北京大学

硕士研究生学位论文

题目：中国上市基金绩效评价
—基于概率图模型（PGM）

姓　　名：	苏熊
学　　号：	1401214422
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

二〇一七 年 四 月

版权声明

摘要

某某问题是……

本文[①]采用了……

研究表明……

关键词：关键词 1，关键词 2，关键词 3……

① 本研究得到某某基金（编号：XXX）资助。

ENGLISH TITLE

Author Name (Major)
Directed by your Supervisor

ABSTRACT

In environmental economics, environmental resources including environmental quality are categorized as amenity resources. Due to its importance to human welfare, the amenity resources theoretical study and valuation is an ongoing issue at the academic frontier in the environmental economics area.

KEY WORDS: Key word 1, Key word 2, Key word 3, ……

目录

第一章　引言

1.1　选题背景及意义

中国公募基金行业自 1998 年以来不断发展壮大，如今成为了除股票和债券等传统二级市场资产外另外一大资产配置方式。实际上，根据 wind 统计数据，截止 2016 年 12 月 31 日，我国公募基金共计 3820 只，资产净值超过 9 万亿。从图 1 也可以看到，从 2010 年上半年至 2016 年底，中国基金净值和总数均快速上升。众所周知，基金是由基金经理及其团队负责运营，该团队往往拥有相对丰富的市场经验、专业知识以及交易技巧，能够在风险可控的情况下获取市场超额收益。另一方面，公募基金形式多样，风险收益水平在不同层级都有相应类型基金对应，能够满足不同风险厌恶程度投资者的需求，成为广大散户和机构投资者的一个重要选择。然而，由于市场上的公募基金往往投资风格迥异，投资经理水平参差不齐，因此在选择基金时，中小投资者一般只能够以基金公司发行的该只基金的推介宣传资料为判断投资的依据，而无法真正得到对于基金的判断。因此，本文认为科学、全面地对公募基金的运营状况进行有效评价，不仅对投资者起到重要的指导作用，也为市场提供了良好的淘汰机制，促进基金行业往更加专业和健康的方向发展。

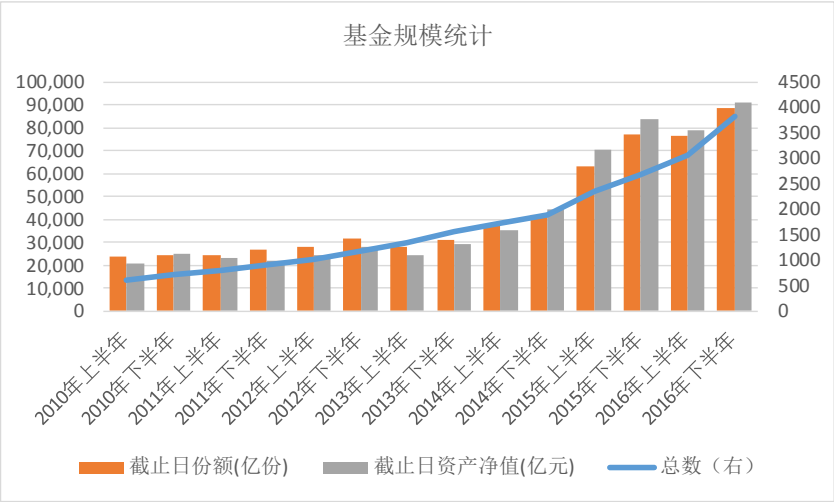

图 1　中国基金规模统计

传统的基金评价方法多采用专家打分法和计量方法，此类方法的一个特点是利用已有的经验对基金的收益率进行归因分析，最终的落脚点在于讨论各因素的影响程度。也即，传统的评价方法均需先验的对基金收益率的各个影响指标有先验的判断，并适当的进行加权或放入回归模型，最终其与收益率的关系在计量模型下往往是线性的（或

