自动选择合适的模型阶数, 以消除人为选择的不确定性的干扰, 如式(8)所示:

$$AIC = 2k - 2\ln L \quad (8)$$

其中, L 表示似然函数, k 是拟合模型中参数的数量. 假设模型误差服从独立正态分布, 设 n 为观测值数目, RSS 为残差平方和, 则式(8)可以改写为:

$$AIC = 2k - n \ln(RSS/n) \quad (9)$$

其中, 第一项表示模型拟合的情况, 第二项表示对模型复杂度的惩罚, 最终达到满足模型有效性和可靠性条件下参数个数最少的目的.

2.3 QFM-CGC 算法描述 根据上述推导和分析过程, 总结 QFM-CGC 算法如下所示.

算法 基于分位数因子模型的条件 Granger 因果分析

输入: 时间序列 $X, X = [X_1, X_2, \dots, X_K]^T \in R^{k \times n}$

输出: 因果关系连接矩阵

(1) ADF 检验:

$$\Delta X_t = \alpha + \beta t + \delta X_{t-1} + \sum_{i=1}^m \beta_i \Delta X_{t-i} + \varepsilon_t$$

(2) for $i = 1: K$

(3) for $j = 1: K$

(4) if $i = j$, 结束本次循环

(5) else

(6) 根据计算得到的最小 AIC 确定模型阶数

(7) for $p = 1: p_{\max}$ (p_{\max} 是时间序列 X 的最大滞后阶数)

$$Z = \text{setdiff}(X_{i,j}, (X_{i,1}, X_{i,2}, \dots, X_{i,p_{\max}}, X_{j,p}))$$

(9) 对条件变量进行分位数因子分析得到降维后的 Z_{NEW}

(10) 对 $X_i, X_{j,p}$ 和条件变量 Z_{NEW} 进行条件 Granger 因果分析建模

(11) if $p_{\text{value}} > 0.9$, 即通过显著性检验

存在 $X_i \rightarrow X_{j,p}$ 的因果关系

(12) else 不存在 $X_i \rightarrow X_{j,p}$ 的因果关系

(13) end

(14) end

(15) end

(16) end

3 仿真实验与分析

在仿真模拟研究中, 比较 QFM-CGC, 经典方法 CGC^[8] 和基准方法 PCA-CGC^[21], mBTS-CGC^[22], PMIME^[23] 的性能. Geweke^[8] 向 VAR 模型中引入条件变量, 提出条件 Granger 因果模型, 改善了传统方法无法判断直接因果关系和间接因果关系的缺陷. Zhou et al^[21] 提出 PCA-CGC 方法, 将 PCA 与条件 Granger 因果模型相结合来处理高维大脑神经网络的计算, 与传统方法相比, 降低了计算成本. Siggiridou and Kugiumtzis^[22] 采用 backward-in-time 方法对每个变量的滞后阶数使用有监督的逐步向前选择, 有效减少 VAR 模型阶数, 并与条件 Granger 因果模型结合, 提出 mBTS-CGC 方法. Kugiumtzis^[23] 将度量混合嵌入的条件互信息 (Conditional Mutual Information from Mixed Embedding, MIME) 拓展到多变量时间序列, 形成可以检测直接耦合的部分 MIME (Partial MIME, PMIME). PMIME 在由非均匀嵌入方案导出的滞后变量 X, Y 和 Z 的联合状态空间的子空间中重构一个点 (向量), 目的是最好地解释 Y 的演化, 得到的混合嵌入向量只包含所有变量中最相关的成分, 避免大维度会恶化估计的情况.

实验考虑的仿真模拟系统是两个不同维度变量的线性仿真系统, 且多项式平稳随机. 生成的多变量时间序列的平稳性要求每个时间序列的数

据在时间函数的合理范围内进行经验评估,实验结果在显著性水平 $\alpha = 0.1$ 下确定^[24].

将 QFM-CGC 方法运用到宏观经济时间序列并建立预测模型对因果分析结果进行验证,最后将仿真结果与 CGC, PCA-CGC, mBTS-CGC 和 PMIME 进行对比.

3.1 多变量线性时间序列 在两个不同维度的线性系统的仿真模拟时间序列上评估了因果关系的集合,共计 19 个,且两个随机系统均为假设. 仿真系统如下所示.

第一组数据是一个 5 维变量的 4 阶线性 VAR 系统 $\text{VAR}_5(5)$ ^[25]. 由式(10)产生:

$$\begin{aligned} X_{1,t} &= 0.3X_{1,t-1} + 0.4X_{2,t-3} + \epsilon_{1,t} \\ X_{2,t} &= 0.4X_{2,t-1} + 0.4X_{5,t-3} + \epsilon_{2,t} \\ X_{3,t} &= 0.4X_{3,t-2} - 0.4X_{1,t-1} + \epsilon_{3,t} \\ X_{4,t} &= 0.4X_{4,t-1} + 0.2X_{4,t-3} - 0.4X_{2,t-1} + \epsilon_{4,t} \\ X_{5,t} &= 0.4X_{3,t-1} + 0.4X_{5,t-2} - 0.4X_{4,t-1} + \epsilon_{5,t} \end{aligned} \quad (10)$$

其中, $\epsilon_i (i=1, \dots, 5)$ 表示高斯白噪声序列. 时间序列长度 $N=500$. 仿真系统 $\text{VAR}_5(5)$ 中真实存在的因果关系为 $X_1 \rightarrow X_3$, $X_2 \rightarrow X_1$, $X_2 \rightarrow X_4$, $X_3 \rightarrow X_5$, $X_4 \rightarrow X_5$ 和 $X_5 \rightarrow X_2$, 共计六个. 其因果关系如图 1 所示.

