北京大学

硕士研究生学位论文

题目：**基于贝叶斯网络的结构学习**

——北京空气污染的主要影响因素实证分析

姓　　名：	张 晗 雨
学　　号：	1701214140
院　　系：	国家发展研究院
专　　业：	西方经济学
研究方向：	经济计量学
导师姓名：	胡大源 马京晶

二〇二〇年四月

版权声明

摘要

本文对图模型进行理论综述探讨，从概率视角下提供一套变量选择和模型设定的规范化流程，遵循“输入-加入先验条件-结构学习-参数学习-调整不合理变量关系-验证变量关系-输出”的步骤重复循环，将专家知识作为先验结构设计，加入图模型的不确定性学习中，提供对引向最终模型设定路径的优化，从而为复杂世界问题提供可供解释的建模方式。

本文将贝叶斯网络用于 PM2.5 影响因素研究，通过先验设计路径的调整来寻找最精简且最合适的模型。结合自然科学领域相关研究，通过黑白名单设计，将环境数据变量中确定性关系加入初始结构设计，再通过算法学习挖掘不确定性关系，得到最终的图模型。在与 Lasso 结果比较中，结合先验知识和结构信息的贝叶斯网络，能够有效进行变量选择，降低模型预测误差，并能提供可视化解释。

关键词：变量选择，人工智能，贝叶斯网络，条件概率，先验设计，PM2.5

Research on Structure Learning Based on Bayesian Network

——Empirical analysis of the main influencing factors of Air Pollution in Beijing

Hanyu Zhang
Directed by Dayuan Hu, Jingjing Ma

ABSTRACT

This paper conducts a theoretical review of the graphic models and provides a set of standardized processes for variable selection and model setting from a probabilistic perspective. Following the repeated steps of "input—setting prior condition—structure learning—parameter learning—adjusting unreasonable variable relationships—verifying variable relationships—Output ", this paper uses the expert knowledge as a priori structural design to add the uncertainty learning into the graphical model, thus provides the optimization of the path leading to the final model and an interpretable modeling method for complex world problems.

In this paper, Bayesian networks are used to explore the influencing factors of PM2.5, and the most simplified and most suitable model is found through the adjustment of the prior design path. Combined with related research in the field of natural sciences, through the design of black and white lists, the deterministic relationships in environmental data variables are added to the initial structural design, and then the uncertain relationships are mined through algorithm learning to obtain the final graph model. In comparison with Lasso results, the Bayesian network combining prior knowledge and structural information can effectively select variables, reduce model prediction errors, and provide visual interpretation.

KEY WORDS: Variable Selection, Artificial Intelligence, Bayesian Network, Conditional Probability, Chain Rule, PM2.5

目录

第一章 引言

1.1 研究背景和目的

通常在熟悉的领域，人们会结合专业领域知识来解决特定问题，但随着海量数据的获得和多学科交叉研究的问题，越来越多的问题已经超出人们经验知识。当跨出熟悉领域，变量选择和模型设定就遇到了新的挑战。过多无用的变量不仅对相互作用关系造成干扰，降低模型预测精度，而且会造成计算的高纬灾难。尽管有一些方法在变量选择上做出尝试，如考察自变量的相关性，加惩罚项的 LASSO，但如何选择合适变量并未形成统一清晰理论。而模型设定也会随复杂度增加带来过度拟合、不可解释和长期预测性差等问题。因此，如何从海量数据中选取出重要变量，是经济学中很重要的问题。

自从 16 年 3 月 AlphaGo 以 4:1 大胜人类顶级棋手李世石之后，以神经网络为代表的机器学习算法再一次引发关注，掀起了新一轮人工智能的热潮。然而尽管以神经网络为代表的人工智能算法在识别和预测上很擅长，但是却无法理解模式和参数的含义，不能解释和理解模式背后的原因。因此，快速发展的机器学习能否给经济学中变量选择与现实验证提供新的启发？人工智能算法可否突破黑箱进行解释？是值得关注的问题。作为人工智能背后的一块，概率图模型结合图和概率论的知识，允许将先验知识整合到模型设计，再将数据中变量间的概率分布对应到网络图中，更形象紧凑地展示各个变量间的依赖关系和条件独立关系，形成一条推理网络。借助概率图模型不同于其他机器学习的特性，从概率和先验设计的角度出发，是否为经济学中变量选择提供了新视角？在从数据中学习图结构的过程，结合交互式和图可视化的特性，是否能为很多存在依赖关系的真实世界任务提供了可解释的建模方式？

1.2 文献综述

1.2.1 经典计量经济学中关于变量选择和模型设定的讨论

随着现代技术发展，诸多领域涌现海量数据，对该选择哪些变量，如何进行模型设定提出了新的挑战。Cuyon 和 Elisseeff（2003）认为变量选择是从大量变量中选取与研究问题所有相关变量的技术，其目的是为了对模型提供更好的解释，给出更有效的估计，并改善预测效果[1]。现代计量经济学更多是在给定模型框架下的数据生成过程中进行估计和推断，却较少地涉及到变量选择和模型构建的问题。而 Kennedy（2008）解释这主要是由于这个设定过程极为不容易，且极具创造力，不存在找到正确设定的公认的最好方式[2]。但是在实际研究中，仍有一些方法可作为指导。

传统的变量选择方法有：平均经济回归（AER）、检验-检验-检验（TTT 法）、AIC 信息准则等。AER 由一个简单的模型并“向上检验”至一种特定更一般的模型。首先从一个认为是正确的设定开始，使用主要的数据来确定少数几个未知参数大小的顺序；如果不能解决问题，再设定新的检验，借助统计量（如正确的符号、高的 R2、系数非零的显著 t 值）的显著值进行判断。TTT 是一个从一般模型“向下检验”到特殊模型的过程。首先从一个更一般化的初始设定开始，进行不同约束检验来进行简化；然后遵循诊断检验，寻找模型被错误设定的迹象；最后模型不断经过设定和检验，得到最终研究结论。但是以上两种方法都因为不能提供对引向最终模型设定路径的充分描述而受到批评。Kennedy（2008）认为在搜寻高 R^2 值和高的 t 统计值的基础上采用的设定是“数据挖掘”的变体，只适用于特定数据集特性设定，而对于其他数据生成过程所给出的信息存在误导。

除了借助回归方法，Hernan 和 Robins（2020）认为现有方法主要包含两类，一是选择可用变量的子集，另一是其替代方法的系数收缩。子集选择的方法，是先用所有可能变量组合来估计模型，最后根据预先指定的标准确定最优变量组合。但是，这种方法对于大量可用变量在计算上变得不可行，有效改进方法就是分为向前、向后和双向筛选的逐步回归（stepwise）。其中，向前逐步回归的基本思路是：首先选定一个标准，然后按自变量的贡献依次进入模型，每选入一个变量进入模型，则重新计算模型外各个自变量的贡献，直到模型外变量均达不到入选标准，没有自变量可被引入为止。但如果一开始变量太少，可能会缺失主要变量，得到有偏的结果。向后逐步回归则先将变量全部引入，再逐步缩减变量，这种方法则有可能会带来多重共线性的问题。因此，逐步回归方法由于其局限性在计量经济学应用中备受争议，正如 Learner（2007）所说：“我们并不依赖于逐步回归或任何自动统计模式识别方法来解释数据，一方面因为目前根本无法将关键的背景信息注入这些方程中，另一方面也因为理解背景是理解混杂实验数据的关键。”然而，近几十年，随着网络技术和海量数据迅速发展，大量缺乏理论指导的新问题涌现，研究者无法利用已有知识对变量关系进行指导，作为数据挖掘主要方法之一的逐步回归也在现实问题中有取得了应用。借助其思路的可用之处，也涌现了一些新的变量选择方法及其衍生算法，我们有必要重新审视逐步回归方法在解决实际问题时的可用之处及其局限性（Efron，2016）。另一类方法是收缩，通过添加惩罚项来修改估计方法，使得模型参数估计值比没有惩罚项的估计更接近零。如 Tibshirani（1996）提出通过构建惩罚项限制系数l_1范数，转化为有约束的线性回归问题，从而将模型系数拉向零，减少系数之间的差异并防止过拟合。与其类似的是限制系数向量的二次l_2范数的岭回归，并在这种收缩方法基础上衍生出来一系列其他升级版本，如 Efron（2004）的最小角回归（LAR）等[3]。LASSO 允许某些参数值精确，可同时进行变量选择和参数估计，由于其在预测方面通常胜于逐步选择，已成为首选回归模型的变量选择方法[4]。

1.2.2 人工智能、机器学习与概率图的讨论

尽管“人工智能”(AI)已成为一个广泛讨论的热点话题，但对其定义还没有达成普遍共识。人工智能是由 McCarthy 于 1956 年在 Dartmouth 学会上首次正式提出，重点是让机器能够处理人脑所能处理的问题。Nilson 认为人工智能是一门研究怎样表示、获得并使用知识的学科。Winston 认为人工智能就是使计算机去做只有人才能做的工作。这些说法均反映人工智能学科的基本思想，即让机器学习人的思考方式和行为，帮助人进行学习和推理的延伸。

人工智能的发展几经沉浮。自 1956 年人工智能概念提出到 20 世纪 60 年代，机器定理证明、西洋跳棋程序、国际象棋计算机程序，图灵机理论模型等主要研究成果，掀起第一个蓬勃发展时期；随后在 20 世纪 60 年代，接二连三的失败和期望落空之后，进入短暂的低谷。20 世纪 70 年代出现能模拟人类的知识和经验解决特定领域问题的专家系统，实现了人工智能从理论研究走向实际应用，在医疗、化学、地质等领域取得成功，引发了第二个高潮。然而 20 世纪 80 年代中，由于其应用领域狭窄、知识获取困难、缺乏分布式功能等局限性逐渐暴露出来，再次进入低迷期。新一轮热潮是计算机视觉、机器学习，自然语言处理、语音识别等技术核心在近几十年的积累与突破，并随着计算能力的提高、和海量网络数据的发展，在诸如图像分类、语音识别、知识问答、人机对弈、无人驾驶等多个领域实现了的应用技术突破。[5] [6]然而尽管人工智能算法擅长识别和预测，但是却无法对多层参数及模式背后的关系提供解释。如果只满足于应用发展和算法改进，在原理上没有突破，则无法打开技术黑箱提供解释，新一轮的热潮仍可能存在像 70 年代专家系统而不能持久。[7]

为了进一步探究人工智能的可解释问题，仍有一批研究者不满足于算法与应用现状，而专注于理论研究，从不同角度在探究人工智能背后的原理本质[8]。Larrañaga 和 Moral（2011）认为比起建立程序来解决特定问题，从长远来看，在现有的科学知识的基础上，如概率和统计，发展理论是更可行的方法[9]。概率图的研究发展是很重要的一部分，它包括所有使用图的语言来表示的模型，以及使用概率作为不确定性表示的复杂问题解决方式，其中最重要的是贝叶斯网络模型，用有向无环图来表示变量间的联合概率分布。Pearl（1988）率先将图模型引入人工智能[10]，其研究认为贝叶斯公式可作为大型，多假设，推理系统中基本信念修改规则，并建议将其作为信念维护和不精确推理的更复杂模型的标准出发点[11]。Michael Jordan（2008）也进一步指出了机器学习与统计学之间的联系，提出的 ELBO(Evidence Lower Bound)，重建了变分贝叶斯的基础框架，解决了概率图模型的计算难题[12]。随着贝叶斯网络的完整知识体系建立（即包含论证和解释基础知识，提出基本的推理算法，用数学推理和示例说明如何将概率用作简单，非单调，连贯且合理的推理系统的基础，并解释了如何使用该模型解决诊断，预测，计划，融合，决

策等问题），在推理算法、推理模型、结构学习、参数学习等诸多方面也有一系列更精细的发展。

作为人工智能的一部分，图模型结合概率理论和图论，可以用来处理不确定性和复杂性，尤其是在机器学习算法的设计和分析起到重要的作用。从概率的角度，概率图模型具有通用机器学习算法所不具备的特性，首先，机器学习算法主要针对单个输出变量的预测，例如二进制结果是正类或负类。而当我们需要尝试对结构化对象进行预测时，例如，在语音识别或自然语言处理中进行序列标记、在图像识别中进行图像分割形成的像素网格等场景下，概率图模型能够利用多个预测变量之间的相关性，显著提高预测性能。其次，概率图模型允许我们将先验知识整合到模型中，这是许多其他算法所不具备的。再者，相比学习特定变量的映射的传统算法，单个概率图模型的学习，也可被多种不同方式使用以回答不同类型的查询。最后，概率图模特别适用于知识发现，因为其结构形式相比其他算法更为直观简洁。已有研究显示，以变量间相互作用为前提、通过结构学习、参数估计以及概率推理来获取条件概率分布信息的贝叶斯网络已广泛用于图像识别、风险分析、可靠性分析、医疗诊断、基因工程等[13][14]，在不确定性概率分析具有很大优势，并能提供一套可解释性的建模方式，然而，利用贝叶斯网络对环境数据的分析和研究却为数不多。

1.3 研究方法与内容

本文拟对图模型进行理论模型探讨，并将其用于 PM2.5 影响因素研究，旨在从概率视角探讨人工智能热潮下背后的原理，提供一套变量选择和模型设定的规范化流程，并将人类先验经验加入图模型的不确定性结构学习中，为交互式学习提供有规则的干预，从而为复杂世界问题提供可供解释的建模方式。

本文其他部分结构安排如下：第二部分是理论与方法。首先阐释了图模型整体框架，包括模型，概率和图形化的关系。其次围绕贝叶斯网络的表示，从贝叶斯网络图上的符号体系、推理模式、变量关系，推理模型的梳理中回答了贝叶斯图是什么；从因子化与独立性的对偶关系的梳理中回答了为什么能从数据分布到图。然后介绍了贝叶斯网络的学习。主要分为参数学习和结构学习，并重点介绍了贝叶斯估计和基于评分搜索的学习。最后提出“输入-结构学习-参数学习-调整变量关系-验证变量关系-输出”来进行变量选择和模型设定的研究框架，从而实现将专家知识与数据学习相结合的交互。第三部分是 PM2.5 影响因素实证分析。首先介绍本文的数据，样本，变量及描述性统计结果；其次使用 R 对 2016 年 PM2.5 数据进行贝叶斯网络初步学习，并进一步讨论在交互式下，贝叶斯网络模型调整与验证；接下来探讨是否存在先验设计的最优路径，可用于其他年份的样本学习的优化？最后将贝叶斯网络与 lasso 进行模型比较，并进一步延伸讨论在其他地区样本的实用性条件。第四部分是结论与讨论。

第二章 理论模型

2.1 图模型总览

在定义什么是概率图模型（PGMs）之前先弄清楚它可以解决什么问题。PGMs 首次进入计算机科学和人工智能领域的应用，就是作为医学诊断。医生在问诊时候可以掌握大量病人的信息，如诱发因素、症状、各种试验结果，根据这些信息，她需要找出病人可能患有的疾病。PGMs 的另一个应用是图像分割。 一张图片有成千上万个像素点，我们想做的就是计算出每个像素对应的是什么，草，天空，牛或者马。这些问题的共同点在于，研究问题包含大量变量，且其正确答案都有很大的不确定性，概率图模型就是一个处理这种应用的框架[15]。

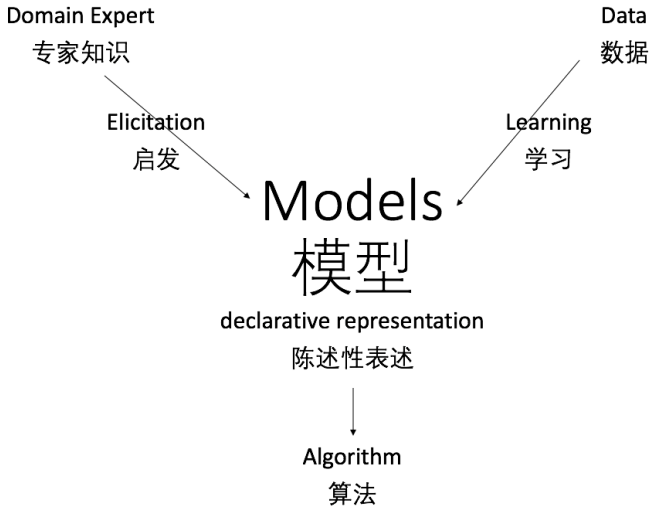

图 2.1:概率图模型框架示意图

2.1.1 模型

模型陈述了我们对世界的理解。 这是计算机内部的一种表示，它捕捉到我们对这些变量是什么以及它们如何相互作用的理解。模型和算法是相互独立的，因为同样的模型表述可以用于不同的情境中的一种算法或多种算法，如能回答任一问题的一种算法 ，或者是回答不同种类问题的其他算法，或者以更有效的方式回答同一个问题的其他算法，或者在准确性和复杂性之间做出不同权衡的其他算法。 拥有独立模型的另一个好处是，我们可以分离开模型的构造与算法推理。因此，如何从现实世界中抽象出模型？一是专家知识，我们可以从人类专家那里，通过知识经验直接抽象出重要关系；二是数据技术导向，从历史数据中，通过统计学机器学习模型；三是两者结合。

2.1.2 概率

概率的出现是因为我们需要设计模型来帮助我们处理大量的不确定性。不确定性有很多形式，也有很多不同的原因。一是我们对于世界存在的状态只有部分的了解，例如医生不对每一种症状或每一项测试结果进行检测，那她对病人的疾病肯定是不确定的。二是现实生活中存在太多的干扰因素和观测，即使我们观察到一些特定的东西，比如血压，这些观察结果经常受到大量噪音的影响；三是归于模型限制，很多现象都没有被我们的模型所包含。最后，现实事件本质就是随机的。

概率论提供了一种处理不确定性的框架。首先，概率模型用清晰明确的符号体系提供了陈述性表示，即一个概率分布对应了我们对世界可能处于不同状态的不确定性。其次，它还为我们提供了强大的推理模式，如在不确定的情况下如何决策。最后，由于概率论和统计学之间错综复杂的联系， 我们可以从统计学习中引入一系列强大的学习方法， 从而能够从历史数据中有效地学习这些模型，避免了需要人工去指定模型的每个方面。

2.1.3 图形化

图形化是计算机科学的词汇，因此概率图模型是从统计学中概率论到计算机科学中图形化的综合，而基本想法是利用计算机科学，特别是图形化，来表示包含大量变量的复杂系统。

我们需要把世界看成是由一组随机变量表示的，X_1到X_n，每个随机变量都代表了世界的某个方面，我们的目标就是通过这些随机变量的概率分布来获取世界可能状态的不确定性，即根据它们的概率分布得到随机变量集合所有可能赋值的联合分布。由此，即使在最简单的情况下，每个随机变量都是二进制值，n个二元值随机变量，就对应2^n个赋值。对于一个变量对应m个状态，处理对象本质上是指数级庞大的。因此，唯一的方法就是利用编码的数据结构，利用计算机科学的思想，利用结构和分布以一种有效的方式来表示和操作。

Koller（2009）给出一个简单的学生推荐信质量的例子[16]。课程难度D和学生智商I同时影响考试成绩G，SAT 成绩S仅取决于学生智商I，推荐信的质量L仅依赖于考试成绩G，该问题中五个随机变量，除了考试成绩可以取三个值，其他均为二值变量，总的来说，联合概率分布的取值总数为$2 \times 2 \times 2 \times 4 \times 3 = 48$。当得知联合概率分布时，就可以根据学生的情况对推荐信质量进行查询和判断。那么，这 48 个取值的概率分布利用图结构进行表达，如图所示，即为贝叶斯网络结构。贝叶斯网络是概率图形模型的两个主要类别之一，它使用有向图作为内在表示，图中的节点表示随机变量，图中的边代表了这些随机变量之间的概率联系。

