

有向图上高维时间序列模型及其在交通网络中的应用

向乔怡¹, 祁辉²

(1. 武汉大学数学与统计学院, 湖北 武汉 430072)

(2. 三明学院信息工程学院, 福建 三明 365004)

摘要: 本文研究了城市道路车辆的平均行驶速度在时间以及空间上的变化规律. 利用道路交通的网络结构特征, 提出了有向图上的高维时间序列模型. 通过引入 ℓ_1 -惩罚项, 本文建立了正则化最小二乘的参数估计方法, 并对此过程中调节参数的选择方式进行了研讨. 进一步, 本文将所提出的方法应用于成都市道路网络的城市链路速度的实例, 并以成都市中心城区和偏远郊区为例, 揭示了其交通运输效率在时间以及空间上的变化规律. 该项研究可为疏导交通堵塞以及缓解交通压力提供有效的统计建议.

关键词: 网络数据; 车辆平均行驶速度; 高维; 交通

MR(2010)主题分类号: 62M10; 62P99

中图分类号: O212

文献标识码: A

文章编号: 0255-7797(2022)04-0345-14

1 引言

随着社会的现代化发展, 人类生活质量不断提高. 因此, 选择公共交通工具出行在日常通勤中的占比日益下降, 城市道路交通网络的运输压力亦会增加. 在居民选择出行路线以及有关部门对道路交通进行管理时, 如何提高城市道路交通网络的运输效率成为社会关注的重要问题之一.

由于城市道路交通符合网络数据集的典型结构特征, 本文将一般网络数据的研究方法, 运用于城市道路交通网络的分析中, 即将道路视为顶点, 路口的每个可行驶方向视为连接顶点的有向边, 构建道路交通网络的有向图结构.

为反映各条道路上的交通运输效率, 本文采用车辆平均行驶速度作为响应变量. 各道路在某一时段中的车辆平均行驶速度随时间的变化而变化, 为典型的时间序列的数据结构, 故可构建多元时间序列模型 [1, 2, 3, 4, 5] 以研究车辆平均行驶速度的变化规律. 多元时间序列模型的经典研究方法是分别对各分量上的时间序列进行建模, 分析与预测 [6, 7, 8]. 然而, 在交通网络数据集中, 各道路在某一时段中的车辆平均行驶速度亦与前一时段中相邻道路的车辆平均行驶速度有关, 即在空间上亦具有变化规律. 故此经典研究方法不足以反映网络数据集中的响应变量在空间上的相关性. Bedrick and Tsai [9] 与 Lütkepohl [10] 提出了向量自回归 (VAR) 模型以反映不同时间序列中响应变量间的相关性. 进一步, 结合网络数据的图结构, Zhu et al. [11] 提出网络向量自回归 (NAR) 模型, 且 Zhu and Pan [12] 提出分组网络向量自回归 (GNAR) 模型, 来解释多元时间序列在空间上的联系. 然而, 在本文所处理的数据

*收稿日期: 2021-03-04 接收日期: 2021-05-11

基金项目: 福建省教育厅中青年基金 (JAT200630).

作者简介: 向乔怡 (1996-), 女, 湖北天门, 硕士研究生, 主要研究方向: 数理统计.

通讯作者: 祁辉 (1981-), 男, 福建三明, 副教授, 主要研究方向: 数理统计.

集中,道路的数量,亦即时间序列模型的维数远大于样本量,因此若直接采用 VAR, NAR 抑或 GNAR 模型会面临因参数维数过大而产生的维数灾难带来的巨大挑战.

因此,本文改进了 NAR 模型并提出有向图上的高维时间序列的线性自回归模型来研究交通运输效率在时间与空间上的变化规律.在参数估计的方法上,基于最小二乘法,通过引入 ℓ_1 -正则项以解决降低参数维数的问题.进一步,数值模拟的结果表明所提出的估计方法在有限样本下具有良好的表现.本文通过处理成都市道路网络的城市链路速度数据集,分别分析了位于成都市中心的交通运输压力较大的春熙路路段,以及位于成都市市郊的交通运输压力较小的犀浦立交桥的各时段交通情况,揭示了不同区域的交通运输效率在时间与空间上的变化规律,并提供了疏导交通堵塞以及缓解交通压力的统计建议.

本文的章节安排如下.在第二节中,介绍了 VAR 模型及其平稳性条件并提出有向图上的高维时间序列模型.在第三节中,建立了有向图上的高维时间序列模型的参数估计方法,并对调节参数的选择进行讨论.在第四节中,对本文提出的有向图上的高维时间序列模型的参数估计方法进行模拟实验,并展示其结果.在第五节中,运用所提出的方法分析成都市道路网络的城市链路速度数据集.第六节对全文进行总结与讨论.

2 模型的建立

本节首先介绍 NAR 模型及其参数的解释,再由此引申出本文提出的有向图上的高维时间序列模型,最后讨论其平稳性条件.

2.1 有向图上的高维时间序列

记网络的有向图结构中顶点的数量为 d ,称之为网络大小.记 $\mathbb{Y} = \{Y_{i,t} : i = 1, \dots, d; t = 1, \dots, T\}$ 为有向图上的时间序列,其中 $Y_{i,t} \in \mathbb{R}^1$ 为 t 时刻于顶点 i 处的连续型响应变量.记 $\mathbf{Y}_{\cdot,t} = (Y_{1,t}, \dots, Y_{d,t})^\top$, $\mathbf{Y} = (\mathbf{Y}_{\cdot,1}, \dots, \mathbf{Y}_{\cdot,T})$. $\mathbf{A} = (a_{ij})_{d \times d}$ 为有向图的邻接矩阵. Zhu et al. [11] 在提出的 NAR 模型假设中,考虑了网络的图结构框架,即响应变量在时间与空间上具有相关性,亦即在时间上 $Y_{i,t}$ 会与该顶点在前一时段的响应变量值 $Y_{i,t-1}$ 具有相关性,在空间上 $Y_{i,t}$ 会与相邻顶点在前一时段的响应变量值 $Y_{j,t-1}, j \in \{j : a_{ij} = 1\}$ 具有相关性.因此,响应变量之间有如下的自回归关系:

$$Y_{i,t} = \beta_0 + \beta_1 Y_{i,t-1} + \beta_2 n_i^{-1} \sum_{j=1}^d a_{ij} Y_{j,t-1} + \varepsilon_{i,t}, \quad (2.1)$$

其中, $\beta_0, \beta_1, \beta_2$ 为待估参数, $n_i = \sum_{j=1}^d a_{ij}$ 为顶点 i 的出度, $\{\varepsilon_{i,t} : i = 1, \dots, d; t = 1, \dots, T\}$ 为一系列独立的随机噪声项.然而,在 Zhu et al. [11] 的 NAR 模型假设中,默认某顶点处某时段的响应变量值在空间上与相邻顶点处在前一时段的相关性是相同的.但是在本文考虑的成都市道路网络的城市链路速度数据集中,此假设不足以反映各条道路因地理位置的不同而导致交通网络提供的运输效率上的差异.因此,考虑空间上的差异性,对 (2.1) 式进行改进,本文提出有向图上的时间序列模型:

$$Y_{i,t} = \beta_0 + \beta_1 Y_{i,t-1} + \sum_{j=1}^d a_{ij} \eta_j Y_{j,t-1} + \varepsilon_{i,t}, \quad (2.2)$$

其中, 与 (2.1) 不同的是, (2.2) 中相邻顶点处响应变量对应的回归系数 $\boldsymbol{\eta} = (\eta_1, \dots, \eta_d)^\top$ 为 d 维待估参数向量.