利用 AIC 算法选择最佳模型阶数, 图 2a~e 分别代表目标变量为 X_1, X_2, X_3, X_4, X_5 的 AIC 算法的实验结果, 最小的 AIC 对应最优的延迟阶数.

表 1 为 $\text{VAR}_5(5)$ 在 100 次蒙特卡洛实验中因

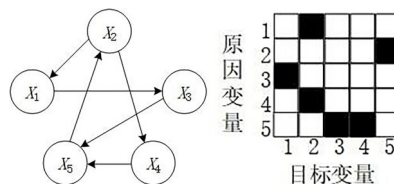


图 1 $\text{VAR}_5(5)$ 真实因果关系 (黑色表示存在因果关系)

Fig.1 The real causality diagram of $\text{VAR}_5(5)$ (Black indicates causality)

果关系的 $p > 0.9$ 的频率, 其中, 选择 QFM-CGC 方法的分位数为 0.5. 由表可见, CGC 和 PCA-CGC 不能完全正确识别式(10)中的因果关系. 其中, CGC 识别正确因果关系 $X_3 \rightarrow X_5$ 的频率仅为 3%, 错误识别直接因果关系 $X_1 \rightarrow X_2$ 的概率高达 98%, 而 PCA-CGC 除了 $X_1 \rightarrow X_3$, 其他识别正确因果关系的概率均未超过 50%. 虽然 PMIME, mBTS-CGC 和 QFM-CGC 都能正确识别所有因果关系, 但仅有 CGC 和 PMIME 受到虚假的因果关系影响, 尤其是 PMIME 受到比 CGC 更多的虚假的因果关系 $X_1 \rightarrow X_2$, $X_2 \rightarrow X_3$, $X_3 \rightarrow X_1$, $X_3 \rightarrow X_2$, $X_4 \rightarrow X_1$, $X_4 \rightarrow X_2$, $X_4 \rightarrow X_3$, $X_5 \rightarrow X_1$ 和 $X_5 \rightarrow X_3$ 的干扰, 假阳性更高. 和 mBTS-CGC 和 QFM-CGC 相比, 综合来看, 本文方法 QFM-CGC 识别正确因果关系的概率更高.

第二组数据是一个 10 维变量的 4 阶线性 VAR 系统 $\text{VAR}_{10}(5)$. 由式(11)产生:

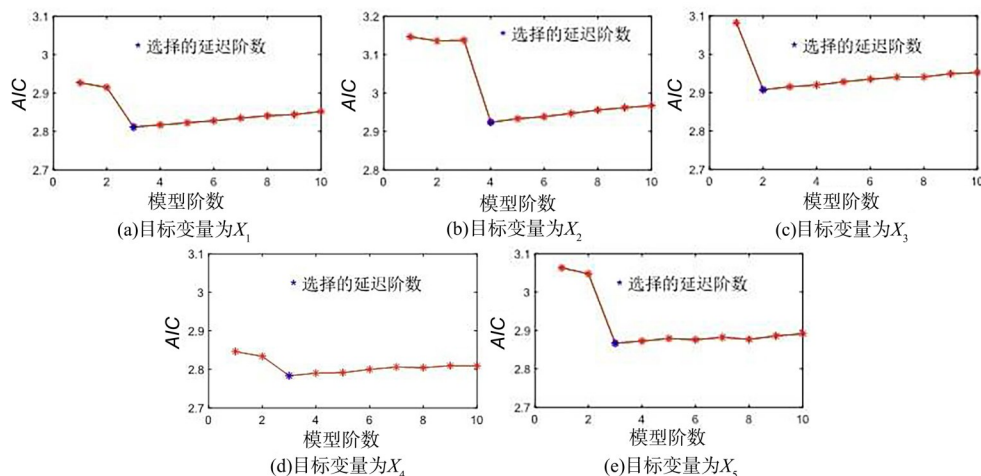


图 2 $\text{VAR}_5(5)$ 模型阶数选择

Fig.2 The order selection of $\text{VAR}_5(5)$

表1 VAR₅(5)的100次蒙特卡洛实验中因果关系结果的频率Table 1 The frequency of causality results in 100 Monte Carlo implementations of VAR₅(5)