者与其对数是线性相关的)。然而，此类方法的问题在于，很难对所有变量给出一种系统性的方法，更加全面的考虑各个影响因素以及各影响因素之间的相互关系，最终在知道基金选择时起到的作用往往也只局限于常识判断。本文在以往文献的基础之上，希望能够找到一种系统性的方法，从众多的影响变量中找到更为明确的相互作用关系，得到基金评价的“网络图”，能够清晰的解释基金业绩来源。实际上，这种多变量关系的方法在医疗数据上已有尝试并取得了较大的成功，即利用概率图模型对变量关系进行图形拟合。基于这一思路，本文试图利用概率图的方式对所有可能影响基金收益率的因素进行紧凑的网络表示，得到各个变量对收益率的直接或间接影响，并得到各个变量之间的相互依赖关系。

概率图模型大致可以分为两种，分别是贝叶斯网络和马尔科夫随机场，这一方法巧妙而结合了图论和概率论，而本文应用的是采用有向网络的贝叶斯网络结构图。从图论的角度，概率图是一个图，包含节点和边；而从概率论的角度，概率图是一个概率分布，图中的节点对应于随机变量，便对应于随机变量之间的依赖关系或因果关系。那么，给定基金业绩的众多影响因素，如果能够利用概率图建立一个图，用观测节点表示观测到的数据，用隐含节点表示潜在的知识，用边来描述知识和数据的相互关系，最后获得一个概率分布，给定概率分布之后，则可以进行推断。这一方法的难点一方面在于，如何结合基金评价已有文献的观点，并应用到数值计算获取的概率图上，简单来说，就是对概率图的解释，以及事后的推断。另一方面，由于基金业绩影响因素复杂多变，关系也错综复杂，利用概率图得到的结果可能错综复杂，网络图解释难以进行。因此，本文会在第一部分介绍本文参考的主要文献后，在第二部分着重解释概率图模型（贝叶斯网络）的基本概念和主要原理，为在第三部分解释概率图模型的结果作铺垫。而文章的第四部分则希望将概率图模型与传统的计量模型进行比较，在本文的实例上讨论其优缺点。

1.2 选题背景及意义

本文按照以下步骤展开研究：

1、概率图模型介绍

依照本文的思路，本文首先给出概率图模型的基础理论部分。概率图模型是图论和概率轮的结合，是多个变量联合分布的一种紧凑表示，其不同的结构表示不同的条件独立关系。本文首先给出一个概率图的基本例子，然后给出概率图模型的图形表示含义，之后给出概率图在给定网络结构下的参数估计方法，最后给出概率图的结构估计方法。需要说明的是，本文主要利用贝叶斯网络这一有向模型，且只介绍了基于极大似然函数的估计方法。前者的原因在于有向模型能够更加直观的展现以来关系，而

后者的原因在于极大似然函数估计方法最易理解。

2、基金业绩评价指标归纳

依照本文的思路，本文最根本的目标是获取基金评价的一个系统性框架，因此，需要对可能影响基金业绩的指标进行归纳和总结。本文根据对文献的综述和梳理，获得的指标主要是这四个方面：其一，被解释变量即基金业绩，主要包括了基金的绝对收益水平、相对收益水平和基金规模；其二，控制变量即经济因素，主要包括了经济环境、市场监管、市场规模和基金投向等指标；其三，解释变量中的基金指标，主要是共同基金规模、基金存续期、基金费率和申购费、资金流动、基金规模和基金经理等指标；其四，则是基金类型、基金经理变更和基金所属公司等其他变量。需要注意的是，本文仍沿用了计量分析中的被解释变量、控制变量和解释变量的说法，是为了突出研究的落脚点和重点变量，而非使用了计量分析方法。

3、基金业绩评价指标初步筛选

为了更加紧凑的表示基金收益率的影响因素，并更加清晰的进行解释，本部分主要利用前面找到的指标进行初步筛选，在本文的前期工作中，利用简单的贝叶斯网络模型，来剔除部分的不相关因素。发现同类基金规模、基金经理数量、 基金经理学历以及基金经理所属公司等并不显著影响基金的业绩。实际上，这一部分利用贝叶斯网络时，加入了一定的先验结构，若先验结构打分函数得分较低，那么就剔除掉这一个变量。

