中国科学 : 数学　2020 年　第 50 卷　第 ? 期 : 1 ∼ 14
SCIENTIA SINICA Mathematica
论　文
《中国科学》杂志社
SCIENCE CHINA PRESS

广义网络向量自回归

王菲菲[1]，朱雪宁[2*]，潘蕊[3]

1. 中国人民大学统计学院, 北京 100872;
2. 复旦大学大数据学院, 上海 200433;
3. 中央财经大学统计与数学学院, 北京 100081

E-mail: feifei.wang@ruc.edu.cn, xueningzhu@fudan.edu.cn, panrui_cufe@126.com

收稿日期: 2018-11-22; 接受日期: 2019-08-12; 网络出版日期: 2020-07-09; * 通信作者
国家自然科学基金 (批准号: 11901105, 11690015, U1811461 和 11971504)、中国人民大学科学研究基金 (中央高校基本科研业务费专项资金) (批准号: 18XNLG02)、上海市科技人才计划 (批准号: 19YF1402700)、复旦新再灵大数据联合实验室、中央财经大学“青年英才”项目 (批准号: QYP1911) 和中央高校基本科研业务费 (批准号: 20190107) 资助项目

摘要　随着社交网络平台的快速发展, 带有网络结构的时序数据越来越多. 为拟合用户行为的动态变化, 网络向量自回归模型被提出. 模型最早研究的是连续型因变量. 然而实际数据常观测到离散型因变量. 由此, 本文提出广义网络向量自回归模型. 模型假设存在一个潜在的连续型变量, 决定了可观测到的离散型因变量的取值. 为了估计和推断模型, 本文提出了 MCMC (Markov chain Monte Carlo) 算法并通过随机模拟进行验证. 最后, 使用某社交网络平台上的两个真实的数据案例作为例证.

关键词　网络数据　MCMC 算法　网络向量自回归　潜在变量

MSC (2010) 主题分类　62F15, 62H11

1　引言

随着互联网的快速发展, 社交网络数据越来越多. 如今, 典型的在线社交网络平台包括 Facebook (www.facebook.com)、Twitter (www.twitter.com) 和新浪微博 (www.weibo.com) 等. 社交网络由许多节点及节点之间的网络关系组成. 例如, 在微博社交关系网络中, 一个用户就是一个节点, 用户之间的相互关注关系就是网络关系 [1]. 在引文网络中, 一篇论文就是一个节点, 论文间的引用关系就是网络关系 [2]. 与传统数据不同, 网络结构数据不再独立, 而是包含了个体之间的关系信息, 在研究社会关系方面具有巨大的商业价值. 因此, 社交网络数据的统计分析在各个学科中具有广泛的应用, 例如, 在组织管理学中研究同伴效应, 在人口学中基于社交网络研究移民特点, 等等 [3–5].

除了网络关系之外, 从每个节点收集的因变量和自变量也可以是时间序列数据. 一般而言, 回归模型被广泛用于研究因变量与自变量的相关关系. 而空间自回归模型的提出既可以将网络关系考虑在内, 又可以拟合时间序列数据的动态变化 [6–8]. 在该模型框架下, 假设因变量 Y_{it} (其中, $i = 1, \ldots, N$,

英文引用格式: Wang F F, Zhu X N, Pan R. Generalized network vector autoregression (in Chinese). Sci Sin Math, 2020, 50: 1–14, doi: 10.1360/SCM-2018-0839

$t = 1, \ldots, T$, N 为节点个数, T 为时间周期) 受到 4 个因素的影响. 首先, Y_{it} 受到滞后期 $Y_{i(t-1)}$ 的影响. 其次, Y_{it} 受到其关注者的影响. 第三, Y_{it} 受到其节点相应自变量的影响. 最后, 独立的随机噪声项也会对因变量产生影响. 基于空间向量自回归模型的应用很多, 例如, Chen 等 [9] 和 Zhou 等 [10] 研究推特类型的在线社交网络, 他们发现用户的发文行为之间存在正向网络效应; Zhu 等 [1] 提出了专门用于研究动态社交行为的网络向量自回归模型; Cohen-Cole 等 [11] 在研究学生的社交活动中发现内部选择和交叉选择的同伴效应对青少年行为有非平凡的影响.

但是, 所有上述研究都关注于连续型因变量, 例如, 社交网络中的用户活跃度. 在实际数据分析中, 经常会遇到离散型因变量. 例如, 网络用户的决策行为被记录为 0-1 变量 (例如, 购买与否); 网络平台上的发布帖子数被记录为计数变量. 在已有的文献中, 状态空间模型 [12] 被广泛用于拟合动态离散型因变量. 模型的基本假定为观察到的因变量是由其潜在状态决定的. 关于状态空间模型估计与应用的相关研究可以参见文献 [13–15]. 本文在空间向量自回归模型和状态空间模型的基础上, 提出了广义网络向量自回归模型.

本文的主要贡献有以下三点. 首先, 借鉴空间自回归模型和状态空间模型的优点提出广义网络向量自回归模型. 因此, 该模型可以在考虑节点间网络关系的同时, 研究网络结构中的动态离散型因变量数据. 其次, 我们提出了 Metropolis Hastings-within-Gibbs MCMC 算法来进行模型估计. 具体而言, 通过设定先验分布并计算似然函数, 可以推导出每个未知参数的完全条件分布, 然后根据其具体形式进行 MCMC 抽样求解. 第三, 以两个真实的社交网络数据集作为例证, 进一步展示广义网络向量自回归模型的应用场景, 其扩展了先前的实证分析框架, 并获得了更为丰富的模型解释.

本文余下内容的结构如下. 第 2 节介绍广义网络向量自回归模型. 第 3 节提出广义网络向量自回归模型的 MCMC 估计方法, 并在第 4 节通过随机模拟进行验证. 第 5 节应用广义网络向量自回归模型对两个实际案例进行分析. 第 6 节进行总结和讨论.

2 广义网络向量自回归模型

假设网络中具有 N 个节点 ($i = 1, \ldots, N$). 使用邻接矩阵 $A = (a_{ij}) \in \mathbb{R}^{N \times N}$ 来刻画网络中节点间的关系. 具体来说, 如果第 i 个节点与第 j 个节点相连, 那么 $a_{ij} = 1$, 否则 $a_{ij} = 0$. 依照惯例, 令 $a_{ii} = 0$, $i = 1, \ldots, N$. 对于每个节点 i, 假设在时间点 t 收集到离散型因变量为 Y_{it} ($t = 1, \ldots, T$).

为描述因变量 Y_{it} 的统计特征, 假设它由一个连续的潜在变量 Z_{it} 决定, 该变量也被称为状态变量. 接下来, 设 $\mathbb{Y}_t = (Y_{1t}, \ldots, Y_{Nt})^\top \in \mathbb{R}^N$, $\mathbb{Z}_t = (Z_{1t}, \ldots, Z_{Nt})^\top \in \mathbb{R}^N$. 令 $\mathcal{F}_t$ 表示由 $\{\mathbb{Y}_s, \mathbb{Z}_s : s \leqslant t\}$ 生成的 σ- 域. 假设 Y_{it} 的条件概率满足

$$\mathrm{P}(Y_{it} \mid \mathcal{F}_t) = \mathrm{P}(Y_{it} \mid Z_{it}). \tag{2.1}$$

可以看出, Y_{it} 的分布具有无记忆性, 只取决于它对应的状态变量 Z_{it}. 本文假定 $\mathrm{P}(Y_{it} \mid Z_{it})$ 具有已知的参数形式. 例如, 如果 $Y_{it} \in \{0, 1\}$ 是一个 0-1 型变量, 那么 Y_{it} 的条件概率可以写成如下形式:

$$\mathrm{P}(Y_{it} = 1 \mid Z_{it}) = \frac{\exp(Z_{it})}{1 + \exp(Z_{it})}. \tag{2.2}$$

再例如, 假设 $Y_{it} \in \{0, 1, 2, \ldots\}$ 是一个计数变量, Y_{it} 的条件分布是 Poisson 分布, 其条件分布为

$$\mathrm{P}(Y_{it} = k \mid Z_{it}) = \frac{\lambda_{it}^k \exp(-\lambda_{it})}{k!}, \quad k = 0, 1, 2, \ldots, \tag{2.3}$$

其中 $\lambda_{it}=\exp(Z_{it})$. 除此之外还可以考虑 Y_{it} 和 Z_{it} 之间其他形式的连接函数, 如负二项分布等.

