doi:10.3772/j.issn.1002-0470.2017.06.011

稀疏组 lasso 罚向量自回归模型的大气污染物预测:京津冀案例研究①

王金甲② 孙梦然 郝 智

(燕山大学信息科学与工程学院 秦皇岛 066004)

摘 要 进行了大气污染物预测研究。针对传统的向量自回归模型方法所面临的过参数化问题,提出了稀疏组 lasso 罚向量自回归模型并应用近邻梯度下降法求解模型参数。为了验证模型的有效性,将其应用于 2015 年京津冀大气污染物数据中并对 2016 年 1 月 1 日北京 6 项大气污染物浓度进行预测。实验数据表明:基于稀疏组 lasso 罚模型的 PM2.5 预测归一化均方误差约为 3.8%,预测精度高于向量自回归 (VAR) 模型、基于各种稀疏结构的向量自回归 (VAR-L)模型、分层向量自回归 (HVAR)模型。此外,京津冀不同城市对北京的空气质量影响程度不同,这可以通过组内稀疏模型参数进行解释。将凸优化概念与向量自回归模型结合应用于大气污染物浓度的预测中,对京津冀大气污染协同治理具有重要意义。

关键词 向量自回归 (VAR)模型,稀疏组 lasso,近邻梯度下降法,凸优化,大气污染

0 引 言

大气污染已经成为京津冀一体化中的区域性难题[1]。空气质量指数是目前反映城市大气环境质量的重要指标,参与空气质量评价的主要大气污染物为细颗粒物(PM10)、可吸入颗粒物(PM2.5)、二氧化硫(SO_2)、二氧化氮(NO_2)、臭氧(O_3)和一氧化碳(CO)等 6 项[2]。频发的霾污染是目前京津冀最严重的环境问题[3]。京津冀地区作为华北地区的主要经济圈,其大气污染问题已经成为公众关注的焦点,也是各级政府亟待解决的问题之一。

目前,对京津冀大气污染的研究主要有以下工作:缪育聪等[4]对京津冀地区频发的霾污染事件进行研究,结果表明京津冀地区独特的地理环境条件加上城市群的快速发展形成的局地大气环流会对局地的污染过程产生重大的影响,京津冀地区的污染控制需要城市群的联动应对治理。安树伟等[5]通过分析京津冀大气污染状况与能源消费情况,明确了京津冀大气污染与河北能源消费总量的快速增加有直接的联系,并认为河北能源消费总量的快速增加与其高耗能高污染的产业结构有关。丁峰等[6]对京津冀地区大气污染现状进行分析,指出三地联防联控需要合理有效利用总量控制下的三种减排模式、实施区域内外多种污染物的协调控制、积极推动以环境空气质量改善为导向的污染物减排模式。

此外,众多学者已经开展了关于大气污染物的统计建模研究工作。滕丽等[7]应用向量自回归(vector autoregression, VAR)模型分析珠江三角洲 9 个城市空气质量的区域影响,数据分析结果显示珠江三角洲 9 个城市间空气质量的区域影响包括自我影响和跨域影响。刘金培等[8]基于向量自回归(VAR)模型研究了西安市大气污染物和气象因素对 PM2.5 影响动态关系,CO、SO_2、O_3 和气温的正向变动会引起 PM2.5 浓度增加,风速和降水量的正向变动则会引起 PM2.5 浓度降低。刘华军等[9]基于

① 国家自然科学基金(61273019,61473339),河北省青年拔尖人才支持计划([2013]17)和中国博士后科学基金(2014M561202)资助项目。

② 男,1978 年生,博士,博士生导师,教授;研究方向:传感器信号处理和模式识别;联系人,E-mail:wjj@ysu.edu.cn

(收稿日期:2017-02-28)

京津冀地区城市空气质量指数(AQI)日报数据,采用 VAR 模型研究了京津冀城市群间大气污染的非线性传导关系,并借助社会网络分析方法揭示其联动网络结构特征。

带有外生变量(exogenous variables)的向量自回归(VAR)模型称为 VARX 模型,通过文献分析可以看出,VARX 模型是对时间序列进行建模以及预测的一种有效方法,但是由于其在参数空间上没有施加约束条件,因而出现严重的过参数化问题[10]。本文提出了采用稀疏组 lasso 罚的 VARX-L 模型和近邻梯度下降法算法用于京津冀大气污染研究,在 VARX 基础上施加结构化稀疏的同时,考虑待估系数矩阵的特点,将组件自己的滞后和其它组件的滞后关系以及内生变量和外生变量潜在的嵌套结构进行描述,有效地缩减了参数空间。比起贝叶斯算法,这种"时滞组"的方法的优势在于:将最小二乘缩减至零的同时以一种高效的方式实现变量选择,同时在给每一个模型系数施加罚函数时,避免赤池信息量准则(Akaike information criterion, AIC)趋于过参数化,贝叶斯信息量准则(Bayesian information criterion, BIC)趋于欠拟合的问题。除此之外,由于 VARX-L 框架引入凸优化的概念使得模型能更好地应用在多变量时间序列中。本研究采用京津冀 13 个城市群的大气污染物 2015 年 1 月 1 日 ~2016 年 1 月 1 日的日报数据,采用 VARX-L 罚模型进行关联分析和预测。从数据驱动的角度分析京津冀城市群对北京大气污染物的影响,对京津冀大气污染协同治理具有重要意义。

1 研究方法

1.1 VARX-L 模型

带有外生变量的向量自回归模型表示为 $\mathrm{VARX}_{k,m}(p,s)$,其中 $\{\boldsymbol{y}_t\}_{t=1}^{\mathrm{T}}$ 作为内生变量,代表 k 维多元时间序列,滞后阶数为 p,$\{\boldsymbol{x}_t\}_{t=1}^{\mathrm{T}}$ 作为外生变量,代表 m 维多元时间序列,滞后阶数为 s。$\mathrm{VARX}_{k,m}(p,s)$ 模型定义如下:

$$\boldsymbol{y}_t = \boldsymbol{v} + \sum_{l=1}^{p} \boldsymbol{\Phi}^{(l)} \boldsymbol{y}_{t-l} + \sum_{j=1}^{s} \boldsymbol{\beta}^{(j)} \boldsymbol{x}_{t-j} + \boldsymbol{u}_t, t = 1, \cdots, T \tag{1}$$