4、利用概率图模型建立图结构

最后一部分是本文的难点所在，概率图模型以往的应用主要集中在工程、电信和医疗等方面，其与经济金融问题的一个重要差别在于其对变量的控制相对容易，容易获得条件概率。而本文所涉及的问题，对概率图的尝试并没有前人文献进行指导，因此，本文会以相关评价类文献作为指引，探讨出此类模型在金融评价问题上的应用。另一难点在于，由于贝叶斯网络结构涉及多个变量，其结果可能变量之间有各种交叉关系，本文需要透过这种交叉关系找到重要变量的影响。

1.3 可能创新与不足

1.3.1　创新之处

本文的主要创新之处在于：

1、引入概率图模型来评价金融问题，国内此类文献几乎没有，国外此类文献也相对缺乏；2、将概率图模型与传统计量模型进行比较，分析概率图模型的优缺点，为概率图模型的进一步应用打下基础；3、试图建立基金评价的“体系”，将影响基金收益率的多个指标均放入模型，经过初次筛选和模型构建，获得基金评价的系统。

1.3.2 创新之处

本文的主要不足之处在于：

1、概率图模型的应用并不成熟，本文将之应用于金融数据属于尝试性工作，文章只对概率图模型的理论进行了简单介绍，并未对其背后公理体系进行探讨，很难形成类似计量分析的一套完整有效体系；2、利用概率图模型进行结构学习时，并未对网络图给出先验的设计，可能导致模型得到的网络结构与常识产生较大差异；3、基金的数据选取的是静态数据而非动态数据，本文还为涉及动态贝叶斯网络模型，因此无法实现时序数据的讨论。

第二章　理论研究

2.1　概率图模型概述

2.1.1 概率图的重要性

概率图模型实际上是一种方法论，如果说机器学习的核心任务在与从观测数据好隐含知识的挖掘，而概率图模型则是实现这一任务的一种基础手段。概率图模型巧妙地结合了图与概率论，无论是多么复杂的问题，都能通过网络图的形式给出表示，并对应到相应的概率分布。可以说，概率图模型能够应用于处理所有的多变量未知关系，是人工智能一个重要的分支。概率图模型在实际中的应用非常成功，除了前文提到的医疗诊断外，隐马尔科夫模型在语音识别中应用十分成功，而条件随机场则在自然语言的处理中得到了广泛应用。除此之外，概率图模型还应用于遗传问题、规划问题、图像识别和游戏中。概率图模型的强大之处在于，能够处理复杂的问题，无论有多少个变量，都只增加了计算的难度，这对于构建大型的人工智能系统来说是十分重要的。