经过以上讨论, 我们知道 Y_{it} 的分布完全取决于 Z_{it}. 因此, 为了研究 Y_{it} 的动态变化, 可以用连续变量 Z_{it} 作为替代进行建模. 为了对 Z_{it} 建模, 考虑 Zhu 等 [1] 提出的网络向量自回归模型, 即

$$Z_{it}=\beta_0+\beta_1 n_i^{-1}\sum_{j=1}^{N}a_{ij}Z_{j(t-1)}+\beta_2 Z_{i(t-1)}+V_i^{\top}\gamma+\varepsilon_{it}, \tag{2.4}$$

其中, $n_i=\sum_j a_{ij}$ 代表个体 i 的出度, 即个体 i 与多少条边相连, 在一定程度上反应个体 i 的活跃度. $V_i=(V_{i1},\ldots,V_{ip})^{\top}\in\mathbb{R}^p$ 是不随时间改变的 p 维自变量. 另外, 噪声项 ε_{it} 相互独立且服从 $N(0,\sigma^2)$ 的正态分布. 参数 $\beta_0\in\mathbb{R}^1$ 是基准效应. 参数 $\beta_1\in\mathbb{R}^1$ 和 $\beta_2\in\mathbb{R}^1$ 分别称为网络效应和自回归效应. 具体而言, β_1 刻画了与其相连节点的影响, β_2 刻画了节点自身滞后期的影响. 最后, $\gamma\in\mathbb{R}^p$ 刻画了不随时间改变的自变量 (如性别、位置) 产生的影响. 本文基于模型 (2.1) 和 (2.4) 提出广义网络向量自回归模型.

记 $W=(w_{ij})\in\mathbb{R}^{N\times N}$ 为行标准化后的邻接矩阵, 其中 $w_{ij}=n_i^{-1}a_{ij}$. 将 $\mathbb{V}=(V_1,\ldots,V_N)^{\top}\in\mathbb{R}^{N\times p}$ 记为节点自变量矩阵. 令 $\mathcal{E}_t=(\varepsilon_{1t},\ldots,\varepsilon_{Nt})^{\top}\in\mathbb{R}^N$, $\mathbb{Z}_t=(Z_{1t},\ldots,Z_{Nt})^{\top}$. 基于上述定义, (2.4) 可以改写为

$$\mathbb{Z}_t=\beta_0+\beta_1 W\mathbb{Z}_{t-1}+\beta_2\mathbb{Z}_{t-1}+\mathbb{V}\gamma+\mathcal{E}_t. \tag{2.5}$$

需要注意的是, 广义网络向量自回归模型也可以被视为状态空间模型 [15,16]. 具体来说, 给定潜在变量 Z_{it}, 离散型因变量 Y_{it} 满足 “观测方程” (2.1); 而 Z_{it} 的动态变化和网络相关性则由 “状态方程” (2.4) 进行描述. 在本文中, 广义网络向量自回归模型的目的是估计参数 $\theta=(\beta_0,\beta_1,\beta_2,\gamma^{\top})^{\top}\in\mathbb{R}^{3+p}$, 并对潜在状态 $\mathbb{Z}_t$ 进行推断.

3 模型估计

本节将进一步讨论广义网络向量自回归模型的参数估计方法. 如果潜在状态 $\{\mathbb{Z}_t:1\leqslant t\leqslant T\}$ 是可观测的, 则模型 (2.4) 中的未知参数可以通过 Zhu 等 [1] 给出的普通最小二乘法进行估计. 然而, 在实际情况中, 潜在状态是不可观测的, 因此使用普通最小二乘法进行估计是不可行的. 为了解决这个问题, 我们采用了 Metropolis Hastings-within-Gibbs MCMC 算法进行估计.

令 $\mathbb{Y}=(\mathbb{Y}_1^{\top},\ldots,\mathbb{Y}_T^{\top})^{\top}\in\mathbb{R}^{NT}$, $\mathbb{Z}=(\mathbb{Z}_1^{\top},\ldots,\mathbb{Z}_T^{\top})^{\top}\in\mathbb{R}^{NT}$. 假设 $t=0$ 时的初始状态变量为 0, 即 $\mathbb{Z}_0=\mathbf{0}$. 然后, 给定 $\mathbb{Y}$、$\mathbb{V}$ 和 $\mathbb{Z}_0$, 状态变量 $\mathbb{Z}$ 和未知参数 $\{\theta,\sigma^2\}$ 的联合后验分布推导如下:

$$\begin{aligned}f(\mathbb{Z},\theta,\sigma^2\mid\mathbb{Y},\mathbb{Z}_0,\mathbb{V})&\propto f(\mathbb{Y}\mid\mathbb{Z})f(\mathbb{Z}\mid\mathbb{Z}_0,\mathbb{V},\theta,\sigma^2)f(\theta)f(\sigma^2)\\&=\left\{\prod_{t=1}^{T}\prod_{i=1}^{N}f(Y_{it}\mid Z_{it})\right\}\left\{\prod_{t=1}^{T}f(\mathbb{Z}_t\mid\mathbb{Z}_{t-1},\mathbb{V},\theta,\sigma^2)\right\}f(\theta)f(\sigma^2),\end{aligned} \tag{3.1}$$

其中符号 $\propto$ 表示成比例, $f(\theta)$ 和 $f(\sigma^2)$ 分别表示 θ 和 σ^2 的先验分布.

给定 (3.1) 中 $\{\mathbb{Z},\theta,\sigma^2\}$ 的联合后验分布, 可以很容易地获得每个状态变量和每个参数的完全条件分布. 随后, 我们设计了 Metropolis Hastings-within-Gibbs 采样算法用于模型估计. 在下文中, 首先在第 3.1 小节中推导状态变量 Z_{it} 的完全条件分布, 进而在第 3.2 小节中推导参数 θ 和 σ^2 的完全条件分布. 整个 MCMC 抽样过程参见算法 1.

算法 1 广义网络向量自回归模型的 MCMC 抽样算法

1. 初始化状态变量 $\{Z_{it}: 1 \leqslant i \leqslant N, 1 \leqslant t \leqslant T\}$ 以及参数 $\{\theta, \sigma^2\}$.
2. 设置迭代器 $m = 1, 2 \ldots$, 重复以下步骤直至收敛:
   (a) 对于 $i = 1, \ldots, N$ 和 $t = 1, \ldots, T$, 使用 Metropolis-Hastings 抽样算法, 根据其后验分布 $\widetilde{f}(Z_{it})$ 更新状态变量 Z_{it};
   (b) 从完全条件分布 $N(\mu_\theta, \Sigma_\theta)$ 中抽样更新参数 θ;
   (c) 从完全条件分布 INV-$\chi^2(NT, s^2)$ (即带尺度的逆 χ^2 分布) 中抽样更新参数 σ^2.