2.2 平稳性条件

在模型的有向图中, 若顶点个数 d 固定, 则 \mathbf{Y} 是多元时间序列. 若其满足平稳性, 则该序列的统计性质不随时间的推移发生改变. 为方便表述, 记 $\Lambda_{\max}(\mathbf{M})$ 为矩阵 \mathbf{M} 的特征值模的最大值, $\mathbf{M}_{l\cdot}$ 与 $\mathbf{M}_{\cdot m}$ 分别为矩阵 \mathbf{M} 的第 l 行与第 m 列, $\mathbf{M}_{-l\cdot}$ 与 $\mathbf{M}_{\cdot -m}$ 分别为矩阵 \mathbf{M} 删去第 l 行与删去第 m 列得到的子矩阵. 记 $\text{diag}(\mathbf{v})$ 为向量 \mathbf{v} 中的所有分量形成的对角阵, $\text{vec}(\cdot)$ 为矩阵的按列拉直算子, \mathbf{I}_d 为 d 阶单位矩阵, $\mathbf{J}_{l \times m}$ 为 $l \times m$ 阶分量全为 1 的矩阵. 记 $\mathbf{H} = \text{diag}(\boldsymbol{\eta})$, $\mathbf{W} = \beta_1 \mathbf{I}_d + \mathbf{A}\mathbf{H}$.

有向图上的时间序列模型 (2.2) 存在唯一严格平稳解的充分必要条件为

$$\Lambda_{\max}(\mathbf{W}) = \Lambda_{\max}(\beta_1 \mathbf{I}_d + \mathbf{A}\mathbf{H}) < 1, \quad (2.3)$$

其解的形式为:

$$\mathbf{Y}_{\cdot,t} = \beta_0 (\mathbf{I}_d - \mathbf{W})^{-1} \mathbf{J}_{d \times 1} + \sum_{j=0}^{\infty} \mathbf{W}^j \boldsymbol{\varepsilon}_{t-j}, \quad (2.4)$$

其中, $\boldsymbol{\varepsilon}_t = (\varepsilon_{1,t}, \dots, \varepsilon_{d,t})^\top$. 特别地, 若 $\varepsilon_{i,t}$ 独立同分布于标准正态分布, 即 $\varepsilon_{i,t} \sim N(0, \sigma^2)$, 严格平稳解服从正态分布, 均值 $\boldsymbol{\mu}$ 和协方差矩阵 $\boldsymbol{\Sigma}$ 分别为:

$$\begin{aligned} \boldsymbol{\mu} &= \beta_0 (\mathbf{I}_d - \mathbf{W})^{-1} \mathbf{J}_{d \times 1} = \beta_0 (\mathbf{I}_d - \beta_1 \mathbf{I}_d - \mathbf{A}\mathbf{H})^{-1} \mathbf{J}_{d \times 1}, \\ \text{vec}(\boldsymbol{\Sigma}) &= \sigma^2 (\mathbf{I}_{d \times d} - \mathbf{W} \otimes \mathbf{W})^{-1} \text{vec}(\mathbf{I}_d). \end{aligned}$$

进一步, 若

$$|\beta_1| + \sum_{j=1}^d |\eta_j| < 1, \quad (2.5)$$

则有向图上的时间序列模型 (2.2) 存在唯一严格平稳解, 称 (2.5) 为 (2.2) 存在唯一严格平稳解的充分条件.

此外, 本文将此性质推广至有向图上的高维时间序列的平稳性条件, 即有向图中, 顶点个数 d 远大于样本量 T , 且二者均趋于无穷的情形. 首先定义有向图上的高维时间序列的平稳性: 令 $\{\mathbf{Y}_{\cdot,t} \in \mathbb{R}^d\}$ 是 $d \rightarrow \infty$ 时的 d 维时间序列.

定义 $\boldsymbol{\Omega} = \{\boldsymbol{\Omega} \in \mathbb{R}^\infty : \sum_{i=1}^{\infty} |\omega_i| < \infty\}$, 其中 $\boldsymbol{\Omega} = (\omega_i \in \mathbb{R}^1 : 1 \leq i \leq \infty)^\top \in \mathbb{R}^\infty$. 对于任意 $\boldsymbol{\Omega} \in \boldsymbol{\Omega}$, 令 $\boldsymbol{\Omega}_d = (\omega_1, \dots, \omega_d)^\top \in \mathbb{R}^d$ 是截断的 d 维向量. 若 $\mathbf{Y}_{\cdot,t}$ 满足下列条件: $\forall \boldsymbol{\Omega} \in \boldsymbol{\Omega}$,

- (i) $Y_t^\omega = \lim_{d \rightarrow \infty} \boldsymbol{\Omega}_d^\top \mathbf{Y}_{\cdot,t}$ 几乎必然存在,
- (ii) $\{Y_t^\omega\}$ 是严格平稳的,
- (iii) $\mathbf{Y}_{\cdot,t}$ 具有有限的一阶矩, 即 $\max_{1 \leq i \leq \infty} E|Y_{it}| < \infty$,

则称 $\mathbf{Y}_{\cdot,t}$ 具有严格平稳性. 在此定义下, 平稳性条件可以由如下命题给出: 若 $\Lambda_{\max}(\mathbf{W}) = \Lambda_{\max}(\beta_1 \mathbf{I}_d + \mathbf{A}\mathbf{H}) < 1$, 则 (2.4) 为一阶矩有限的有向图上的高维时间序列模型 (2.2) 的唯一严格平稳解.

3 参数估计

本节介绍有向图上的高维时间序列模型 (2.2) 的参数估计方法. 由于在成都市道路网络的城市链路速度数据集中, 顶点个数 d 远大于观测时间段数 T , 即模型参数的维数远大于样本量, 因此选用正则化方法进行参数估计.