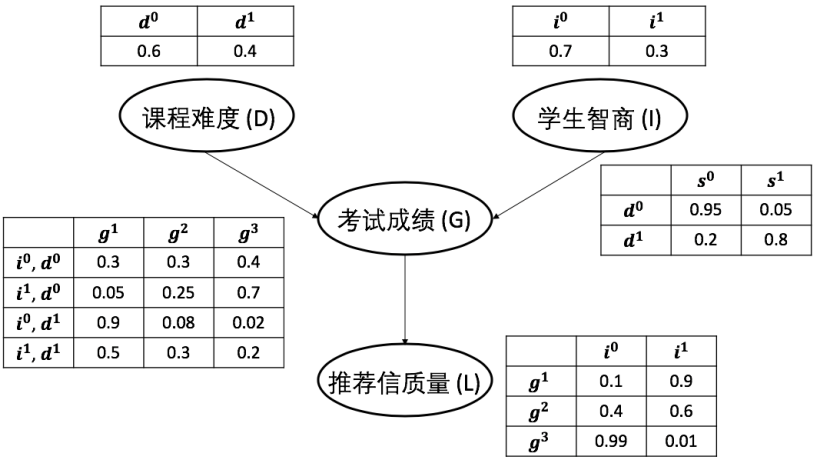

图 2.2：学生示例

因此，图形表示为我们提供了一个直观而紧凑的数据结构来捕捉这些高维概率分布。与此同时，它利用图形结构的通用算法，为我们提供了一套高效推理的方法。并且，由于图形结构编码参数少，使我们可以从专家知识中得到启发，也可以从数据中自动学习。

2.2 贝叶斯网络表示

2.2.1 贝叶斯网络基础

2.2.1.1 概率分布、因子化与链式法则

对贝叶斯网络的理解建立在概率分布，因子，链式法则的基础上。贝叶斯网络是一个有向无环图，图中的节点用随机变量$X_1 \cdots X_n$表示，并且每一个X_i都有对应的给定其父节点条件概率分布$P(X_i \mid Par_G(X_i))$。例如，基于课程难度D和学生智商I的考试成绩变量G组合，对应的X_i就是考试成绩，相应的父节点就是D和I。因此，贝叶斯网络就可以看作，通过链式法则（即将所有随机变量的条件概率相乘），计算得到的联合概率分布。贝叶斯图本质是表达了一个联合概率分布：一个概率分布 P 是基于图 G 的因子化，即这个图 G 节点是随机变量$X_1 \cdots X_n$，且这个概率分布能够通过链式法则，用这个概率图 G 来表示的时候。

2.2.1.2 变量之间基本依赖关系和结构

变量之间的相互关系，可以形象的在贝叶斯图上表示为概率从一个节点到另一个节点的流动。假设有一系列观察结果，用Z来表示， X与 Y的概率影响关系可分为以下情况：

如果影响通过X和Y直接相连，不管是因果还是逆因果，任何其中一个变化就会影响到另一个的变化。那Z的信息不会对X和Y产生任何影响。例如，$I \rightarrow G$，无论推荐信结果如何，学生的智商会影响到学生课程成绩。

如果影响通过中间变量W，X和Y不直接相连，那X通过W对Y的影响，可根据中间变量能否被观测，以及变量关系的结构，可分为图示几种结构。

因果迹，证据迹，共同的原因三种结构的影响情况相似，当W不是我们的已知事实，无法被观测时，影响可以被传递；当W属于已知事实，能够被观测时，影响被阻断。例如，$D \rightarrow G \rightarrow L$，当没有观测到学生成绩的时候，可以推测课程难度是会影响到推荐信的好坏；但是如果观测到学生成绩，因为推荐信直接受到成绩的影响，那么课程难度的作用被阻断了，就不再影响推荐信的结果。$I \rightarrow G \rightarrow L$，反过来看，当没有观测到学生成绩的时候，可以从推荐信的结果反推出对智商的影响；但如果观测到学生成绩，推荐信对智商的推测的作用就被阻断了，智商完全只受到成绩的推测。$G \leftarrow I \rightarrow S$，智商共同影响课程成绩和 SAT 成绩，当智商无法被观测到，我们可以通过课程成绩来推测智商，从而影响 SAT 成绩；但如果观测到智商，那么课程成绩和 SAT 的作用就被阻断了，分别只受到智商的影响。

共同的作用（V 结构）影响情况与之相反，当W不是我们的已知事实，无法被观测时，影响被阻断；当W属于已知事实，能够被观测时，影响可以被传递。例如，$D \rightarrow G \leftarrow I$，当学生成绩无法被观测到，我们可以认为课程难度与学生智商无关；但如果观测到学生成绩，就可以共同推测出课程难度和智商的信息，因此影响就可以被传递。

关于贝叶斯网络的变量关系，即概率流动，Koller 形象地比喻作某种水流，当水闸的结构不同，水流的表现不同。在 V 结构里，如果有一个变量关闭了水闸，就会使得水流往上游走。同时 Koller 给出有效迹（即影响能在贝叶斯网上传递）严谨的描述：假设$X_1 \rightleftharpoons \cdots \rightleftharpoons X_n$是贝叶斯网络上的一条迹，给定可观测变量的一个子集$Z$，那么需要满足（1）一旦有 V 结构$X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$，则$X_i$或其子节点在$Z$中；（2）迹上其他节点都不在$Z$中。

	$W \in Z$	$W \in Z$
• $X \rightarrow Y$		
• $X \leftarrow Y$		
• $X \rightarrow W \rightarrow Y$	✓	X
• $X \leftarrow W \leftarrow Y$	✓	X
• $X \leftarrow W \rightarrow Y$	✓	X
• $X \rightarrow W \leftarrow Y$	X	✓

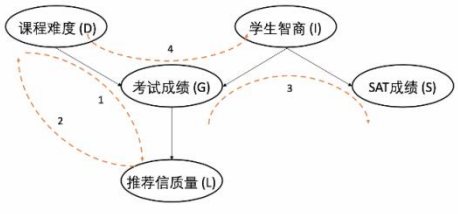

图 2.3:有效迹

2.2.2 贝叶斯网络独立性

以上，我们把图模型定义为一种来编码概率分布的数据结构，那么重点是讨论如何拿到一个概率分布，然后使用一套与图结构有联系的参数，进而用一种因子形式来表达

一个高维空间中的概率分布。 事实证明，可以用一种完全互补的视角看待图模型，也就是概率分布必须满足的，一个表达独立性的集合。

2.2.2.1 条件独立：因子和信息的双重视角

独立：对于事件A和B，$P(A,B)=P(A)P(B)$，那么A和B相互独立。对于独立性的理解，一种从信息流，在有A的信息下B的概率，不会对B本身的概率有影响；另一种从因子角度理解，A和B的联合分布是两个低维因子乘积。

条件独立：对于事件A和B，在给定Z的条件下，$P(A,B|Z)=P(A|Z)P(B|Z)$。同样有两种理解方式，一是因子的表达，即在给定Z的条件下A和B的概率，等于给定Z条件下A的概率与给定Z条件下B的概率相乘。二是信息流的理解，给定Z的情况下，B不会给A提供额外的信息来改变A的概率分布，或者给定Z，A也不会对B的概率分布提供额外知识。

同样值得注意的，给定一个条件，有时候可以得到变量间的条件独立性，有时候也可能丢失变量之间的条件独立性。在学生示例中，$D\rightarrow G\leftarrow I$，学生智力和课程难度两个因素都会影响成绩。在原来的分布里，通过计算 I 和 D 的边际概率是相互独立的；但如果给定成绩的话，它们的独立性就不再满足了。

2.2.2.2 贝叶斯网络的独立性：分布 P 因子化与图 G 独立性

给定一个分布P，如何能构建图G，使得图G的独立性成为概率分布P的独立性的合理替代？贝叶斯网络上最重要的性质就是用低维因子相乘来表示高维概率分布的因子化过程，并满足其独立性条件。换言之描述了图中独立性和分布因子化的本质关系是一种对偶性：即概率分布的因子化，体现了概率分布中的独立性，如果一个概率分布P能在图G上因子化表示，我们便直接从图G结构中读出独立性条件；反过来，如果概率分布P满足图G中显示出来的独立性，我们就可以把分布用贝叶斯图来表示。

（1）图的独立性：d-分离与 I-等价

Pearl（1988）首次将贝叶斯网的概念作为定型表示独立性关系的一种数据结构提出，并阐释了 d-分离概念。上述我们讨论过贝叶斯网络上的概率流动，取反，我们就可以得到 d-分离的概念：即给定Z的情况下，X和Y在图中没有有效迹。例如，$D\rightarrow G\leftarrow I\rightarrow S$，在给定学生成绩的情况下，课程难度和 SAT 成绩不存在有效迹，因此就是 d-分离。推而广之，在概率图中可以清晰的体现所有条件独立性：在给定其父节点，任何一个节点都与其非子节点满足 d-分离。例如，推荐信好坏L的子节点是能否找到工作J和是否幸福H，其父节点是课程成绩G，那么给定观测到的课程成绩，推荐信L与所有非子节点的 SAT 成绩S，智力I，课程程度D等都是 d-分离。

I-等价的概念由 Verma 和 Pearl（1990，1992）提出，I-map：$I(G)=\{(X\perp Y|Z)\}:d-sep_G(X\perp Y|Z)$ 满足图中所有 d-分离体现的条件独立性的集合。因此，如果概率分布P

满足$I(G)$，那么G就是该概率分布P的 I-map。I-等价其在识别网络，尤其是从数据中学习网络具有重大的作用

（2）分布与图

从分布到图，如果概率分布P能在图G上因子分解，且图上存在$d-sep_G(X \perp Y|Z)$的条件独立性，那么概率分布P满足其$(X \perp Y|Z)$条件独立性，即图G是概率分布的一个 I-map。这个定理意味着，无论参数如何，我们都可以直接从图中读出概率分布中的独立性。从图到分布，如果G是概率分布的一个 I-map，那么概率分布P能在图G上因子分解。这个定理意味着，如果分布满足图中显示出来的独立性，我们就可以把概率分布用贝叶斯图来表示。

综上，我们可以有双重视角来看图结构：一是因子化。作为一种数据结构，告诉我们一个概率分布如何被分解成一系列因子或者条件概率分布的集合，被图所表示。二是独立性。图结构及其图中的独立性如何可以被概率分布所满足，而这两种视角是相互转化的。换言之，如果有一个能被某个贝叶斯网络表示的概率分布，我们可以直接通过了解图的参数来知道分布中的独立性，进而可以知道分布的结构，分布的变化，以及不同观测所带来的变量关系的影响。

2.3 贝叶斯网络学习

2.3.1 总述

如何从数据中学习到概率图模型，是学习最重要的问题。我们需要先建立起一些设定：假设世界上有真实分布P^*，但因为通常是未知，所以假设真实分布P^*由一个概率图模型M^*中生成。从P^*分布中生成一个样本，得到一个数据集D。除了数据，我们还有一些可以放入模型作为先验信息的专家知识，最终，结合专家知识和数据学习，我们可以学习得到贝叶斯网络。那么学习的任务就分为：根据新的观测回答关于一般概率查询问题，即变量间的概率分布？根据新的观测回答关于特定预测任务，如用观测到的变量来预测某个目标变量？关于贝叶斯网络的知识发现，如区分随机变量是直接关系还是间接关系，可能存在影响的方向，以及是否存在潜在变量？

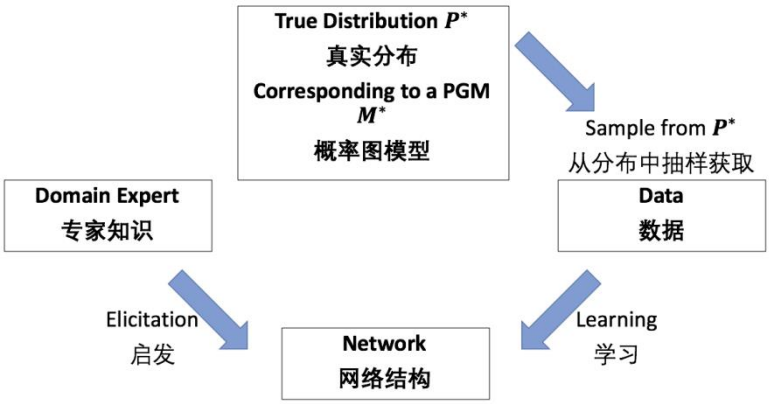

图 2.4：模型设定

2.3.2 参数学习

参数学习过程是在网络结构已知的情况下，从训练数据中学习随机变量的条件概率分布。条件概率分布的参数模型已预先指定，只需估计其中的参数，而极大似然估计和贝叶斯方法是最常用的两种参数学习方法。极大似然估计把待估参数看作取值未知的确定性量，依据参数与数据集的似然程度，来选择使似然函数值最大的参数值作为学习的结果。贝叶斯估计是基于贝叶斯公式，根据样本信息修正先验信息，由先验知识和观察到的数据集共同决定不确定性的概率参数。

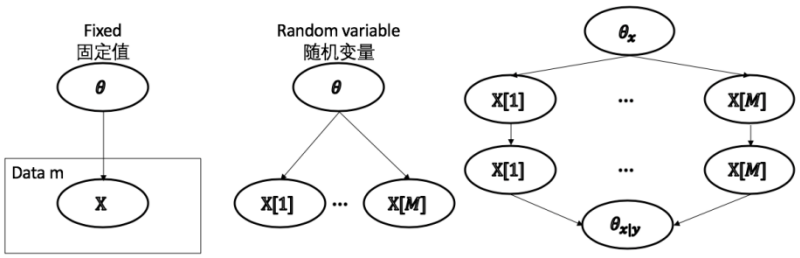

图 2.5：极大似然估计和贝叶斯估计

2.3.2.1 贝叶斯网上的极大似然估计

极大似然估计是给定数据估计参数中很重要的方法，其目的是找到一组最有可能使得样本数据出现的参数，即在给定参数的情况下，样本数据先出现的概率最大。因此，给定数据分布假设，就可以计算出数据的联合概率分布，转化成最优化问题。

我们可以借助极大似然估计对一般贝叶斯网络参数进行估计。如图所示，假设X Y服从多项式分布，给定参数$\{\theta_x: x \in Val(X)\}$和$\{\theta_{y|x}: x \in Val(X),\ y \in Val(Y)\}$，那么全局似然函数可通过链式法则化简，最终转化成如下局部似然函数的乘积，即我们可以通过对局部似然函数最大化得到参数估计：

$$L(\Theta:D)=\prod_{m=1}^{M}P(x[m],y[m]:\theta)=\prod_{m=1}^{M}P(x[m]:\theta)P(\ y[m]\ |\ x[m]:\theta)$$
$$=\prod_{m=1}^{M}P(x[m]:\theta)\prod_{m=1}^{M}P(\ y[m]\ |\ x[m]:\theta)$$
$$=\prod_{m=1}^{M}P(x[m]:\theta_x)\prod_{m=1}^{M}P(\ y[m]\ |\ x[m]:\theta_{y|x})$$

由此，借助极大似然函数对贝叶斯网络估计框架如下，D为已知数据，Θ为要估计参数集合，$x_i[m]$为随机变量第i个观测，$U_i[m]$为随机变量x_i的父节点，那么贝叶斯网的似然函数如下：

$$L(\Theta:D)=\prod_{m}P(x[m]:\Theta)$$
$$=\prod_{m}^{M}\prod_{i}P(x_i[m]\ |\ U_i[m]:\Theta_{\mathrm{i}})$$
$$=\prod_{i}\prod_{m}^{M}P(x_i[m]\ |\ U_i[m]:\Theta_{\mathrm{i}})\quad=\prod_{i}L_i(D:\Theta_{\mathrm{i}})$$

因此，只要参数是独立分开的，那么最大化全局似然函数就可以简化成最大化每一个局部似然函数。

2.3.2.2 贝叶斯网上的贝叶斯估计

MLE 中把参数当作固定的，目的就是找到使得数据最有可能出现的参数数值，因此无论硬币投掷 10 次正面朝上 7 次，还是投掷 10000 次朝上 7000 次，其正面朝上的概率都是 0.7，然而第一种场景下给出这个推断显然是不合理的。与 MLE 相比，贝叶斯估计最核心的是把不确定的事物，包括参数，都看作随机变量，其分布会随着我们获取的数据而变化和更新。

如图所示，在 MLE 中，参数是固定的，因此随机变量不同观测之间是相互独立的。在贝叶斯估计中，由于参数也是未知的随机变量，那么$X_1\cdots X_m$之间就不再是相互独立的，每一个样本都为参数估计提供了信息。因此，在贝叶斯估计就是，首先有参数的先验概率，再通过样本信息迭代，来计算参数后验概率的过程。其中，通常以 Dirichlet 分布作为先验概率，因为其后验概率具有同样多项式分布的形式便于计算，并且能够使用样本数据中的充分统计量进行快速迭代。

$$P(\ x[1],\cdots,x[M],\theta) = p(x[1],\cdots,x[M]\ \mid\ \theta)P(\theta) = P(\theta)\prod_{i=1}^{M} P(x[i]|\theta)$$

$$P(\theta\ \mid\ x[1],\cdots,x[M]) = \frac{p(x[1],\cdots,x[M]\ \mid\ \theta)P(\theta)}{p(x[1],\cdots,x[M])}$$

因此，我们可以用贝叶斯估计来进行贝叶斯网络的参数学习。给定参数$\{\theta_x: x\ \in Val(X)\}$和$\{\theta_{y|x}: x\ \in Val(X),\ y\ \in Val(Y)\}$。我们可以结合贝叶斯图的条件独立性质，得到以下结论：一是，给定θ_X，成对$(X_i\ Y_i)$和$(X_j\ Y_j)$彼此是 d-分离。二是，给定样本数据观测，参数的先验概率彼此之间是独立的,从而参数的先验概率可以被简化成各个参数先验概率的乘积，$P(\theta) = \prod_i P(\theta_{x_i|Ps(x_i)})$ 。三是，给定完整的样本数据，参数的后验概率也是相互独立的，从而$P(\theta_X, \theta_{Y|X}|D) = P(\theta_X|D)P(\theta_{Y|X}|D)$，那么后验概率就简化成每一个参数的条件概率分布相乘。

2.3.3 结构学习

除了给定网络结构进行参数学习，很多情况下，没有足够的经验知识提前给定一个足够好的网络结构，就需要结合样本数据学习其中的独立性。或者在海量数据和高维变量中，研究目的就是为了找到网络结构来推测变量之间的关系从而更好的理解该领域的知识。

基于评分搜索的方法、基于约束的方法和混合学习方法，是进行结构学习的三类算法。基于约束的学习算法视为约束满足问题，主要是通过检验随机变量间的条件独立性来构建结构。其低阶数据处理简单，但是高阶的条件独立性检验很复杂而且结果不一定可靠。而基于评分搜索的学习算法视为结构优化问题，主要是利用得分函数评价网络结构优劣，然后用搜索算法来寻找出分数最高的最优结构以优化。其可以把专家经验知识以结构先验概率分布的形式融入到过程中，但是算法收敛速度慢，计算复杂，容易陷入局部最优问题。混合算法结合两种方法，一般先用各变量间条件独立性检验减少搜索空间或构造便利序列，然后再通过评分搜索算法对数据进行结构学习。本文主要对评分搜索方法进行详细说明。本文只介绍基于评分搜索的结构学习方法。