3.1 潜在状态变量的完全条件分布

本节将计算 Z_{it} 的完全条件分布. 令 $X_{it} = (1, w_i^\top \mathbb{Z}_{t-1}, Z_{i(t-1)}, V_i^\top)^\top \in \mathbb{R}^{3+p}$ 和 $\mathbb{Z}_{-(it)} = \{Z_{js} : j \neq i, s \neq t\}$. 用 $\widetilde{f}(Z_{it}) = f(Z_{it} \mid \mathbb{Z}_{-(it)}, \mathbb{Y}, \mathbb{Z}_0, \mathbb{V}, \theta, \sigma^2)$ 表示 Z_{it} 的完全条件分布. 根据联合后验分布 (3.1), 可以得到 Z_{it} 的完全条件分布

$$\begin{aligned}
\widetilde{f}(Z_{it}) &\propto f(\mathbb{Z}, \theta, \sigma^2 \mid \mathbb{Y}, \mathbb{Z}_0, \mathbb{V}) \\
&\propto \mathrm{P}(Y_{it} \mid Z_{it}) f(Z_{it} \mid \mathbb{Z}_{t-1}, \mathbb{V}, \theta, \sigma^2) f(\mathbb{Z}_{t+1} \mid \mathbb{Z}_t, \mathbb{V}, \theta, \sigma^2) \\
&\propto \mathrm{P}(Y_{it} \mid Z_{it}) \exp\left\{ -\frac{1}{2\sigma^2}(Z_{it} - X_{i(t-1)}^\top \theta)^2 \right\} \prod_{j \neq i} \exp\left\{ -\frac{1}{2\sigma^2}(Z_{j(t+1)} - X_{jt}^\top \theta)^2 \right\}. \qquad (3.2)
\end{aligned}$$

(3.2) 中的第二行来自于 (2.4) 中假定的动态关系. 由于 (3.2) 中的完全条件分布不是标准分布形式, 因此, 直接根据 $\widetilde{f}(Z_{it})$ 对 Z_{it} 进行抽样是很困难的. 为此, 我们使用 Metropolis-Hastings 抽样算法来更新 Z_{it}. 在 Metropolis-Hastings 算法中, 需要找到一个提议分布, 它需要近似于真实的完全条件分布 $\widetilde{f}(Z_{it})$ 并且易于抽样.

为了找到合适的提议分布, 我们深入研究 $\widetilde{f}(Z_{it})$ 的性质. 首先, 可以写出 $\widetilde{f}(Z_{it})$ 的对数形式:

$$\begin{aligned}
\log\{\widetilde{f}(Z_{it})\} &= \log\{\mathrm{P}(Y_{it} \mid Z_{it})\} + \log\{f(Z_{it} \mid \mathbb{Z}_{t-1}, \mathbb{V}, \theta, \sigma^2) f(\mathbb{Z}_{t+1} \mid \mathbb{Z}_t, \mathbb{V}, \theta, \sigma^2)\} + C_1 \\
&= \log\{\mathrm{P}(Y_{it} \mid Z_{it})\} - \frac{1}{2\sigma^2}\{(Z_{it} - E_{i(t-1)})^2 + (F_{i(t+1)} - \beta_2 Z_{it})^2\} \\
&\quad - \frac{1}{2\sigma^2}\left\{ \sum_{j \neq i}^{N} (G_{j(t+1)} - \beta_1 w_{ji} Z_{it})^2 \right\} + C_1, \qquad (3.3)
\end{aligned}$$

其中

$$\begin{aligned}
E_{i(t-1)} &= \beta_0 + \beta_1 \sum_j w_{ij} Z_{j(t-1)} + \beta_2 Z_{i(t-1)} + \mathbb{V}_i \gamma, \\
F_{i(t+1)} &= Z_{i(t+1)} - \beta_0 - \beta_1 \sum_j w_{ij} Z_{jt} - \mathbb{V}_i \gamma, \\
G_{j(t+1)} &= Z_{j(t+1)} - \beta_0 - \beta_1 \sum_{s \neq i} w_{js} Z_{st} - \beta_2 Z_{jt} - \mathbb{V}_i \gamma,
\end{aligned}$$

C_1 是与 Z_{it} 无关的项. 进一步地, 通过省略与 Z_{it} 无关的项, 可以将 $\log\{\widetilde{f}(Z_{it})\}$ 进一步化简为

$$\log\{\widetilde{f}(Z_{it})\} = \log\{\mathrm{P}(Y_{it} \mid Z_{it})\} - \frac{1}{2\widetilde{\sigma}_{it}^2}(Z_{it} - \widetilde{\mu}_{it})^2 + C_2, \qquad (3.4)$$

其中 C_2 是与 Z_{it} 无关的项, 对于 $1\leqslant t\leqslant T-1$, 推导得到 $\widetilde{\mu}_{it}$ 和 $\widetilde{\sigma}_{it}^2$ 分别为

$$\widetilde{\mu}_{it}=\frac{E_{i(t-1)}+\beta_2F_{i(t+1)}+\sum_{j\neq i}\beta_1w_{ji}G_{j(t+1)}}{1+\beta_2^2+\sum_{j\neq i}\beta_1^2w_{ji}^2},$$
$$\widetilde{\sigma}_{it}^2=\frac{\sigma^2}{1+\beta_2^2+\sum_{j\neq i}\beta_1^2w_{ji}^2}.$$

当 $t=T$ 时, 推导得到 $\widetilde{\mu}_{iT}=E_{i(T-1)}$ 和 $\widetilde{\sigma}_{iT}^2=\sigma^2$.

由于 (3.4) 右侧的第二项是正态分布形式, 所以使用正态分布函数 $N(\widetilde{\mu}_{it},\widetilde{\sigma}_{it}^2)$ 作为提议分布. 最后, 在抽样过程中, 可以调整 $\widetilde{\sigma}_{it}^2$ 以使 Metropolis-Hastings 算法的接受率在 15% 到 40% 之间.

3.2 未知参数的完全条件分布

本节首先推导 θ 的完全条件分布, 然后推导 σ^2 的完全条件分布. 假设 θ 的先验分布为 $f(\theta)\propto 1$, 其完全条件分布可以推导如下:

$$\begin{aligned}\widetilde{f}(\theta)&\propto f(\mathbb{Z},\theta,\sigma^2\mid\mathbb{Y},\mathbb{Z}_0,\mathbb{V})\\&\propto\prod_{t=1}^{T}f(\mathbb{Z}_t\mid\mathbb{Z}_{t-1},\mathbb{V},\theta,\sigma^2)\\&\propto\prod_{t=1}^{T}\exp\Big\{-\frac{1}{2\sigma^2}(\mathbb{Z}_t-\mathbb{X}_{t-1}\theta)^\top(\mathbb{Z}_t-\mathbb{X}_{t-1}\theta)\Big\}\\&\propto\exp\Big\{-\frac{1}{2\sigma^2}(\mathbb{Z}-\mathbb{X}\theta)^\top(\mathbb{Z}-\mathbb{X}\theta)\Big\},\end{aligned}\tag{3.5}$$

其中

$$\mathbb{X}_t=(X_{1t},\ldots,X_{Nt})^\top\in\mathbb{R}^{N\times(3+p)},\quad \mathbb{X}=(\mathbb{X}_1^\top,\ldots,\mathbb{X}_T^\top)^\top\in\mathbb{R}^{(NT)\times(3+p)}.$$

给定 (3.5), θ 的完全条件分布是正态分布 $N(\mu_\theta,\Sigma_\theta)$, 其中 $\mu_\theta=(\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\mathbb{Z}$, $\Sigma_\theta=(\mathbb{X}^\top\mathbb{X})^{-1}$.

依照常规, 假设 σ^2 的先验分布为 $f(\sigma^2)\propto\sigma^{-2}$. 然后, 根据联合后验分布 (3.1), 可以推导 σ^2 的完全条件分布如下:

$$\begin{aligned}\widetilde{f}(\sigma^2)&\propto f(\mathbb{Z},\theta,\sigma^2\mid\mathbb{Y},\mathbb{Z}_0,\mathbb{V})\\&\propto f(\sigma^2)\prod_{t=1}^{T}f(\mathbb{Z}_t\mid\mathbb{Z}_{t-1},\mathbb{V},\theta,\sigma^2)\\&\propto\sigma^{-2}\sigma^{-NT}\prod_{t=1}^{T}\exp\Big\{-\frac{1}{2\sigma^2}(\mathbb{Z}_t-\mathbb{X}_{t-1}\theta)^\top(\mathbb{Z}_t-\mathbb{X}_{t-1}\theta)\Big\}\\&\propto(\sigma^2)^{-(\frac{NT}{2}+1)}\exp\Big\{-\frac{1}{2\sigma^2}(\mathbb{Z}-\mathbb{X}\theta)^\top(\mathbb{Z}-\mathbb{X}\theta)\Big\}.\end{aligned}\tag{3.6}$$

给定 (3.6), σ^2 的完全条件分布是带尺度的逆 χ^2 分布, 即 INV-$\chi^2(NT,s^2)$, 其中

$$s^2=\frac{(\mathbb{Z}-\mathbb{X}\theta)^\top(\mathbb{Z}-\mathbb{X}\theta)}{NT}.$$

4 数值模拟

4.1 模型设定

为了验证广义网络向量自回归模型在有限样本下的表现, 本节设计了 3 个数值模拟实验. 在每个实验中, 考虑两种典型的连接函数, 即 (2.2) 中的逆 logit 连接函数 (简称为 Logit), 以及 (2.3) 中的 Poisson 连接函数 (简称为 Poisson). 每种实验设定的主要差异在于邻接矩阵 A 的生成机制和参数的设置.