其中 $v \in \mathbb{R}^k$ 是 k 维的截距常向量,$\{\boldsymbol{\Phi}^{(l)} \in \mathbb{R}^{k\times k}\}_{l=1}^{p}$ 是时滞为 l 的内生变量系数矩阵,$\{\boldsymbol{\beta}^{(j)} \in \mathbb{R}^{k\times m}\}_{j=1}^{s}$ 是时滞为 j 的外生变量系数矩阵,$\{\boldsymbol{u}_t \in \mathbb{R}^k\}_{t=1}^{T}$ 是独立同分布的 k 维白噪声,均值为 0,协方差矩阵为 $\boldsymbol{\Sigma}_u$,即 $\boldsymbol{u}_t$ 满足以下两个条件:

$$① E(\boldsymbol{u}_t) = 0; ② E(\boldsymbol{u}_t \boldsymbol{u}_\tau) = \begin{cases} 0 & t \neq \tau \\ \boldsymbol{\Sigma}_u & t = \tau \end{cases} \tag{2}$$

$\{\boldsymbol{\beta}^{(j)} \in \mathbb{R}^{k\times m}\}_{j=1}^{s} = 0$ 时的 $\mathrm{VARX}_{k,m}(p,s)$ 即为 VAR 模型。

VARX 模型参数 $\{\hat{\boldsymbol{v}}, \hat{\boldsymbol{\Phi}}, \hat{\boldsymbol{\beta}}\}$ 通过求解以下优化问题进行更新:

$$\min_{v,\Phi,\beta} \sum_{t=1}^{T} \left\| \boldsymbol{y}_t - \boldsymbol{v} - \sum_{l=1}^{p} \boldsymbol{\Phi}^{(l)} \boldsymbol{y}_{t-l} - \sum_{j=1}^{s} \boldsymbol{\beta}^{(j)} \boldsymbol{x}_{t-j} \right\|_F^2 \tag{3}$$

上述问题是一个简单的最小二乘问题,使用最小二乘需要估计的参数个数为 $k(1 + kp + ms)$。当内生变量和外生变量的维数增高、时滞增大时,传统 VARX 模型出现严重的过参数化问题。因此通过在系数矩阵内部增加各种稀疏结构的方式(即 VARX-L 模型)来压缩 VARX 模型的参数空间。

VARX-L 模型参数估计描述为以下形式:

$$\min_{v,\Phi,\beta} \sum_{t=1}^{T} \left\| \boldsymbol{y}_t - \boldsymbol{v} - \sum_{l=1}^{p} \boldsymbol{\Phi}^{(l)} \boldsymbol{y}_{t-l} - \sum_{j=1}^{s} \boldsymbol{\beta}^{(j)} \boldsymbol{x}_{t-j} \right\|_F^2 + \lambda [P_y(\boldsymbol{\Phi}) + P_x(\boldsymbol{\beta})] \tag{4}$$

其中,$\| \boldsymbol{A} \|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2}$ 是 $m \times n$ 的矩阵 $\boldsymbol{A}$ 的 F-范数,$\lambda \geqslant 0$ 代表正则化罚参数,$Py(\boldsymbol{\Phi})$ 代表内生待估矩阵系数的罚函数,$P_x(\boldsymbol{\beta})$ 代表外生待估矩阵系数的罚函数。表 1 中给出了 VARX-L 模型的几种罚结构。由于所有的模型系数均使用了同一个罚参数,因此将所有时间序列进行均值为 0 方差为 1 的标准化处理。表 1 中 $\boldsymbol{\Phi}_{on}^{(l)}$ 和 $\boldsymbol{\Phi}_{off}^{(l)}$ 分别代表系数矩阵 $\boldsymbol{\Phi}^{(l)}$ 的对角线和非对角线元素。从表中可以看出式(5)是式(7) $\alpha = 0$ 时的特例,式(6)是式(8) $\alpha = 0$ 时的特例,式(9)是 VARX 的 lasso 罚模型。下面重点介绍式(7)和式(8)的稀疏组 lasso 罚函数 VARX-L 模型。Endogenous-First VARX-L 模型和算法参考文献[11]。

表 1　VARX-L 模型罚函数

罚函数	$Py(\boldsymbol{\Phi})$	$Px(\boldsymbol{\beta})$
(5) Lag	$\sqrt{k^2}\sum_{l=1}^{p}\|\boldsymbol{\Phi}^{(l)}\|_F$	$\sqrt{k}\sum_{j=1}^{s}\sum_{i=1}^{m}\|\boldsymbol{\beta}_{\cdot,i}^{(j)}\|_F$
(6) Own/Other	$\sqrt{k}\sum_{l=1}^{p}\|\boldsymbol{\Phi}_{on}^{(l)}\|_F+\sqrt{k(k-1)}\sum_{l=1}^{p}\|\boldsymbol{\Phi}_{off}^{(l)}\|_F$	$\sqrt{k}\sum_{j=1}^{s}\sum_{i=1}^{m}\|\boldsymbol{\beta}_{\cdot,i}^{(j)}\|_F$
(7) Sparse Lag	$(1-\alpha)(\sqrt{k^2}\sum_{l=1}^{p}\|\boldsymbol{\Phi}^{(l)}\|_F)+\alpha\|\boldsymbol{\Phi}\|_1$	$(1-\alpha)(\sqrt{k}\sum_{j=1}^{s}\sum_{i=1}^{m}\|\boldsymbol{\beta}_{\cdot,i}^{(j)}\|_F)+\alpha\|\boldsymbol{\beta}\|_1$
(8) Sparse Own/Other	$(1-\alpha)(\sqrt{k}\sum_{l=1}^{p}\|\boldsymbol{\Phi}_{on}^{(l)}\|_F+\sqrt{k(k-1)}\sum_{l=1}^{p}\|\boldsymbol{\Phi}_{off}^{(l)}\|_F)+\alpha\|\boldsymbol{\Phi}\|_1$	$(1-\alpha)(\sqrt{k}\sum_{j=1}^{s}\sum_{i=1}^{m}\|\boldsymbol{\beta}_{\cdot,i}^{(j)}\|_F)+\alpha\|\boldsymbol{\beta}\|_1$
(9) Basic	$\|\boldsymbol{\Phi}\|_1$	$\|\boldsymbol{\beta}\|_1$