基于评分搜索学习的核心是定义一个评分函数，该函数针对每种候选结构评估该结构与数据的匹配程度。当我们有数据集 D，N 个示例网络结构，评分函数 S，该函数将告诉我们，这些候选的网络结构中的每一个与样本数据的匹配性能好坏，学习的目标就是寻找一种最大化评分函数得分的网络结构，从而将学习问题转化为优化问题，对组合空间（网络结构空间）的优化。因此，基于评分搜索的学习可分为两个部分，一是定义评分函数，对不同的网络结构进行打分；二是搜索算法，帮助在网络结构候选空间里选取得分函数最高的网络结构。

2.3.3.1 得分函数

极大似然函数得分$score(G:D)=l\left(\ (\hat{\theta},G):D\ \right)$是用极大似然估计的参数来计算与图结构相关的样本数据似然函数值，即找到给定样本数据和图结构，使得似然函数取值最大的参数，再计算得到的似然函数得分作为对该图结构的评价。通过下述简单的例子来进一步理解极大似然得分本质。

图 2.6：示例

$$score_L(G_0:D)=\sum_m(log\hat{\theta}_{x[m]}+log\hat{\theta}_{y[m]})\quad score_L(G_1:D)=\sum_m(log\hat{\theta}_{x[m]}+log\hat{\theta}_{y[m]|x[m]})$$

$$score_L(G_1:D)-score_L(G_0:D)=\sum_m(\ log\hat{\theta}_{y[m]|x[m]}-log\hat{\theta}_{y[m]})$$

$$=\sum_{x,y}M[x,y]log\hat{\theta}_{y|x}$$

$$-\sum_y M[y]log\hat{\theta}_y=M\sum_{x,y}\hat{P}(x,y)log\hat{P}(y|x)-M\sum_y\hat{P}(y)log\hat{P}(y))$$

$$=M(\sum_{x,y}\hat{P}(x,y)log\hat{P}(y|x)-\sum_{x,y}\hat{P}(x,y)log\hat{P}(y))$$

$$=M(\sum_{x,y}\hat{P}(x,y)\frac{log\hat{P}(y|x)}{\hat{P}(x)\hat{P}(y)})=M\cdot I_{\hat{P}}(X\ ;Y)$$

其中，$I_{\hat{P}}(X\ ;Y)$是互信息，衡量了在实际分布中，随机变量间联合概率分布和其边缘概率分布乘积之间的距离。因此一般意义上，似然函数可以被分解成两部分：

$$score_L(G{:}D)=M\sum_{i=1}^{n}I_{\hat{P}}\left(X_i\ ;Pa_{X_i}^G\right)-M\sum_i H_{\hat{P}}(X_i)$$

$$I_P(X;Y)=\ \sum_{x,y}P(x,y)log\frac{P(x,y)}{P(x)P(y)}$$

$$H_P(X)=\ -\sum_x P(x)logP(x)$$

互信息从信息解释的视角阐释了图结构中变量之间的独立和相关。也就是说，当随机变量X_i与其父节点越相关，其网络结构的似然函数得分越高，与现实直觉相符，即其

父节点与变量越相关，越应该把其作为父节点纳入网络结构，所以纳入网络结构作为父节点都是那些有最高互信息的节点。但同时也有缺陷：由于$I_{\hat{P}}(X;Y) \geq 0 \ \& = 0 \ iff \ X \perp Y$，只有在随机变量完全独立的时候互信息才有可能为零，但是在现实数据分布中，由于噪声干扰，样本数据中很难出现绝对的独立，因此互信息几乎总是大于 0，那么就会导致在图结构中加边总是会对得分有帮助，因此得分最大化总是会倾向于建立全连接边的图结构。为了避免对训练数据的过度拟合，一个策略是，可以加入假设来限制父节点的个数或者参数的个数。另一个策略是，在得分函数中加入一些惩罚项来限制模型的复杂程度，只要加边带来的收益足够大，又不会完全限制加边的行为，但是可以有效的筛除一些相关性小的边。

BIC 得分$score_{BIC}(G:D) = l\left(\ (\hat{\theta},G):D\ \right) - \frac{\log M}{2}DIM[G]$，通过加入惩罚项，限制参数的数量增，在训练数据拟合和模型复杂度中平衡以避免过度拟合。随着样本数据趋于无穷，如果样本数据由真实网络结构中产生，其图结构上的独立条件的 BIC 分数最高，即满足一致性。

贝叶斯得分$score_B(G:D) = logP(D|G) + logP(G)$，由贝叶斯法则而来，把所有不确定的都当做随机变量对待。第一项是给定图结构样本数据的的边缘概率分布，第二项是图结构参数的先验分布。极大似然得分是找到使得极大似然函数最高的参数来计算最大概率，而贝叶斯得分，根据$logP(D|G) = \int P(D\ |G,\theta_G)P(\theta_G|G)d\theta_G$，则是利用所有可能的参数取值取得正在计算所有可能参数设置上的平均概率，从而对给定特定结构数据的概率的乐观估计大大降低，因为不仅要考虑适用于数据集的参数，还要使用先验的所有可能的参数设置。因此直觉上，这种评估对训练数据可能没有那么适合，但可以有效降低过度拟合问题。除了样本信息迭代过程，还需要设定两个先验分布。通常，结构的先验可被忽略掉，当作常数处理$P(G) \propto constant$。参数的先验分布可用 BDe 先验，其包含两个超参数，α是样本大小，B_0是体现我们对事件概率理解的初始网络图结构。BDe 的好处是，一个初始B_0网络结构先验可用于所有候选的网络图，并且使得同样具有 I-map 的图结构有相同的贝叶斯得分。综上，贝叶斯得分在大样本下和 BIC 得分等价，并且是一致的，而且可以有效避免，小样本下，BIC 得分所导致的对训练数据拟合不够。

2.3.3.2 搜索算法

综上，基于评分搜索的学习需要定义一个评分函数，然后在所有的候选结构中通过搜索算法找到一个可以使得得分最高的，这就转化成了一个最优化问题。输入是训练数据集，评分函数包括先验分布，以及一组候选网络结构；输出就是一个最大化的得分的网络结构；最重要的性质是可分解性，$score(G:D) = \sum_i score\left(\ X_i|Pa_{X_i}^G:D\ \right)$，即全局得分可以被分解成一些局部得分数值和，从而简化的计算的复杂度。

寻找使得得分函数最高的网络结构，当随着图中随机变量父节点增加，算法将呈现多项式增加，出现 NP-Hard 的困境，因此需要重新在算法的设计选择上进行改进。第一

组设计选择是搜索操作对象，即一系列调整网络结构的步骤。例如采取在网络中很小变动的局部步骤，加边，删边，或者反转边，或者采取会带来较大变化的全局步骤，将整个节点从网络中移出并其放置在其他位置。第二组设计选择是搜索技术。例如试图爬升更快的贪婪爬山法，或者其他类型的搜索技术。

贪婪爬山算法是贝叶斯网络网络的基础，很多算法是以此为基础进行改进。首先，从一个给定的初始图出发，可以是空的网络，最好的树结构，随机结构，或者利用先验知识建立起来的初始图。其次，在每一次的迭代中，都在设定好的全部搜索操作对象，计算所有可能的改变值，从中选取最高得分的操作，一直迭代到没有任何的操作可以再对网络结构得分进行提高。但贪婪爬山算法也会有缺陷。一是可能是局部最大值，而非全局最大值。如果我们以图形方式画出可能的网络的空间，绘制成一条连续的有高低起伏的曲线，如果我们从起点开始，采取小小的爬坡步骤，最终可能会困在一个局部极大值的坡，而无法达到需要跨越较多步才能抵达全局极大值。二是平稳期困境。在贝叶斯网络结构学习中，有很多网络结构一开始都是相同的分数，然而在某些方向，经过一定数量的步骤，可能会带来提升，问题目前尚不清楚其中哪些将最终引发使我们脱离高原的举动，进入一个实际上得分更高的网络。为避免这两种缺陷，可采取两种简单的策略：第一种策略是随机重启，即如果陷入困境，可采取一定数量的随机步骤，然后再次开始爬升。如果处于一个相当浅的局部最大值，那么少量的随机步长将使我们进入一个更好的最大值搜索区域，并且继续攀升到一个更好的最优值。第二种策略是禁忌表。禁忌清单是一种避免一遍又一遍踩踏相同步骤的方法，即保留了最近执行的 k 个步骤的列表，并限制了搜索，以使其无法逆转这些步骤中的任何一个。这就使得搜索算法继续向前迈进，而不是走相同的步骤。

综上，学习问题可以从三个不同的方面来看。一是假设空间，这是我们感兴趣的经典模型的一个候选集合；二是应用学习算法时要尝试优化的目标函数；三是优化算法本身，旨在优化目标函数以选择好的模型。几乎所有学习算法都由这三个不同的成分组合。

假设空间是我们要搜索的空间，实际上我们可以搜索参数，结构或者同时搜索。并且为了提高计算效率，减少模型搜索空间，或者加入先验知识，我们对要搜索的内容施加约束条件，从而将学习算法导向比完全不受约束的模型更合理的模型。

第二部分是目标函数。最常用的是带惩罚项的似然函数$l\big((G,\theta_G):D\big)+R(G,\theta_G)$，其具有对数似然分量，测量了我们的图结构和参数与训练数据集的相似度；同时第二项有作为正则化，用于引导学习模型向不太可能过度拟合数据的简单模型。对于正则化项，一种形式是可以包含先验参数中的一个或两个，这通常倾向于使参数平滑并避免过度拟合训练集的统计信息，如在贝叶斯网络里，先验设定为 Dirichlet。第二种形式是搜索结构时的结构复杂度惩罚，目的是将我们推向参数较少或边较少的模型。另一种是贝叶斯得分$\log P(G|D)=logP(D|G)+logP(G)+constant$。贝叶斯得分（即给定数据的图的概

率的对数）等于边际似然函数（即给定图的数据的概率的对数）和图结构的先验的对数。其中，图结构的先验，以及边际似然函数中都含有正则化，用来减少模型的过拟合。

最后是优化算法，它取决于优化的空间以及优化的目标。如果在连续空间上进行优化，有些情况下可以封闭形式优化似然函数，并找到唯一的最优值；其他情况下，可使用梯度上升或某种其他可能的二阶方法，通过爬升法来逼近似然函数最优值。如果在离散空间上进行优化，有最大的权重扩展树和爬山法等。如果同时搜索连续和离散空间，情况会变得复杂，计算成本上升，但是也有一些策略来降低成本。

2.4 贝叶斯网络下的交互式探讨

实践中应用贝叶斯学习时使用的典型学习循环如下：首先，设计模型模板，如涉及的随机变量集以及模型是有向还是无向的。其次，通过对训练集进行交叉验证来选择超参数，并按照所选的超参数在训练集进行训练。然后，我们在一个留存数据集合上评估性能，如错误率等。接下来，探讨错误分析的过程，并重新设计模型、目标函数或优化算法来解决问题。以上步骤重复循环，直到得到一个稳定不需要修正的模型，再在不同的单独测试集上报告结果。值得注意的是，概率图模型允许我们将先验知识整合到模型中，这是许多其他算法所不具备的。而其中基于评分搜索的学习算法正是把专家经验知识以结构先验概率分布的形式融入到过程中，于是可由“输入-加入先验条件-结构学习-参数学习-调整不合理变量关系-验证变量关系-输出”的研究框架收敛到最终图模型。

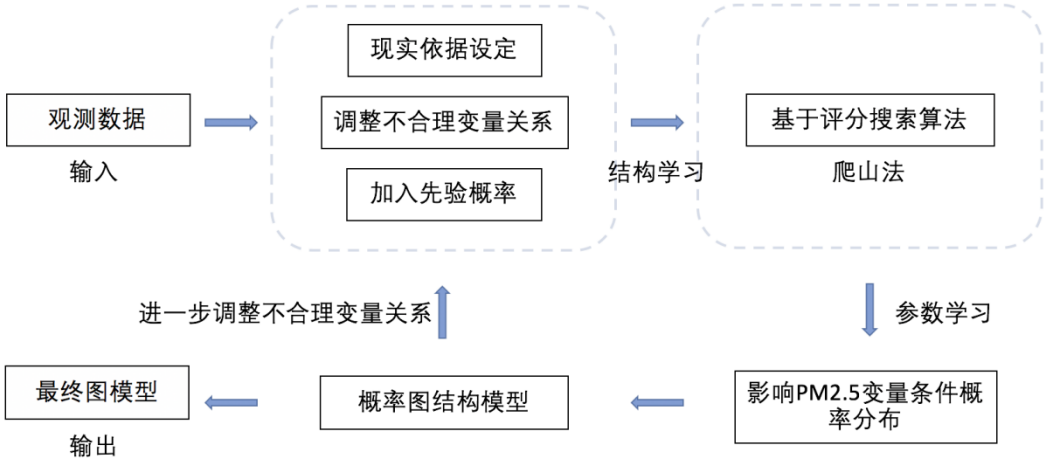

图 2.7：贝叶斯网络模型学习框架

综上，贝叶斯网网络的应用主要是从不确定性中找到变量的可能关系，根据专家知识和数据学习建立“正确”的模型，并根据建立好的模型和已观测到的数据进行参数估计和因果推理。在这个过程中，并不是依赖于纯数据挖掘的自动统计模式选择，而需要将关键的背景信息注入这些模型结构中，重复在测试样本上测试性能，并识别和调整不

合理变量关系，避免数据过拟合问题，以得到最终图模型。因此，最重要的是需要具备两点：一是要求在建立模型前，研究者已经对研究的问题有一定的认识，根据前人文献和专家知识设定一定的条件独立关系，对已经确定的事实给出先验设计；二是研究问题中不确定的部分需要根据数据信息决定模型，即采取某种方法来构造网络，并在已有观测值中，根据极大似然估计或贝叶斯估计来选择最优结构图。在从数据中学习图结构的过程，可通过“黑名单和白名单”设置，实现人工先验知识和机器学习结合，为交互式提供了可能。

第三章 北京 PM2.5 影响因素的贝叶斯网络分析

经济学的发展也逐渐涉及生态环境领域，空气质量与经济活动密不可分，特别是近些年受到广泛关注的细颗粒物（粒径小于 2.5 微米，PM2.5）。随着数据收集技术的进步，近年来大量气象观测数据和环境质量监测数据得以发布，也为研究提供了数据基础。环境领域中，许多自然科学研究成果是在特定实验室条件下，来研究变量之间化学转化的确定性关系，如刘保献等研究发现 PM2.5 化学主要组分为 OM、EC、SO42-、NO3-、NH4+、CI-、地壳元素、微量元素，SO42-[24]。但是由于其结果所依赖的实验条件苛刻，现实数据情况错综复杂，各变量之间相互影响，如气象条件、光照、O3 浓度、SO2 浓度、NO2 浓度、湿度等[25][26]，在观测数据中的关系需要有待于进一步验证。因此正好提供了一个贝叶斯网络学习的场景，既可借助专家确定性知识进行先验结构设计，又可利用数据学习不确定性关系来进行人际交互，不断调整不合理关系，验证现实情况[17][18]。

3.1 PM2.5 数据及样本选择

3.1.1 数据库介绍（调查/收集方式）

本文使用的是 2015 年 1 月-2018 年 12 月北京市农展馆监测点的大气浓度小时数据及相应时间段的气象小时数据，并用机场气象记录对缺失值进行插补[1][2]。每天共 24 个观测值，去除缺失值，观测值数 2015 年、2016 年、2017 年和 2018 年分别对应为 8448、8241、8732 和 8432，总观测值数为 33853 个，具体变量如下表所示：

其中， PM2.5 为主要被解释变量，CO，SO2，NO，NO2 等污染气体，相对湿度、温度、气压、风速等气象条件为解释变量，并考虑各个变量一期滞后项影响，同时控制风向的影响作用，根据风向 16 方位图，共有 17 种风（包括静风和不定向风），由于涉及虚拟变量过多，因此将相邻方向合并，最终合并成八种风向，每一种类型作为一个虚拟变量。

表 3.1：变量定义

变量	解释	编码
time*	年-月-日 时刻	
weekend	是否为周末	0/1 虚拟变量
season	季节	1-4 编码

[1] 北京环境与气象数据主要来源网站：http://zx.bjmemc.com.cn/；http://zx.bjmemc.com.cn/getAqiList.shtml?timestamp=1586650096480；http://www.weather.com.cn/air/；http://www.weather.com.cn/weather/101010100.shtml；

[2] 感谢王敏老师，徐晋涛老师，中国气象局朱定真老师和北京市气象台在数据整理过程中的帮助。

NZGPM25	农展馆 PM2.5 的浓度，单位μg/ m3	
CO	一氧化碳质量浓度，单位 mg/ m3	
NO2	氮氧化物质量浓度，单位μg/ m3	
O3	臭氧质量浓度，单位μg/ m3	
SO2	SO2 质量浓度，单位μg/ m3	
WindSpeed	风速，单位 m/s	
RelaHumi	相对湿度，百分比	
TempC	温度，单位 摄氏度	
Pressure	气压，单位 帕斯卡	
N_NNE	风向，北+北东北	0/1 虚拟变量
NE_ENE	风向，东北+东东北	0/1 虚拟变量
E_ESE	风向，东+东东南	0/1 虚拟变量
SE_SSE	风向，东南+南东南	0/1 虚拟变量
S_SSW	风向，南+南西南	0/1 虚拟变量
SW_WSW	风向，西南+西西南	0/1 虚拟变量
W_WNW	风向，西+西西北	0/1 虚拟变量
NW_NNW	风向，西北+北西北	0/1 虚拟变量
C	风向，不定和无风	0/1 虚拟变量
NZGPM25_lag1	滞后一期:农展馆 PM2.5 的浓度，单位μg/ m3	
CO_lag1	滞后一期:一氧化碳质量浓度，单位 mg/ m3	
NO2_lag1	滞后一期:氮氧化物质量浓度，单位μg/ m3	
O3_lag1	滞后一期:臭氧质量浓度，单位μg/ m3	
SO2_lag1	滞后一期:SO2 质量浓度，单位μg/ m3	
WindSpeed_lag1	滞后一期:风速，单位 m/s	
RelaHumi_lag1	滞后一期:相对湿度，百分比	
TempC_lag1	滞后一期:相对湿度，百分比	
Pressure_lag1	滞后一期:温度，单位 摄氏度	

3.1.2 数据描述性统计

表 3.2 是文中主要被解释变量及解释变量的描述性统计结果，PM2.5、PM10、NO2、SO2、O3 的单位均为 μg/m3，CO 样本均值较小是由于其单位为 mg/m3，与前几类气体污染物选取的衡量单位有所差异。

全时段数据来看，2015 年到 2018 年期间 PM2.5 的均值为 68.10μg/m3，最大值为 835μg/m3。分年数据[3]来看，在 2015 年、2016 年、2017 年和 2018 年，PM2.5 的全年均值分别为 85.24μg/m3、74.78μg/m3、60.55μg/m3 和 52.22 μg/m3，呈现稳步下降趋势。除了臭氧，其他污染性气体浓度在 2015 年到 2018 年期间均有下降，如 CO 的全年均值 2015 年、2016 年、2017 年和 2018 年分别为 1.38 mg/m3，1.20 mg/m3，0.99 mg/m3 和 0.89 mg/m3。风速、湿度、温度、气压等气象变量在 2015 年到 2018 年期间全年均值比