在每个实验设定中, 令 $p=5$, $\gamma=(-0.1,-0.2,0.3,0,0)^\top$、并从均值为 0、协方差为 $\Sigma_v=(\sigma_{j_1j_2})$ 的多元正态分布生成 $V_i=(V_{i1},\ldots,V_{i5})^\top$, 其中 $\sigma_{j_1j_2}=0.5^{|j_1-j_2|}$. 接下来, 从 $N(0,1)$ 生成 ε_{it}, 并且将 $\mathbb{Z}_0$ 设置为 $\mathbf{0}$. 给定 A 和 $(\beta_0,\beta_1,\beta_2)^\top$, 根据 (2.4) 可以生成 $\mathbb{Z}$. 最后, 分别使用 (2.2) 和 (2.3) 表示的连接函数生成 Y_{it}. 在下文中, 考虑了邻接矩阵 A 的 3 种不同的生成机制. 前两种生成机制借鉴自 Zhu 等 [1] 的研究, 最后一个是来自于真实数据.

例 1 (随机分块模型) 随机分块模型是一种常用的网络结构, 主要用于社区发现 [17–19]. 在这个实验设计中, 我们依照 Nowicki 和 Snijders[18] 使用随机分块模型来生成邻接矩阵 A. 具体而言, 假设网络中有 K 个社区. 每个节点都以相同的概率被随机分配到某一个社区中. 如果节点 i 和 j 属于同一个社区, 则令 $\mathrm{P}(a_{ij}=1)=0.3N^{-0.3}$, 否则 $\mathrm{P}(a_{ij}=1)=0.3N^{-1}$. 设定参数 $(\beta_0,\beta_1,\beta_2)^\top$ 为 $(0,0.1,-0.2)^\top$. 最后, 考虑两种不同的社区个数, 即 $K=5$ 和 $K=10$.

例 2 (幂律分布模型) 在这个模型设定中, 考虑网络节点的入度 (即 $d_i=\sum_j a_{ji}$) 服从幂律分布. 幂律分布是一种常见的网络现象, 它表示网络中只有少量节点拥有大量的关注者, 而大多数节点的关注者非常少. 在这个实验设计中, 参考 Clauset 等 [20] 用幂律分布模型生成邻接矩阵 A. 对于每个节点 i $(1\leqslant i\leqslant N)$, 首先根据离散型幂律分布产生其入度, 即 $\mathrm{P}(d_i=h)=ch^{-\alpha}$, 其中 c 是归一化常数, α 是指数参数. 然后, 在所有 N 个节点中随机寻找 d_i 个节点作为节点 i 的关注者. 设定参数 $(\beta_0,\beta_1,\beta_2)^\top$ 为 $(0.3,-0.1,0.5)^\top$. 最后, 考虑两种指数参数的取值, 分别为 $\alpha=1.2$ 和 $\alpha=2$.

例 3 (真实网络结构) 基于第 5.1 小节中某社交网络平台的真实数据设定这个实验中的网络结构. 具体而言, 该网络中有 $N=1{,}735$ 个用户. 根据用户之间的关注关系生成邻接矩阵 A. 如果用户 i 关注了用户 j, 则令 $a_{ij}=1$, 否则 $a_{ij}=0$. 在这种情形下, 网络结构可能是非对称的. 定义 $\sum_{i\neq j}a_{ij}/n(n-1)$ 为网络密度, 表示网络中边的密集程度. 计算可知, 该网络结构的实际网络密度为 2.76%. 最后设定参数 $(\beta_0,\beta_1,\beta_2)^\top$ 为 $(-0.1,-0.2,0.3)^\top$.

将每个数值模拟实验重复 $R=500$ 次以获得可靠的估计结果. 在前两个实验设定中, 将时间长度固定为 $T=10$, 节点的数量 N 分别取 100、500 和 1,000. 最后一个基于实际数据的实验设定中, 将节点数固定为 $N=1{,}735$, 考虑两种不同的时间长度, 即 $T=10$ 和 $T=30$. 在每个生成的数据集上使用 MCMC 算法进行模型估计. 在使用 MCMC 算法时, 我们将选取 3 个不同的初始点, 因此会得到 3 条 MCMC 链. 每条 MCMC 链会首先进行 300 次迭代; 随后, 在 Metropolis-Hastings 步中对每条链再进行 700 次迭代, 通过调整参数 $\tilde{\sigma}_{it}^2$ 来得到合适的接受率; 当 MCMC 链平稳后, 再进行 1,000 次迭代, 从每 5 次迭代中抽取一个样本用以进行模型估计. 在 MCMC 的运行过程中, 我们会监控潜在状态 $\mathbb{Z}$, 参数 (θ,σ^2) 的抽样曲线, 并通过计算潜在尺度缩减因子 [21,22] 来确保其收敛.

4.2 评价指标和模拟结果

设 $\hat{\theta}^{(r)}=(\hat{\theta}_j^{(r)})^\top=(\hat{\beta}_0^{(r)},\hat{\beta}_1^{(r)},\hat{\beta}_2^{(r)},\hat{\gamma}^{(r)})^\top$ 为第 r 个数据集中参数的后验均值估计. 为了评估模型

表现, 计算每个给定参数 θ_j $(1 \leqslant j \leqslant 3+p)$ 的均方根误差, 即

$$\mathrm{RMSE}_j = \left\{ R^{-1} \sum_{r=1}^{R} (\hat{\theta}_j^{(r)} - \theta_j)^2 \right\}^{-1/2}.$$

对于每个 θ_j, 构建其可信度为 95% 的可信区间. 可信区间为

$$\mathrm{CI}_j^{(r)} = (\hat{\theta}_j^{(2.5,r)}, \hat{\theta}_j^{(97.5,r)}),$$

其中 $\hat{\theta}_j^{(2.5,r)}$ 和 $\hat{\theta}_j^{(97.5,r)}$ 是所有后验样本的 2.5% 和 97.5% 分位数. 进一步地, 定义

$$\mathrm{CC}_j = R^{-1} \sum_{r=1}^{R} I(\theta_j \in \mathrm{CI}_j^{(r)})$$

为 θ_j 的可信区间的覆盖范围, 其中 $I(\cdot)$ 代表示性函数.

表 1–3 总结了 3 种模拟实验的估计结果. 由于 3 种实验的估计结果类似, 我们重点对例 1 的估计结果进行详细说明. 例 1 的具体结果见表 1. 可以看出, 对于两个不同的连接函数 (Logit 和 Poisson), 所有的均方根误差均随着样本量 N 的增加而减小. 例如, 在连接函数为 Logit、社区数 $K = 5$ 的情形下, 在 N 从 100 增加到 1,000 时, 网络效应的均方根误差从 8.7% 下降到 0.9%; 而自回归效应的均方根误差从 6.8% 下降到 1.2%. 该现象与所提出的广义网络向量自回归模型的估计结果是一致的. 其次, 不同情形下可信区间的覆盖率都接近 95%, 这保证了参数推断的准确性. 最后, 该估计方法对社区数的改变是稳定的. 表 2 和 3 汇报了例 2 和 3 的估计结果, 该结果与表 1 得出的结论一致. 综上所述, 以上结果表明我们提出的估计方法是稳定的.