式(7)对应的 Sparse Lag Group VARX-L 模型讨论如下。在实际应用中,式(5)和式(6)的组罚函数结构可能太过于严格,它要求整组系数都为零或者整组系数都非零,并且组罚结构在一个组的内部不能产生稀疏性。Song 等[12]试图通过增加额外的 lasso 罚来避免这一限制,但是这种方法需要一个多维网格来选择罚参数。Sim 等[13]首次提出的稀疏组 lasso 方法,将 lasso 罚和组 lasso 罚优化结合产生了组内稀疏结构。Sparse Lag Group VARX-L 在稀疏组 lasso 的基础上纳入内在的滞后结构,其罚结构如下:

$$P_y(\boldsymbol{\Phi})=(1-\alpha)(\sqrt{k^2}\sum_{l=1}^{p}\|\boldsymbol{\Phi}^{(l)}\|_F)+\alpha\|\boldsymbol{\Phi}\|_1 \tag{10}$$

$$P_x(\boldsymbol{\beta})=(1-\alpha)(\sqrt{k}\sum_{j=1}^{s}\sum_{i=1}^{m}\|\boldsymbol{\beta}_{\cdot,i}^{(j)}\|_F)+\alpha\|\boldsymbol{\beta}\|_1 \tag{11}$$

其中 $\alpha\in[0,1]$ 是额外控制组内稀疏的罚参数,可以由交叉验证的方法进行估计,本文根据相关组的大小对 α 进行设置,令 $\alpha=\frac{1}{k+1}$,其系数矩阵稀疏结构如图 1 所示。

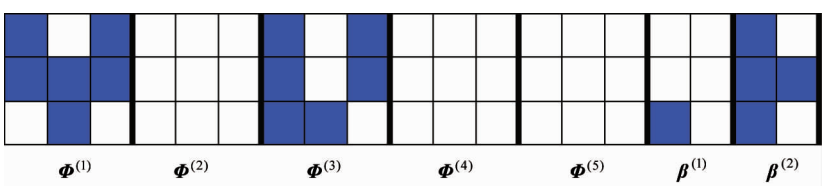

图 1　Sparse Lag Group VARX-L$_{(3,2)}$(5,2) 稀疏模式示意图

Sparse Own/Other Group VARX-L 模型讨论如下。$\boldsymbol{\Phi}^{(l)}$ 对角线上的系数代表序列自己的滞后回归,在很多应用中这些系数都是非零的,非对角线上的系数代表与其他组件的滞后交叉依赖关系。在 Sparse Lag Group VARX-L 的基础之上,Sparse Own/Other Group VARX-L 将每一个内生滞后系数矩阵 $\boldsymbol{\Phi}^{(l)}$ 加入到不同的分组里,其罚结构如下:

$$P_y(\boldsymbol{\Phi})=(1-\alpha)(\sqrt{k}\sum_{l=1}^{p}\|\boldsymbol{\Phi}_{on}^{(l)}\|_F+\sqrt{k(k-1)}\sum_{l=1}^{p}\|\boldsymbol{\Phi}_{off}^{(l)}\|_F)+\alpha\|\boldsymbol{\Phi}\|_1 \tag{12}$$

$$P_x(\boldsymbol{\beta})=(1-\alpha)(\sqrt{k}\sum_{j=1}^{s}\sum_{i=1}^{m}\|\boldsymbol{\beta}_{\cdot,i}^{(j)}\|_F)+\alpha\|\boldsymbol{\beta}\|_1 \tag{13}$$

系数矩阵稀疏结构如图 2 所示。

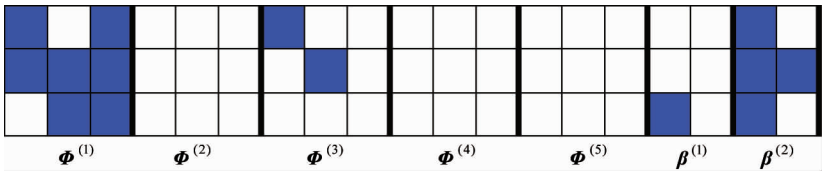

图 2 Sparse Own/Other Group VARX-L$_{(3,2)}$(5,2)稀疏模式示意图

1.2 近邻梯度下降法求解模型参数

为求解方便,我们定义 VARX-L 模型中的紧矩阵符号: $\boldsymbol{Y} = [\boldsymbol{y}_1,\cdots,\boldsymbol{y}_T]$, $\boldsymbol{Z}_t = [\boldsymbol{y}_t^T,\ldots,\boldsymbol{y}_{t-p}^T, \boldsymbol{x}_t^T,\ldots,\boldsymbol{x}_{t-p}^T]^T$, $\boldsymbol{Z} = [\boldsymbol{Z}_2,\ldots,\boldsymbol{Z}_{T-1}]$, $\boldsymbol{\Phi} = [\boldsymbol{\Phi}^{(1)}, \boldsymbol{\Phi}^{(2)},\ldots,\boldsymbol{\Phi}^{(p)}]$, $\boldsymbol{\beta} = [\boldsymbol{\beta}^{(1)},\ldots,\boldsymbol{\beta}^{(s)}]$, $\boldsymbol{B} = [\boldsymbol{\Phi}, \boldsymbol{\beta}]$。

首先给出 Sparse Lag Group VARX-L 模型参数估计方法。由于组内稀疏不能产生可分离目标函数,因此传统的组 lasso 解决方法例如坐标下降法将不再适用,本文使用近邻梯度下降法求解此种结构。首先考虑时滞矩阵系数 $\boldsymbol{\Phi}^{(q)}$, $q = 1,2,\cdots,p$ 的子问题:

$$\min_{\boldsymbol{\Phi}^{(q)}} \frac{1}{2k} \| \boldsymbol{R}_{-q} - \boldsymbol{\Phi}^{(q)} \boldsymbol{Z}_q \|_F^2 + (1-\alpha)\lambda \| \boldsymbol{\Phi}^{(q)} \|_F + \alpha\lambda \| \boldsymbol{\Phi}^{(q)} \|_1 \tag{14}$$

其中 $\boldsymbol{R}_q = \boldsymbol{\Phi}^{(-q)} \boldsymbol{Z}_{-q} - \boldsymbol{Y} \in \mathbb{R}^{k \times T}$ 代表部分残差。将式(14)看做一个带有利普希茨梯度的可微函数与一个不可微函数之和。