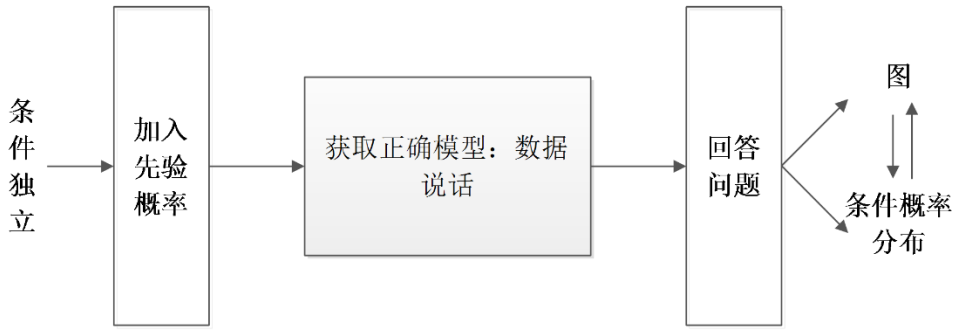

图 2 概率图模型概述

前面已经提到，概率图模型的应用主要是从未知中找到变量的可能关系，并从找到的可能关系中回答需要探讨的问题。那么，有三个问题需要提出：其一，如何获取未知关系并且如何加入先验关系；其二，如何得知什么样的模型是最“正确”的；其三，如何通过最终的模型来回答问题。关于第一个问题，实际上，在建立模型前，研究者对研究的问题已经有一定的认识，一个简单的例子是，运动过量是疲惫和肌肉疼痛的共同原因，而疲惫和肌肉疼痛在给定这一条件之后就没有直接关系了。这是条件独立的一个典型例子，也即是说，给定运动过量，疲惫和肌肉疼痛是条件独立的。在建立概率图模型之前，很多问题要根据过往文献和专家评价提前给定一定的条件独立关系，然后才能够获得符合直觉的模型。关于第二个问题，网络图的构建都需要通过观测到的数据情况根据一定的方法来构造网络，也即是根据数据情况决定模型。在已有每个变量的观测值下，根据极大似然估计或贝叶斯估计方法，来得到每个标准下的最优模

型。关于第三个问题，这个问题实质上在讨论概率图模型的目的，概率图模型的主要目的即找到变量的关系并最终形成变量的联合分布，而最终回答模型问题也即通过其他变量的观测情况，结合所学习的联合分布，给出探讨变量的概率分布。也即是说，研究概率图模型遵循“目的是获得概率分布，方法是利用数据进行估计，而设定依据现实经验”这一逻辑。

概率图模型的最终目的是获得概率分布，那么图的作用是什么呢？实际上，图形的表示是概率图的一种紧凑表示，这种紧凑表示能够直观的展示条件独立关系，并形成一条清晰的判断网络。除此之外，选用概率图模型有一个另外的好处——减少参数的估计量并加入先验设计。一组观测变量只要能够获取其观测数据，即能估计其联合分布，不直接采用联合分布的估计一方面因为需要估计的参数相对较多，另一方面因为无法加入先验的条件独立性。而概率图满足了这两个要求，在 2.2.1 的例子中本文会具体说到。

2.1.2 概率图在金融中的应用

前面已经谈到，概率图模型在医疗等行业中应用较为成熟，在金融中的应用实际上处于起步阶段。概率图模型在金融分析中主要被应用于风险分析、价格预测以及投资决策中，且此类方法主要出现在国外文献中。Abramson 和 Finizza(1995)利用贝叶斯网络对石油价格进行预测，不仅加入了各项经济变量，且对贝叶斯网络结构依照专家评价进行了先验结构设定，主要分为核心储藏能力、核心生产能力和政策三类变量，同时加入了市场预期，底层变量包括各个海湾国家以及南美国家的生产能力和储藏能力，以及各个合作组织的政策导向变量，最终得到的预测结果稳健，对当年的价格预测几乎都落在中间水平。Shenoy 和 Shenoy(2001)认为传统的金融模型只强调组合收益和市场收益的历史关系，且依赖变量之间很强的假设，而文章利用贝叶斯网络对收益和风险相关变量进行了处理，并不依赖任何假设，且能够利用贝叶斯网络给出的定量关系对未来的风险和收益进行概率预测。Gemela(2001)对捷克公司 1993-1997 的各项财务指标利用贝叶斯网络进行财务分析，得到的贝叶斯网络图形成两个主要的影响渠道，一方面是确定的依赖关系，即股份到流动性、债务水平和应收款周转；另一方面的关系相对不明确，即应付款周转到流动性、应收款周转到应付款周转等。文章在进行贝叶斯网络分析时，并未给出先验的依赖关系，网络结构由模型自动计算得到。Kemmerer (2002）将贝叶斯网络方法应用到风险投资的决策问题上，文章结合了因果图和贝叶斯网络方法得到了贝叶斯因果图。其主要目的在于建立一个能够帮助风险投资人实际决策的决策过程，文章利用高科技行业的天使投资轮数据对模型进行了测试，发现贝叶斯因果图能够通过减少偏误、减少非系统性失误、提前假设、条件分析以及系统学习等方式来提高风险投资者的项目成功概率。Demirer, Mau 和 Shenoy(2006)同