表 1 模拟实验 (例 1) 的估计结果. 表格中汇报了均方根误差 ($\times 10^2$) 和可信度覆盖率 (%)

连接函数	θ	$K=5$			$K=10$		
		$N=100$	$N=500$	$N=1{,}000$	$N=100$	$N=500$	$N=1{,}000$
Logit	β_0	7.6 (95.1)	2.8 (95.6)	0.8 (94.6)	8.1 (95.7)	4.8 (95.2)	2.1 (95.0)
	β_1	8.7 (94.7)	1.6 (95.6)	0.9 (94.3)	8.7 (94.3)	2.1 (95.8)	1.5 (95.3)
	β_2	6.8 (94.3)	3.9 (94.6)	1.2 (94.7)	8.4 (94.6)	5.7 (94.6)	2.0 (94.3)
	γ_1	7.0 (94.4)	4.3 (95.8)	2.7 (94.4)	7.4 (94.2)	5.5 (94.6)	3.3 (95.7)
	γ_2	4.7 (94.6)	2.1 (94.6)	1.4 (95.9)	5.5 (94.1)	2.9 (94.8)	1.2 (95.6)
	γ_3	7.2 (94.8)	5.1 (94.3)	2.1 (94.6)	6.1 (95.5)	3.3 (95.3)	1.3 (94.9)
	γ_4	6.5 (94.8)	4.6 (94.4)	1.1 (95.8)	8.6 (94.3)	6.4 (95.7)	3.6 (95.2)
	γ_5	8.5 (94.7)	4.6 (95.4)	2.3 (94.6)	8.5 (94.7)	4.6 (95.4)	2.3 (94.6)
Poisson	β_0	5.7 (94.8)	1.1 (94.9)	0.8 (94.2)	5.9 (94.7)	2.3 (94.5)	1.1 (94.1)
	β_1	2.5 (95.3)	1.4 (95.1)	0.9 (94.1)	6.7 (95.9)	3.1 (94.5)	1.9 (94.4)
	β_2	3.8 (95.6)	2.0 (94.5)	1.5 (94.8)	6.8 (94.0)	4.3 (94.3)	2.2 (94.6)
	γ_1	8.1 (95.9)	5.5 (95.0)	2.7 (94.1)	8.0 (95.9)	4.1 (94.4)	3.0 (94.4)
	γ_2	9.8 (94.2)	4.8 (95.2)	2.6 (94.5)	8.8 (94.4)	5.6 (94.2)	2.2 (95.2)
	γ_3	3.9 (95.3)	2.5 (95.2)	1.0 (94.3)	8.8 (95.3)	6.9 (94.8)	3.9 (94.4)
	γ_4	3.1 (95.3)	2.1 (94.1)	1.1 (94.3)	6.6 (95.0)	3.8 (95.9)	1.3 (95.0)
	γ_5	3.5 (94.4)	3.6 (95.6)	1.1 (95.7)	3.5 (94.4)	3.6 (95.6)	1.1 (95.7)

表 2 模拟实验 (例 2) 的估计结果. 表格中汇报了均方根误差 ($\times 10^2$) 和可信度覆盖率 (%)

连接函数	θ	$\alpha = 1.2$			$\alpha = 2$		
		$N = 100$	$N = 500$	$N = 1{,}000$	$N = 100$	$N = 500$	$N = 1{,}000$
Logit	β_0	5.1 (95.5)	2.1 (94.7)	1.2 (94.1)	4.0 (94.5)	2.1 (94.4)	1.3 (94.4)
	β_1	5.0 (94.1)	2.7 (95.3)	1.3 (95.5)	4.7 (95.3)	2.5 (94.7)	1.7 (95.8)
	β_2	8.2 (95.5)	4.8 (94.4)	2.4 (94.2)	8.4 (95.3)	5.8 (94.3)	3.4 (95.9)
	γ_1	8.1 (94.3)	6.5 (94.2)	3.9 (94.9)	9.8 (94.4)	5.9 (94.5)	2.8 (95.7)
	γ_2	7.0 (94.5)	4.0 (95.7)	2.3 (94.7)	8.6 (95.5)	5.3 (95.7)	3.0 (94.3)
	γ_3	9.3 (95.3)	7.0 (94.2)	4.7 (94.7)	6.4 (95.2)	4.9 (94.7)	1.9 (94.2)
	γ_4	9.9 (94.8)	5.9 (94.1)	2.5 (95.2)	9.8 (95.8)	6.8 (95.8)	2.6 (94.6)
	γ_5	4.2 (95.1)	2.7 (95.2)	1.5 (95.6)	6.1 (95.8)	3.9 (94.5)	1.5 (95.9)
Poisson	β_0	8.3 (95.3)	6.7 (95.2)	2.6 (95.1)	8.2 (95.2)	3.3 (95.3)	1.5 (95.1)
	β_1	2.2 (95.6)	1.6 (94.4)	0.8 (94.4)	2.8 (95.9)	1.2 (95.8)	0.4 (95.4)
	β_2	5.8 (95.2)	1.9 (94.6)	0.9 (95.0)	7.8 (95.4)	4.6 (94.1)	1.1 (95.3)
	γ_1	4.6 (95.1)	2.3 (94.4)	1.2 (94.3)	3.2 (94.6)	2.1 (94.3)	1.1 (95.9)
	γ_2	5.5 (95.6)	3.8 (95.6)	1.1 (94.3)	6.9 (94.5)	3.8 (94.9)	1.6 (94.9)
	γ_3	4.1 (95.5)	2.7 (95.7)	1.1 (95.5)	7.3 (94.6)	5.8 (94.8)	1.2 (95.4)
	γ_4	2.6 (95.7)	1.9 (95.5)	0.8 (95.8)	3.4 (94.4)	1.9 (94.7)	1.1 (95.1)
	γ_5	3.9 (94.8)	2.2 (95.9)	1.6 (94.4)	5.0 (94.3)	2.0 (94.4)	1.7 (95.9)

表 3 模拟实验 (例 3) 的估计结果. 表格中汇报了均方根误差 ($\times 10^2$) 和可信度覆盖率 (%)

θ	Logit		Poisson	
	$T = 10$	$T = 30$	$T = 10$	$T = 30$
β_0	4.2 (95.9)	3.1 (95.8)	6.7 (95.9)	3.8 (94.5)
β_1	3.2 (94.3)	2.1 (95.5)	3.1 (95.0)	1.9 (94.6)
β_2	2.9 (94.0)	2.1 (94.1)	2.3 (95.8)	1.2 (94.4)
γ_1	5.7 (94.9)	4.5 (95.8)	7.3 (94.9)	5.6 (94.3)
γ_2	1.9 (94.9)	1.1 (95.4)	1.6 (94.3)	1.0 (94.2)
γ_3	2.7 (95.9)	1.6 (94.9)	2.6 (94.9)	1.6 (94.1)
γ_4	5.2 (94.4)	4.0 (94.5)	6.2 (94.6)	4.2 (95.3)
γ_5	5.8 (94.5)	4.4 (95.5)	6.4 (94.1)	3.9 (95.9)

5 案例分析

5.1 某社交网络平台用户发帖数据

本节使用广义网络向量自回归模型对某社交网络平台收集的真实数据进行分析. 所使用的数据集包含了某明星官方账号 $N = 1{,}735$ 名粉丝的日常发帖行为. 我们对这些用户进行了连续 $T = 30$ 天的观察. 在观察期间, 记录了每个用户每天发布的微博数量, 以及用户个人信息, 即性别和标签. 其中, 标签是由用户自己创建的一条文本, 用以描述自身的生活方式和职业状态.

这里, 将 N 个用户之间的网络结构 (即 A) 定义为用户之间的关注 - 被关注关系. 记 $a_{ij}=1$ 为节点 i 和 j 之间的一条连边. 在此网络中, 总共存在 1,735 个节点和 41,550 个连边. 该网络的密度为 2.76%. 图 1 展示了网络的入度和出度直方图. 可以看出, 入度分布比出度分布的右偏程度更明显. 这表明该网络中可能存在 “超级明星”.