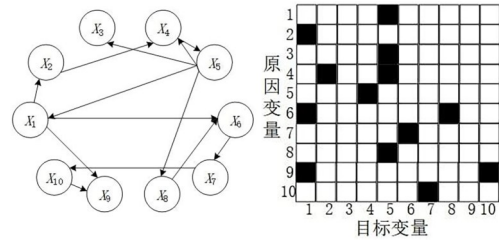
方法	CGC	PMIME	PCA- CGC	mBTS- CGC	QFM- CGC
$X_1 \rightarrow X_2$	98%	1%	0	0	0
$X_1 \rightarrow X_3$	99%	95%	100%	94%	99%
$X_2 \rightarrow X_1$	98%	99%	40%	72%	99%
$X_2 \rightarrow X_3$	0	4%	0	0	0
$X_2 \rightarrow X_4$	100%	98%	50%	98%	100%
$X_3 \rightarrow X_1$	0	8%	0	0	0
$X_3 \rightarrow X_2$	0	1%	0	0	0
$X_3 \rightarrow X_5$	3%	86%	43%	96%	100%
$X_4 \rightarrow X_1$	0	4%	0	0	0
$X_4 \rightarrow X_2$	0	1%	0	0	0
$X_4 \rightarrow X_3$	0	2%	0	0	0
$X_4 \rightarrow X_5$	100%	97%	26%	100%	99%
$X_5 \rightarrow X_1$	0	2%	0	0	0
$X_5 \rightarrow X_2$	100%	99%	18%	100%	99%
$X_5 \rightarrow X_3$	0	1%	0	0	0

$$\begin{aligned}
 X_{1,t} &= 0.4X_{1,t-3} - 0.4X_{1,t-4} + 0.4X_{5,t-3} + \epsilon_{1,t} \\
 X_{2,t} &= 0.4X_{2,t-1} - 0.3X_{2,t-5} + 0.5X_{1,t-1} + \epsilon_{2,t} \\
 X_{3,t} &= 0.4X_{3,t-5} + 0.3X_{3,t-4} - 0.4X_{5,t-3} + \epsilon_{3,t} \\
 X_{4,t} &= 0.6X_{4,t-3} + 0.3X_{5,t-3} - 0.4X_{2,t-4} + \epsilon_{4,t} \\
 X_{5,t} &= 0.3X_{5,t-1} + 0.4X_{5,t-3} - 0.3X_{4,t-1} + \epsilon_{5,t} \\
 X_{6,t} &= 0.4X_{1,t-3} + 0.4X_{6,t-2} - 0.6X_{8,t-2} + \epsilon_{6,t} \\
 X_{7,t} &= 0.3X_{7,t-1} - 0.4X_{6,t-4} + 0.3X_{7,t-3} + \epsilon_{7,t} \\
 X_{8,t} &= 0.5X_{8,t-4} - 0.2X_{8,t-2} + 0.4X_{5,t-3} + \epsilon_{8,t} \\
 X_{9,t} &= 0.5X_{9,t-3} - 0.5X_{10,t-3} + 0.4X_{1,t-3} + \epsilon_{9,t} \\
 X_{10,t} &= 0.5X_{7,t-2} - 0.5X_{10,t-3} - 0.3X_{10,t-1} + \epsilon_{10,t}
 \end{aligned} \quad (11)$$

其中, $\epsilon_i (i=1, \dots, 10)$ 表示高斯白噪声序列. 仿真系统 VAR₁₀(5) 中真实存在的因果关系为 $X_1 \rightarrow X_2$, $X_1 \rightarrow X_6$, $X_1 \rightarrow X_9$, $X_2 \rightarrow X_4$, $X_4 \rightarrow X_5$, $X_5 \rightarrow X_1$, $X_5 \rightarrow X_3$, $X_5 \rightarrow X_4$, $X_5 \rightarrow X_8$, $X_6 \rightarrow X_7$, $X_7 \rightarrow X_{10}$, $X_8 \rightarrow X_6$ 和 $X_{10} \rightarrow X_9$, 共计 13 个. 其因果关系如图 3 所示.

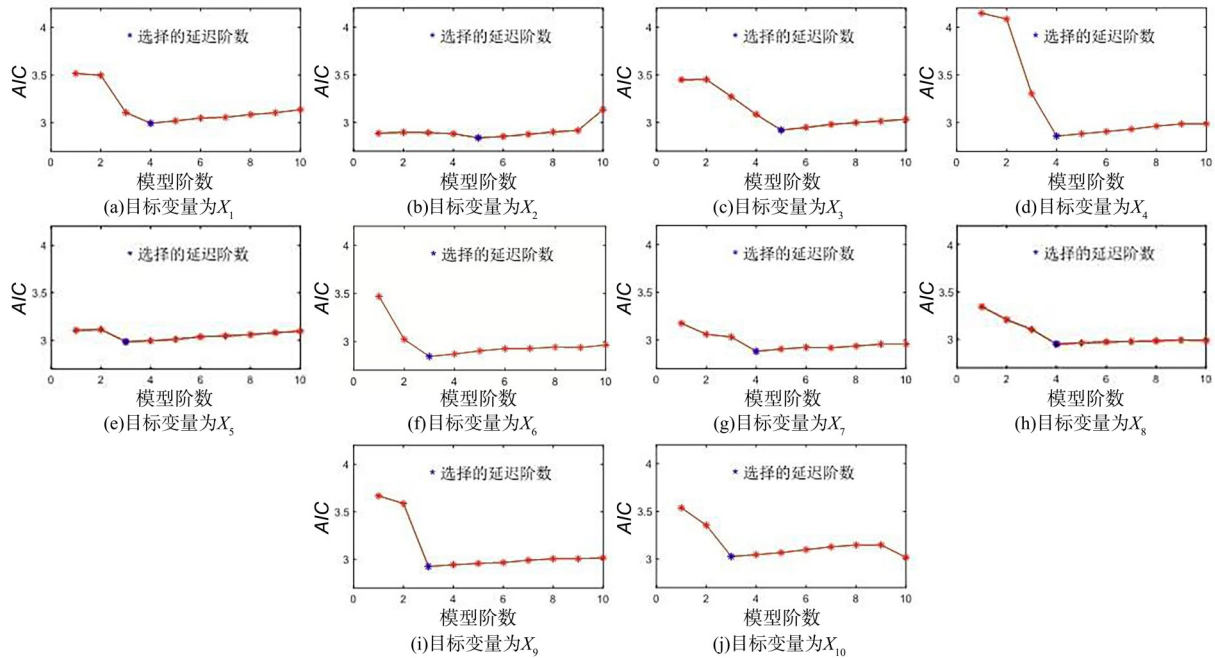
图 4 是利用 AIC 算法选择最佳模型阶数, 图 4a~j 分别代表目标变量为 $X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}$ 的 AIC 算法的实验结果, 最小的 AIC 对应最优的延迟阶数.