[3] 见附录分年描述性统计表

较稳定，没有明显变化，如风速从 0m/s 到 17m/s 不等，均值为 2.91m/s。此外，在 2015 年到 2018 年期间频率最高的风向是北和北东北风，占 20%左右，最低的是西南和西西南，占 2%左右。

表 3.2：主要被解释变量及解释变量的描述性统计结果

Var	N	Mean	Sd	Min	Max
Year	33853	2016.51	1.12	2015	2018
weekend	33853	1.29	0.45	1	2
season	33853	2.49	1.12	1	4
NZGPM25	33853	68.10	74.35	1	835
CO	33853	1.12	1.10	0	17
NO2	33853	51.45	33.34	2	265
O3	33853	59.97	57.78	1	393
SO2	33853	10.97	14.57	1	257
WindSpeed	33853	2.91	2.22	0	17
RelaHumi	33853	0.55	0.26	0.05	1
TempC	33853	13.28	12.08	-16	41
Pressure	33853	1015.77	10.62	990.9	1046
NZGPM25_lag1	33853	68.09	74.39	1	835
CO_lag1	33853	1.12	1.10	0	17
NO2_lag1	33853	51.44	33.34	2	265
O3_lag1	33853	60.04	57.85	1	393
SO2_lag1	33853	10.97	14.57	1	257
WindSpeed_lag1	33853	2.91	2.22	0	17
RelaHumi_lag1	33853	0.55	0.26	0.05	1
TempC_lag1	33853	13.28	12.09	-16	41
Pressure_lag1	33853	1015.77	10.62	990.9	1046
N_NNE	33853	0.20	0.40	0	1
NE_ENE	33853	0.06	0.25	0	1
E_ESE	33853	0.10	0.30	0	1
SE_SSE	33853	0.12	0.32	0	1
S_SSW	33853	0.10	0.30	0	1
SW_WSW	33853	0.02	0.13	0	1
W_WNW	33853	0.06	0.23	0	1
NW_NNW	33853	0.11	0.32	0	1

3.1.3 变量关系初探

下面将对被解释变量 PM2.5 及不同类型的解释变量进行详细介绍，并探讨 PM2.5 与主要解释变量可能存在的关系，并对接下来的贝叶斯网络先验设计提供环境领域知识支持。

3.1.3.1 被解释变量 PM2.5

PM2.5 是本文所关注的被解释变量，其浓度值在 2015 年到 2018 年的变化如下图所示，纵轴为 PM2.5 浓度值，单位是 μg/m3，横轴为时间。从图中可以看出，PM2.5 浓度值在 2015 年到 2018 年整体呈现下降趋势。根据中国空气质量等级标准，PM2.5 浓度值在 75 μg/m3 以下为良以上，PM2.5 浓度在 250 μg/m3 以上严重污染，从 2015 年到 2018 年空气质量良以上的观测逐步增加，严重污染的观测逐步减少。此外，各个年份中 PM2.5 浓度值在 12-2 月整体相对较高，波动幅度大，且有较高的极端数值，而 6-8 月整体相对较低，波动比较稳定。

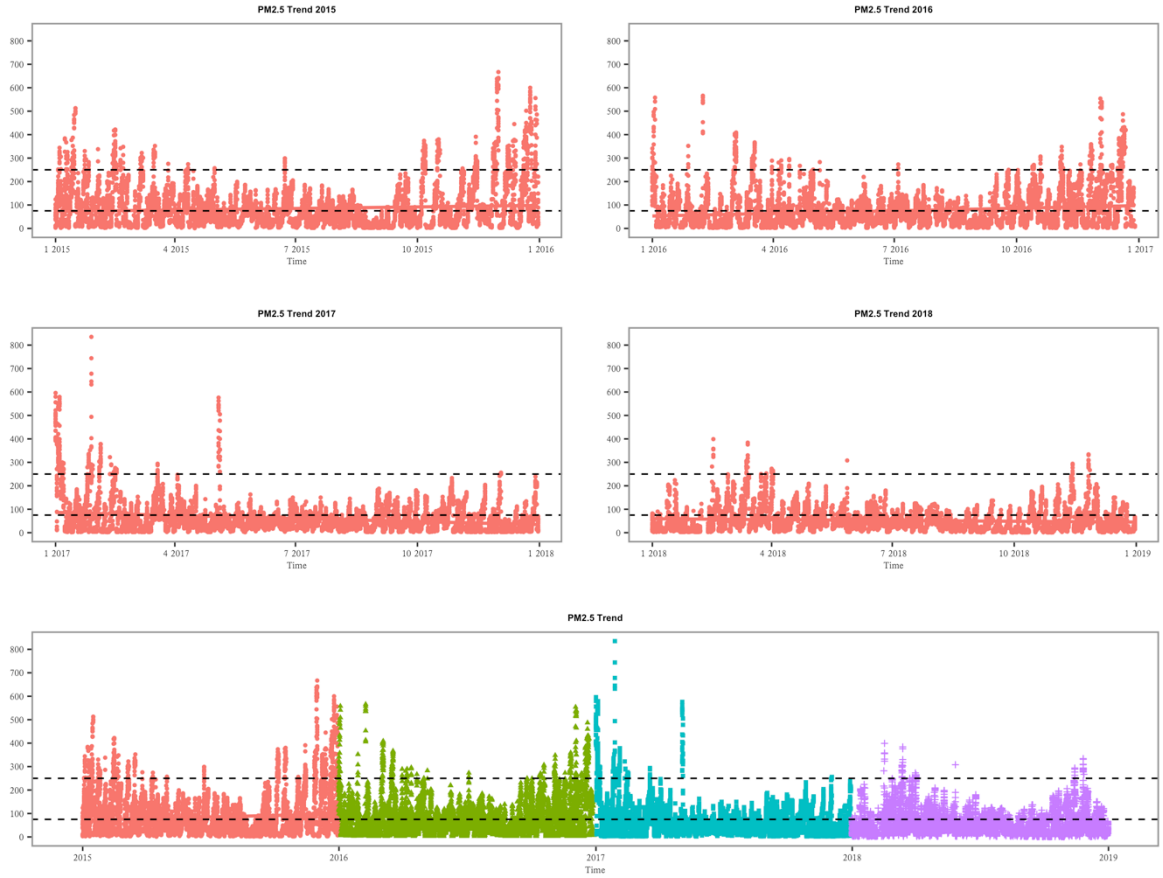

图 3.1：2015 年到 2018 年 PM2.5 浓度变化趋势

3.1.3.2 解释变量

（1）污染物气体变量

从下图可以看出，一氧化碳、二氧化氮和二氧化硫气体浓度在 2015 年到 2018 年期间均有下降，且其浓度随时间变化趋势与 PM2.5 浓度整体相似，并在 12-2 月整体相对较高，波动幅度大，且有较高的极端数值，而 6-8 月整体相对较低，波动比较稳定。而臭氧浓度在 2015 年到 2018 年期间没有明显变化，其浓度随时间变化趋势与 PM2.5 浓度关系不确定，并在 5-9 月整体相对较高，波动幅度大，在 11-3 月整体相对较低，波动稳定。

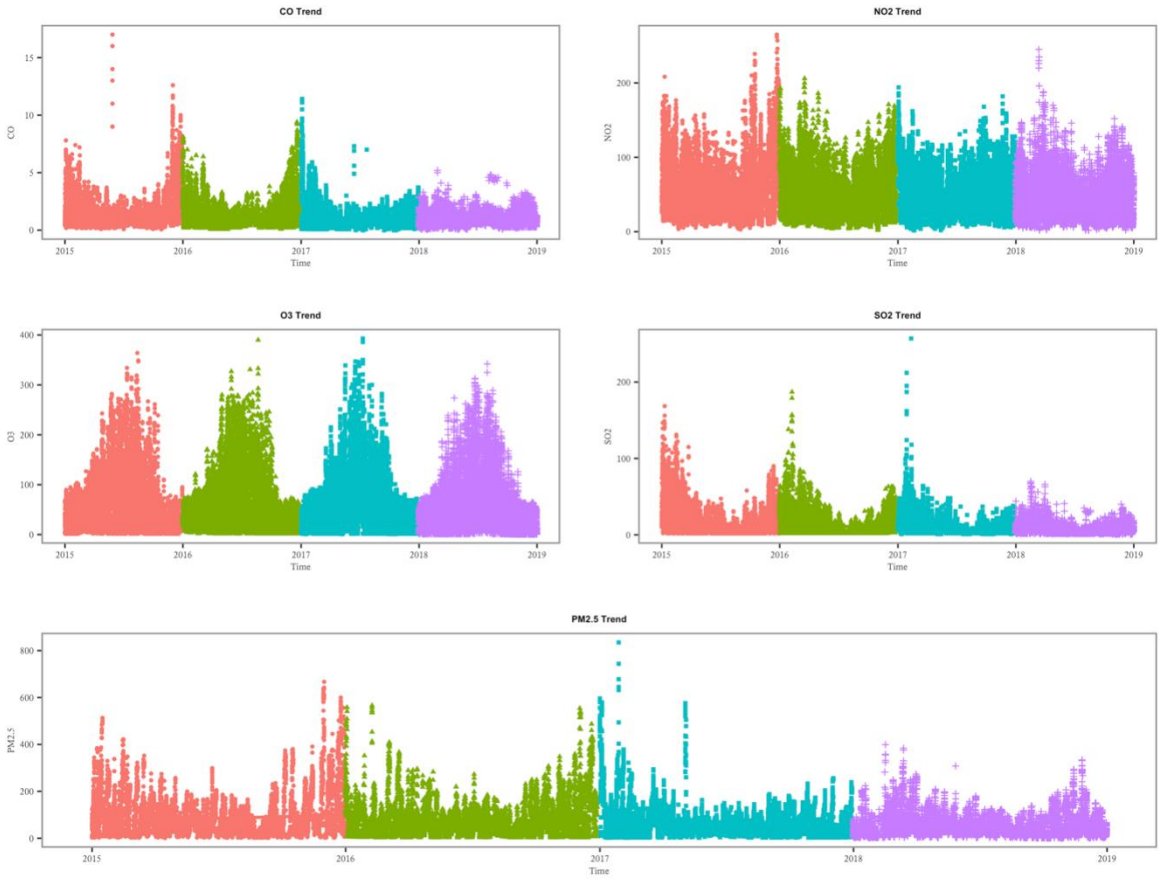

图 3.2：2015 年到 2018 年污染物气体浓度变化趋势

从图 3.3 可以看出， NO2 与 PM2.5 整体呈正相关关系，但是波动范围较大。在 NO2 浓度小于 100 μg/m3 时，PM2.5 的浓度值增加速率慢；在 NO2 浓度 100-200 μg/m3 时，正向增长速率加快，PM2.5 的浓度值陡然增加，且波动较大；在 NO2 浓度超过 200 μg/m3 时，正向增长速率减缓，PM2.5 的浓度值基本上围绕在 200-400 μg/m3 左右波动。

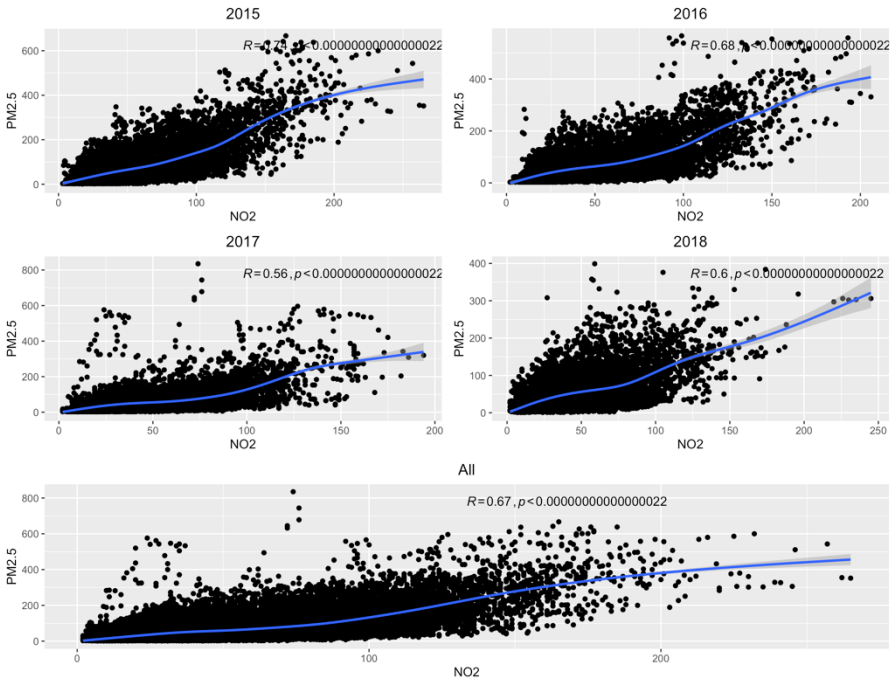

图 3.3：PM2.5 浓度与 NO2 气体浓度关系

PM2.5 浓度与 SO2 浓度的关系不如与 NO2 明显，在 SO2 浓度小于 45μg/m3 时，正向增长速率减慢，关系的波动愈发明显；在 SO2 浓度在 45~100μg/m3 之间时，增长速率几乎为零，PM2.5 的浓度值基本不变，围绕在 200μg/m3 左右；在 SO2 浓度值超过 100μg/m3 时，PM2.5 的浓度值陡然增加。

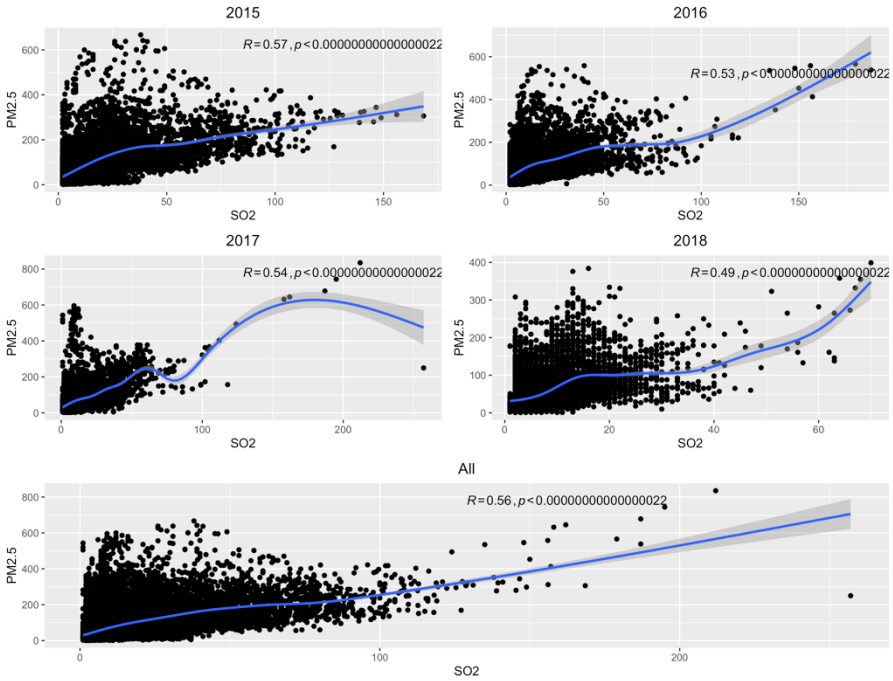

图 3.4：PM2.5 浓度与 SO2 气体浓度关系

PM2.5 浓度与 CO 浓度的关系，在 CO 浓度小于 10mg/m3 时，与 PM2.5 有比较明显的正相关关系，而当 CO 浓度超过 10mg/m3 时，不再具有明显的正相关关系，这可能是由于高浓度观测值数量较少的缘故。

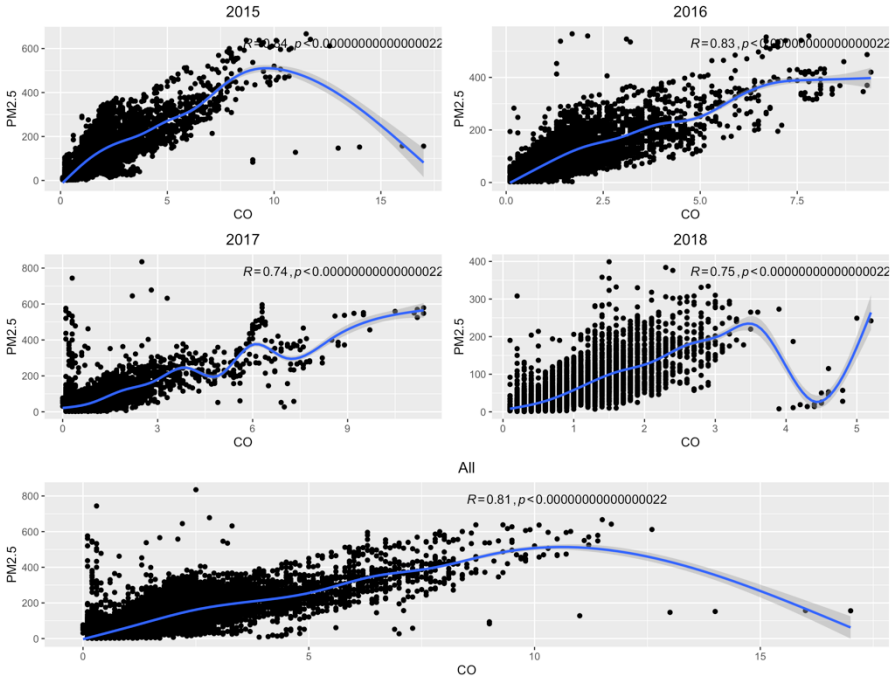

图 3.5：PM2.5 浓度与 CO 气体浓度关系

而 PM2.5 浓度与 O3 浓度在下图中并不能看到明显的线性关系。

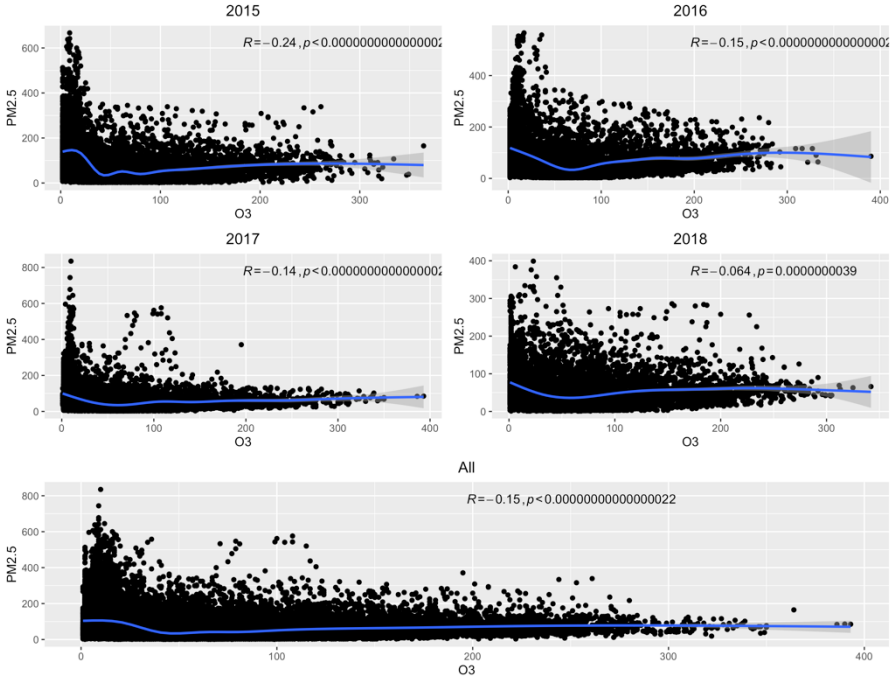

图 3.6：PM2.5 浓度与 O3 气体浓度关系

（2）气象变量

从图 3.7 中，可以看到 PM2.5 浓度与风速的关系，二者整体上呈负向的线性关系，在风速小于 14m/s 的时候，PM2.5 的浓度随风速的增大而降低，但大于 15m/s 时，出现了反向的提升。根据 PM2.5 的作用原理，在现实情况中，小风速对空气起搅拌作用，有利于 PM2.5 的聚沉；大风速下空气呈湍流特性，有利于 PM2.5 吹起，并带来其他地区的污染物导致 PM2.5 浓度的提升。此外，PM2.5 浓度与相对湿度总体上呈正向关系。PM2.5 浓度与温度总体上无明显的线性关系。在 0 摄氏度以下，PM2.5 浓度与温度整体是正向关系，在 0 摄氏度以上，PM2.5 浓度与温度整体是负向关系，开始随着温度的升高缓慢下降。