定义因变量 Y_{it} 为用户 i 在第 t 天所发微博数量, 它能够反映用户的活跃度. Y_{it} 的直方图和用户平均发帖量的时间序列图 ($N^{-1}\sum_t Y_{it}$) 如图 2 所示. 可以观察到, Y_{it} 的直方图是右偏的, 这表明大多数人很少发帖子, 但是仍存在 “高度活跃” 的用户, 例如, 一些媒体账户可以发布数百条微博. 每天平

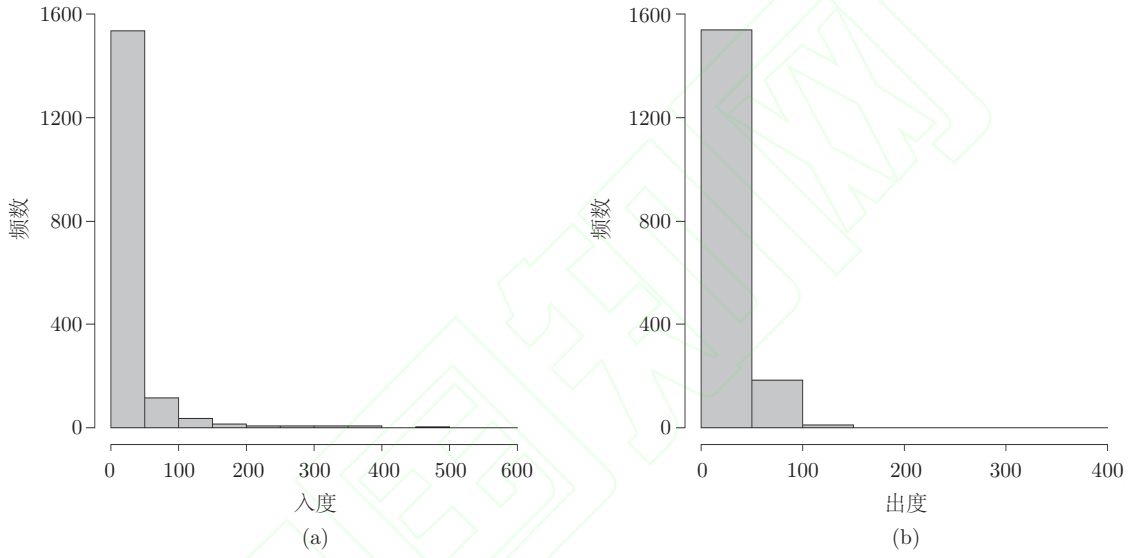

图 1 用户发贴行为分析. (a) $N=1{,}735$ 个节点的入度直方图, 高度右偏表明网络中存在 “超级明星”; (b) 出度直方图

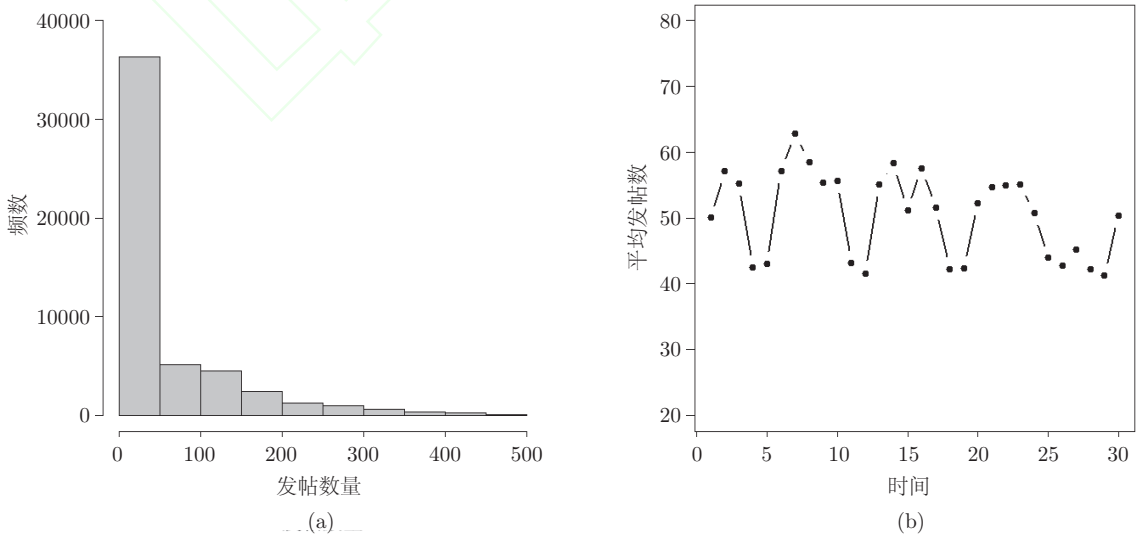

图 2 用户发帖行为分析. (a) 发帖数量 (Y_{it}) 的分布直方图, 高度右偏说明存在少量活跃用户; (b) 用户平均发帖量 ($T^{-1}\sum_t Y_{it}$) 的时间序列图, 表明用户具有稳定的发帖行为

均发帖数量的整体趋势是稳定的, 这说明用户的发帖行为是稳定的. 此外, 我们还考虑了两个不随时间变化的自变量, 即性别 (男性 = 1, 女性 = 0) 和标签数量.

由于因变量 Y_{it} 是计数变量, 我们在使用广义网络向量自回归模型进行估计时, 选择使用 Poisson 连接函数 (2.3). 表 4 展示了相应的模型估计结果. 对于每个参数, 表 4 分别展示了其后验均值、后验标准差和 95% 的可信区间. 如表 4 所示, 所有参数的 95% 置信区间均不包含零. 具体而言, 网络效应 ($\hat{\beta}_1$) 的后验均值为 0.11, 95% 可信区间为 $[0.09, 0.12]$. 这说明用户发帖的活跃程度与其邻居呈现正相关关系. 模型所估计的自回归效应 ($\hat{\beta}_2$) 也对发帖数量产生正向影响, 这表明之前发帖活跃程度较高 (或较低) 的用户, 可能之后的发帖活跃性仍较高 (或较低). 模型估计中两个自变量对应的估计值 $(\hat{\gamma}_1, \hat{\gamma}_2)$ 均为正值, 表明在此数据集中具有更多自创标签的男性用户表现更为活跃.

值得注意的是, Zhu 等 [1] 将该因变量视为连续型变量, 所得模型估计结果与表 4 大体一致. 但是, 将因变量视为离散型计数变量, 并采用本文提出的广义网络向量自回归模型进行估计时, 可以发现网络效应有所提高, 但是自回归效应有所降低.

最后, 我们探索了模型估计得到的潜在状态变量 $\mathbb{Z}$ 的情况. 图 3 显示了每个用户的平均潜在状态数 (即 $T^{-1}\sum_t Z_{it}$) 的分布直方图和每天平均潜在状态数 (即 $N^{-1}\sum_i Z_{it}$) 的时间序列图. 可以看到,

表 4　基于某社交平台用户发贴行为数据的广义网络向量自回归模型估计结果

连接函数	系数估计	均值	标准差 ($\times 10^2$)	置信区间上限	置信区间下限
Poisson	基准效应 ($\hat{\beta}_0$)	−0.32	1.99	−0.36	−0.28
	网络效应 ($\hat{\beta}_1$)	0.11	0.79	0.09	0.12
	自回归效应 ($\hat{\beta}_2$)	0.28	0.35	0.27	0.29
	性别 ($\hat{\gamma}_1$)	0.17	1.82	0.14	0.21
	标签数 ($\hat{\gamma}_2$)	0.04	0.18	0.04	0.05