首先在当前估计值 $\boldsymbol{\Phi}_0^{(q)}$ 处将二次项近似值线性化,d 代表步长,去除与 $\boldsymbol{\Phi}^{(q)}$ 独立的项,目标函数变为:

$$\begin{aligned} &\operatorname{argmin}_{\boldsymbol{\Phi}^{(q)}} \frac{1}{2k} \| \boldsymbol{R}_{-q} - \boldsymbol{\Phi}_0^{(q)} \boldsymbol{Z}_q \|_F^2 + \langle \boldsymbol{\Phi}^{(q)} - \boldsymbol{\Phi}_0^{(q)}, \\ &(\boldsymbol{\Phi}_0^{(q)} \boldsymbol{Z}_q - \boldsymbol{R}_{-q}) \boldsymbol{Z}_q^T \rangle + \frac{1}{2d} \| \boldsymbol{\Phi}^{(q)} - \\ &\boldsymbol{\Phi}_0^{(q)} \|_F^2 + P(\boldsymbol{\Phi}^{(q)}) = \operatorname{argmin}_{\boldsymbol{\Phi}^{(q)}} \frac{1}{2d} || \boldsymbol{\Phi}^{(q)} \\ &- (\boldsymbol{\Phi}_0^{(q)} - d(\boldsymbol{\Phi}_0^{(q)} \boldsymbol{Z}_q - \boldsymbol{R}_{-q}) Z_q^T) ||_F^2 + \\ &P(\boldsymbol{\Phi}^{(q)}) \end{aligned} \tag{15}$$

根据文献[19],应用 Nesterov 加速更新,在第 j 步时,有

$$\hat{\boldsymbol{\Phi}}_j^{(q)} \leftarrow \hat{\boldsymbol{\Phi}}_{j-1}^{(q)} + \frac{j}{j+3}(V(\boldsymbol{\Phi}^{(q)}) - \hat{\boldsymbol{\Phi}}_{j-1}^{(q)}) \tag{16}$$

收敛率为 $1/j^2$,其中更新函数用 $V(\boldsymbol{\Phi})$ 表示为

$$\begin{aligned} V(\boldsymbol{\Phi}^{(q)}) = &\left(1 - \frac{d(1-\alpha)\lambda}{\| ST(\boldsymbol{\Phi}_0^{(q)} - d(\boldsymbol{\Phi}_0^{(q)} \boldsymbol{Z}_q - \boldsymbol{R}_{-q}) \boldsymbol{Z}_q^T, d\alpha\lambda) \|_F}\right) \\ &+ ST(\boldsymbol{\Phi}_0^{(q)} - d(\boldsymbol{\Phi}_0^{(q)} \boldsymbol{Z}_q - \boldsymbol{R}_{-q}) \boldsymbol{Z}_q^T, d\alpha\lambda) \end{aligned} \tag{17}$$

根据利普希茨常量,我们引入一个不变的步长 H, H 必须满足

$$\| \nabla_X l(\boldsymbol{X}) - \nabla_Y l(\boldsymbol{Y}) \| \leqslant H \| \boldsymbol{X} - \boldsymbol{Y} \| \tag{18}$$

现考虑两个子矩阵 $\boldsymbol{A}^{(q)}$ 和 $\boldsymbol{C}^{(q)}$,有

$$\begin{aligned} &\nabla_{A^{(q)}} l(\boldsymbol{A}^{(q)}) = \boldsymbol{A}^{(q)} \boldsymbol{Z}_q \boldsymbol{Z}_q^T - \boldsymbol{R}_{-q} \boldsymbol{Z}_q^T \\ &\nabla_{C^{(q)}} l(\boldsymbol{C}^{(q)}) = \boldsymbol{C}^{(q)} \boldsymbol{Z}_q \boldsymbol{Z}_q^T - \boldsymbol{R}_{-q} \boldsymbol{Z}_q^T \\ &\Rightarrow \nabla_{A^{(q)}} l(\boldsymbol{A}^{(q)}) - \nabla_{C^{(q)}} l(\boldsymbol{C}^{(q)}) \\ &\quad = (\boldsymbol{A}^{(q)} - \boldsymbol{C}^{(q)}) \boldsymbol{Z}_q \boldsymbol{Z}_q^T \\ &\Rightarrow \| (\boldsymbol{A}^{(q)} - \boldsymbol{C}^{(q)}) \boldsymbol{Z}_q \boldsymbol{Z}_q^T \|_2 \\ &\quad \leqslant \| \boldsymbol{A}^{(q)} - \boldsymbol{C}^{(q)} \|_2 \| \boldsymbol{Z}_q \boldsymbol{Z}_q^T \|_2 \end{aligned} \tag{19}$$

因此可以得到结论,利普希茨常量 $\| \boldsymbol{Z}_q \boldsymbol{Z}_q^T \|_2 = \sqrt{\sigma_1(\boldsymbol{Z}_q)}$ 是 $\boldsymbol{Z}_q$ 的最大特征值的均方根,可以利用功率方法求得最大特征值,其最大特征向量可以作为本方法的热启动值,这样可以节省运算量。

求解该模型的过程如下:

对所有的组进行迭代,对于每一个组,有下述过程:

(a)通过条件 $|| (\boldsymbol{\Phi}^{(q)} Z_q - R_{-q}) Z_q^T ||_F \leqslant (1-\alpha)\lambda$,检查该组是否被选中。

(b)如果是被选中的,执行内部循环算法 1,反之,将组内所有值置为零。

(c)重复上述过程直到收敛。

内部循环算法 1 如下所示:

输入:当前估计值 $\boldsymbol{\Phi}_0$,内外生时间序列的组合 $\boldsymbol{Z}_q$,部分残差 $\boldsymbol{R}_{-q}$

输出:内生变量系数矩阵的最优化更新结果: $\boldsymbol{\Phi}^{j+1}$

步骤 1:求解 $\boldsymbol{Z}_q$ 最大特征值的倒数,即步长: h

$\leftarrow \frac{1}{\sigma_1(\boldsymbol{Z}_q)}$

步骤 2 :求解内生系数矩阵的第 j 个估计值 $\boldsymbol{\Phi}^j$: $\boldsymbol{\Phi}_0 \leftarrow \boldsymbol{\Phi}^1$,repeat, $j \leftarrow 1$, $\gamma^j \leftarrow \boldsymbol{\Phi}^j$