3.1 估计方法

在高维时间序列的框架下, 本文通过对最小二乘损失函数加 ℓ_1 - 惩罚项, 得到正则化估计. 具体如下,

$$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\eta}}) \in \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^2, \boldsymbol{\eta} \in \mathbb{R}^d} Q(\boldsymbol{\beta}, \boldsymbol{\eta}), \quad (3.1)$$

其中

$$Q(\boldsymbol{\beta}, \boldsymbol{\eta}) = \sum_{i=1}^d \sum_{t=2}^T \left(Y_{i,t} - \beta_0 - \beta_1 Y_{i,t-1} - \sum_{j=1}^d a_{ij} \eta_j Y_{j,t-1} \right)^2 + \lambda |\boldsymbol{\eta}|_1, \quad (3.2)$$

$\boldsymbol{\beta} = (\beta_0, \beta_1)^\top \in \mathbb{R}^2$, 且 $\lambda > 0$ 为调节参数, $|\cdot|_r$ 表示向量的 ℓ_r - 范数, $1 \leq r \leq \infty$. 根据优化问题 (3.1) 的 Karush–Kuh–Tucker 条件, 可以得到如下等式:

$$\frac{\partial Q(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\eta}})}{\partial \boldsymbol{\beta}} = 2 \sum_{t=2}^T \mathbf{X}_{\cdot, t-1}^\top \mathbf{X}_{\cdot, t-1} \hat{\boldsymbol{\beta}} - 2 \sum_{t=2}^T \mathbf{X}_{\cdot, t-1}^\top (\mathbf{Y}_{\cdot, t} - \mathbf{A} \hat{\mathbf{H}} \mathbf{Y}_{\cdot, t-1}) = \mathbf{0}, \quad (3.3)$$

$$\frac{\partial Q(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\eta}})}{\partial \boldsymbol{\eta}} = -2 \sum_{t=2}^T \text{diag}(\mathbf{Y}_{\cdot, t}) \mathbf{A}^\top (\mathbf{Y}_{\cdot, t-1} - \mathbf{X}_{\cdot, t} \hat{\boldsymbol{\beta}} - \mathbf{A} \hat{\mathbf{H}} \mathbf{Y}_{\cdot, t-1}) + \lambda \text{sign}(\hat{\boldsymbol{\eta}}) = \mathbf{0}, \quad (3.4)$$

其中 $\mathbf{X}_{i,t} = (1, Y_{i,t})^\top$, $\mathbf{X}_{\cdot, t} = (\mathbf{X}_{1,t}, \dots, \mathbf{X}_{d,t})^\top$, 且 $\hat{\mathbf{H}} = \text{diag}(\hat{\boldsymbol{\eta}})$. 由于 (3.1) 中的目标函数是凸函数, 通过 (3.3) 和 (3.4) 中的关系, 本文运用交替迭代的方式结合坐标下降算法以求优化问题的最优解. 记 $\Pi_{\theta, m}(\mathbf{v}) = (v_1, \dots, v_{m-1}, 0, v_{m+1}, \dots, v_d)^\top$ 为作用在向量 $\mathbf{v} = (v_1, \dots, v_d)^\top$ 上的投影算子. 所提算法的迭代过程如下:

3.2 调节参数的选择

在高维模型中, 调节参数的大小控制了模型参数的稀疏度, 因而如何选择是一个非常重要的环节. 本节对于有向图上高维时间序列的参数估计中涉及到的调节参数 λ 的不同选择方法进行比较与讨论. 一般而言, K -折交叉验证法和 *hold-out* 交叉验证法是两种最常见的选择调节参数的方法. 然而, 对于本文提出的有向图上的高维时间序列模型而言, 之后发生的事件与之前的事件有关, 因此须保证测试集数据在时间顺序上是位于训练数据之后的, 故在此情况下 K -折交叉验证不能用于本文的调节参数选择. 至于 *hold-out* 交叉验证法, 它是将数据集根据时间顺序分割为训练集和测试集两个子集. 然而, 经典的 *hold-out* 交叉验证法对于测试集的选择具有不确定性, 会导致在测试集上预测能力不足. 为了克服这些不足, 本文提出日向前链交叉验证法, 即按时间进行分割, 将该天的数据作为测试集, 并将此前的所有时段的数据分配到训练集中. 其与 *hold-out* 交叉验证的区别如下图所示:

Algorithm 1 优化问题 (3.1) 的迭代流程图

Require: $k = 0$, $\widehat{\boldsymbol{\eta}}^{(0)} = \left(\widehat{\eta}_1^{(0)}, \dots, \widehat{\eta}_d^{(0)}\right)^\top$, $\text{tol} = 10^{-2}$.

(i) 由 (3.3) 式更新参数 $\widehat{\boldsymbol{\beta}}^{(k)} = \left(\widehat{\beta}_0^{(k)}, \widehat{\beta}_1^{(k)}\right)^\top$ 的值:

$$\widehat{\boldsymbol{\beta}}^{(k)} = \left(\sum_{t=2}^T \mathbf{X}_{\cdot, t-1}^\top \mathbf{X}_{\cdot, t-1} \right)^{-1} \sum_{t=2}^T \mathbf{X}_{\cdot, t-1}^\top \left(\mathbf{Y}_{\cdot, t} - \mathbf{A} \text{diag} \left(\widehat{\boldsymbol{\eta}}^{(k)} \right) \mathbf{Y}_{\cdot, t-1} \right);$$

(ii) 计算矩阵

$$\widetilde{\mathbf{G}}^{(k)} = \mathbf{A}_{\cdot, j} \mathbf{Y}_{j, -T} \odot \left(\mathbf{Y}_{\cdot, -T} - \widehat{\beta}_0^{(k)} \mathbf{J}_{N \times (T-1)} - \widehat{\beta}_1^{(k)} \mathbf{Y}_{\cdot, -T} - \mathbf{A} \text{diag} \left\{ \Pi_{0, j} \left(\widehat{\boldsymbol{\eta}}^{(k)} \right) \right\} \mathbf{Y}_{\cdot, -T} \right);$$

其中 \odot 为矩阵的 Hadamard 乘积;

(iii) 计算数值

$$G^{(k)} = \mathbf{J}_{1 \times d} \widetilde{\mathbf{G}}^{(k)} \mathbf{J}_{(T-1) \times 1};$$

(iv)

for $j = 1, \dots, p$ **do**

由 (3.4) 式更新参数 $\widehat{\boldsymbol{\eta}}^{(k+1)} = \left(\widehat{\eta}_1^{(k+1)}, \dots, \widehat{\eta}_d^{(k+1)}\right)^\top$ 的值:

if $|G^{(k)}| < \lambda/2$, **then**

$$\widehat{\eta}_j^{(k+1)} = 0;$$

else

$$\widehat{\eta}_j^{(k+1)} = \{G^{(k)} - \lambda \text{sign}(G^{(k)})/2\} / \{|\mathbf{Y}_{j, -T}|_2^2 |\mathbf{A}_{\cdot, j}|_2^2\};$$

end if

end for

(v)

if $\left| \widehat{\boldsymbol{\eta}}^{(k+1)} - \widehat{\boldsymbol{\eta}}^{(k)} \right|_\infty < \text{tol}$, **then**

stop;

else

$$k \leftarrow k + 1,$$

return to (i);

end if

Output

$$\widehat{\boldsymbol{\eta}} = \widehat{\boldsymbol{\eta}}^{(k+1)} \text{ 且 } \widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^{(k)};$$

return $\widehat{\boldsymbol{\eta}}$ 与 $\widehat{\boldsymbol{\beta}}$.