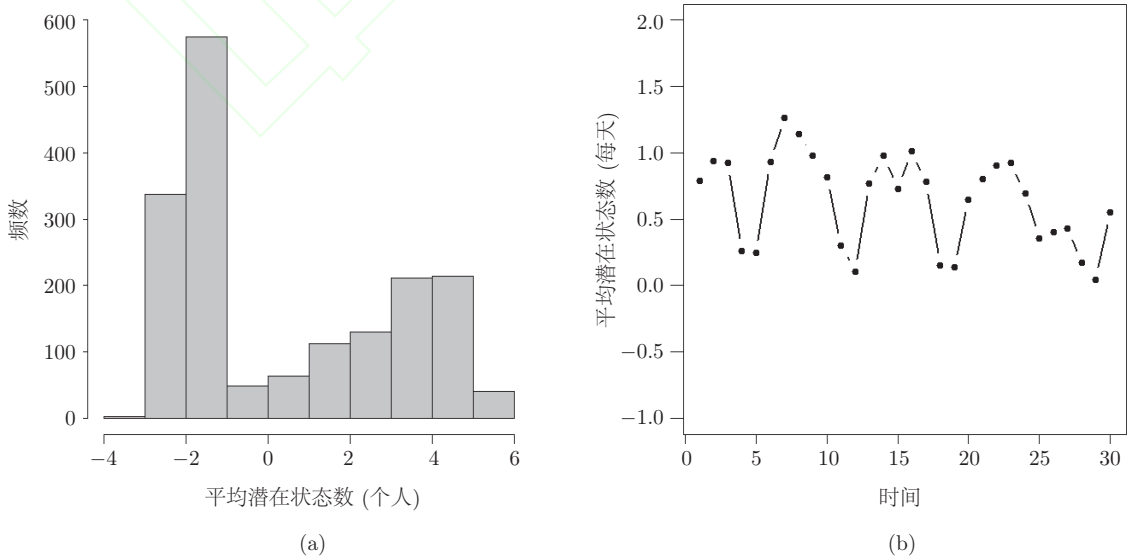

图 3　用户发帖行为分析. (a) 每个用户的平均潜在状态直方图 ($T^{-1}\sum_t Z_{it}$), 两个峰值表示了两类具有不同的发帖习惯的人群; (b) 每天平均潜在状态 ($N^{-1}\sum_i Z_{it}$)

直方图 (见图 3(a)) 有两个峰值, 表示了两类具有不同发帖习惯的人群. 每天平均潜在状态的趋势 (见图 3(b)) 与图 2 中所示的每天平均帖子数的趋势相一致.

5.2　某社交网络平台地震新闻传播数据

在第二个实例研究中, 我们通过某社交网络平台上的地震新闻数据集来验证广义网络向量自回归模型的表现. 该数据集包含了 2013 年 4 月 20 日中国四川省雅安市发生 7 级地震相关的帖子. 地震发生后, 公众非常关注该事件的相关新闻, 许多用户都在社交平台上发帖表示他们的关心和关注. 我们收集了某社交平台上 $N = 6{,}541$ 名用户的发帖数据, 并记录他们在地震后 $T = 7$ 天的发帖行为. 数据中的因变量为用户对地震相关新闻的反应, 具体来说, 如果第 i 个用户在第 t 天发布有关地震的信息, 则令 $Y_{it} = 1$, 否则 $Y_{it} = 0$. 图 4 中展示了发帖用户数量 (即 $\sum_i Y_{it}$) 的时间趋势图. 从图中可以看出, 总的发帖用户数呈现下降趋势.

与第 5.1 小节的实例研究类似, 将 N 个用户之间的网络结构 (即 A) 定义为网络中用户之间的关注 - 被关注关系. 在此网络中, 总共存在 6,541 个节点和 96,338 个连边. 该网络的密度为 2.25%. 入度和出度的直方图如图 5 所示, 均为右偏分布. 由于该实例研究中, 因变量是 0-1 型变量, 我们在使用广义网络向量自回归模型进行估计时, 选择逆 logit 连接函数 (2.2). 表 5 展示了模型估计结果. 与第 5.1 小节的估计结果相似, 可以发现正向的网络效应和自回归效应. 此外, 尽管模型估计的自回归效应大于网络效应, 但它们的数量级是可比较的.

最后, 图 6 展示了每个用户或每天的平均潜在状态数的分布情况. 每个用户的平均每天的潜在状态数分布直方图 (见图 6(a)) 显示, 大多数用户的平均潜在状态数小于 0, 因为采用逆 logit 连接函数, 因此意味着, 大多数用户关于地震微博的发帖概率低于 0.5, 即发帖意愿不强. 每天的平均潜在状态图 (见图 6(b)) 显示了下降趋势, 表明有关地震的信息正在降温.

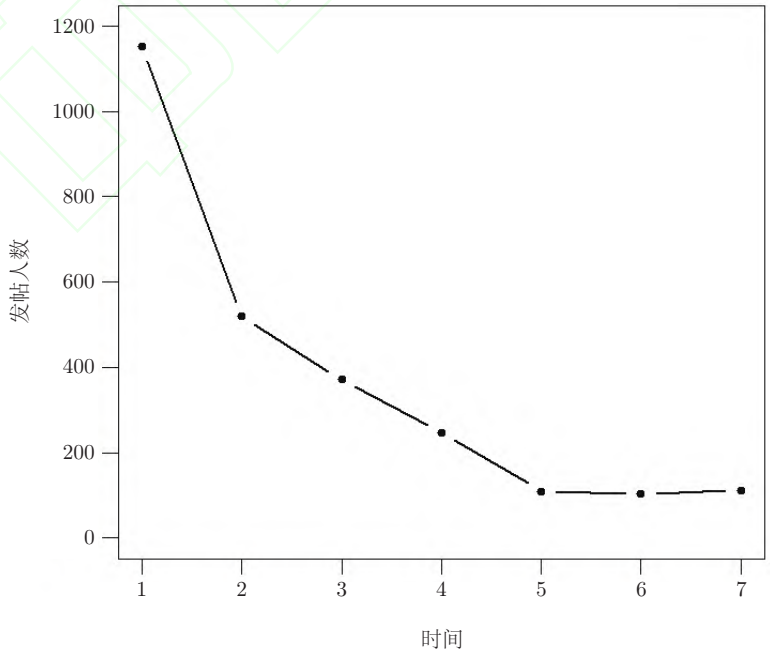

图 4　地震新闻传播分析. 发布有关地震消息帖子用户数量 ($\sum_i Y_{it}$ 的时间趋势图, 图中呈现下降趋势)

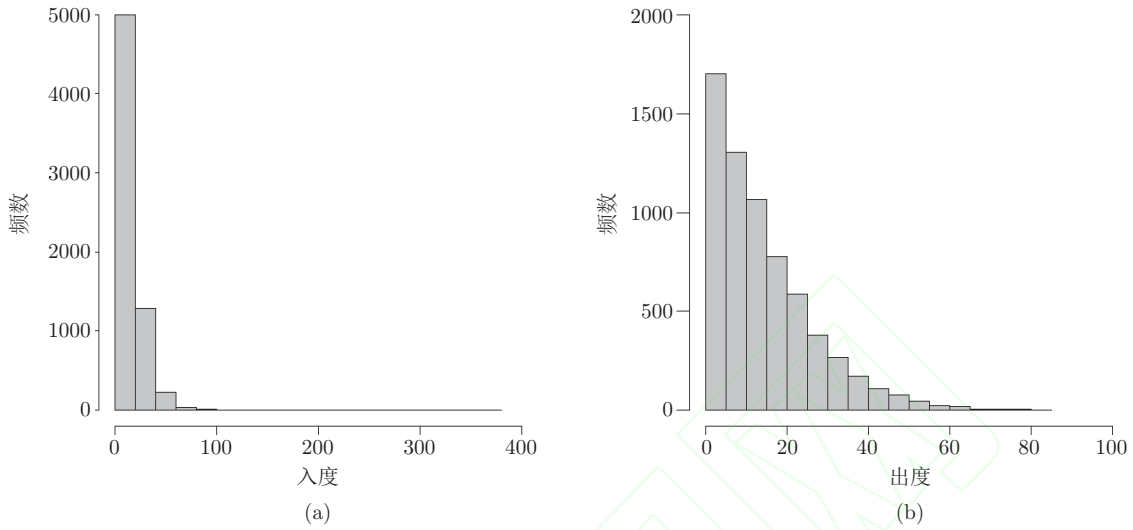

图 5 地震新闻传播分析. (a) $N = 6{,}541$ 个节点的入度直方图; (b) 出度直方图

表 5 基于某社交平台地震新闻传播数据的广义网络向量自回归模型估计结果

联系函数	变量	均值	标准差 ($\times 10^2$)	置信区间上限	置信区间下限
Logit	基准效应 ($\hat{\beta}_0$)	−2.37	7.61	−2.54	−2.07
	网络效应 ($\hat{\beta}_1$)	0.11	1.40	0.09	0.14
	自回归效应 ($\hat{\beta}_2$)	0.17	6.01	0.08	0.26
	性别 ($\hat{\gamma}_1$)	0.10	1.22	0.08	0.13
	标签数 ($\hat{\gamma}_2$)	0.10	0.21	0.10	0.11