表 2 为 VAR₁₀(5) 的 100 次蒙特卡洛实验中因

图3 VAR₁₀(5)真实因果关系(黑色表示存在因果关系)Fig.3 The real causality diagram of VAR₁₀(5) (Black indicates causality)

果关系的 $p > 0.9$ 的频率, 选择 QFM-CGC 的分位数为 0.5. 与线性 VAR 系统 VAR₅(5) 的实验结果相似, CGC 和 PCA-CGC 无法完全正确识别式 (11) 中的因果关系, 其中 CGC 识别正确因果关系 $X_2 \rightarrow X_4$, $X_6 \rightarrow X_7$ 的频率为 0, PCA-CGC 识别正确因果关系 $X_1 \rightarrow X_9$ 的概率未超过 50%. 本仿真系统中 PMIME 仍受许多虚假的因果关系干扰, 识别正确因果关系 $X_5 \rightarrow X_4$ 的概率仅为 69%. mBTS-CGC 和 QFM-CGC 都能正确识别所有因果关系, 但 mBTS-CGC 正确识别的概率不高. 虽然本文方法错误识别了虚假因果关系 $X_6 \rightarrow X_8$, 概率为 8%, 但综合比较, 本文方法识别正确因果关系的概率更高, 表现更好.

虽然 CGC 理论上能正确区分直接因果关系和间接因果关系, 但实验结果表明, CGC 在仿真系统 VAR₅(5) 中仍然错误识别了 $X_1 \rightarrow X_2$ 因果关系, 这可能是受到间接因果关系 $X_1 \rightarrow X_3 \rightarrow X_5 \rightarrow X_2$ 的影响. 由于 PMIME 是基于 KNN 算法的, 该算法受维度灾难的影响, 对于多维度的数据处理不准确, 因为随着维度的增加, “看似相近”的两个点的距离越来越大, 就会越来越“不像”, 对于高度依赖距离的 KNN 算法其结果会影响准确率. PCA-CGC 和 QFM-CGC 虽然方法类似, 但 PCA 在降维时没有捕获隐藏的因素, 尤其是这些因素可能改变时间序列的分布特征, 造成 PCA-CGC 不能准确识别因果关系. mBTS-CGC 对每个变量的滞后阶数使用有监督的逐步向前选择, 有效地减少 VAR 模型阶数, 但在噪声的干扰下, 其中一个条件变量选择错误会引起其他因果关系的判断不准确, 最终造成该方法的假阴性较高.

图 4 $\text{VAR}_{10}(5)$ 模型阶数选择Fig.4 The order selection of $\text{VAR}_{10}(5)$ 表 2 $\text{VAR}_{10}(5)$ 的 100 次蒙特卡洛实验中因果关系结果的频率Table 2 The frequency of causality results in 100 Monte Carlo implementations of $\text{VAR}_{10}(5)$

方法	CGC	PMIME	PCA-CGC	mBTS-CGC	QFM-CGC
$X_1 \rightarrow X_2$	99%	99%	100%	99%	100%
$X_1 \rightarrow X_6$	32%	95%	75%	57%	91%
$X_1 \rightarrow X_9$	35%	98%	40%	55%	100%
$X_2 \rightarrow X_4$	0	100%	85%	99%	100%
$X_2 \rightarrow X_5$	0	1%	0	0	0
$X_4 \rightarrow X_5$	88%	99%	0	99%	100%
$X_5 \rightarrow X_1$	51%	85%	100%	57%	100%
$X_5 \rightarrow X_3$	36%	100%	100%	99%	100%
$X_5 \rightarrow X_4$	95%	69%	100%	97%	100%
$X_5 \rightarrow X_8$	23%	99%	100%	80%	94%
$X_6 \rightarrow X_7$	0	100%	99%	89%	100%
$X_6 \rightarrow X_8$	0	0	0	0	8%
$X_7 \rightarrow X_{10}$	100%	100%	100%	99%	100%
$X_8 \rightarrow X_6$	100%	96%	100%	99%	100%
$X_9 \rightarrow X_8$	0	1%	0	0	0
$X_{10} \rightarrow X_8$	0	1%	0	0	0
$X_{10} \rightarrow X_9$	100%	100%	100%	99%	100%