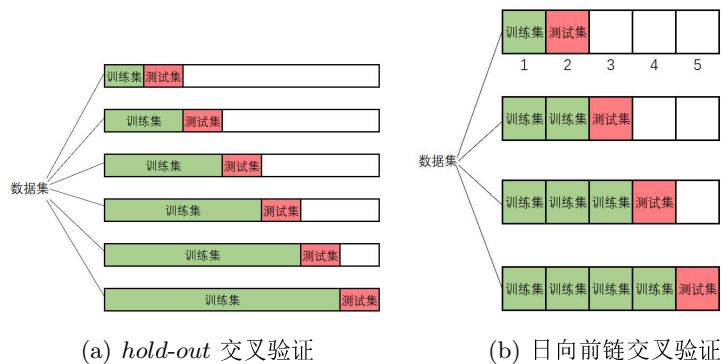


图 1

若数据集有 T 天, 则将生成 $T - 1$ 个不同的训练集和测试集分割, 如图 (b) 所示. 这样既可以保证测试集的时间顺序在训练子集后面, 又避免了分割测试集时的不确定性. 因此, 在本文的模拟中采用日向前链交叉验证法对调节参数进行选择.

将时间序列数据集按图 (b) 方法进行 $T - 1$ 次分割, 结合式子 (3.2), 记 $\hat{\beta}^{(m)}(\lambda), \hat{\eta}^{(m)}(\lambda)$ 为第 m 次分割 ($m = 1, \dots, T - 1$) 求解的参数估计值. 给出交叉验证得分的定义如下:

$$CV(\lambda) = \frac{1}{T-1} \sum_{m=1}^{T-1} L^{(m)}\left(\hat{\beta}^{(m)}(\lambda), \hat{\eta}^{(m)}(\lambda)\right),$$

其中 $L^{(m)}(\cdot)$ 为第 m 次分割的损失函数

$$L^{(m)}\left(\hat{\beta}^{(m)}(\lambda), \hat{\eta}^{(m)}(\lambda)\right) = \sum_{i=1}^d \sum_{t=m+1}^T (Y_{i,t} - \hat{\beta}_0^{(m)}(\lambda) - \hat{\beta}_1^{(m)}(\lambda) Y_{i,t-1} - \sum_{j=1}^d a_{ij} \hat{\eta}_j^{(m)}(\lambda) Y_{j,t-1})^2.$$

本文通过求取 $CV(\lambda)$ 的最小值点以得到调节参数 λ 的选取值.

4 模拟计算

本节对有向图上的高维时间序列模型进行模拟实验, 用于评估所提出方法的实用性. 利用 2.2 节中提到的充分必要条件 (2.3) 和充分条件 (2.5), 可控制生成时间序列数据的平稳性. 分别考虑 (2.3) 和 (2.5) 成立时, $d = 50, 100$ 和 200 以及 $T = 30, 100, 150$ 和 200 时参数的估计, 模拟次数设定为 500 次. 结果的展示中省略了 $\eta_j = 0$ 的估计值 $\hat{\eta}_j$, 缺失项 NA 表示该 η_j 无法估计. 采用日向前链时间序列交叉验证法选取 λ 为 0.01.

表 1, 表 2, 表 3 分别是在 (2.3) 成立时, $d = 50, 100$ 和 200 时参数 β 和 η 的估计值. 在网络大小 d 不变的条件下, 随着时间序列 T 的增大, 估计值的标准差越来越小. 在时间序列 T 相同的条件下, 随着 d 的增大, 参数 η 的维数也随之增加, 尽管如此, 所提的估计方法依然表现良好.

表 4, 表 5, 表 6 分别是在 (2.5) 成立时, $d = 50, 100$ 和 200 时参数 β 和 η 的估计值, 从中可以得到类似结论, 不做赘述.

5 交通网络实例分析

本节将本文所提出的方法应用于分析网络的都市链路速度数据集 (<https://www.nature.com/articles/s41597-019-0060-3>). 该数据集包含成都市交通网络的 $d=5943$ 条道路, 所观测为 2015 年 6 月 1 日 08:00–10:00, 12:00–14:00, 17:00–19:00, 21:00–23:00 四个不同时间段内每条道路的汽车平均行驶速度, 每个时间段中以两分钟为一个时间间隔, 即 $T = 60$. 将本文提出的有向图上的高维时间序列模型分别应用于成都市交通网络的四个不同时间段, 以研究成都市交通网络在时间以及空间上的变化规律.

我们首先选取成都市的繁华区域春熙路附近进行分析, 如图 2 所示. 以道路 4 为例, 它在四个时间段的交通平均速度的估计函数分别为:

$$\begin{aligned}\hat{Y}_{4t} &= 12.588 + 0.594\hat{Y}_{4(t-1)} - 0.004\hat{Y}_{5(t-1)} - 0.086\hat{Y}_{7(t-1)} - 0.090\hat{Y}_{10(t-1)}, t \in (08:00, 10:00), \\ \hat{Y}_{4t} &= 14.552 + 0.595\hat{Y}_{4(t-1)} - 0.075\hat{Y}_{5(t-1)} - 0.116\hat{Y}_{7(t-1)} + 0.058\hat{Y}_{10(t-1)}, t \in (12:00, 14:00), \\ \hat{Y}_{4t} &= 11.427 + 0.607\hat{Y}_{4(t-1)} - 0.017\hat{Y}_{5(t-1)} - 0.069\hat{Y}_{7(t-1)} - 0.078\hat{Y}_{10(t-1)}, t \in (17:00, 19:00), \\ \hat{Y}_{4t} &= 13.641 + 0.596\hat{Y}_{4(t-1)} - 0.024\hat{Y}_{5(t-1)} - 0.081\hat{Y}_{7(t-1)} - 0.013\hat{Y}_{10(t-1)}, t \in (21:00, 23:00).\end{aligned}$$