样利用贝叶斯网络方法对投资组的风险进行了分析，文章指出，行为金融模型认为购买和卖出的投资行为体现了有偏的决策过程，但只能给出描述性的说法，而贝叶斯网络则能够利用条件概率的定量分析，对投资这的决策有偏行为进行精确的纠正，控制组合风险，同时能够系统性的提高风险收益预测能力。值得指出的是，文章在进行贝叶斯网构建前，先验地给出了贝叶斯网络结构，对贝叶斯网络结构进行了一定的限制。Sanford 和 Moosa（2012）利用贝叶斯网络衡量银行的管理风险，文章利用一家澳洲银行的数据，并结合了人力因素对银行的风险网络进行构造，最终得到一个复杂的网络结构，对多种风险环境都能够有效刻画。

实际上，通过文献总结可以发现，经济和金融中应用概率图模型解决的问题主要集中于预测和因果探寻，除此之外就是对风险结构的网络刻画。此类问题均要求探索前对网络结构有一定的先验了解，对模型进行一定的设定后才能够获得相对符合直觉的结果。另一方面，此类文章的应用并不多见，2012 年以来的文章几乎难以见到。也就是说，利用概率图模型研究经济金融问题并未形成一套成熟的体系。实际上，在本文后续的探讨中，也可以发现，如果不给定任何先验的假设，概率图模型寻找到的网络结构很难得到解释，原因在于此类模型只能关注数据上的最优，而并未排除不可能的父子节点关系。因此，本文在后续实证探讨时，也将遵循“不给定任何假设——根据情况给定假设——做出判断”这一逻辑线条进行梳理。

2.2　概率图模型理论

2.2.1 简单示例

概率图模型是图模型理论和概率论相互结合的产物，旨在构建一种能够将各种复杂关系利用图形和概率分布的形式给出的一种复杂系统模型。一个常规的例子是在医疗诊断中，患者可能罹患的多种可能性疾病，几十种或者几百种症状以及诊断检测，经常形成疾病诱发因素的个体特征，以及其他许多需要考虑的因素。可以利用一组随机变量来刻画每个因素，概率图的任务是概率性地推断出一个或者多个变量的可能值，并采用规则化的概率推理来完成这一任务，最后形成一种紧凑表示的机制。

考虑一个非常简单的医疗诊断问题，关注流感和花粉热，这两种疾病之间并不排斥，因此得到了两个二值的随机变量：流感和花粉热（得或不得）。同时还有一个与流感和花粉热均有关系的 4 值随机变量：季节（春夏秋冬）。此外，还有两个二值随机变量症状：充血和肌肉疼痛（有或无）。总的来说，概率空间有 2*2*4*2*2=64 个值，对应上述 5 个变量的所有可能取值。当得知这个概率空间上的联合分布式，就可以根据患者出现的情况对问题进行查询和判断。比如，假定现在是秋季，并且患者有鼻窦充血的症状，但没有肌肉疼痛的症状，那么患者患流感的几率是多少？利用概率表述既可以

表达为:

$$P(Flu = true \mid Season = fall, Congestion = true, Muscle = false)$$

那么，这 64 个取值的概率空间利用图结构进行表达，既可以表示成图 3 所示的情况。实际上，这个示例也是本文应用的贝叶斯网络结构。在这种图结构中，可以推出这样的独立关系：

$$(F \perp H \mid S), (C \perp S \mid F, H), (M \perp H, C \mid F), (M \perp C \mid F)$$

根据这种独立关系进行因子分解，可以判断前述所需要询问事件“秋季、有流感、没有花粉热、充血且没有肌肉疼痛”的概率，可以简化到只利用如下五个数值计算得到：

$$P(Season = fall)$$
$$P(Flu = true \mid Season = fall)$$
$$P(Hayfever = false \mid Season = fall)$$
$$P(Congestion = true \mid Hayfever = false, Flu = true)$$
$$P(Pain = false \mid Flu = true)$$