3.2 宏观经济时间序列 使用宏观经济时间序列进行因果分析并建立预测模型,对因果分析结果进行验证,主要目标是从高维宏观经济变量面板中找寻与实际 GDP 变化趋势有因果关系的时间序列. 该数据集由 1960 年第一季度至 2019 年第二季度的 211 个美国宏观经济变量组成 ($N=211, T=238$), 其中的数据会及时更新, 可以在网站 (<http://research.stlouisfed.org/econ/mccracken/>) 免费下载. 计算之前, 对每个序列进行平稳性处理, 代码也可以在 FRED-QD 数据网站上获得. 利用因果关系的方法找出影响宏观经济变量的主要因素, 剔除无关变量, 保留相关变量, 并将该相关变量作为预测模型的输入进行建模预测, 根据预测误差反向验证因果分析方法的有效性.

与 Chen et al^[13] 相同, 设置估计量的最大因子数 $k=8$. 使用秩最小化估计器^[13] 估计分位数为 $\{0.01, 0.05, 0.1, 0.25, 0.75, 0.9, 0.95, 0.99\}$ 时的因子估计数如表 3 所示. 由表可见, QFA 因子的数量在不同分位数之间存在显著差异, 表明该数据集存在非标准因子结构. 为了比较 QFA 因子和 PCA 因子, 将 QFA 因子的每个元素与选择的八个 PCA 因子进行回归并计算这些回归中的 R^2 , 结

果如表4^[13]所示.很明显,当 τ 接近0.5时,QFA因子与PCA因子高相关, R^2 均在0.9以上.相比之下, $\tau=0.9$ 时的第一个QFA因子(表示为 $\hat{F}_{QFA}^{0.90}$)和 $\tau=0.95, 0.99$ 时的第一个QFA因子(分别表示为 $\hat{F}_{QFA}^{0.95}$ 和 $\hat{F}_{QFA}^{0.99}$)与PCA因子的相关性较低, R^2 低于0.4.因此, $\hat{F}_{QFA}^{0.90}, \hat{F}_{QFA}^{0.95}$ 和 $\hat{F}_{QFA}^{0.99}$ 包含可能有助于预测宏观经济变量的额外信息,在此应用程序中有使用QFA的空间.由表4可得,由于 $\hat{F}_{QFA}^{0.90}, \hat{F}_{QFA}^{0.95}$ 和 $\hat{F}_{QFA}^{0.99}$ 的 R^2 分别为0.316, 0.261和0.266,与其他QFA因子相比, $\hat{F}_{QFA}^{0.95}$ 与 $\hat{F}_{QFA}^{0.90}$ 和 $\hat{F}_{QFA}^{0.99}$ 有非常高的相关性,它们具有类似的捕获额外信息的能力,这些信息能够帮助预测宏观经济变量.因此,在后续分析中重点关注 $\hat{F}_{QFA}^{0.90}$ 和 $\hat{F}_{QFA}^{0.99}$ 的预测能力.

使用不同方法进行因果关系分析后,选出与目标变量具有因果关系的原因变量作为模型的输

表3 不同分位数下的因子估计数

Table 3 Estimation of factors at different quantiles

分位数 τ	因子个数
0.01	1
0.05	1
0.10	2
0.25	4
0.50	5
0.75	5
0.90	2
0.95	1
0.99	1

表4 \hat{F}_{QFA} 和 \hat{F}_{PCA} 的比较结果

Table 4 Comparison of \hat{F}_{QFA} and \hat{F}_{PCA}

分位数 τ	\hat{F}_{QFA}^τ 的元素个数				
	1	2	3	4	5
0.01	0.657				
0.05	0.733				
0.10	0.796	0.871			
0.25	0.952	0.932	0.939	0.890	
0.50	0.993	0.976	0.964	0.945	0.923
0.75	0.906	0.945	0.943	0.903	0.882
0.90	0.316	0.911			
0.95	0.261				
0.99	0.266				

入进行预测,并对分析结果进行进一步的验证.采用CNN-LSTM预测模型来分析每一种方法得出的因变量进行建模的预测效果,进行30次实验,取平均值来消除偶然因素对实验结果的影响.最后,采用均方根误差(RMSE)、平均绝对百分误差(MAPE)和对称平均绝对百分比误差(SMAPE)三个指标来定量评价预测精度,三个评价指标的定义如下:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (12)$$

$$MAPE = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{\hat{y}_i - y_i}{y_i} \right| \quad (13)$$

$$SMAPE = \frac{100\%}{n} \sum_{i=1}^n \frac{|\hat{y}_i - y_i|}{\frac{|y_i| + |\hat{y}_i|}{2}} \quad (14)$$

其中, y_i 和 \hat{y}_i 分别是真实值和预测值, n 是样本个数.