可以看出, 道路 4 在 t 时刻的平均速度受道路 4 在 $t-1$ 时刻的平均速度, 以及其相邻道路 5, 7, 10 在 $t-1$ 时刻的平均速度的影响. 不同时间段内, 道路 4 在 $t-1$ 时刻的平均速度对道路 4 在 t 时刻的平均速度影响都是正相关, 即道路 4 在 $t-1$ 时刻的平均速度越快, 道路 4 在 t 时刻的平均速度也越快. 不同时间段内, 道路 5, 7 在 $t-1$ 时刻的平均速度负向影响道路 4 在 t 时刻的平均速度, 这可能是因为道路 5, 7 周围是春熙路附近的商业街, 每个时间段内车流量都相对较大. 道路 5, 7 在 $t-1$ 时刻车辆较少时, 车辆几乎都拥堵在道路 4 上, 于是道路 5, 7 在 $t-1$ 时刻平均速度越快, 道路 4 在 t 时刻的平均速度相对较慢. 但不同时间段内, 道路 10 在 $t-1$ 时刻的平均速度对道路 4 在 t 时刻的平均速度的影响不同. 这可能与道路 10 旁侧商店较少且位于居民楼附近有一定关系, 故中午 12:00–14:00 车流量较小, 其余时间段附近车流量大. 因此中午 12:00–14:00 道路 10 在 $t-1$ 时刻速度越快, 道路 4 在 t 时刻速度也相对较快. 但在其余时间段, 道路 10 在 $t-1$ 时刻车辆较少时, 车辆均拥堵在道路 4 上, 于是道路 10 在 $t-1$ 时刻速度越快, 道路 4 在 t 时刻的平均速度也变慢.

进一步, 我们选取了成都市的郊区犀浦立交桥附近路段进行分析, 如图 3 所示. 以道路 16 为例, 它在四个时间段的交通平均速度的估计函数分别为:

$$\begin{aligned}\hat{Y}_{16t} &= 12.588 + 0.594\hat{Y}_{16(t-1)} - 0.044\hat{Y}_{12(t-1)} + 0.099\hat{Y}_{13(t-1)}, t \in (08:00, 10:00), \\ \hat{Y}_{16t} &= 14.552 + 0.595\hat{Y}_{16(t-1)} + 0.070\hat{Y}_{12(t-1)} - 0.002\hat{Y}_{13(t-1)}, t \in (12:00, 14:00), \\ \hat{Y}_{16t} &= 11.427 + 0.607\hat{Y}_{16(t-1)} + 0.017\hat{Y}_{12(t-1)} - 0.009\hat{Y}_{13(t-1)}, t \in (17:00, 19:00), \\ \hat{Y}_{16t} &= 13.641 + 0.596\hat{Y}_{16(t-1)} - 0.031\hat{Y}_{12(t-1)} + 0.246\hat{Y}_{13(t-1)}, t \in (21:00, 23:00).\end{aligned}$$

同样可以看出, 道路 16 在 t 时刻的平均速度受道路 16 在 $t-1$ 时刻的平均速度, 以及其相邻道路 12, 13 在 $t-1$ 时刻的平均速度的影响. 不同时间段内, 道路 16 在 $t-1$ 时刻的平均速度对道路 16 在 t 时刻的平均速度影响都是正相关, 即道路 16 在 $t-1$ 时刻的平均速度越快, 道路 16 在 t 时刻的平均速度也越快. 但不同时间段内, 道路 12, 13 在 $t-1$ 时刻的平均速