步骤 3:计算最优化更新式: $vec(\gamma^{(j+1)}) \leftarrow (1 - \frac{h(1-\alpha)\lambda}{\| ST(\boldsymbol{\Phi}^j - h(\boldsymbol{\Phi}^j\boldsymbol{Z}_q - \boldsymbol{R}_{-q})\boldsymbol{Z}_q^{\mathrm{T}}, h\alpha\lambda) \|_F})_+$

$ST(vec(\boldsymbol{\Phi}^j) - hvec(\frac{(\boldsymbol{\Phi}^j\boldsymbol{Z}_q - \boldsymbol{R}_{-q})\boldsymbol{Z}_q^{\mathrm{T}}}{k}), h\alpha\lambda)$

步骤 4:利用 Nesterov 加速更新: $\boldsymbol{\Phi}^{j+1} \leftarrow \gamma^{j+1} + \frac{j}{j+3}(\gamma^{j+1} - \gamma^j)$

然后给出 Sparse Own/Other Group VARX-L 模型参数估计方法。$\boldsymbol{\Phi}^{(q)}$ 在分组时考虑到“组件自己的滞后”和“其它组件的滞后”,必须将等式(4)转化成最小二乘问题。为了实现这一转化,我们做如下定义:

$$r_{-qq} = vec(\boldsymbol{R}_{-qq}), \boldsymbol{\Phi}_{qq} = vec(\boldsymbol{\Phi}_{on}^{(q)}), \boldsymbol{M}_{qq} = (\boldsymbol{Z}^{\mathrm{T}} \otimes \boldsymbol{I}_k)_{qq} \tag{20}$$

每一个“自己的滞后”组的子问题就可以描述为

$$\min_{\boldsymbol{\Phi}_{qq}} \frac{1}{2} \| \boldsymbol{M}_{qq}\boldsymbol{\Phi}_{qq} + \boldsymbol{r}_{-qq} \|_F^2 + \lambda \| \boldsymbol{\Phi}_{qq} \|_F = \min_{\boldsymbol{\Phi}_{qq}} \frac{1}{2}\boldsymbol{\Phi}_{qq}^{\mathrm{T}}\boldsymbol{M}_{qq}^{\mathrm{T}}\boldsymbol{M}_{qq}\boldsymbol{\Phi}_{qq} + \boldsymbol{r}_{-qq}^{\mathrm{T}}\boldsymbol{M}_{qq}\boldsymbol{\Phi}_{qq} + \lambda \| \boldsymbol{\Phi}_{qq} \|_F \tag{21}$$

子梯度求解如下:

$$\frac{\partial}{\partial \boldsymbol{\Phi}_{qq}} = \boldsymbol{M}_{qq}^{\mathrm{T}}\boldsymbol{M}_{qq}\boldsymbol{\Phi}_{qq} + \boldsymbol{M}_{qq}^{\mathrm{T}}\boldsymbol{r} + \lambda\omega(\boldsymbol{\Phi}_{qq}) \tag{22}$$

ω 定义如下:

$$\omega(s) \in \begin{cases} \left\{\frac{s}{\| s \|_F}\right\}, & s \neq 0 \\ \{u: \| u \|_F \leqslant 1\}, & s = 0 \end{cases} \tag{23}$$

然后可根据 Sparse Lag Group VARX-L 的算法求解模型数据。

2 实验数据说明及预处理

2.1 实验数据说明

实验中用到的所有数据都收集自天气后报官网,详见网址 http://www.tianqihoubao.com/,该网站的数据来源于当天的天气后报并且实时的记录,数据集中的变量包含北京、保定、廊坊和唐山 2015 年 1 月 1 日 ~2015 年 12 月 31 日各 6 项大气污染物浓度,其中这 6 项大气污染物分别是可吸入颗粒物 PM2.5、PM10 以及 NO_2、CO、SO_2、O_3。

2.2 实验数据预处理

非结构化 VARX-L 模型要求时间序列数据平稳,因此首先对时间序列进行单位根检验,本文采用增广迪基-福勒(augmented Dickey-Fuller, ADF)方法进行单位根检验,表 2 给出了 ADF 检验结果。

表 2 ADF 检验

变量名	最大滞后阶	检验值	临界值		
			1%	5%	10%
PM2.5	16	-2.717	-3.98	-3.42	-3.13
PM10	16	-3.426	-3.98	-3.42	-3.13
SO_2	16	-2.334	-3.98	-3.42	-3.13
NO_2	16	-2.226	-3.98	-3.42	-3.13
CO	16	-1.810	-3.98	-3.42	-3.13
O_3	16	-1.647	-3.98	-3.42	-3.13
DPM2.5	16	-7.634	-3.98	-3.42	-3.13
DPM10	16	-7.367	-3.98	-3.42	-3.13
DSO_2	16	-7.035	-3.98	-3.42	-3.13
DNO_2	16	-7.315	-3.98	-3.42	-3.13
DCO	16	-8.104	-3.98	-3.42	-3.13
DO_3	16	-7.687	-3.98	-3.42	-3.13

从表 2 的检验结果来看,在显著性水平为 1%、5% 和 10% 下,原序列不完全是平稳的,而它们的一阶差分都近似平稳,因此采用这 6 个变量的差分序列进行分析,分别记为 DPM2.5、DPM10、DSO_2、DNO_2、DCO 和 DO_3。后续实验中将采用一阶差分后的数据作为实验变量。