表5是不同方法预测GDP的精度比较,表中黑体字表示最优值.图5~10展示了不同方法预测的GDP变化趋势.由表5可见,本文方法在0.90分位数条件下的RMSE,MAPE和SMAPE都是最小的,并在预测对比图中具有明显的优越性拟合效果,能更精准地追踪GDP的变化趋势.

从表5还可以看出,QFM-CGC识别出对GDP具有因果关系的变量主要与个人消费支出、私人固定投资、生产制造和消费有关.内需、投资和出口俗称拉动经济增长的“三驾马车”,尤其是消费需求是生产的目的,消费可以创造出生产的动力,并刺激投资需求,以此促进经济发展.然而,CGC未能识别出投资与GDP具有因果关系,

表5 GDP预测结果

Table 5 The prediction of GDP

对比方法	因变量(编号)	RMSE	MAPE	SMAPE
CGC	28,64,74,104, 116,162	2.53558	2.30354	1.06974
mBTS-CGC	4,9,11,16,18,26,36, 60,66,86,89,102, 138,141,148,203	2.37887	2.01709	1.06333
PCA-CGC	6,70,71,77,141,148	2.24680	2.06530	1.03592
PMIME	70,137,161,163	1.90992	1.54485	0.95669
QFM-CGC ($\tau=0.90$)	2,4,7,10,21, 79,163,171	1.70379	1.46841	0.93114
QFM-CGC ($\tau=0.99$)	2,4,7,10,21,26, 79,160,163,189	1.80852	1.58768	0.95876

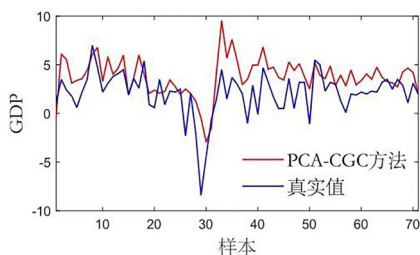


图 5 PCA-CGC 的 GDP 预测图

Fig. 5 The prediction of GDP by PCA-CGC

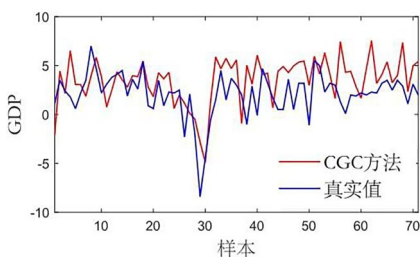
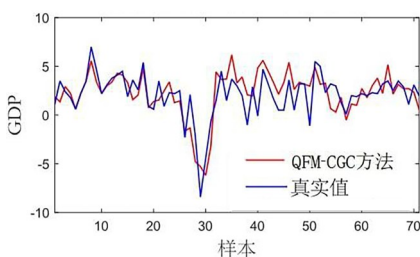


图 7 CGC 的 GDP 预测图

Fig. 7 The prediction of GDP by CGC

图 9 QFM-CGC ($\tau = 0.90$) 的 GDP 预测图Fig. 9 The prediction of GDP by QFM-CGC ($\tau = 0.90$)

因此使用 CGC 来预测 GDP 造成的误差最大. 虽然 mBTS-CGC 可以识别出许多因变量, 但其中可能包含错误的因变量(如货币存量、国库券等), 这些因变量会干扰预测结果, 导致预测误差较大. PCA-CGC 和 PMIME 识别出制造业和非监督员工的平均每周工作小时数与 GDP 之间有 Granger 因果关系, 加班和额外的工作时间可能会增加生产和服务活动, 对 GDP 产生积极影响. 然而, 过度的长时间工作可能导致劳动力疲劳, 影响效率或产生健康问题, 可能减少 GDP. 尽管制造业和非监督员工的平均每周工作小时数可能与 GDP 相关, 但它们之间不一定存在 Granger 因果关系, 因为过高或过低都会对 GDP 产生负面效应. 实验结果与 GDP 影响关系是一致的, 进一步验证了本文方法的有效性.

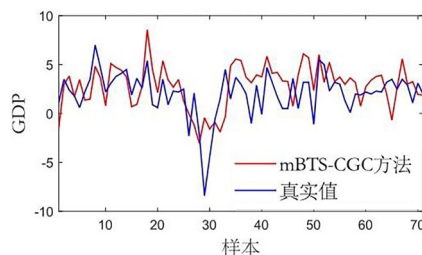


图 6 mBTS-CGC 的 GDP 预测图

Fig. 6 The prediction of GDP by mBTS-CGC

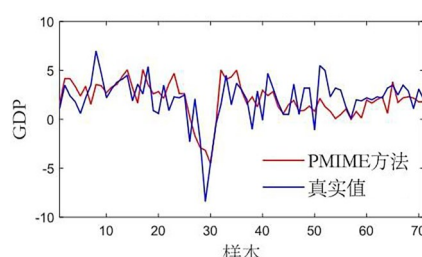
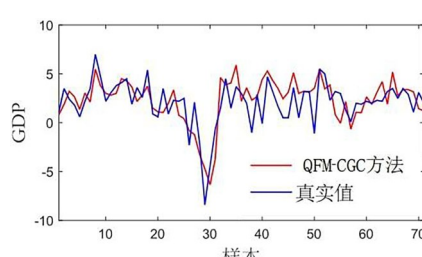


图 8 PMIME 的 GDP 预测图

Fig. 8 The prediction of GDP by PMIME

图 10 QFM-CGC ($\tau = 0.99$) 的 GDP 预测图Fig. 10 The prediction of GDP by QFM-CGC ($\tau = 0.99$)

3.3 北京 AQI 及气象时间序列 使用北京 AQI 及气象数据集进行因果分析, 并建立预测模型, 对因果分析的结果进行验证. 该数据集选用 2016 年 1 月 1 日到 2016 年 6 月 15 日的每小时数据, 共 4008 个样本, 每个样本包括 6 维 AQI 时间序列和 5 维气象时间序列, 详细描述见表 6.