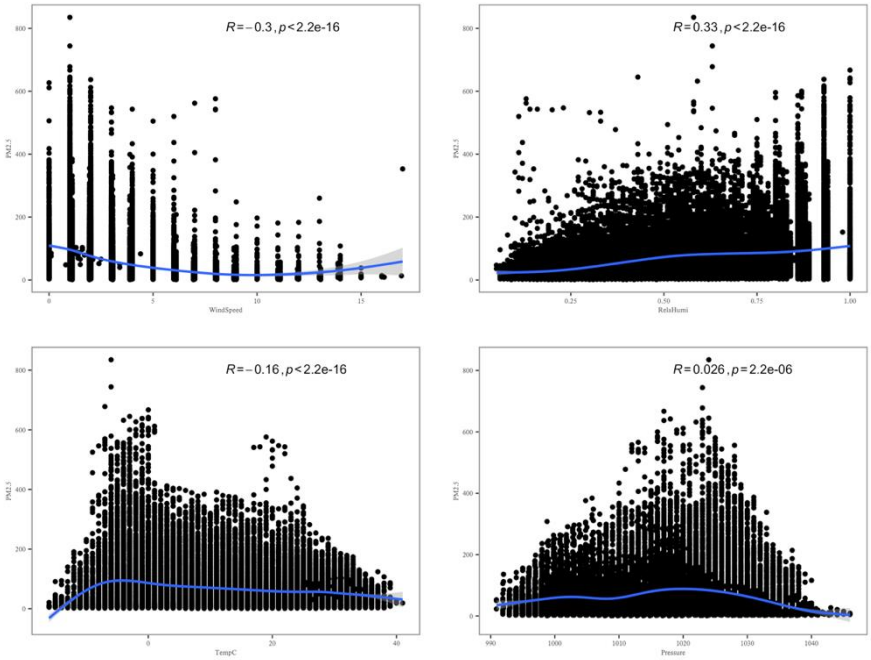

图 3.7：PM2.5 浓度与气象要素关系

（3）控制变量

风向也是影响 PM2.5 浓度的重要变量，对于北京来说，不同方向的风对 PM2.5 的扩散作用不同。从描述统计表可知，从 2015 年到 2018 年期间，全时段来看，各个风向的频率是:N_NNE>SE_SSE>NW_NNW>E_ESE=S_SSW>W_WNW=NE_ENE>SW_WSW，最常出现的是无风、不定风向和北_北西北。下图是 2015 年到 2018 年全时段和分年份不同风向下 PM2.5 浓度均值的比较。其中，全时段，各个风向下 PM2.5 浓度均值排序是：E_ESE>SE_SSE>NE_ENE>S_SSW>N_NNE>NW_NNW>W_WNW，所以西_西西北、西北_北西北、北_北西北可以起到降低 PM2.5 浓度的作用，而东_东东南、东南_南东南等对 PM2.5 浓度降低作用不大。考虑到北京北部和西部没有污染源，而南部的工厂较多，污染程度较高，因此西风和北风更可能起到净化作用。

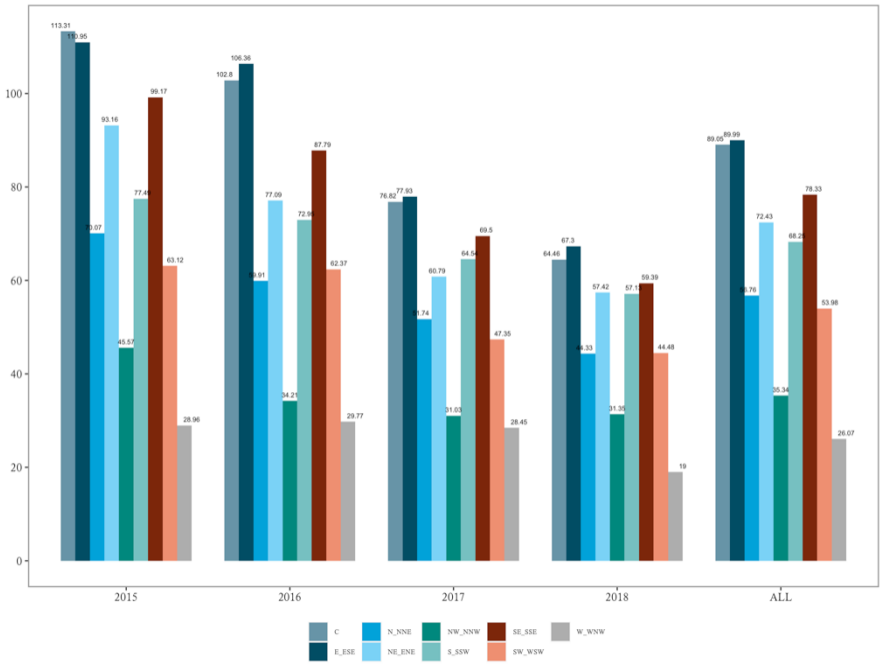

图 3.8：不同风向下 PM2.5 浓度均值比较

在 PM2.5 浓度趋势图中可以看出在不同季节间的 PM2.5 浓度均值有非常明显的差异，总体趋势为上半年不断降低，下半年再不断升高的抛物线状。从下图不同季节下 PM2.5 浓度均值的比较中更清晰地看到，PM2.5 浓度均值的排序为：冬季>秋季>春季>夏季。考虑到秋冬季节居民有燃煤供暖需求，从而导致 PM2.5 前体物排放增加，因此 PM2.5 浓度值也会增加。

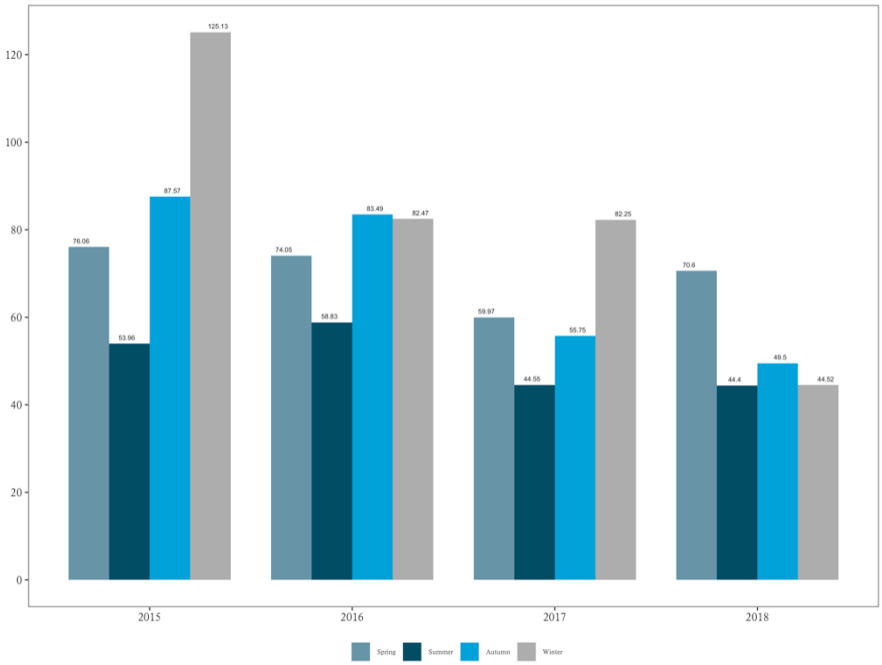

图 3.9：不同季节下 PM2.5 浓度均值比较

3.2 交互式下贝叶斯网络结果

在构建贝叶斯网络的过程中，采取专家知识和数据学习相结合的办法，为了考虑模型的可视性与计算速度，在这里使用爬山法（Hill-climbing） 算法构造贝叶斯网络。借助 R 中 bnlearn 包[21][22]，首先，通过黑白名单设计，将先验变量关系作为初始图结构设计。其次，使用算法对训练数据上进行结构和参数学习。然后，在一个测试数据集合上评估性能，如均方误差等，并根据影响 PM2.5 变量条件概率分布结果，探讨并重新设计不合理的变量关系。以上步骤重复循环，直到得到一个稳定不需要修正的模型。

3.2.1 贝叶斯网络初始图

首先使用 2016 年训练样本，通过改变先验条件和模型约束，记录下每次模型改变约束后的结果，并在测试样本评估性能，尝试建立一条贝叶斯网络学习的“路径”。在构造初始化模型前，有一些基本模型设定和假设：

本文主要关注的被解释变量:

$$Y = \ NZGPM25。$$

当期解释变量集合:

$$Var = \{CO, NO2, O3, SO2,$$
$$WindSpeed, RelaHumi, TempC, , Pressure,$$
$$N\_NNE, NE\_ENE, E\_ESE, SE\_SSE, S\_SSW, SW\_WSW, W\_WNW, NW\_NNW\}$$

滞后一期变量集合：

$$Lag = \{NZGPM25\_lag1,$$
$$CO\_lag1, NO2\_lag1, O3\_lag1, SO2\_lag1,$$
$$WindSpeed\_lag1, RelaHumi\_lag1, TempC\_lag1, Pressure\_lag1 \ \ \}$$

以及时间控制变量：

$$\text{Control} \ = \{season，\ weekend\}$$

初始设定有：

- 当期变量不会影响滞后一期变量，即$\Pr(X_i|Z_i) = 0, X_i \in Var, Z_i \in Lag$。
- 本文关注的主要被解释变量是 PM2.5 浓度，因此在考虑其他变量对 PM2.5 浓度影响时有$\Pr(X_i|NZGPM25) = 0, X_i \in \{Var, Lag, Control\}$。
- 时间变量外生的，即$\Pr(X_i|Z_i) = 0, X_i \in Control, Z_i \in \{Var, Lag, NZGPM25\}$

可以构造初始化模型 BN1 如下图所示：

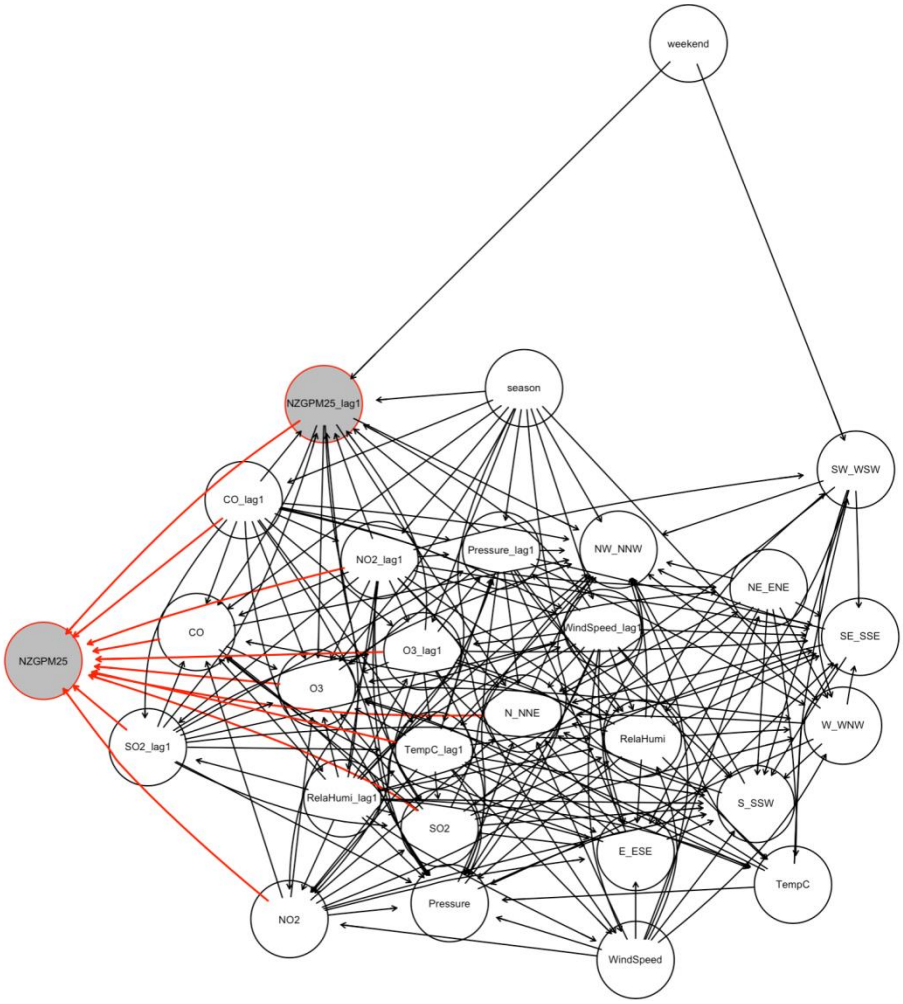

图 3.10：贝叶斯网络初始图 BN1

由上图可以发现，共有 11 个变量作为父节点指向被解释变量 PM2.5：$[NZGPM25|CO:NO2:O3:SO2:N\_NNE:NZGPM25\_lag1:CO\_lag1:NO2\_lag1:O3\_lag1:SO2\_lag1:TempC\_lag1]$。与回归不同，表 3.4 是从数据和图中学习到的 PM2.5 参数估计的条件概率分布：在 BN1 列中，PM2.5 浓度会随着污染物气体浓度，PM2.5 一期滞后项等增加而增加，北、北东北风对 PM2.5 浓度起降低作用。模型结构和参数学习结果在测试样本上得到的均方误差为 113.5584。但是根据图中结果，可以看出存在一些明显不符合实际结果的变量关系，如污染物气体浓度变量指向了风向变量，但是根据调查结果显示，风向主要与所属的半球和纬度位置，气流由高压地区吹向低压地区，再由所属的半球决定气流是向左偏转还是向右偏转。同时，也缺乏和研究调查结果相符的一些明显的变量关系，如风速的大小是影响到 PM2.5 的一个重要因素，随着风速的增大，PM2.5 的浓度会有一个较为明显的下降；东北和西北风向没有 PM2.5 的污染源，其中包含的颗粒物浓度比较低，因此会带来 PM2.5 的下降等；相对湿度较大的时候，可能会增加颗粒物的吸湿性，因此 PM2.5 的质量浓度应当和相对湿度呈现正比关系。

3.2.2 贝叶斯网络交互式下修正图

因为贝叶斯网络的构建可以通过“输入-加入先验概率-结构学习-参数学习-调整不合理变量关系-验证变量关系-输出”来进行变量选择和模型设定，实现将专家知识与数据学习相结合的交互，从而输出更稳定的最终图模型。上述初始图一共有 222 条有向边，且明显存在不合理的变量关系，因此我们进一步通过改变先验条件和模型约束，来探索一条逐步增加约束条件的变量选择优化路径。

（1）风速的大小，相对湿度会对 PM2.5 浓度产生较显著的影响。

$$\Pr(NZGPM25|Z_i) = 1, Z_i \in \{WindSpeed, RelaHumi\}$$

（2）风速、风向变量不会受到污染物气体变量的影响，相对于农展馆监测点是外生的。

$$\Pr(X_i|Z_i) = 0,$$

$$X_i \in \{WindSpeed, N_{NNE}, NE_{ENE}, E_{ESE}, SE_{SSE}, S_{SSW}, SW_{WSW}, W_{WNW}, NW_{NNW}, C\}$$

$$Z_i \in \{CO, NO2, O3, SO2, CO\_lag1, NO2\_lag1, O3\_lag1, SO2\_lag1, NZGPM25\_lag1\}$$

得到修正图 BN2 如下，共有 28 个节点，188 条边，其中 PM2.5 的父节点有 16 个：[NZGPM25|CO:NO2:O3:SO2:WindSpeed:RelaHumi:Pressure:N_NNE:NZGPM25_lag1:CO_lag1:NO2_lag1:O3_lag1:SO2_lag1:RelaHumi_lag1:TempC_lag1:Pressure_lag1]。从数据和图中学习到的 PM2.5 参数估计的条件分布为表 3.3 的 BN2 列：其中新增变量，相对湿度对 PM2.5 浓度影响为正，相对湿度滞后一期对 PM2.5 影响为负。气压对 PM2.5 浓度影响为负，气压滞后一期对 PM2.5 影响为负。且在测试样本上均方误差降为 107.8268。但是也有一些显然不符合之前所描述的变量关系，如风速的回归系数为正，这也就说明风速每增加 1m/s，会导致 PM2.5 的浓度增加 0.276 $\mu g/m^3$，这种现象与实际情况相悖，因此需要进一步调整概率图模型的结构特征，因为从图中可以看出，依然存在一些和实际情况不符合的变量关系，如 NO2,SO2 浓度对于温度的影响；SO2_lag1 对于气压的影响等。