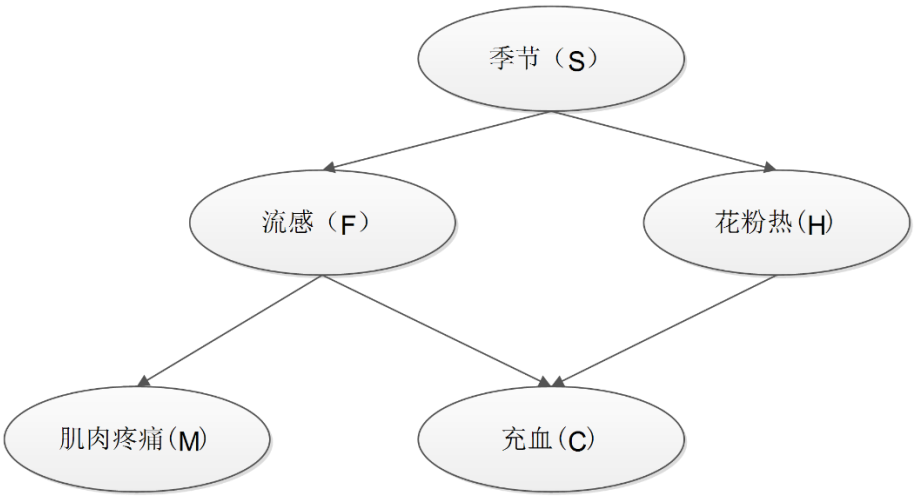

图 3 概率图示例

相对于原始联合分布的 63 个必要的参数，经过简化之后结构显然更加紧凑，在引入了条件独立之后，只需要 17 个必要的参数既可以估计这一空间的概率分布。也就是说，分布的独立特性使得分布能够更加紧凑地以引资形势对分布进行表示。从这里可以看出，概率图模型的任务在于找到独立关系，将复杂的关系简化，在此基础之上进行诊断和判断。以上是概率图模型（贝叶斯网络图结构）应用的一个简单例子，可以看到的是，图结构刻画了随机变量间的条件依赖和独立关系，其中结点代表了各个随机变量，而结点间的边则代表了随机变量之间的统计关系，这种图结构能够从系统的角度揭示变量之间的不确定性，并且为不确定性的传递提供手段。基于概率论，概率图模型算法利用图结构有效的计算边缘概率或其他条件概率。下面两个部分给出概率

图模型的理论介绍，分别包括图和信息论两个部分。其中，图定义了概率图的概率表示规范，而信息论则阐释如何通过得分函数来获取最为优化的概率图结构。

2.2.2 理论阐述：图

考虑一个包含 n 个随机变量的联合分布 $P(X_1, X_2, \cdots, X_n)$，利用链式法则，则可以将其写成

$$P(X_1, X_2, \cdots, X_n) = \prod_{i=1}^{n} P(X_i \mid X_1, X_2, \cdots, X_{i-1})$$

对于任意 X_i，如果存在 $\pi(X_i) \subseteq \{X_1, X_2, \cdots, X_{i-1}\}$，使得给定 $\pi(X_i)$，X_i 与 $\{X_1, X_2, \cdots, X_{i-1}\}$ 中的其他随机变量条件独立，即