对每个对比模型的因果分析结果进行预测,

表 6 北京 AQI 及气象时间序列编号及变量对照表

Table 6 The number and variable comparison table of Beijing AQI and meteorological time series

编号	1	2	3	4	5	6
变量	PM _{2.5}	PM ₁₀	SO ₂	NO ₂	CO	O ₃
编号	7	8	9	10	11	
变量	气温	气压	露点	降雨量	风速	

根据预测的准确性来判断不同因果分析方法的有效性. 与上节相同, 设置估计量的最大因子数 $k=8$. 使用 Chen et al^[13] 的秩最小化估计器在分位数为 $\{0.01, 0.05, 0.1, 0.25, 0.75, 0.9, 0.95, 0.99\}$ 时进行因子估计, 不同分位数条件下的因子估计数同为 1. 此外, 将 QFA 因子的每个元素与选择的八个 PCA 因子进行回归并计算这些回归中的 R^2 , 分位数在 $\{0.01, 0.05, 0.1, 0.25, 0.75, 0.9, 0.95, 0.99\}$ 时的 R^2 分别为 0.794, 0.786, 0.999, 0.999, 0.999, 0.984, 0.945, 0.961. $\tau=0.01$ 时的 QFA 因子 (表示为 $\hat{F}_{QFA}^{0.01}$) 和 $\tau=0.05$ 时的 QFA 因子 (表示为 $\hat{F}_{QFA}^{0.05}$) 与 PCA 因子的相关性较低, R^2 低于 0.8. 因此, $\hat{F}_{QFA}^{0.01}$ 和 $\hat{F}_{QFA}^{0.05}$ 包含可能有助于预测宏观经济变量的额外信息, 在此应用程序中有使用 QFA 的空间.

使用不同的方法进行因果关系分析后, 选出与目标变量具有因果关系的原因变量作为模型的输入进行预测, 并对分析结果进行进一步的验证. 表 7 是不同方法对 NO_2 的因果关系分析结果并使用得到的因变量采用 CNN-LSTM 预测模型预测 NO_2 的精度比较, 表中黑体字表示最优值.

从表 7 可以发现, 本文方法在低分位数 0.01 和 0.05 时的 $RMSE$, $MAPE$ 和 $SMAPE$ 都是最小的. 此外, 结合气象学等领域的实际背景, 进一步分析因果关系分析结果的应用价值. 露点是空气中水蒸气达到饱和和所需的温度, 而 SO_2 是一种气体, 在大气中可能对人类健康和环境造成负面影响, 虽然 SO_2 的排放会随着湿度升高而下降, 但这不是因为露点与 SO_2 之间存在因果关系, mBTS-CGC 和 PCA-CGC 错误判断了露点与 SO_2 之间的因果关系. $\text{PM}_{2.5}$, PM_{10} 和 SO_2 都是大气污染物, SO_2 在大气中与水蒸气、氧气等物质相互作用, 形成硫酸盐颗粒物, 并与大气中的其他颗粒物结合生成复合颗粒物. 这些复合颗粒物包括 $\text{PM}_{2.5}$ 和 PM_{10} 等, 因此 SO_2 与 $\text{PM}_{2.5}$ 和 PM_{10} 之间可能存在一定的因果关系, 然而其他方法没有捕获到因变量 $\text{PM}_{2.5}$ 和 PM_{10} . 因此, 通过从理论出发分析所提模型和对比模型的因果关系分析的结果, 进一步证明了本文方法的有效性.

表 7 NO_2 的预测结果Table 7 The prediction of NO_2

对比方法	因变量(编号)	$RMSE$	$MAPE$	$SMAPE$
mBTS-CGC	5, 9, 10, 11	5.89589	0.49675	0.02104
PCA-CGC	2, 5, 9	3.37085	0.56704	0.02329
PMIME, CGC	6, 7, 11	5.37264	0.49821	0.02043
QFM-CGC ($\tau=0.01, 0.05$)	1, 2, 3, 5, 7	2.46499	0.42411	0.01817

4 结论

传统的 Granger 因果模型由 VAR 模型推导, 但当 VAR 模型中涉及大量待估计系数时容易产生维度灾难, 因此, 使用一般的因果分析方法在高维时间序列中难以准确地发现因果关系, 目前对于高维和超高维的时间序列因果分析缺少有效的处理方法. 针对高维线性时间序列因果关系的识别问题, 本文提出 QFM-CGC 方法, 使用分位数因子模型对条件 Granger 因果分析方法的条件变量进行降维, 解决了 VAR 模型中存在大量待估计系数的问题, 可以有效识别高维线性时间序列的因果关系, 尤其是在数据包含异常值时, 使用分位数因子模型降维具有更明显的优势. 大量实验证明, 将降维技术与条件 Granger 因果分析相结合, 能准确识别直接因果关系和间接因果关系.