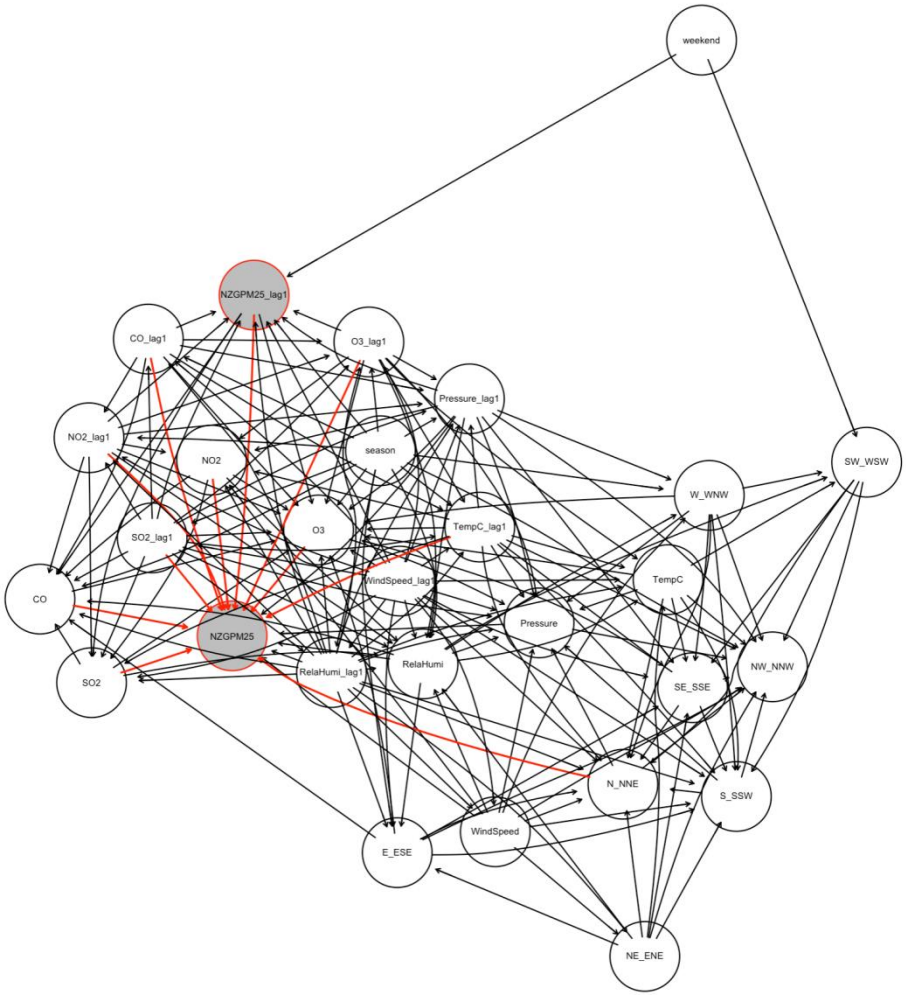

图 3.11：贝叶斯网络初始图 BN2

（3）温度、气压、相对湿度不会受到污染物气体和风向的影响。

$$\Pr(X_i|Z_i) = 0, X_i \in \{TempC, Pressure, RelaHumi\}$$

$$Z_i \in \left\{ \begin{matrix} CO, NO2, O3, SO2, CO_{lag1}, NO2_{lag1}, O3_{lag1}, SO2_{lag1}, NZGPM25_{lag1}, \\ N\_NNE, NE\_ENE, E\_ESE, SE\_SSE, S\_SSW, SW\_WSW, W\_WNW, NW\_NNW, C \end{matrix} \right\}$$

（4）温度、相对湿度与风速相对独立。

$$\Pr(X_i|Z_i) = 0, X_i \in \{TempC, RelaHumi\}$$

$$Z_i \in \{WindSpeed, WindSpeed_{lag1}\}$$

（5）风速滞后一期和温度对 PM2.5 浓度有影响。

$$\Pr(NZGPM25|Z_i) = 1, Z_i \in \{WindSpeed_{lag1}, TempC\}$$

得到修正图 BN3 如下，184 条边，其中 PM2.5 的父节点有 17 个：[NZGPM25|CO:NO2:O3:SO2:WindSpeed:RelaHumi:TempC:Pressure:N_NNE:NZGPM25_lag1:CO_lag1:NO2_lag1:O3_lag1:SO2_lag1:WindSpeed_lag1:RelaHumi_lag1:Pressure_lag1]。模型预测误差有轻微降低。其中新增变量，温度对 PM2.5 浓度影响为正，温度滞后一期被移除模型。但是风速及其滞后一期的回归系数为正，即风速增加会导致 PM2.5 浓度增

加，这种现象与实际观察情况相违背。考虑到以往研究中，O3 与 PM2.5 关系方向不明，其是一种与 PM2.5 并列的污染物，且其形成也会受到光照和温度的影响，而本身并非 PM2.5 的组成部分或是前体物，因此尝试阻隔 O3 变量的影响，考察一下概率图模型的结构特征变化。

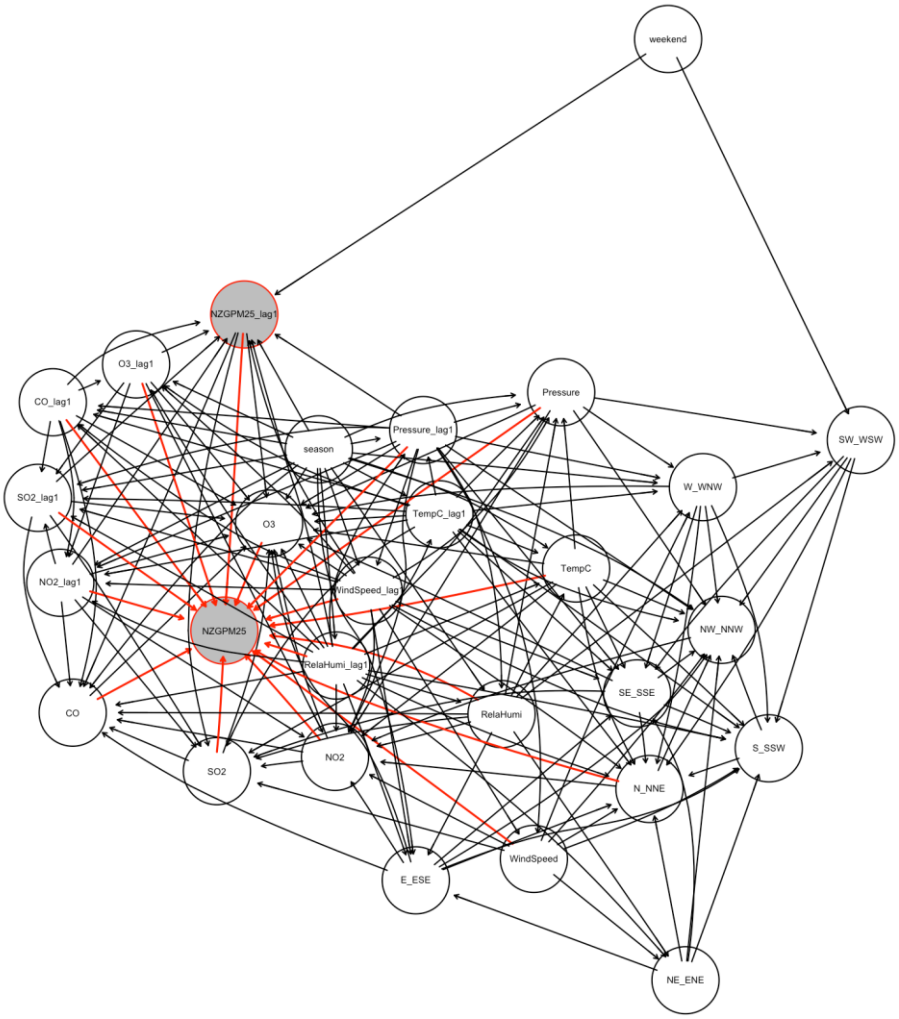

图 3.12：贝叶斯网络初始图 BN3

（6）尝试阻隔 O3 对 PM2.5 的影响

$$\Pr(NZGPM25|O3) = 0$$

得到修正图 BN4 如下，181 条边，其中 PM2.5 的父节点变成 14 个：[NZGPM25|CO:NO2:SO2:WindSpeed:RelaHumi:TempC:Pressure:NZGPM25_lag1:CO_lag1:NO2_lag1:SO2_lag1:WindSpeed_lag1:RelaHumi_lag1:Pressure_lag1]。模型预测误差也随之降低。其中风速滞后一期的回归系数为负，即这也就说明滞后一期的风已开始逐渐对 PM2.5 起到驱散作用，滞后风速每增加 1m/s，会导致 PM2.5 的浓度降低 0.081 $\mu g/m^3$。但是，之前的 N_NNE 被剔除模型，根据实际情况，东北和西北风向没有 PM2.5 的污染

源，其中包含的颗粒物浓度比较低，因此会带来 PM2.5 的下降等。因此尝试加入特定风向变量的影响，让数据和算法进一步学习，考察一下概率图模型的结构特征变化。

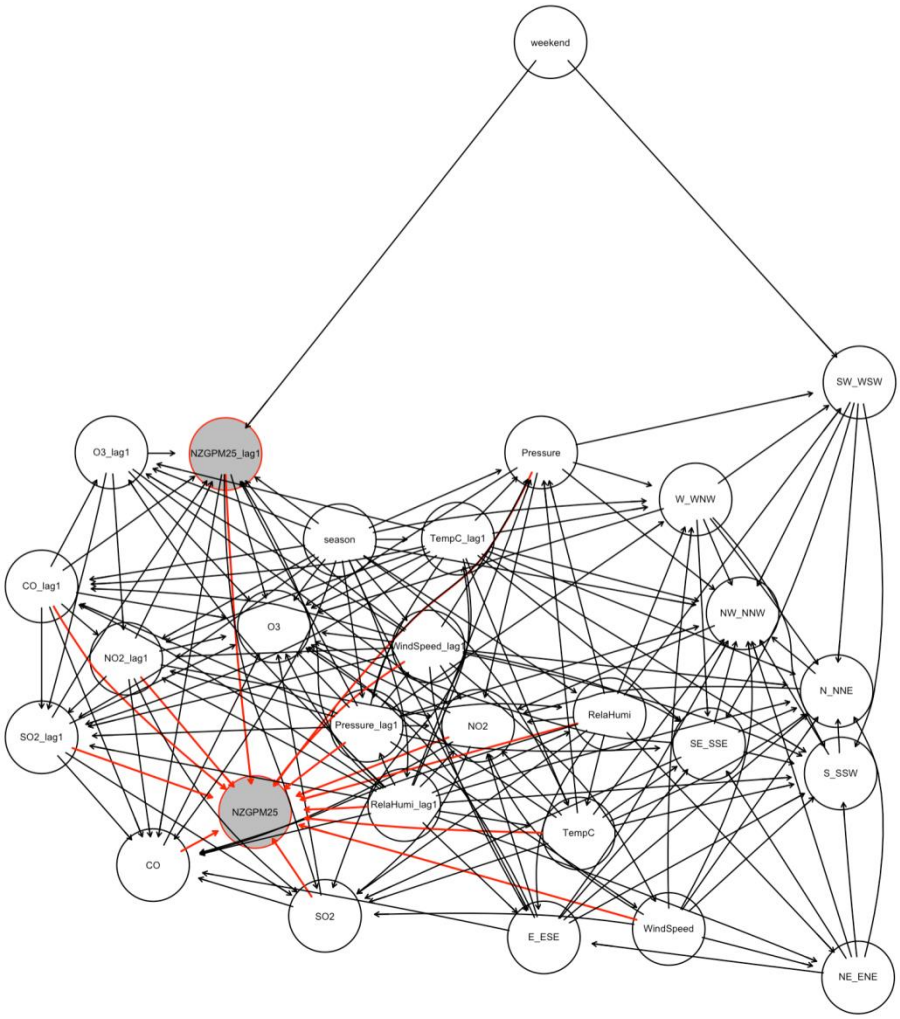

图 3.13：贝叶斯网络初始图 BN4

（7）考虑到东北和西北风向对 PM2.5 浓度产生较显著的影响，因此考虑增加约束：

$$\Pr(NZGPM25|Z_i) = 1, Z_i \in \{N\_NNE, W\_WNW, NW\_NNW\}$$

得到的最终概率图结构模型 BN5 如下图所示[4]：

[4] 最终图所有变量父节点见附录 B

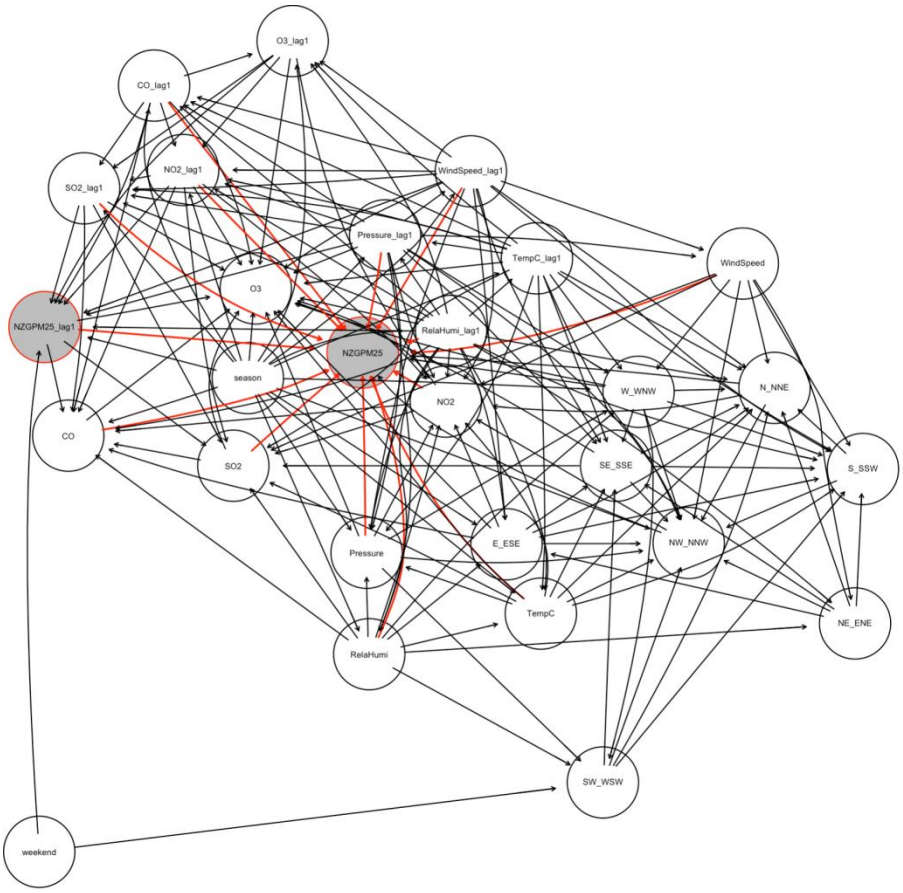

图 3.14：贝叶斯网络初始图 BN5

从最终贝叶斯网络图的结构和参数结果可以看出，修正不合理变量关系，从数据中学习得到 28 个节点，184 条边，其中 PM2.5 有 17 个父节点，新增加的东北和西北风向对 PM2.5 浓度起降低作用。并且经过先验调整，其学习的结构和参数，在测试样本上均方误差降低至 104.3824，预测能力比之前模型有提高。

由最终贝叶斯网络结构发现，在考虑影响 PM2.5 的因素时，主要有以下几种因素：SO2、NO2、CO、湿度、温度等都是 PM2.5 形成的前体物或催化条件，因此对 PM2.5 起正向作用。如相对湿度较大的时候，可能会增加颗粒物的吸湿性，从而有利于 PM2.5 的聚集，然而其滞后项起反作用。风速对 PM2.5 影响为正，但滞后一期影响为负。在现实观察中，在小风速的条件下，风速不仅利于 PM2.5 的扩散，同时小风速对空气有搅拌作用，利于气溶胶中 PM2.5 聚沉。而当风速过大时，空气容易呈现湍流特性，从而地面上的 PM2.5 更可能被吹起，同时更可能带来其他地区的沙土导致 PM2.5 的增高。与其他变量相比，风速的影响滞后性明显，滞后一期风速更容易起到驱散作用。东北和西北方向的风向对 PM2.5 起降低作用。对于北京来说，西部和北部没有污染源，污染程度较低，而京南为工厂聚居地，污染源较多，不同方向的风对 PM2.5 的扩散作用不同，因此从西北方来的风会更好地起到净化作用。

表 3.3：贝叶斯网络 PM2.5 节点参数估计的条件分布结果

	BN1	BN2	BN3	BN4	BN5
(Intercept)	-5.460	-87.063	-85.828	-118.340	-105.482
CO	27.609	27.182	27.170	29.064	28.999
NO2	0.398	0.409	0.407	0.222	0.220
O3	0.181	0.186	0.184	NA	NA
SO2	0.679	0.706	0.706	0.763	0.756
N_NNE	-1.320	-1.202	-1.200	NA	-1.427
NZGPM25_lag1	0.927	0.924	0.924	0.929	0.927
CO_lag1	-25.303	-25.067	-25.064	-26.799	-26.697
NO2_lag1	-0.292	-0.293	-0.291	-0.124	-0.120
O3_lag1	-0.169	-0.172	-0.170	NA	NA
SO2_lag1	-0.605	-0.615	-0.616	-0.661	-0.661
TempC_lag1	0.106	0.158	NA	NA	NA
WindSpeed	NA	0.276	0.271	0.241	0.342
RelaHumi	NA	11.103	13.577	8.483	7.580
Pressure	NA	-1.085	-1.053	-1.616	-1.582
RelaHumi_lag1	NA	-7.554	-10.046	-7.766	-6.715
Pressure_lag1	NA	1.161	1.129	1.725	1.679
TempC	NA	NA	0.156	0.238	0.213
WindSpeed_lag1	NA	NA	0.009	-0.081	-0.083
W_WNW	NA	NA	NA	NA	-1.521
NW_NNW	NA	NA	NA	NA	-0.910

表 3.4：交互式下贝叶斯网络

模型	假设	先验图设计	测试集 MSE
BN1		当期变量不影响滞后一期变量 关注主要被解释变量 PM2.5 浓度 时间变量外生	113.5584
BN2	1	风速、湿度会影响 PM2.5	107.8268
	2	风速、风向不受污染物气体的影响	
	3	温度、气压、湿度不受污染物气体和风向的影响	
BN3	4	湿度，温度与风速独立	107.6031
	5	风速滞后一期、温度会影响 PM2.5	
BN4	6	阻隔臭氧对 PM2.5 影响	106.2406
BN5	7	加入特定风向的影响	104.3824

3.3 先验设计路径讨论

3.3.1 延伸验证 1：其他时间点

根据 PM2.5 形成原理，借助先验知识和数据学习的交互式过程，可建立起 2016 年样本集的贝叶斯网络图，识别出 PM2.5 的主要影响因素，并将整个模型结构可视化表示。那么根据 2016 年样本学习得到的先验设计路径是否同样适用于其他年份？其他年份和样本学习是否可借助同样的路径得到优化？

将其用于 2015、2017 和 2018 年，结果如下：对于 2015 年和 2017 年，先验设计路径总体是有助于调整不合理关系，并降低测试集均方误差。但其中有关于风速和臭氧的设定需要进一步讨论，如 2015 年风速滞后一期的加入（BN3）使得 PM2.5 条件概率更符合现实设定，但是会短暂使得测试集的预测性能降低；同时，如果阻隔臭氧对 PM2.5 的影响（BN4），也会降低对 PM2.5 的预测准确性；2017 年中更是需要在贝叶斯网络图中允许臭氧对 PM2.5 的影响（BN4），以更好的调整图结构。由此发现，关于风速和臭氧的设计在 2016 年贝叶斯网络学习中属于探索性尝试，则在其他年份中不一定适用，而在 2016 年贝叶斯网络中属于确定性知识的设计，能够有效避免数据拟合，学习更合理的图结构，则在变量选择上能起到优化的作用。

对于 2018 年，先验设计路径虽然能够调整最终图结构体现的变量关系，但是最终图模型（BN5）整体预测性能相比于初始图（BN1）会降低。这是因为调整的先验设计，是加入专家知识来避免过拟合，当现实情况发生改变的时候，算法学习能够更快的调整参数，如与之前年份相比，2018 年 PM2.5 的父节点在风向、气压和温度上均有变化。2017 年 11 月 18 日，大兴区西红门镇新建二村发生重大火灾事故，造成 19 人死亡，8 人受伤。火灾发生后，2017 年 11 月 20 日起，北京市部署开展为期 40 天的安全隐患大排查大清理大整治专项行动。因此，2018 年数据会受到政策的冲击，与前几年可能存在系统性差异。