差分后的 6 元时间序列图如图 3 所示。

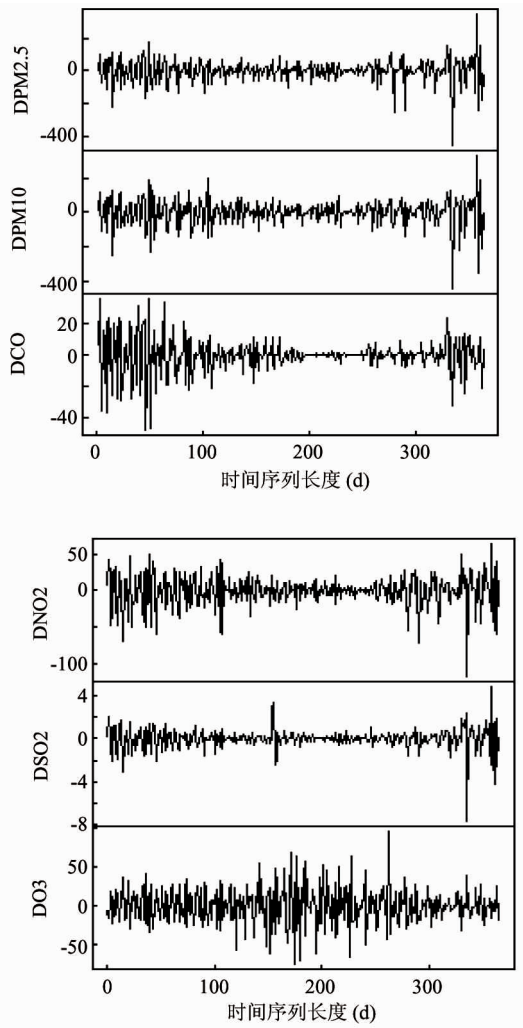

图 3 差分后的时间序列图

3 实验过程与结果分析

本部分共有两个实验,第一个实验对模型本身的预测性能进行研究,将 h 步向前预测均方误差(mean-squared h-step-ahead forecast error, $MSFE(T_1, T_2) = 1/k(T_2 - T_1)\sum_{i=1}^{k}\sum_{t=T_1}^{T_2-h-1}(\hat{\boldsymbol{y}}_{i,t+h} - \boldsymbol{y}_{i,t+h})^2$ 作为判定模型好坏的标准,MSFE 越小表明模型预测性能越好。第二个实验用该模型对未来一天的大气污染物做预测,并求解各种方法的归一化均方误差(normalized mean square error, $NMSE = \sum_{t=1}^{T}|\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t|^2 / \sum_{t=1}^{T}|\boldsymbol{y}_t|^2$)。

由于时间序列自身的相关性,传统的交叉验证已经不再适用,此处选择滚动交叉验证选择最优罚参数。将长度为 T 的时间序列分为近似相等大小的三部分,定义 $T1 = \lfloor T/3 \rfloor$, $T2 = \lfloor 2T/3 \rfloor$。如图 4 所示,1 到 $T1$ 时刻的数据用来初始化协变量矩阵 $\boldsymbol{Z}$;$T1$ 到 $T2$ 时刻的数据用来进行罚参数选择;$T2$ 到 T 的数据用来预测评估。为便于理解,此处定义预测间隔为 1。在时刻 $T1$,预测 $\hat{y}_{T_1+1}^{\lambda_i}$,根据公式:$MSFE(\lambda_i) = 1/(T_2 - T_1)\sum_{t=T_1}^{T_2-1}\|\hat{\boldsymbol{y}}_{t+1}^{\lambda_i} - \boldsymbol{y}_{t+1}\|_F^2$,计算相应的 MSFE。每顺序地增加一个观测值就做一次预测,选择最小的 MSFE 对应的 λ 值即为所求。

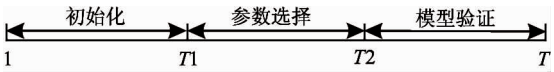

图 4 滚动交叉验证示意图

原始数据的时间分布是 2015 年 1 月 1 日 ~ 2015 年 12 月 31 日,总长度为 365,一阶差分后,时间序列长度变为 364,取 $T1 = 121$, $T2 = 242$,1 ~ $T1$ 的数据用作模型初始化,$T1$ ~ $T2$ 的数据用作罚参数选择,$T2$ ~ T 的数据用作模型验证。

实验 1:以北京的 6 项大气污染物为内生变量,以其周围城市唐山、廊坊和保定的各 6 项大气污染物作为外生变量,做 1 步和 2 步向前预测,求解 VARX-L 模型的 MSFE,实验中所有网格均取(30, 10)。当内生变量滞后阶为 2 阶($P = 2$),外生变量滞后阶为 1 阶($s = 1$)时,模型预测性能相对较好。为便于比较,将表中各 MSFE 值写成各种方法与 Conditional Mean 的比值形式。本实验中加入 VAR-L(不加入外生变量)模型作为对比,实验结果如表 3 所示。

从表 3 中,可以得到以下结论:(1)在两种预测范围内,VARX-L 框架的预测效果超过了 BIC、样本均值和随机游走等基准方法,其中,在 $h = 1$ 时,Sparse Own/Other Group VARX-L 的预测效果最佳,在 $h = 2$ 时,Sparse Own/Other Group VAR-L 的预测效果最佳,说明在天气情况的预测中,将组件自己的滞后与其他组件的滞后的关系进行描述可以提高模型的预测能力。(2)在预测步长为 1 时,VARX-L 模型的预测误差均比 VAR-L 的预测误差小,这意味着

表 3　1 步和 2 步向前均方预测误差 MSFE

Model/VARX-L Penalty Structure	$h=1$	$h=2$
Basic	0.859	0.954
Lag Group	0.854	0.976
Own/Other Group	0.856	0.954
Sparse Lag Group	0.858	0.960
Sparse Own/Other Group	0.847	0.965
Endogenous-First	0.854	0.953
Conditional Mean	1.000	1.000
VARX with lags selected by AIC	0.757	1.007
VARX with lags selected by BIC	1.000	1.000
Random Walk	2.205	2.504
Model/VAR-L Penalty Structure	$h=1$	$h=2$
Basic	0.916	0.949
Lag Group	0.900	0.948
Own/Other Group	0.906	0.943
Sparse Lag Group	0.905	0.948
Sparse Own/Other Group	0.912	0.942
Conditional Mean	1.000	1.000
VAR with lags selected by AIC	0.892	1.003
VAR with lags selected by BIC	1.000	1.000
Random Walk	2.205	2.504

本文提到的几种方法可以有效地从外生变量中提取有用信息,同时也代表唐山、廊坊、保定三个城市的空气质量对北京有直接影响。在多步预测中,外生变量经过更复杂的运算导致部分信息不完整,对预测结果产生负面影响,其结果随着预测时间的加长而使预测误差逐渐增大。(3)在两种预测范围内,BIC、随机游走和样本均值这三种基准方法在 VAR-L 和 VARX-L 两种模型中的预测效果相同。当 $h=1$ 时,在四种基本准则里,AIC 的预测效果超过了其他三种准则,但是随着预测范围增大,AIC 的预测效果下降,这是由于在滞后阶数增高时,AIC 所施加的罚结构太弱而有可能导致过参数化。

$h=1$ 时 Own/Other Group VARX-L 和 Sparse Own/Other Group VARX-L 对应的系数矩阵稀疏图如图 5、图 6 所示:

图中颜色越深表示系数值越大,即时间序列对彼此之间的相互作用越强。首先从两种模型的系数矩阵稀疏图可以很直观的看出在主对角线上的值明显比同组内其他的数大,这说明序列自身的时滞包含的信息要比其他时间序列的时滞包含的信息多,即主效应比较强。

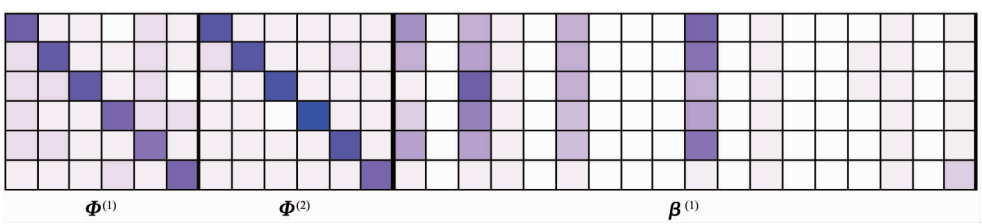

图 5　Own/Other Group VARX-L$_{(6,18)}$(2,1) 的系数矩阵稀疏图

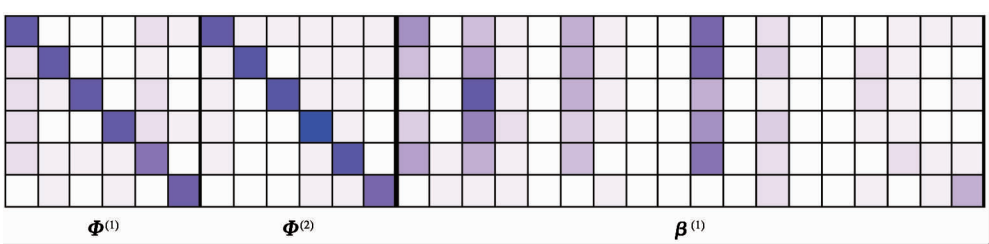

图 6　Sparse Own/Other Group VARX-L$_{(6,18)}$(2,1) 的系数矩阵稀疏图

从图中可以获悉各模型进行特征选择的方式和结果：(1)Own/Other Group VARX-L 是按照组的方式进行分类，在一个组之内的系数要么全为零，要么全非零。从图 5 来看，所有内生变量即滞后一天和滞后两天的北京的 6 项大气污染物全部被选中，外生变量中廊坊的 PM10、CO，保定的 PM2.5、PM10、SO_2、CO 和唐山的 PM2.5、PM10、CO 没有被选中。从表 4 的结论中就已得知廊坊、保定、唐山的空气质量对北京的空气质量有直接影响，从本例中可以进一步推断，这三个城市中对北京空气质量影响较为严重的指标依次为保定的 NO_2、廊坊的 SO_2、O_3 和 PM2.5，没有被选中的变量代表对北京的空气质量影响小一些。(2)Sparse Own/Other Group VARX-L 在组稀疏的基础上产生了组内稀疏，例如在图 6 的内生滞后系数矩阵中，一个组的内部出现了很多系数为 0 的情况。外生变量中既有组稀疏也有组内稀疏，其中廊坊的 CO、保定的 SO_2、CO 和唐山的 PM2.5、PM10 整组都没有被选中。

综合来看，如果在一个组内部只有少数几个数非零，Sparse Own/Other Group VARX-L 的优势在于：在组稀疏的基础上进一步缩减模型参数，更好的解决待估参数过多的问题。

实验 2：本实验采用实验 1 的数据对未来一天的各项大气污染物浓度进行预测（由于外生变量的系数矩阵无法被估计，所以 VARX-L 框架只能做一步向前预测），并且加入分层向量自回归模型(HVAR)和向量自回归模型(VAR)作对比，求解各种方法的归一化均方误差，NMSE 越小，代表预测结果越好，其结果如表 4 和表 5 所示。

表 4 2016 年 1 月 1 日北京市各污染物浓度预测结果

方法名称 \ 污染物名称		PM2.5	PM10	CO	NO_2	SO_2	O_3
	实际值	167	185	32	102	3.37	7
VARX-L	Basic	128.717	175.180	25.373	89.001	2.704	10.914
	Lag Group	129.609	178.909	25.503	90.163	2.718	11.301
	Own/Other Group	130.217	173.205	24.419	89.027	2.661	11.972
	Sparse Lag Group	130.510	180.468	25.667	90.877	2.736	11.312
	Sparse Own/Other Group	134.271	179.187	25.132	90.778	2.706	11.946
	Endogenous-First	130.555	181.401	25.064	90.642	2.700	11.312
VAR-L	Basic	108.322	154.995	22.292	84.676	2.429	11.505
	Lag Group	109.376	161.023	23.072	85.890	2.431	11.963
	Own/Other Group	109.058	157.684	22.665	85.345	2.422	11.531
	Sparse Lag Group	108.399	158.161	22.924	85.260	2.429	11.487
	Sparse Own/Other Group	109.526	154.694	22.287	84.698	2.418	11.326
HVAR	Componentwise	107.034	156.693	22.675	83.911	2.379	12.404
	Own-other	107.837	154.391	22.167	83.414	2.347	12.859
	Elementwise	108.438	155.248	22.367	83.322	2.403	12.257
VAR		115.894	174.984	23.818	87.602	2.519	12.975

表 5　2016 年 1 月 1 日北京市各大气污染物浓度预测归一化均方误差

方法名称 \ 污染物名称		PM2.5	PM10	CO	NO_2	SO_2	O_3
	实际值	167	185	32	102	3.37	7
VARX-L	Basic	0.052550	0.002818	0.042889	0.016242	0.039034	0.312661
	Lag Group	0.050132	0.001084	0.041225	0.013468	0.037378	0.377504
	Own/Other Group	0.048515	0.004065	0.056132	0.016177	0.044320	0.504593
	Sparse Lag Group	0.047743	0.000600	0.039166	0.011891	0.035372	0.379398
	Sparse Own/Other Group	0.038408	0.000998	0.046062	0.012105	0.038876	0.499279
	Endogenous-First	0.047627	0.000378	0.046986	0.012310	0.039564	0.379371
VAR-L	Basic	0.123457	0.026306	0.092037	0.028846	0.078042	0.414217
	Lag Group	0.119063	0.016797	0.077848	0.024945	0.077599	0.502681
	Own/Other Group	0.120381	0.021802	0.085103	0.026663	0.079120	0.418956
	Sparse Lag Group	0.123134	0.021047	0.080442	0.026936	0.078024	0.410960
	Sparse Own/Other Group	0.118443	0.026835	0.092125	0.028773	0.079854	0.381980
HVAR	Componentwise	0.128937	0.023413	0.084911	0.031452	0.086396	0.596062
	Own-other	0.125507	0.027375	0.094418	0.033201	0.092131	0.700499
	Elementwise	0.122969	0.025864	0.090615	0.033534	0.082368	0.564053
VAR		0.093651	0.002931	0.065383	0.019926	0.063735	0.728488