$$P(X_i \mid X_1, X_2, \cdots, X_{i-1}) = P(X_i \mid \pi(X_i))$$

则有

$$P(X_1, X_2, \cdots, X_n) = \prod_{i=1}^{n} P(X_i \mid \pi(X_i))$$

这样，就可以得到一个分布的分解。在此分解的基础上，可以利用贝叶斯网络图对其进行紧凑表示。

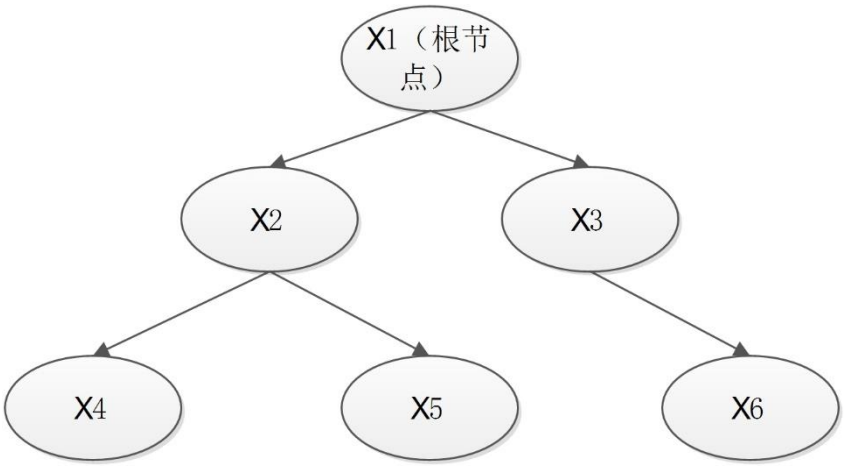

图 4 完整贝叶斯网络

Pearl(1986)提出用折中方法来构造一个有向图来表示每个变量之间的依赖和独立关系：其一，每个变量都表示为一个节点；其二，对每个节点 X_i，都从 $\pi(X_i)$ 的每个节点画一条有向边到 X_i。也即是说，贝叶斯网是一个有向的无圈图，其中节点代表随机变量，节点间的边代表变量之间的直接依赖关系。每个节点都附有一个概率分布，根节点 X 所附的是它的边缘分布，而其他节点 X 所附的则是其条件概率分布 $P(X \mid \pi(X))$。上述例子中，X_1 为该贝叶斯网络中的根节点，而 X_2 和 X_3 为其子节点，依次类推。由贝叶斯网的构造，可以得到该图的条件独立性为

$$(X_2 \perp X_3 \mid X_1),(X_4 \perp X_5 \mid X_2),(X_4 \perp X_1 \mid X_2),(X_5 \perp X_1 \mid X_2),(X_6 \perp X_1 \mid X_3)$$

可以从定性和定量两个方面来理解贝叶斯网络图。定性的理解即其利用一个有向无圈图描述了变量之间的以来和独立关系；定量的理解即其利用条件概率分布刻画了变量对其父节点的依赖关系。联合概率分布降低了概率模型的复杂程度，贝叶斯网络的引入尽管没有进一步降低复杂度，但它为概率推理提供了很大的方便。主要是因为，贝叶斯网一方面是严格的数学语言，计算机处理起来相对方便；另一方面也是因为网络图形结构直观易懂，方便讨论和建模。

2.2.3 理论阐释：参数学习

贝叶斯网络学习过程指的是通过分析数据而获得贝叶斯网的过程，它包括参数学习和结构学习两种情况。参数学习指的是一直网络结构，确定网络参数的问题；而结构学习则是既要确定网络参数，又要网络结构。本部分讨论参数学习的过程。前面已经讨论，一个贝叶斯网 N 包括定性和定量两个方面的内容：定性的内容为变量之间的依赖关系即网络结构，可以记为 g；而定量内容则是指各个变量的概率分布，可以记为 θ。也就是说，根据一组数据，需要得到贝叶斯网 N 的二元组 (g,θ) 的形式。当模型已知时，只需要确定参数 θ，这一过程叫做参数学习。参数学习的基本方法包括最大似然估计和贝叶斯估计两种。本部分只介绍最大似然估计方法。

从单参数贝叶斯网开始，最大似然估计遵循数据拟合程度最高的原则，设数据 D 由样本 $(D_1, D_2, \cdots, D_m)$ 组成，则参数 θ 的最大似然估计，是另似然函数 $L(\theta \mid D)$ 达到最大的取值 θ^*，即

$$\theta^* = \arg\sup_{\theta} L(\theta \mid D)$$

$$L(\theta \mid D) = P(D \mid \theta) = \Pi_{i=1}^{m} P(D_i \mid \theta)$$

这一表示遵循了传统的最大似然估计方法的假设，即每个样本在给定参数 θ 时相互独立。若已知贝叶斯网络，在观测到每个节点的数据后，对每个观测值，均能够通过联合分布的分解获得该数据下的联合分布概率值，将所有的概率值相乘，即得到了极大似然函数。值得注意的是，每一个节点的所有条件概率均将成为一个参数进入到似然函数中，尽管相较于原始的参数估计量有所降低，但其估计参数仍然较多，大多时候难以得到解析解。