综上，这条先验设计路径在一定程度上起到优化作用，特别是确定性知识，但是需要根据样本的情况进行个性化的优化调整，这也正是贝叶斯的思路，即参数是可变的。

表 3.5：2015、2017 和 2018 年交互式下贝叶斯网络调整路径结果

样本	模型	先验图设计	测试集 MSE
2015	BN1	初始图	392.00
	BN2	风速、湿度会影响 PM2.5 风速、风向不受污染物气体的影响 温度、气压、湿度不受污染物气体和风向的影响	388.64
	BN3	湿度，温度与风速独立 风速滞后一期、温度会影响 PM2.5	388.91
	BN4	阻隔臭氧对 PM2.5 影响	410.67

	BN5	加入特定风向的影响	388.26
2017	BN1	初始图	145.90
	BN2	风速、湿度会影响 PM2.5 风速、风向不受污染物气体的影响	145.17
	BN3	温度、气压、湿度不受污染物气体和风向的影响 湿度，温度与风速独立 风速滞后一期、温度会影响 PM2.5	145.70
	BN4	臭氧会影响 PM2.5 影响	134.58
	BN5	加入特定风向的影响	132.78
2018	BN1	初始图	17.98
	BN2	风速、湿度会影响 PM2.5 风速、风向不受污染物气体的影响	18.44
	BN3	温度、气压、湿度不受污染物气体和风向的影响 湿度，温度与风速独立 风速滞后一期、温度会影响 PM2.5	18.96
	BN4	阻隔臭氧对 PM2.5 影响	18.96
	BN5	加入特定风向的影响	18.55

表 3.6：2015-2018 年贝叶斯网络最终图 PM2.5 节点参数估计的条件分布结果

	BN2015	BN2016	BN2017	BN2018
(Intercept)	-7.306	-105.482	-2.491	-2.243
CO	19.950	28.999	6.248	10.623
NO2	0.612	0.220	0.511	0.264
O3	0.262	NA	0.125	NA
SO2	0.756	0.756	0.731	0.925
WindSpeed	0.492	0.342	0.193	0.151
RelaHumi	26.186	7.580	1.377	12.468
TempC	0.041	0.213	0.020	0.413
N_NNE	-2.476	-1.427	-2.138	-1.005
W_WNW	-1.864	-1.521	-1.181	NA
NW_NNW	-2.943	-0.910	-1.670	-0.726
NZGPM25_lag1	0.934	0.927	0.944	0.927
CO_lag1	-18.403	-26.697	-5.334	-7.891
NO2_lag1	-0.508	-0.120	-0.450	-0.198
O3_lag1	-0.254	NA	-0.122	NA
SO2_lag1	-0.713	-0.661	-0.605	-0.811
WindSpeed_lag1	-0.009	-0.083	-0.084	-0.191
RelaHumi_lag1	-21.682	-6.715	NA	-13.545
Pressure	NA	-1.582	NA	NA
Pressure_lag1	NA	1.679	NA	NA
E_ESE	NA	NA	NA	1.076
S_SSW	NA	NA	NA	0.865

TempC_lag1	NA	NA	NA	-0.370

3.3.2 模型预测比较

Lasso（the least absolute shrinkage and selection operator）是一种压缩估计方法。它通过构造惩罚函数压缩一些系数，从而避免过度拟合，实现变量选择。因此我们以 Lasso 的结果与贝叶斯网络进行对比。首先在训练样本上，通过交叉检验选择使得均方误差最小的λ，得到估计的系数和节点，然后将结果在 100 个测试样本上计算其均方误差。

结果如下，与贝叶斯网络结果相似，在考虑影响 PM2.5 的因素时，主要有 SO2、NO2、CO、湿度、温度等 PM2.5 形成的前体物或催化条件，也存在 O3 的影响较小，当期风速影响为正，但滞后一期影响为负。尽管 Lasso 可以通过惩罚项来压缩系数，但是不能更好的结合先验知识和结构设计，因此会将更多的变量纳入模型，如过多风向变量。过多的变量对相互作用关系造成干扰，对训练数据过度拟合，降低模型预测精度，如表 3.8 所示，在交互式中，各个年份最终贝叶斯图在测试集上的均方误差更小。

表 3.7：2015-2018 年 LASSO 结果：系数和节点

	Lasso2015	Lasso2016	Lasso2017	Lasso2018
(Intercept)	-41.907	-70.466	-78.4417	0.6642
CO	20.327	27.436	6.2222	10.4343
NO2	0.556	0.355	0.4604	0.3385
O3	0.219	0.152	0.0897	0.0847
SO2	0.718	0.671	0.7167	0.859
WindSpeed	0.36	0.241	0.0823	0.1093
RelaHumi	21.014	5.595	2.3059	6.196
TempC	0.043	.	.	.
Pressure	.	.	.	-0.0024
N_NNE	-1.989	-1.573	-1.6644	-0.9151
NE_ENE	-0.076	.	-0.3536	.
E_ESE	1.632	0.173	0.8384	1.2741
SE_SSE	1.191	0.446	0.875	0.3757
S_SSW	.	0.129	0.5898	0.8476
SW_WSW	-0.058	-2.247	0.0022	0.5568
W_WNW	-1.033	-0.935	-0.3877	-0.021
NW_NNW	-2.347	-0.923	-0.8789	-0.4755
NZGPM25_lag1	0.93	0.922	0.9418	0.9217
CO_lag1	-18.62	-25.2	-5.1826	-7.4712
NO2_lag1	-0.449	-0.241	-0.3949	-0.2698
O3_lag1	-0.209	-0.14	-0.0877	-0.0819
SO2_lag1	-0.675	-0.582	-0.5907	-0.7534
WindSpeed_lag1	0.027	.	-0.0441	-0.1183
RelaHumi_lag1	-16.894	-3.515	-1.0899	-6.5027
TempC_lag1	.	0.155	0.0594	0.0033

Pressure_lag1	0.035	0.061	0.0749	.
weekend	-0.354	0.036	-0.1517	0.1566
season	-0.322	0.175	-0.3739	-0.3669

表 3.8：2015-2018 年 LASSO 与最终 BN 测试性能比较

	LASSO MSE	BN MSE
2015	392.0011	388.2615
2016	110.3558	104.3824
2017	135.8668	132.7823
2018	18.5553	18.5459

3.3.3 延伸验证 2：亦庄观测点

接下来，我们改变观测地点进行延伸讨论，探讨农展馆观测点样本数据学习得到的先验设计路径是否同样适用于其他观测点分析建模？能否提供对引向最终模型设定路径的优化？首先对 2015-2018 年亦庄观测点训练样本，根据先验设计路径依次进行贝叶斯结构和参数学习；其次，使用 lasso 算法做对比，通过交叉检验选择使得均方误差最小的λ_{min}和一个标准误的λ_{1se}，得到分别估计的系数和节点；最后将所有模型结果在 100 个测试样本上计算其均方误差作比较。

结果如下：（1）除了 2015 年之外，在其他年份，先验设计路径在一定程度上能够调整不合理变量关系，验证现实情况，降低预测的均方误差，如 2016 年均方误差从 716.204 降低到 712.282，2017 年均方误差从 367.059 降低到 361.536。但在先验设计调整的过程中，均方误差会存在小幅度的震荡，且一些不确定性关系可变，导致条件概率参数也会发生变化。例如即时风速和滞后一期风速，在 2015 年到 2017 年 PM2.5 会随其增加而增加，而在 2018 年 PM2.5 会随其增加而减少。由于 2017 年 12 月底开展的北京市部署开展为期 40 天的安全隐患大排查大清理大整治专项行动，2018 年受到政策冲击可能整体环境发生变化，北京周边大量污染企业外迁，此时条件变化已经发生变化，即参数是可变的，洁净的风更可能驱散 PM2.5。因此，在贝叶斯网络学习交互式过程中，加入专家先验知识，能更好的识别和调整变量关系，有助于及时察觉数据发生条件的变化。（2）与 Lasso 预测相比，在一些样本下，贝叶斯初始图学习结果可能不如λ_{min}下的 lasso 结果，但借助专业知识，调整先验设计后，贝叶斯最终图的学习结果明显更优，且贝叶斯学习过程更直观可解释。例如 2017 年中，初始图（BN1）均方误差为 367.059，高于 lasso 的 364.945，但经过先验设计路径调整，最终图结构（BN5）均方误差降低为 361.536。从表 3.10 可见，λ_{1se}下的 lasso 压缩系数能力更强，但可包含信息较少，预测误差较大；λ_{min}下的 lasso 压缩系数能力相对较弱，可能包含一些无关变量，也会对预测性能产生影响；而贝叶斯网络对于变量的选择介于两者之间，可利用实验室确定结论进行结构先验设计，既可避免无关变量，又可以根据数据学习来发现不确定性关系。

表 3.9：2015-2018 年亦庄观测点 LASSO 与 BN 测试性能比较

	Lasso		BN				
	λ_{min}	λ_{1se}	BN1	BN2	BN3	BN4	BN5
2015	478.827	534.935	438.449	490.695	478.703	478.703	478.409
2016	738.905	880.201	716.204	716.204	714.123	714.123	712.282
2017	364.945	447.741	367.059	367.109	363.345	363.345	361.536
2018	15.013	16.649	15.022	15.181	13.805	13.655	13.804

表 3.10：2015-2018 年亦庄观测点 LASSO 与 BN 最终图结果：系数和节点

	BN2015	Lasso2015		BN2016	Lasso2016		BN2017	Lasso2017		BN2018	Lasso2018	
		λ_{min}	λ_{1se}		λ_{min}	λ_{1se}		λ_{min}	λ_{1se}		λ_{min}	λ_{1se}
(Intercept)	-2.90	-13.65	-3.87	-8.24	-56.09	1.08	-3.79	-56.81	-3.06	-4.36	1.55	-3.67
CO	33.74	33.49	30.86	22.64	22.32	14.83	3.43	3.53	2.15	27.94	28.90	23.45
NO2	0.27	0.35	0.19	0.29	0.37	0.14	0.48	0.47	0.26	0.19	0.14	0.06
O3	.	0.11	0.00	.	0.11	.	0.13	0.13	.	0.06	0.03	.
SO2	0.39	0.35	0.14	0.80	0.77	0.53	1.03	1.02	0.65	0.83	0.79	0.33
WindSpeed	0.32	0.15	.	0.76	0.57	.	0.16	0.07	-0.07	-0.06	-0.09	-0.24
RelaHumi	54.19	18.40	2.85	65.97	33.97	6.46	60.16	41.70	4.47	42.46	12.21	2.06
TempC	2.67	0.03	.	2.39	0.13	.	1.59	0.53	.	1.41	0.00	.
Pressure	.	.	.	.	.	-0.01	.	-0.07	.	.	-0.01	.
N_NNE	-4.28	-0.27	-0.63	-1.51	-0.53	-0.07	-1.82	-0.17	-1.32	-0.09	-0.67	-0.28
NE_ENE	-4.21	.	-0.03	-2.68	-1.71	-0.51	-1.90	-0.20	-0.81	.	-0.60	.
E_ESE	-2.94	1.03	.	.	-0.49	.	.	1.27	.	-0.03	-0.33	.
SE_SSE	.	3.29	2.05	.	0.42	.	.	1.71	.	.	0.19	.
S_SSW	.	4.23	3.63	.	1.94	1.56	1.42	3.15	1.93	1.21	0.84	0.78
SW_WSW	.	4.07	3.55	.	1.21	0.98	.	2.13	1.07	2.18	1.79	1.78
W_WNW	-4.32	-0.02	.	-0.99	0.01	.	-0.80	0.91	.	.	-0.62	.
NW_NNW	-4.75	-0.34	-0.68	-1.20	.	.	-1.87	.	-1.07	0.50	.	.
NZGPM25_lag1	0.94	0.93	0.92	0.92	0.92	0.89	0.94	0.94	0.92	0.93	0.92	0.90
CO_lag1	-31.80	-31.26	-27.81	-22.02	-21.64	-13.86	-2.73	-2.67	-0.75	-25.48	-26.05	-19.35
NO2_lag1	-0.18	-0.26	-0.10	-0.18	-0.25	0.00	-0.42	-0.41	-0.20	-0.14	-0.09	.
O3_lag1	.	-0.11	.	.	-0.11	0.01	-0.11	-0.12	0.00	-0.05	-0.02	0.01
SO2_lag1	-0.37	-0.32	-0.11	-0.62	-0.58	-0.34	-0.92	-0.90	-0.50	-0.64	-0.57	-0.12
WindSpeed_lag1	0.31	0.18	.	0.26	0.23	.	0.08	0.00	.	-0.06	-0.04	.
RelaHumi_lag1	-50.91	-14.68	.	-59.36	-25.76	.	-54.80	-36.04	.	-39.28	-9.28	.
TempC_lag1	-2.65	-0.02	.	-2.27	.	0.04	-1.62	-0.54	.	-1.41	.	.
Pressure_lag1	.	0.01	.	.	0.05	.	.	0.12	.	.	.	.
weekend	.	-0.13	.	.	0.36	.	.	-0.06	.	.	0.13	.
season	.	-0.32	-0.22	.	-0.29	-0.11	.	-0.37	-0.04	.	0.01	.

第四章 结论与讨论

本文对图模型进行理论综述探讨，并将其用于北京 PM2.5 影响因素研究，遵循“输入-加入先验条件-结构学习-参数学习-调整不合理变量关系-验证变量关系-输出”的步骤重复循环，来寻找最精简且最合适的模型，来验证自然科学领域中的理论关系，并降低模型预测误差。

在一个完整的贝叶斯网络中，节点表示随机变量，节点之间的有向边表示随机变量的相互依赖关系，且每一个节点都附有一个条件概率分布。数据中随机变量的联合分布，可通过链式法则，简化成条件概率连乘的形式，从而得到分布的分解，实现贝叶斯网络的紧凑表示。因此，贝叶斯网络可以反应因子化和独立性的双重视角：一是因子化。作为一种数据结构，告诉我们一个概率分布如何被分解成一系列因子或者条件概率分布的集合，被图所表示。二是独立性。图结构及其图中的独立性如何可以被概率分布所满足。换言之，如果有一个能被某个贝叶斯网络表示的概率分布，我们可以直接通过了解图的参数来知道分布中的独立性，进而可以知道分布的结构，分布的变化，以及不同观测所带来的变量关系的影响。

从数据中学习到贝叶斯网络结构主要包含参数学习和结构学习。参数学习过程是在网络结构已知的情况下，从数据中学习随机变量的条件概率分布。条件概率分布的参数模型已预先指定，只需估计其中的参数，而极大似然估计和贝叶斯方法是最常用的两种参数学习方法。极大似然估计把待估参数看作取值未知的确定性量，依据参数与数据集的似然程度，来选择使似然函数值最大的参数值作为学习的结果。贝叶斯估计是基于贝叶斯公式，参数是可变的随机变量，根据样本信息修正先验信息，由先验知识和观察到的数据集共同决定不确定性的概率参数。在网络结构未知情况下，用贝叶斯网络进行结构学习，目的在于构建一个有向无环图，用于表示随机变量间概率依赖关系和条件独立关系。基于评分搜索的学习算法视为结构优化问题，主要是利用得分函数评价网络结构优劣，然后用搜索算法来寻找出分数最高的最优结构以优化，其可以把专家经验知识以结构先验概率分布的形式融入到过程中。概率图模型允许我们将先验知识整合到模型中，这是许多其他算法所不具备的。因此，基于概率图交互式特性，我们可以建立起一条规范化学习路径：输入观测数据，结合专家经验对确定性部分进行先验结构设计，利用统计算法对不确定性部分进行学习，在训练集进行交叉验证来选择图结构和超参数，在测试集上评估性能，探讨错误分析和不合理变量关系，并重新设计模型、目标函数或优化算法来解决问题的过程。在这个过程中，并不是依赖于纯数据挖掘的自动统计模式选择，而需要将其他已有信息加入模型结构中，给出更好的学习和估计。

在北京 PM2.5 影响因素的贝叶斯网络结构学习过程中，首先使用农展馆观测点 2016 年训练样本，通过改变先验条件和模型约束，记录每次模型改变约束后的结果，并测试样本评估性能，尝试建立一条贝叶斯网络学习的“路径”。在构造初始化图时，存在不合理的变量关系对参数估计造成干扰，如风速符号为正，污染物影响风向变量等。这也是本文逐步加入先验设计，避免数据过拟合，探讨优化模型的动因。因此，在结合自然科学领域相关研究和已有成果的基础上，对环境数据变量中确定性关系，通过黑白名单设计加入初始结构设计，再通过算法学习不确定性关系，得到最终图模型, 预测性能得到了提高。在 PM2.5 的影响因素中，风向变量从九种筛选到三种；且删去了无关解释变量 O3；瞬时风速为正，滞后一期风速为负，说明 PM2.5 的驱散需要一定时间，瞬时风速会先带来其他方位的污染物，造成其浓度的上升，随后带来的洁净空气才能起到净化作用；SO2、NO2、CO、湿度、温度等 PM2.5 形成的前体物或催化条件有利于 PM2.5 的聚集。此外，图中其他节点的变量关系也符合实际情况，如湿度受到其滞后项，气压和季节等影响；温度受到湿度，季节等影响；风向也受到其他类型风向、风速等影响，同时风速会影响到氮氧化物的浓度，且风速越大，氮氧化物的浓度越低。

在先验设计路径讨论中，本文首先在时间维度上验证，将农展馆观测点 2016 年的先验设计路径运用于其他年份样本。对于 2015 年和 2017 年，先验设计路径总体有助于调整不合理关系，并降低测试集均方误差。但受到政策的冲击的 2018 年，与前几年相比可能存在系统性差异，先验设计路径虽然能够调整最终图结构体现的变量关系，但是最终图模型整体预测性能会低于初始图。进而本文在地点维度上验证，将先验设计路径用于亦庄观测点的分析，发现在先验设计调整的过程中，均方误差会存在小幅度的震荡，且一些不确定性关系可变，导致条件概率参数发生变化。与 Lasso 预测相比，在一些样本下，贝叶斯初始图学习结果可能不如λ_{min}下的 lasso 结果，但借助专业知识，调整先验设计后，贝叶斯最终图的学习结果明显更优，且贝叶斯学习过程更直观可解释。综上，这条先验设计路径在一定程度上起到优化作用，特别是确定性知识，但是需要根据样本的情况进行个性化的优化调整，这也正是贝叶斯的思路，即参数是可变的。同时在与 Lasso 结果比较中，结合先验知识和结构信息的贝叶斯网络，能够有效进行变量选择，提高模型预测性能，并能提供可视化解释。贝叶斯网可以将模型和算法分离开，从领域专家那里，通过知识经验直接抽象出重要关系，加入先验结构设计；从数据中，通过统计学机器学习算法模型，学习数据中不确定性部分；两者结合，这种不依赖于全自动选择而融合已有知识的方式，能给出更好的学习和估计，使我们能更准确地把握现实世界的规律。

参考文献

[1] Guyon I, Elisseeff A. An introduction to variable and feature selection[J]. Journal of machine learning research, 2003, 3(Mar): 1157-1182.

[2] Kennedy P. A guide to econometrics[M]. MIT press, 2008.

[3] Efron, Trevor Hastie. Computer Age Statistical Inference: Algorithms, Evidence, and Data Science, 2016

[4] Hernan M A, Robins J M. Causal inference[J]. 2010.

[5] 崔雍浩,商聪,陈锶奇,郝建业. 人工智能综述:AI 的发展[J]. 无线电通信技术,2019,45(03):225-231.