从以上两表可以看出,VARX-L 模型的预测结果要优于 VAR-L 模型、分层向量自回归模型(HVAR)和向量自回归模型(VAR)。其中,Endogenous-First VARX-L 的预测效果最佳,其对 PM10 的预测值为 181.401,而实际值为 185,归一化均方误差为 0.000378,预测值与实际值非常接近,达到了比较理想的预测效果,说明在此方案中考虑内外生变量的优先性问题可以提高预测精度。雾霾的首要污染物 PM2.5 的最佳预测值为 134.271,实际值为 167,归一化均方误差为 0.0384080,最佳预测值由 Sparse Own/Other Group VARX-L 得到,说明在组稀疏的基础上加入组内稀疏并且考虑序列自身的滞后与其它序列的滞后之间的关系可以改善预测性能。以上结果说明 VARX-L 模型用在空气质量的预测中是可行的方案选择。

4　结　论

本文将正则化结构与 VARX 模型结合所得到的稀疏组 lasso 罚参数模型,不仅能够解决传统向量自回归模型所面临的过参数化问题,而且算法简单有效。稀疏组 lasso 罚模型能广泛应用于包含各种动态结构的背景中,本文将此模型用于大气污染物的预测中,达到了较为理想的结果。从模型本身来看,VARX-L 框架的预测效果要优于 AIC、BIC、样本均值和随机游走等基准模型以及不含有外生变量的 VAR-L 模型。在对未来 1 天的大气污染物浓度做一步预测时,稀疏组 lasso 罚模型的预测结果也要优于传统的向量自回归模型(VAR),分层向量自回归模型(HVAR)。此外,从模型参数进一步分析,可以得知京津冀不同城市对北京空气质量的影响程度不同,这一结论对京津冀大气污染协同治理具有重要意义。VARX-L 框架将稀疏组 lasso 罚推广至时间依赖问题,采用凸优化算法计算效率高,这适合变量数等于或超过序列长度的时间序列建模问题。本文工作都是基于稀疏组 lasso 向量自回归模型在预测方面的应用,未来将侧重研究该模型在变量选择方

面的应用。

参考文献

[1] Pui D Y H, Chen S C, Zuo Z. PM2.5 in China: measurements, sources, visibility and health effects, and mitigation. *Particuology*, 2014, 13(2):1-26

[2] Marc M, Tobiszewski M, Zabiegała B, et al. Current air quality analytics and monitoring: a review. *Analytica Chimica Acta*, 2015, 853(1):116-126

[3] 王跃思,张军科,王莉莉. 京津冀区域大气霾污染研究意义、现状及展望. 地球科学进展,2014,29(3):388-396

[4] 缪育聪,郑亦佳,王姝等. 京津冀地区霾成因机制研究进展与展望. 气候与环境研究, 2015, 20(3):356-368

[5] 安树伟,郁鹏,母爱英. 基于污染物排放的京津冀大气污染治理研究. 城市与环境研究, 2016,2:17-30

[6] 丁峰, 张阳,李鱼. 京津冀大气污染现状及防治方向探讨. 环境保护, 2014, 42(21):55-57

[7] 滕丽,卢君. 珠江三角洲城市空气质量的区域影响分析. 云南地理环境研究,2015, 27(6):1-7

[8] 刘金培,汪官镇,陈华友. 基于 VAR 模型的 PM2.5 与其影响因素动态关系研究. 干旱区资源与环境, 2016, 30(5):78-84

[9] 刘华军,刘传明. 京津冀地区城市间大气污染的非线性传导及其联动网络. 中国人口科学,2016,30(2):84-95,128

[10] Davis R A, Zang P, Zheng T. Sparse Vector Autoregressive Modeling. *Journal of Computational and Graphical Statistics*, 2016, 25(4):1077-1096

[11] Nicholson W, Matteson D, Bien J. VARX-L: structured regularization for large vector autoregressions with exogenous variables. https://arxiv. org/abs/1508. 07497: Arxiv, 2015

[12] Song S. Bickel P J. Large vector auto regressions. https://arxiv. org/abs/1106. 3915: Arxiv, 2011

[13] Simon N, Friedman J, Hastie T. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 2013, 22(2):231-245

Sparse group lasso VARX for prediction of atmospheric pollutants: a case study of Beijing-Tianjin-Hebei

Wang Jinjia, Sun Mengran, Hao Zhi

(School of Information Science and Engineering, Yanshan University, Qinhuangdao 066004)

Abstract

The prediction of atmosphere pollutants was studied. Aiming at the problem of over parameterization of the traditional vector autoregressive (VAR) model, a sparse group lasso penalized VAR model for atmosphere pollutant prediction was proposed. The model parameters are solved by the proximal gradient descent method. In order to prove the validity of the model, this model was applied to prediction of 6 indexes of air quality of January 1, 2016 in Beijing region by using the Beijing-Tianjin-Hebei air pollutant data of 2015. The experimental results show that the normalized mean square error of PM2.5 of the model based on the sparse group lasso penalty is about 3.8%, and its prediction accuracy is higher than that of the VAR model, the large VAR based on various sparse structures (VAR-L), and the hierarchical vector autoregression (HVAR). In addition, the impacts of different cities in Beijing-Tianjin-Hebei on the air quality of the Beijing region can be explained by the parameters of sparse lag group penalized VARX-L and Sparse Own/Other Group penalized VARX-L model. The application of the combination of the convex optimization and the VAR model to the predication of atmospheric pollutant concentration is of great significance to the synergistic control of air pollution in Beijing-Tianjin-Hebei.

Key words: vector autoregressive (VAR) model, sparse group lasso, proximal gradient descent method, convex optimization, air pollution