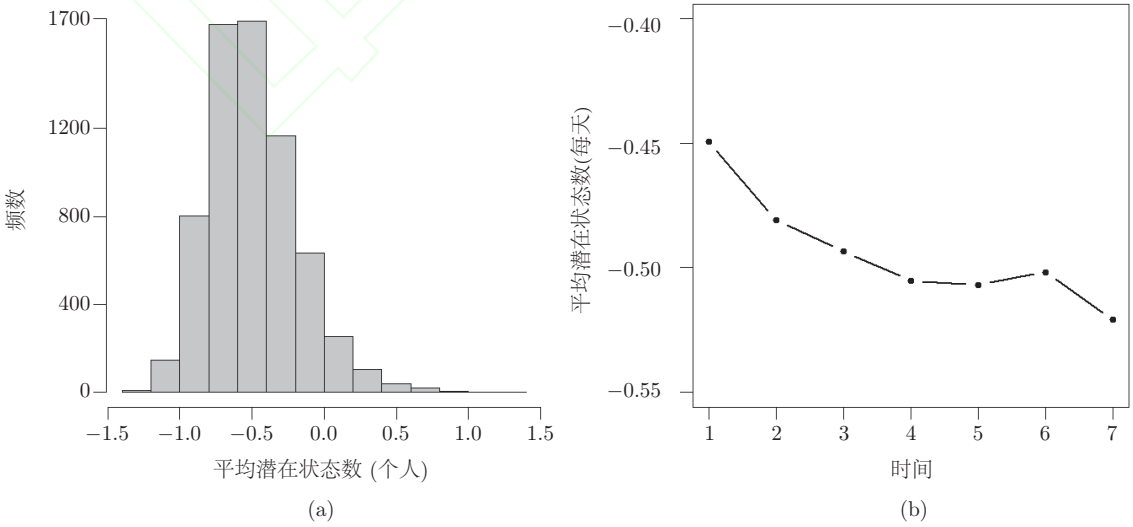

图 6 地震新闻传播分析. (a) 每个用户的平均每天潜在状态数 ($T^{-1}\sum_t Z_{it}$) 的分布直方图; (b) 每天平均潜在状态数 ($N^{-1}\sum_i Z_{it}$) 的时间趋势图

6 结论

本文提出了一种广义网络向量自回归模型, 该模型对动态过程进行建模, 并允许因变量为离散型变量. 本文提出的模型借鉴了空间自回归模型和状态空间模型的优点. 假设每个观测到的因变量由其潜在状态确定, 并且使用网络向量自回归模型对潜在状态之间的关系进行建模. 在模型估计方面, 我们提出 Metropolis Hastings-within-Gibbs MCMC 抽样算法, 并通过随机模拟和两个实证研究进行了验证.

最后, 我们对未来工作提出一些可行的研究方向. 第一, 本文分别使用逆 logit 连接函数和 Poisson 连接函数探讨了广义网络向量自回归模型对于 0-1 型变量和计数变量的表现. 接下来可以应用模型对其他离散型变量和连接函数展开研究. 第二, 在广义网络向量自回归模型中, 假定网络效应 (β_1) 和自回归效应 (β_2) 均为固定值. 而在实际情形中, 这两类效应可以是与时间相关的或与节点相关的, 因此, 未来可以更加灵活地设置这些参数. 第三, 本文假定广义网络向量自回归模型中的网络结构是静态的, 然而实际中网络结构可以随着时间缓慢变化, 因此, 在未来的研究中可以将动态变化的网络结构纳入广义网络向量自回归模型.

致谢 感谢中央财经大学统计与数学学院高天辰对本文修改及文字翻译方面做出的贡献. 感谢主编及审稿人对本文提出的建设性意见.

参考文献

1 Zhu X, Pan R, Li G, et al. Network vector autoregression. Ann Statist, 2017, 45: 1096–1123
2 Ji P, Jin J. Coauthorship and citation networks for statisticians. Ann Appl Stat, 2016, 10: 1779–1812
3 Goldsmith-Pinkham P, Imbens G W. Social networks and the identification of peer effects. J Bus Econom Statist, 2013, 31: 253–264
4 Dunbar R I M, Arnaboldi V, Conti M, et al. The structure of online social networks mirrors those in the offline world. Soc Networks, 2015, 43: 39–47
5 Ryan L, D'Angelo A. Changing times: Migrants' social network analysis and the challenges of longitudinal research. Soc Networks, 2018, 53: 148–158
6 Lee L F. Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. Econometrica, 2004, 72: 1899–1925
7 Anselin L. Spatial Econometrics: Methods and Models, Vol. 4. New York: Springer Science & Business Media, 2013
8 Banerjee S, Carlin B P, Gelfand A E. Hierarchical Modeling and Analysis for Spatial Data. Boca Raton: CRC Press, 2014
9 Chen X, Chen Y, Xiao P. The impact of sampling and network topology on the estimation of social intercorrelations. J Marketing Res, 2013, 50: 95–110
10 Zhou J, Tu Y, Chen Y, et al. Estimating spatial autocorrelation with sampled network data. J Bus Econom Statist, 2017, 35: 130–138
11 Cohen-Cole E, Liu X, Zenou Y. Multivariate choices and identification of social interactions. J Appl Econometrics, 2018, 33: 165–178
12 De Jong P. The likelihood for a state space model. Biometrika, 1988, 75: 165–169
13 Geweke J, Tanizaki H. Bayesian estimation of state-space models using the Metropolis-Hastings algorithm within Gibbs sampling. Comput Statist Data Anal, 2001, 37: 151–170
14 Davis R A, Rodriguez-Yam G. Estimation for state-space models based on a likelihood approximation. Statist Sinica, 2005, 15: 381–406
15 Meyn S, Tweedie R L. Markov Chains and Stochastic Stability. Cambridge: Cambridge University Press, 2012
16 Koller D, Friedman N. Probabilistic Graphical Models: Principles and Techniques. Cambridge: MIT Press, 2009
17 Wang Y J, Wong G Y. Stochastic blockmodels for directed graphs. J Amer Statist Assoc, 1987, 82: 8–19
18 Nowicki K, Snijders T A B. Estimation and prediction for stochastic blockstructures. J Amer Statist Assoc, 2001, 96: 1077–1087
19 Zhao Y, Levina E, Zhu J. Consistency of community detection in networks under degree-corrected stochastic block models. Ann Statist, 2012, 40: 2266–2292

20 Clauset A, Shalizi C R, Newman M E J. Power-law distributions in empirical data. SIAM Rev, 2009, 51: 661–703
21 Gelman A, Rubin D B. Inference from iterative simulation using multiple sequences. Statist Sci, 1992, 7: 457–472
22 Gelman A, Carlin J B, Stern H S, et al. Bayesian Data Analysis. Boca Raton: Chapman & Hall/CRC, 2014

Generalized network vector autoregression

Feifei Wang, Xuening Zhu & Rui Pan

Abstract With the rapid development of social network platforms, the time series of network data is becoming increasingly available. To model the dynamic user behaviours, a network vector autoregression (NAR) model is developed, which targets at the continuous type responses. In practice, discrete type of data (e.g., number of posts, user decisions) are frequently collected from the network users. To model such type of data, we propose a generalized network vector autoregression (GENAR) model in this work. It assumes that a latent continuous variable exists for each node at each time point, which determines the observed response variable. The dynamic and network dependence is assumed based on the latent variables (states). To estimate and make a valid inference of the model, an MCMC (Markov chain Monte Carlo) algorithm is designed and verified by extensive numerical studies. Two real data examples are presented using datasets from a social network platform for illustration purpose.

Keywords **network data, MCMC algorithm, network vector autoregression, latent variable**
MSC(2010) **62F15, 62H11**
doi: 10.1360/SCM-2018-0839