[6] 李世锋. 大数据时代人工智能在计算机网络技术中的应用[J]. 电子技术与软件工程,2017(23):259.

[7] 万赟. 从图灵测试到深度学习:人工智能 60 年[J]. 科技导报,2016,34(07):26-33.

[8] Sejnowski T J. The deep learning revolution[M]. MIT Press, 2018.

[9] Larrañaga P, Moral S. Probabilistic graphical models in artificial intelligence[J]. Applied soft computing, 2011, 11(2): 1511-1528.

[10] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann, San Mateo, California, 1988.

[11] J. Pearl, Reverend Bayes on inference engines: a distributed hierarchical approach, AAAI, 1982, pp. 133–136.

[12] Wainwright M J, Jordan M I. Graphical models, exponential families, and variational inference[J]. Foundations and Trends® in Machine Learning, 2008, 1(1–2): 1-305.

[13] Madigan D. Bayesian data mining for health surveillance[J]. Spatial and syndromic surveillance for public health, 2005: 203-221.

[14] Lucas P. Bayesian analysis, pattern analysis, and data mining in health care[J]. Current opinion in critical care, 2004, 10(5): 399-403.

[15] 李小琳，何湘东.面向大数据决策的概率图模型研究与应用[M]. 南京大学出版社, 2017.

[16] Koller D, Friedman N. Probabilistic graphical models: principles and techniques[M]. MIT press, 2009.

[17] Scutari M, Denis J B. Bayesian networks: with examples in R[M]. CRC press, 2014.

[18] Nagarajan R, Scutari M, Lèbre S. Bayesian networks in R[J]. Springer, 2013, 122: 125-127.

[19] 北京空气质量: http://zx.bjmemc.com.cn/;

[20] 中国天气: http://www.weather.com.cn/air/

[21] Scutari M. Learning Bayesian networks with the bnlearn R package[J]. arXiv preprint arXiv:0908.3817, 2009.

[22] Bellot D. Learning probabilistic graphical models in R[M]. Packt Publishing Ltd, 2016.

[23] 冯婧. 空气污染的统计和计量经济学实证研究——数据挖掘北京 PM2.5 浓度的主要影响因素[D].北京大学,2019.

[24] 刘保献, 杨懂艳, 张大伟等. 北京城区大气 PM2.5 主要化学组分构成研究[J]. 环境科学, 2015, 36(7): 2346-2352.

[25] 朱光磊，张远航，曾立民等.北京市大气细颗粒物 PM2.5 的来源研究[J]. 环境科学研究，2005，18(5):1-5.

[26] 陈添，华蕾，金蕾等.北京市大气 PM10 源解析研究[J]. 中国环境监测，2006，22(6)：59-63

[27] Tibshirani R. Regression shrinkage and selection via the lasso[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1996, 58(1): 267-288.

附录 A 描述性统计结果

表 A.1：主要被解释变量及解释变量的描述性统计结果（2015 年）

Year	Var	N	Mean	Sd	Min	Max
	weekend	8448	1.29	0.45	1	2
	season	8448	2.48	1.12	1	4
	NZGPM25	8448	85.24	90.77	1	667
	CO	8448	1.38	1.43	0.1	17
	NO2	8448	58.73	38.57	3	265
	O3	8448	61.13	59.12	2	364
	SO2	8448	15.87	20.04	2	168.5
	WindSpeed	8448	2.83	2.14	0	16
	RelaHumi	8448	0.58	0.26	0.06	1
	TempC	8448	13.39	11.54	-12	40
	Pressure	8448	1016.75	10.36	992	1041
	N_NNE	8448	0.21	0.41	0	1
	NE_ENE	8448	0.07	0.25	0	1
	E_ESE	8448	0.10	0.30	0	1
2015	SE_SSE	8448	0.11	0.31	0	1
	S_SSW	8448	0.11	0.31	0	1
	SW_WSW	8448	0.02	0.14	0	1
	W_WNW	8448	0.05	0.22	0	1
	NW_NNW	8448	0.10	0.31	0	1
	C	8448	0.23	0.42	0	1
	NZGPM25_lag1	8448	85.19	90.79	1	667
	CO_lag1	8448	1.38	1.43	0.1	17
	NO2_lag1	8448	58.68	38.57	3	265
	O3_lag1	8448	61.26	59.29	2	364
	SO2_lag1	8448	15.87	20.04	2	168.5
	WindSpeed_lag1	8448	2.83	2.14	0	16
	RelaHumi_lag1	8448	0.58	0.26	0.06	1
	TempC_lag1	8448	13.39	11.54	-12	40
	Pressure_lag1	8448	1016.74	10.36	992	1041

表 A.2：主要被解释变量及解释变量的描述性统计结果（2016 年）

Year	Var	N	Mean	Sd	Min	Max
	weekend	8241	1.29	0.46	1	2
	season	8241	2.49	1.12	1	4
	NZGPM25	8241	74.78	77.13	3	566
	CO	8241	1.20	1.16	0.1	9.4
	NO2	8241	52.49	32.72	2	206
	O3	8241	59.94	57.97	2	390
	SO2	8241	11.72	14.62	2	187
	WindSpeed	8241	2.99	2.23	0	14
	RelaHumi	8241	0.56	0.25	0.08	1
	TempC	8241	12.99	12.13	-16	37
	Pressure	8241	1016.86	10.58	994	1046
	N_NNE	8241	0.20	0.40	0	1
	NE_ENE	8241	0.06	0.24	0	1
	E_ESE	8241	0.09	0.29	0	1
2016	SE_SSE	8241	0.12	0.33	0	1
	S_SSW	8241	0.10	0.30	0	1
	SW_WSW	8241	0.01	0.12	0	1
	W_WNW	8241	0.05	0.22	0	1
	NW_NNW	8241	0.13	0.33	0	1
	C	8241	0.23	0.42	0	1
	NZGPM25_lag1	8241	74.74	77.12	3	566
	CO_lag1	8241	1.20	1.16	0.1	9.4
	NO2_lag1	8241	52.49	32.72	2	206
	O3_lag1	8241	60.08	58.04	2	390
	SO2_lag1	8241	11.72	14.62	2	187
	WindSpeed_lag1	8241	2.99	2.23	0	14
	RelaHumi_lag1	8241	0.56	0.25	0.08	1
	TempC_lag1	8241	13.00	12.13	-16	37
	Pressure_lag1	8241	1016.86	10.58	994	1046

表 A.3：主要被解释变量及解释变量的描述性统计结果（2017 年）

Year	Var	N	Mean	Sd	Min	Max
	weekend	8732	1.29	0.45	1	2
	season	8732	2.49	1.12	1	4
	NZGPM25	8732	60.55	69.06	2	835
	CO	8732	0.99	0.99	0	11.4
	NO2	8732	48.89	29.92	2	194
	O3	8732	57.85	57.08	1	393
	SO2	8732	9.28	12.13	1	257
	WindSpeed	8732	2.96	2.30	0	17
	RelaHumi	8732	0.54	0.27	0.05	1
	TempC	8732	13.54	11.92	-14	39
	Pressure	8732	1016.81	10.29	993	1041
	N_NNE	8732	0.19	0.39	0	1
	NE_ENE	8732	0.06	0.24	0	1
	E_ESE	8732	0.10	0.30	0	1
2017	SE_SSE	8732	0.12	0.32	0	1
	S_SSW	8732	0.11	0.31	0	1
	SW_WSW	8732	0.02	0.15	0	1
	W_WNW	8732	0.06	0.23	0	1
	NW_NNW	8732	0.11	0.31	0	1
	C	8732	0.23	0.42	0	1
	NZGPM25_lag1	8732	60.60	69.22	2	835
	CO_lag1	8732	1.00	0.99	0	11.4
	NO2_lag1	8732	48.90	29.92	2	194
	O3_lag1	8732	57.85	57.08	1	393
	SO2_lag1	8732	9.28	12.13	1	257
	WindSpeed_lag1	8732	2.96	2.29	0	17
	RelaHumi_lag1	8732	0.54	0.27	0.05	1
	TempC_lag1	8732	13.54	11.92	-14	39
	Pressure_lag1	8732	1016.81	10.29	993	1041

表 A.4：主要被解释变量及解释变量的描述性统计结果（2018 年）

Year	Var	N	Mean	Sd	Min	Max
	weekend	8432	1.28	0.45	1	2
	season	8432	2.49	1.11	1	4
	NZGPM25	8432	52.22	50.36	1	399
	CO	8432	0.89	0.59	0.1	5.2
	NO2	8432	45.81	30.13	2	245
	O3	8432	61.02	56.91	1	342
	SO2	8432	7.08	6.80	1	70
	WindSpeed	8432	2.87	2.20	0	16.9444444
	RelaHumi	8432	0.53	0.26	0.06	1
	TempC	8432	13.16	12.72	-15	41
	Pressure	8432	1012.64	10.65	990.9	1042.7
	N_NNE	8432	0.19	0.40	0	1
	NE_ENE	8432	0.06	0.24	0	1
	E_ESE	8432	0.10	0.30	0	1
2018	SE_SSE	8432	0.13	0.33	0	1
	S_SSW	8432	0.09	0.29	0	1
	SW_WSW	8432	0.01	0.11	0	1
	W_WNW	8432	0.07	0.25	0	1
	NW_NNW	8432	0.11	0.31	0	1
	C	8432	0.24	0.43	0	1
	NZGPM25_lag1	8432	52.22	50.36	1	399
	CO_lag1	8432	0.89	0.59	0.1	5.2
	NO2_lag1	8432	45.80	30.14	2	245
	O3_lag1	8432	61.05	56.94	1	342
	SO2_lag1	8432	7.08	6.80	1	70
	WindSpeed_lag1	8432	2.87	2.20	0	16.9444444
	RelaHumi_lag1	8432	0.53	0.26	0.06	1
	TempC_lag1	8432	13.16	12.72	-15	41
	Pressure_lag1	8432	1012.64	10.65	990.9	1042.7

附录 B 贝叶斯网络结果

表 B. 1：贝叶斯网络最终学习结构（2016 年）

Variable	NZGPM25	CO	NO2	O3	SO2	NZGPM25_lag1	NW_NNW	S_SSW	N_NNE
Parents	CO	NO2	WindSpeed	CO	NO2	CO_lag1	WindSpeed	WindSpeed	WindSpeed
	NO2	SO2	RelaHumi	NO2	WindSpeed	NO2_lag1	TempC	TempC	TempC
	SO2	RelaHumi	TempC	SO2	RelaHumi	O3_lag1	Pressure	NE_ENE	NE_ENE
	WindSpeed	E_ESE	Pressure	TempC	TempC	SO2_lag1	N_NNE	E_ESE	E_ESE
	RelaHumi	NZGPM25_lag1	N_NNE	Pressure	SE_SSE	WindSpeed_lag1	NE_ENE	SE_SSE	SE_SSE
	TempC	CO_lag1	E_ESE	E_ESE	NZGPM25_lag1	RelaHumi_lag1	E_ESE	SW_WSW	S_SSW
	Pressure	NO2_lag1	NW_NNW	SE_SSE	NO2_lag1	Pressure_lag1	SE_SSE	W_WNW	SW_WSW
	N_NNE	SO2_lag1	NO2_lag1	W_WNW	SO2_lag1	weekend	S_SSW	WindSpeed_lag1	W_WNW
	W_WNW	RelaHumi_lag1	O3_lag1	NZGPM25_lag1	TempC_lag1	season	SW_WSW	RelaHumi_lag1	WindSpeed_lag1
	NW_NNW	season	WindSpeed_lag1	CO_lag1			W_WNW	TempC_lag1	RelaHumi_lag1
	NZGPM25_lag1		RelaHumi_lag1	NO2_lag1			TempC_lag1		TempC_lag1
	CO_lag1		Pressure_lag1	O3_lag1			Pressure_lag1		
	NO2_lag1			SO2_lag1			season		
	SO2_lag1			WindSpeed_lag1					
	WindSpeed_lag1			RelaHumi_lag1					
	RelaHumi_lag1			TempC_lag1					
	Pressure_lag1			Pressure_lag1					
				season					

Variable	NE_ENE	E_ESE	SE_SSE	W_WNW	Pressure	CO_lag1	NO2_lag1	O3_lag1	SO2_lag1

Parents	WindSpeed RelaHumi RelaHumi_lag1	RelaHumi NE_ENE WindSpeed_lag1 RelaHumi_lag1 season	TempC NE_ENE E_ESE SW_WSW W_WNW WindSpeed_lag1 TempC_lag1 Pressure_lag1	WindSpeed RelaHumi Pressure TempC_lag1 Pressure_lag1 season	RelaHumi TempC WindSpeed_lag1 RelaHumi_lag1 TempC_lag1 Pressure_lag1 season	WindSpeed_lag1 RelaHumi_lag1 TempC_lag1 Pressure_lag1 season	CO_lag1 O3_lag1 WindSpeed_lag1 RelaHumi_lag1 TempC_lag1 Pressure_lag1 season	CO_lag1 WindSpeed_lag1 RelaHumi_lag1 TempC_lag1 season	CO_lag1 NO2_lag1 O3_lag1 WindSpeed_lag1 RelaHumi_lag1 TempC_lag1 Pressure_lag1 season

Variable	SW_WSW	WindSpeed	RelaHumi	TempC	WindSpeed_lag1	RelaHumi_lag1	TempC_lag1	Pressure_lag1
Parents	RelaHumi Pressure W_WNW weekend	WindSpeed_lag1 Pressure_lag1	RelaHumi_lag1 Pressure_lag1 season	RelaHumi RelaHumi_lag1 TempC_lag1 season	Pressure_lag1 season	season	RelaHumi_lag1 season	RelaHumi_lag1 TempC_lag1 season

附录 C 其他模型预测比较

表 C.1：2015-2018 年 BN 和 LASSO 联合结果：系数和节点

	BNLasso2015	BNLasso2016	BNLasso2017	BNLasso2018
(Intercept)	-6.806	-83.174	-2.054	-2.092
CO	20.316	28.623	6.217	10.441
NO2	0.549	0.197	0.454	0.243
O3	0.214		0.085	
SO2	0.723	0.669	0.718	0.842
WindSpeed	0.402	0.128	0.095	
RelaHumi	20.958	0.945	0.735	0.778
TempC	0.034	0.198	0.025	
Pressure				
N_NNE	-2.514	-1.714	-2.074	-1.101
NE_ENE				
E_ESE				1.068
SE_SSE				
S_SSW				0.777
SW_WSW				
W_WNW	-1.72	-1.224	-0.967	
NW_NNW	-2.915	-1.107	-1.464	-0.756
NZGPM25_lag1	0.931	0.923	0.942	0.922
CO_lag1	-18.661	-26.117	-5.209	-7.503
NO2_lag1	-0.444	-0.101	-0.393	-0.175
O3_lag1	-0.205		-0.083	
SO2_lag1	-0.682	-0.572	-0.592	-0.725
WindSpeed_lag1	.		-0.053	-0.069
RelaHumi_lag1	-16.95	-1.107		-1.927
TempC_lag1				0.039
Pressure_lag1		0.076		
weekend				
season				

表 C.2 2015-2018 年不同模型测试性能比较

	LASSO MSE	BN MSE	BN+LASSO MSE	Ramdom Forest MSE
2015	392.0011	388.2615	394.6028	625.0908
2016	110.3558	104.3824	111.4929	228.0438
2017	135.8668	132.7823	138.7527	144.5595
2018	18.5553	18.5459	18.1774	13.7194

致谢

丁酉年春，本科别离在即，自叹七幸以记之。三载过后，今硕士学业将结，顾往昔，余虽鲁钝，尚以勤勉，故得稍有累积。离日且近，多有不舍，提笔小抒，念初心未改，校理旧文，以赋新绪。

一幸择道。三年所学唯悟“研究”之真味，学问如浩瀚宇宙，终其本心，方有存得。吾念之真学者，乃兼具冷眼热心。时疏离于世，勿以情利压理智；亦心系社稷，愿以纸笔照人心。世事洞明，人情练达，梦寐思之，心向往之。是年秋将赴美求学，愿终有所成。

二幸家人相伴。爱之亲之，养之育之，责之教之。吾家倡和而不同，虽有争辩也求同存异。自高中起，母之书信六十有余，言词真切乃日常小悟，不盼几多成功，但念女真实自由，寻心之所幸。去年十月抵美，遇疫归期未定，父母虽顾念，然言语怯怯，恐增吾之虑，甚每日手书十六福，以祈平安。

三幸得与良师。恩师之提携，授东坡八面受敌之法，学术亦如领军作战，重在择何城而为，攻则兵集一点可破，守则八方受敌不败。愿吾勿生余念，低调谦逊，以板凳坐得十年冷之心，日积月累，切磋琢磨，未尝无可期之来日。学术之趣始于京晶吾师，幸蒙其教，得见识之敏锐，学问之意趣，生活之雅致。海外求学之困，其惦念备至，闻其言之鼓励，吾泣不成声，甚冬假赴美，亲临教导，乃今同困于此处。其关爱至致，今思惭之，愧不能对。宏举老师之恩言传身教，求学之事，事无巨细，皆言语谆谆。留学之时，求教于亦青老师，常获益良多，暗自叹服。岔口抉择，又得屹天老师之提点，少弯路。三年遇数位良师大家，均记之念之，奈无一一列举，愿愚生终有所成，不负厚恩。

四幸新知旧雨存长久。幸与吴君，相识相知十余年，暂别重逢仍牵挂，两眼相看无需答。谈天下论时事，聊人情说世故，虽相隔万里，情仍在意长存。幸与波崽，虽重洋远隔，学业之难，未来之思，生活之惑，感情之困，皆伴之扶之。幸与雍钊，识于求学途，志趣相投，共历申请关，终待花开，其神思敏捷，善规划决断，乃同辈之师。幸遇青，悲不离，喜不妒，学术之难，海外之途，疫期之哀，得其伴之，幸甚何求？

五幸难时终有助。师门之下，皆为忠正纯良之辈，但有所托，均施援手。舍友佩瑜，亚杰，陶源相伴三载，病榻关怀，远洋惦念，意难忘之。五层之谊，乃三高砥砺，共伴曦光至夜半。人生磕绊，幸得师兄姐之提点，同窗之善待，友人之鼓励，方能回首向来萧瑟处，也无风雨也无晴，心之感念，不胜言表。

六幸觅得良人归。齐君之呵护，不辞琐碎，虽远洋相隔，所历一二事，亦如影相随。吾之决定，虽路漫漫，皆全力支持，与吾同求索之。碌碌一生，得一知己，一佳偶，俱为上苍垂怜之恩赐。

七幸千帆过尽，归来仍是少年心。燕园情，千千结，问少年心事。对知识的追求，对爱的渴望，对未知的敬畏，愿始终保持着对生命单纯而强烈的激情。

离别在即，不胜感慨，然点滴在心，终不悔三年路。

北京大学学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名：　　日期：　　年　　月　　日

学位论文使用授权说明

（必须装订在提交学校图书馆的印刷本）

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

- 按照学校要求提交学位论文的印刷本和电子版本；
- 学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务，在校园网上提供服务；
- 学校可以采用影印、缩印、数字化或其它复制手段保存论文；
- 因某种特殊原因需要延迟发布学位论文电子版，授权学校□一年/□两年/□三年以后，在校园网上全文发布。

（保密论文在解密后遵守此规定）

论文作者签名：　　导师签名：

日期：　　年　　月　　日