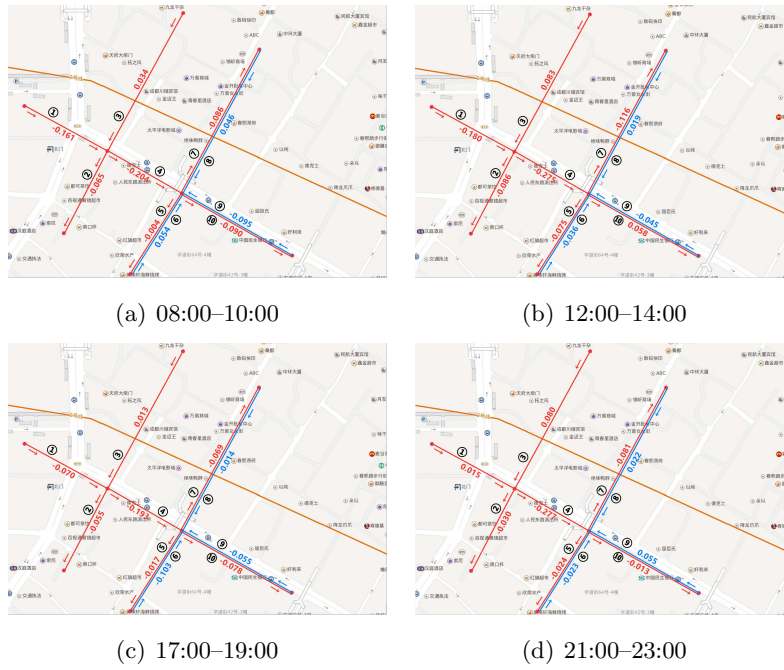


图 2 2015 年 6 月 1 日四个时间段春熙路附近道路网络及参数估计

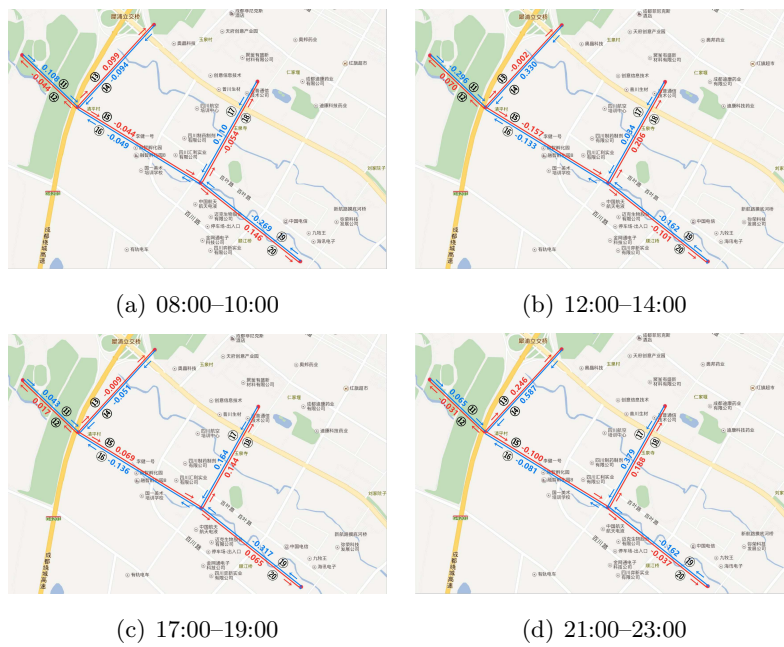


图 3 2015 年 6 月 1 日四个时间段犀浦立交附近道路网络及参数估计

度对道路 16 在 t 时刻的平均速度的影响不同. 在早上 08:00–10:00 和晚上 21:00–23:00, 道路 12 在 $t-1$ 时刻的平均速度负向影响道路 16 在 t 时刻的平均速度. 道路 12 前往成都电子科技大学, 可能与教职工在早上 08:00–10:00 上班和晚上 21:00–23:00 下班有一定关系, 故道路 12 在这两个时间段车流量大, 道路 12 在 $t-1$ 时刻车辆较少时, 车辆均拥堵在道路 16 上. 因此道路 12 在 $t-1$ 时刻平均速度越快, 道路 16 在 t 时刻的平均速度相对较慢. 而道路 13 前往博雅体育俱乐部, 人们运动和学习的时间往往是相反的, 于是在这两个时间段, 道路 13 在 $t-1$ 时刻的平均速度正向影响道路 16 在 t 时刻的平均速度. 其余两个时间范围则恰恰相反.

值得注意的是, 由于正则化方法只给出了参数的估计值, 而对于其假设检验问题未曾涉及. 显然, 该问题具有更大的挑战性. 鉴于此, 这些统计结论均为描述性的, 其显著性如何尚未考究.

6 总结

本文提出了有向图上的高维时间序列模型, 建立其正则化统计方法. 数值的结果表明所提出的方法具有良好的表现, 在成都市交通网络数据的应用上, 也得到一些可解释的统计结论. 进一步, 若考虑网络的组群效应, 可将 group lasso [13] 作为正则项引入, 相应的估计方法与算法实现均需进一步研究.

参 考 文 献

- [1] Box G E P, Jenkins G M, Reinsel G C. Time series analysis forecasting and control [M]. New York: Prentice Hall, 1994.
- [2] Fan Jianqing, Yao Qiwei. Nonlinear Time Series: Nonparametric and parametric methods [M]. New York: Springer Science and Business Media, 2003.
- [3] Hamilton James. Time series analysis [M]. New Jersey: Princeton University Press, 1994.
- [4] McQuarrie D R, Tsai C L. Regression and time series model selection [M]. Singapore: World Scientific, 1998.
- [5] Shumway R H, Stoffer D S. Time series analysis and its application [M]. New York: Springer, 2000.
- [6] Brown L D, Hagerman R, Griffin P A, Zmijewski M. Security analyst superiority relative to univariate time-series models in forecasting quarterly earnings [J]. Journal of Accounting and Economics, 1987, 9(1): 61–87.
- [7] Cuaresma J C, Hlouskova J, Kossmeier S, Obersteiner M. Forecasting electricity spot-prices using linear univariate time-series models [J]. Applied Energy, 2004, 77(1): 87–106.
- [8] Newbold P, Granger C W. Experience with forecasting univariate time series and the combination of forecasts [J]. Journal of the Royal Statistical Society Series B, 1974, 137(2): 131–165.
- [9] Bedrick E J, Tsai C L. Model selection for multivariate regression in small samples [J]. Biometrics, 1994, 50(1): 226–231.
- [10] Lütkepohl H. New introduction to multiple time series analysis [M]. New York: Springer Science and Business Media, 2007.
- [11] Zhu Xuening, Pan Rui, Li Guodong, Liu Yuewen, Wang Hansheng. Network vector autoregression [J]. Annals of Statistics, 2017, 45(3): 1096–1123.
- [12] Zhu Xuening, Pan Rui. Grouped network vector autoregression [J]. Statistica Sinica, 2020, 30(3): 1437–1462.

- [13] Yuan Ming, Lin Yi. Model selection and estimation in regression with Grouped variables [J]. Journal of the Royal Statistical Series B, 2006, 68(1): 49–67.

HIGH-DIMENSIONAL TIME SERIES MODEL ON DIRECTED GRAPH AND ITS APPLICATION IN TRAFFIC NETWORK

XIANG Qiao-yi¹, QI Hui²

(1. School of Mathematics and Statistics, Wuhan University, Wuhan 430072, China)

(2. Institute of Information Engineering, Sanming University, Sanming 365004, China)

Abstract: In this paper, we study the variation of the average speed of urban road vehicles in time and space. Based on the characteristics of road traffic network structure, the higher dimensions on the directed graph are proposed. By introducing ℓ_1 -penalty term, the regularized least squares parameter estimation method is established, and the selection method of adjusting parameters in this process is studied. The method is applied to the urban link speed data set of Chengdu road network. The temporal and spatial variation rules of transportation efficiency in downtown and remote suburbs of Chengdu are obtained. This study can provide effective statistical suggestions for relieving traffic jams and alleviating traffic pressure.

Keywords: network data; average vehicle speed; high dimensions; transportation

2010 MR Subject Classification: 62M10; 62P99

表 1 当 (2.3) 成立时, $d = 50$ 时参数 β 和 η 的估计值

真值	$T = 30$		$T = 100$		$T = 150$		$T = 200$	
	估计值	标准差	估计值	标准差	估计值	标准差	估计值	标准差
$\beta_1 = 1$	1.052	0.032	1.050	0.017	1.050	0.014	1.050	0.012
$\beta_2 = 0.2$	0.201	0.006	0.202	0.003	0.202	0.003	0.202	0.002
$\eta_1 = 1$	0.977	0.137	0.974	0.070	0.976	0.058	0.974	0.050
$\eta_2 = 6$	5.962	0.142	5.966	0.069	5.965	0.056	5.965	0.050
$\eta_{16} = 3$	NA	NA	NA	NA	NA	NA	NA	NA
$\eta_{22} = 5$	4.972	0.189	4.972	0.091	4.971	0.077	4.971	0.067
$\eta_{40} = 2$	1.974	0.086	1.967	0.046	1.966	0.037	1.966	0.032
$\eta_{42} = 0.8$	0.797	0.007	0.796	0.004	0.796	0.003	0.796	0.003
$\eta_{49} = 3$	2.990	0.022	2.986	0.013	2.986	0.011	2.987	0.009