本文提出的方法能在平稳和线性的时间序列下进行建模, 并实现较好的因果分析结果, 但现实中部分数据具有非线性特征, 因此, 未来将对高维非线性时间序列因果关系的分析展开研究. 此外, 当时间序列非平稳时, 处理时间序列平稳化的过程可能会使原时间序列的结果和意义发生变化, 导致因果分析方法的解释意义发生变化.

参考文献

- [1] Granger C W J. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 1969, 37(3): 424—438.
- [2] Wismüller A, Vosoughi M A, DSouza A, et al. Exploring directed network connectivity in complex systems using large-scale augmented Granger causality // *Proceedings of SPIE 12033, Medical Imaging 2022: Computer-Aided Diagnosis*. San Diego, CA, USA: SPIE, 2022: 168—177.

- [3] Maradana R P, Pradhan R P, Dash S, et al. Innovation and economic growth in European Economic Area countries: The Granger causality approach. *IIMB Management Review*, 2019, 31(3): 268—282.
- [4] Billio M, Getmansky M, Lo A W, et al. Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of Financial Economics*, 2012, 104(3): 535—559.
- [5] Chang T, Gupta R, Inglesi-Lotz R, et al. Renewable energy and growth: Evidence from heterogeneous panel of G7 countries using Granger causality. *Renewable and Sustainable Energy Reviews*, 2015 (52): 1405—1412.
- [6] Hlinka J, Hartman D, Vejmelka M, et al. Reliability of inference of directed climate networks using conditional mutual information. *Entropy*, 2013, 15(6): 2023—2045.
- [7] Blinowska K J, Kuś R, Kamiński M. Granger causality and information flow in multivariate processes. *Physical Review E*, 2004, 70(5): 050902.
- [8] Geweke J. Measurement of linear dependence and feedback between multiple time series. *Journal of the American Statistical Association*, 1982, 77(378): 304—313.
- [9] 李松, 胡晏铭, 郝晓红, 等. 基于维度分组降维的高维数据近似 k 近邻查询. *计算机研究与发展*, 2021, 58(3): 609—623. (Li S, Hu Y M, Hao X H, et al. Approximate k -nearest neighbor query of high dimensional data based on dimension grouping and reducing. *Journal of Computer Research and Development*, 2021, 58(3): 609—623.)
- [10] 刘淑伟, 陈威, 赵伟, 等. 基于簇内乘积量化的最近邻检索方法. *计算机学报*, 2020, 43(2): 303—314. (Liu S W, Chen W, Zhao W, et al. Nearest neighbor search based on product quantization in clusters. *Chinese Journal of Computers*, 2020, 43(2): 303—314.)
- [11] Wold S, Esbensen K, Geladi P. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 1987, 2(1—3): 37—52.
- [12] Abdi H, Williams L J. Principal component analysis. *WIREs Computational Statistics*, 2010, 2(4): 433—459.
- [13] Chen L, Dolado J J, Gonzalo J. Quantile factor models. *Econometrica*, 2021, 89(2): 875—910.
- [14] Mooney C Z. Monte Carlo simulation. Thousand Oaks: Sage Publications, 1997, 103.
- [15] Pankratz A. Forecasting with dynamic regression models. Hoboken: John Wiley & Sons, 2012, 400.
- [16] Brandt P T, Williams J T. Multiple time series models. Sage Publications, 2006, 120.
- [17] Guo S X, Seth A K, Kendrick K M, et al. Partial Granger causality—eliminating exogenous inputs and latent variables. *Journal of Neuroscience Methods*, 2008, 172(1): 79—93.
- [18] Dickey D A, Fuller W A. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 1979, 74(366): 427—431.
- [19] Barber R F, Drton M. High-dimensional Ising model selection with Bayesian information criteria. *Electronic Journal of Statistics*, 2015, 9(1): 567—607.
- [20] Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 1974, 19(6): 716—723.
- [21] Zhou Z Y, Chen Y H, Ding M Z, et al. Analyzing brain networks with PCA and conditional Granger causality. *Human Brain Mapping*, 2009, 30(7): 2197—2206.
- [22] Siggiridou E, Kugiumtzis D. Granger causality in multivariate time series using a time - ordered restricted vector autoregressive model. *IEEE Transactions on Signal Processing*, 2016, 64(7): 1759—1773.
- [23] Kugiumtzis D. Direct-coupling information measure from nonuniform embedding. *Physical Review E*, 2013, 87(6): 062918.
- [24] Quiroga R Q, Kraskov A, Kreuz T, et al. Performance of different synchronization measures in real data: A case study on electroencephalographic signals. *Physical Review E*, 2002, 65(4): 041903.
- [25] Jia Z Y, Lin Y F, Liu Y X, et al. Refined nonuniform embedding for coupling detection in multivariate time series. *Physical Review E*, 2020, 101(6): 062113.

(责任编辑 杨可盛)