表 2 当 (2.3) 成立时, $d = 100$ 时参数 β 和 η 的估计值

真值	$T = 30$		$T = 100$		$T = 150$		$T = 200$	
	估计值	标准差	估计值	标准差	估计值	标准差	估计值	标准差
$\beta_1 = 1$	1.049	0.031	1.045	0.017	1.045	0.014	1.045	0.012
$\beta_1 = 0.2$	0.199	0.002	0.199	0.001	0.200	0.001	0.200	0.001
$\eta_1 = 1$	0.979	0.106	0.982	0.053	0.983	0.045	0.983	0.040
$\eta_2 = 6$	5.981	0.081	5.983	0.043	5.983	0.035	5.984	0.031
$\eta_{16} = 3$	2.989	0.173	2.995	0.088	2.994	0.071	2.995	0.058
$\eta_{22} = 5$	4.985	0.083	4.984	0.042	4.985	0.033	4.986	0.030
$\eta_{40} = 2$	1.992	0.073	1.990	0.038	1.990	0.032	1.989	0.027
$\eta_{42} = 0.8$	0.780	0.128	0.779	0.062	0.779	0.052	0.778	0.046
$\eta_{49} = 3$	2.986	0.175	2.985	0.096	2.987	0.072	2.989	0.062
$\eta_{52} = 1$	0.993	0.101	0.992	0.050	0.990	0.040	0.991	0.034
$\eta_{64} = 9$	9.003	0.022	9.004	0.011	9.003	0.008	9.003	0.007
$\eta_{68} = 4$	3.999	0.015	3.999	0.007	3.998	0.006	3.998	0.005
$\eta_{71} = 1$	0.981	0.141	0.985	0.075	0.985	0.060	0.985	0.054
$\eta_{75} = 6$	5.982	0.189	5.990	0.091	5.989	0.070	5.991	0.061
$\eta_{79} = 2$	1.983	0.087	1.985	0.045	1.985	0.035	1.984	0.030
$\eta_{82} = 3$	2.993	0.047	2.995	0.024	2.995	0.020	2.994	0.017
$\eta_{90} = 0.9$	0.892	0.050	0.894	0.026	0.894	0.022	0.892	0.019
$\eta_{97} = 7$	6.994	0.089	6.992	0.048	6.991	0.037	6.991	0.032
$\eta_{100} = 2$	1.997	0.022	1.998	0.010	1.998	0.009	1.998	0.008

表 3 当 (2.3) 成立时, $d = 200$ 时参数 β 和 η 的估计值

真值	$T = 30$		$T = 100$		$T = 150$		$T = 200$	
	估计值	标准差	估计值	标准差	估计值	标准差	估计值	标准差
$\beta_1 = 1$	1.049	0.022	1.048	0.012	1.049	0.010	1.048	0.008
$\beta_2 = 0.2$	0.200	0.001	0.200	0.001	0.200	0.0004	0.200	0.0004
$\eta_1 = 1$	0.981	0.066	0.982	0.035	0.981	0.029	0.982	0.024
$\eta_2 = 6$	NA	NA	NA	NA	NA	NA	NA	NA
$\eta_{16} = 3$	2.999	0.017	2.998	0.008	2.998	0.007	2.998	0.006
$\eta_{22} = 5$	4.989	0.087	4.993	0.044	4.992	0.036	4.993	0.031
$\eta_{40} = 2$	1.993	0.157	1.984	0.077	1.983	0.062	1.981	0.055
$\eta_{42} = 0.8$	0.770	0.105	0.765	0.055	0.766	0.045	0.766	0.037
$\eta_{49} = 3$	2.990	0.095	2.993	0.048	2.992	0.039	2.992	0.033
$\eta_{52} = 1$	0.980	0.081	0.981	0.041	0.980	0.034	0.980	0.028
$\eta_{64} = 9$	8.993	0.034	8.990	0.018	8.990	0.015	8.990	0.013
$\eta_{68} = 4$	3.997	0.006	3.996	0.003	3.996	0.003	3.996	0.002
$\eta_{71} = 1$	0.987	0.074	0.989	0.038	0.989	0.031	0.989	0.027
$\eta_{75} = 6$	5.976	0.144	5.981	0.078	5.979	0.061	5.979	0.052
$\eta_{79} = 2$	2.019	0.153	2.027	0.084	2.023	0.067	2.025	0.059
$\eta_{82} = 3$	2.979	0.077	2.981	0.041	2.980	0.034	2.980	0.030
$\eta_{90} = 0.9$	0.891	0.057	0.890	0.031	0.890	0.026	0.889	0.023
$\eta_{97} = 7$	NA	NA	NA	NA	NA	NA	NA	NA
$\eta_{100} = 2$	1.996	0.028	1.996	0.014	1.996	0.011	1.997	0.009
$\eta_{111} = 4$	3.991	0.113	3.982	0.059	3.982	0.049	3.981	0.041
$\eta_{118} = 1$	0.987	0.088	0.983	0.047	0.984	0.038	0.984	0.033
$\eta_{125} = 2$	1.993	0.017	1.994	0.009	1.994	0.007	1.994	0.006
$\eta_{132} = 8$	NA	NA	NA	NA	NA	NA	NA	NA
$\eta_{139} = 5$	4.974	0.124	4.969	0.068	4.969	0.054	4.972	0.048
$\eta_{141} = 3$	2.978	0.067	2.979	0.036	2.980	0.027	2.980	0.023
$\eta_{144} = 1$	0.977	0.076	0.978	0.037	0.978	0.030	0.979	0.027
$\eta_{145} = 4$	3.998	0.004	3.997	0.003	3.997	0.002	3.997	0.002
$\eta_{160} = 3$	2.991	0.033	2.992	0.017	2.992	0.014	2.992	0.012
$\eta_{163} = 7$	6.968	0.119	6.974	0.061	6.975	0.048	6.975	0.043
$\eta_{167} = 0.7$	NA	NA	NA	NA	NA	NA	NA	NA
$\eta_{172} = 1$	0.983	0.111	0.982	0.057	0.983	0.046	0.983	0.040
$\eta_{176} = 2$	NA	NA	NA	NA	NA	NA	NA	NA
$\eta_{181} = 5$	4.980	0.085	4.977	0.046	4.978	0.037	4.977	0.032
$\eta_{183} = 1$	0.988	0.087	0.985	0.042	0.985	0.034	0.985	0.030
$\eta_{191} = 6$	NA	NA	NA	NA	NA	NA	NA	NA
$\eta_{194} = 2$	1.986	0.106	1.980	0.059	1.984	0.047	1.983	0.041
$\eta_{199} = 1$	1.007	0.053	1.010	0.029	1.009	0.024	1.009	0.022

表 4 当 (2.5) 成立时, $d = 50$ 时参数 β 和 η 的估计值

真值	$T = 30$		$T = 100$		$T = 150$		$T = 200$	
	估计值	标准差	估计值	标准差	估计值	标准差	估计值	标准差
$\beta_1 = 1$	1.096	0.125	1.071	0.073	1.065	0.062	1.063	0.055
$\beta_2 = 0.2$	0.169	0.026	0.187	0.013	0.189	0.011	0.190	0.009
$\eta_1 = 0.1$	0.098	0.031	0.096	0.016	0.097	0.013	0.097	0.011
$\eta_2 = 0.06$	0.055	0.029	0.057	0.014	0.057	0.011	0.057	0.010
$\eta_{16} = 0.03$	0.030	0.023	0.029	0.012	0.028	0.010	0.028	0.008
$\eta_{22} = 0.05$	0.050	0.028	0.049	0.014	0.049	0.011	0.049	0.009
$\eta_{40} = 0.2$	0.198	0.025	0.196	0.013	0.195	0.011	0.196	0.009
$\eta_{42} = 0.08$	0.077	0.031	0.077	0.014	0.077	0.012	0.077	0.010
$\eta_{49} = 0.03$	0.028	0.032	0.028	0.017	0.028	0.014	0.028	0.012

表 5 当 (2.5) 成立时, $d = 100$ 时参数 β 和 η 的估计值

真值	$T = 30$		$T = 100$		$T = 150$		$T = 200$	
	估计值	标准差	估计值	标准差	估计值	标准差	估计值	标准差
$\beta_1 = 1$	1.083	0.116	1.065	0.077	1.073	0.066	1.072	0.057
$\beta_2 = 0.2$	0.169	0.018	0.186	0.010	0.190	0.008	0.191	0.007
$\eta_1 = 0.01$	0.009	0.022	0.008	0.011	0.009	0.009	0.009	0.007
$\eta_2 = 0.06$	0.059	0.024	0.059	0.012	0.059	0.010	0.059	0.008
$\eta_{16} = 0.03$	0.030	0.018	0.029	0.009	0.029	0.008	0.029	0.006
$\eta_{22} = 0.05$	0.048	0.022	0.048	0.010	0.049	0.008	0.049	0.007
$\eta_{40} = 0.02$	0.020	0.021	0.020	0.010	0.020	0.008	0.020	0.007
$\eta_{42} = 0.08$	0.079	0.024	0.078	0.012	0.079	0.010	0.078	0.009
$\eta_{49} = 0.03$	0.031	0.020	0.029	0.009	0.029	0.008	0.029	0.007
$\eta_{52} = 0.01$	0.009	0.023	0.009	0.011	0.009	0.009	0.009	0.008
$\eta_{64} = 0.09$	0.091	0.019	0.089	0.009	0.088	0.008	0.088	0.007
$\eta_{68} = 0.04$	0.041	0.019	0.040	0.010	0.039	0.008	0.039	0.007
$\eta_{71} = 0.01$	0.008	0.026	0.010	0.013	0.010	0.010	0.010	0.009
$\eta_{75} = 0.06$	0.060	0.022	0.058	0.012	0.058	0.009	0.058	0.008
$\eta_{79} = 0.02$	0.019	0.022	0.020	0.011	0.019	0.009	0.019	0.008
$\eta_{82} = 0.03$	0.030	0.021	0.029	0.010	0.028	0.008	0.028	0.007
$\eta_{90} = 0.09$	0.089	0.024	0.089	0.012	0.088	0.010	0.088	0.009
$\eta_{97} = 0.07$	0.070	0.025	0.068	0.012	0.067	0.010	0.067	0.009
$\eta_{100} = 0.02$	0.019	0.023	0.018	0.010	0.019	0.009	0.019	0.008

表 6 当 (2.5) 成立时, $d = 200$ 时参数 β 和 η 的估计值

真值	$T = 30$		$T = 100$		$T = 150$		$T = 200$	
	估计值	标准差	估计值	标准差	估计值	标准差	估计值	标准差
$\beta_1 = 1$	1.200	0.173	1.089	0.112	1.069	0.099	0.986	0.091
$\beta_2 = 0.2$	0.172	0.013	0.189	0.007	0.191	0.006	0.191	0.005
$\eta_1 = 0.01$	0.011	0.015	0.010	0.008	0.009	0.006	0.009	0.005
$\eta_2 = 0.06$	0.062	0.017	0.060	0.008	0.059	0.007	0.058	0.006
$\eta_{16} = 0.03$	0.032	0.017	0.030	0.008	0.029	0.007	0.028	0.006
$\eta_{22} = 0.05$	0.051	0.017	0.050	0.008	0.049	0.006	0.048	0.005
$\eta_{40} = 0.02$	0.019	0.016	0.019	0.008	0.019	0.006	0.020	0.005
$\eta_{49} = 0.03$	0.030	0.016	0.029	0.008	0.029	0.006	0.029	0.005
$\eta_{52} = 0.01$	0.009	0.015	0.010	0.008	0.010	0.006	0.010	0.005
$\eta_{68} = 0.04$	0.039	0.016	0.039	0.008	0.039	0.006	0.039	0.005
$\eta_{71} = 0.01$	0.011	0.017	0.010	0.008	0.010	0.006	0.010	0.005
$\eta_{75} = 0.06$	0.060	0.014	0.059	0.007	0.059	0.006	0.059	0.005
$\eta_{82} = 0.03$	0.031	0.017	0.029	0.008	0.029	0.006	0.029	0.005
$\eta_{100} = 0.02$	0.019	0.016	0.019	0.008	0.020	0.006	0.020	0.005
$\eta_{111} = 0.04$	0.039	0.015	0.039	0.007	0.039	0.006	0.039	0.005
$\eta_{118} = 0.01$	0.004	0.018	0.009	0.009	0.008	0.007	0.011	0.006
$\eta_{125} = 0.02$	0.016	0.019	0.019	0.009	0.019	0.007	0.020	0.006
$\eta_{139} = 0.05$	0.046	0.017	0.048	0.008	0.049	0.006	0.050	0.005
$\eta_{141} = 0.03$	0.027	0.016	0.029	0.008	0.029	0.006	0.030	0.005
$\eta_{160} = 0.03$	0.029	0.018	0.029	0.008	0.030	0.007	0.031	0.006
$\eta_{163} = 0.07$	0.068	0.018	0.068	0.008	0.070	0.007	0.071	0.006
$\eta_{172} = 0.01$	0.009	0.016	0.009	0.007	0.010	0.006	0.011	0.005
$\eta_{176} = 0.02$	0.017	0.017	0.018	0.008	0.019	0.006	0.020	0.005
$\eta_{181} = 0.05$	0.050	0.017	0.049	0.008	0.050	0.007	0.051	0.006
$\eta_{191} = 0.06$	0.062	0.017	0.059	0.009	0.060	0.007	0.059	0.006
$\eta_{194} = 0.02$	0.019	0.022	0.018	0.009	0.020	0.007	0.021	0.006
$\eta_{199} = 0.01$	0.009	0.017	0.009	0.008	0.010	0.007	0.010	0.006