将单参数的最大似然估计推广到多参数的估计。考虑一个由 n 个变量 $X = \{X_1, X_2, \cdots, X_n\}$ 组成的贝叶斯网 N。不是一般性的假设其中的节点 X_i 共有 r_i 个取值 $1, 2, \cdots, r_i$，其父节点 $\pi(X_i)$ 的取值共有 q_i 个组合 $1, 2, \cdots, q_i$。若 X_i 无父节点，则 $q_i = 1$。那么，网络的参数为

$$\theta_{ijk} = P(X_i = k \mid \pi(X_i) = j)$$

其中，i 的取值范围是 $1,2,\cdots,n$，而对一个固定的 i，j,k 的取值范围分别是 $1,2,\cdots,q_i$ 和 $1,2,\cdots,r_i$。另外，由于概率分布具有规范性，因此

$$\sum_{k=1}^{r_i} \theta_{ijk} = 1, \forall i, j$$

因此，贝叶斯网 N 中独立参数的个数共有 $\sum_{i=1}^{n} q_i(r_i - 1)$ 个。进一步地，对数似然函数可以表示为

$$l(\theta \mid D) = \ln \Pi_{i=1}^{m} P(D_i \mid \theta) = \sum_{i=1}^{m} \ln P(D_i \mid \theta)$$

为了得到关于 $\ln P(D_i \mid \theta)$ 的表达式，定义样本 D_i 的特征函数为

$$\chi(i, j, k : D_l) = \begin{cases} 1 & if\ X_i = k\ and\ \pi(X_i) = j \\ 0 & if\ not \end{cases}$$

则有

$$\ln P(D_l \mid \theta) = \sum_{i=l}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} \chi(i, j, k : D_l) \ln \theta_{ijk}$$

本节解决了贝叶斯网络的参数估计问题，即利用已有数据对联合分布的似然函数最大化求解每个条件概率。下一节着重解决贝叶斯网络结构的确定，这也是此类问题的关键所在，即父节点和子节点之间依赖关系的确定。

2.2.4 理论阐释：结构学习

前面对贝叶斯网的讨论，都是基于一个较强的假设——事先知道网络结构，或者不管正确与否，至少制定了某个结构。实际上，网络结构的确定是此类模型发挥作用的最重要部分——通过探索网络中的依赖关系，能够获知领域中变量之间的依赖关系，并对直接和间接的以来关系进行区分，而不是仅仅给出相关性。结构学习满足了网络结构构造的要求，一般分为两个部分，模型选择部分和模型优化部分。其中，模型选择部分回答了用什么准则来评价模型结构的优劣，而模型优化过程则寻找最优的模型结构。

模型选择主要利用得分函数来对模型结构进行评价，主要的得分函数分为最优参数对数似然得分函数、Cooper-Herskovits 得分函数（CH 得分函数）和贝叶斯信息准则得分函数（BIC 得分函数）。这里只介绍对数似然得分函数的处置方法。同样，首先给出设定。设 $X = \{X_1, X_2, \cdots, X_n\}$ 是一组随机变量，$(D_1, D_2, \cdots, D_m)$ 是关于这些变量的一组数据，而 g 是一个以 $X_1, X_2, \cdots, X_n$ 为节点的贝叶斯网络，g 对应参数集合 θ_g，二者

共同组成贝叶斯网络(g,θ_g)，在此贝叶斯网中，可以计算每一个样本D_l的概率$P(D_l|g,\theta_g)$，其对应的极大似然函数估计为

$$l(g,\theta_g|D)=\sum_{l=1}^{m}\ln P(D_l|g,\theta_g)$$

需要注意的是，这是一个二元组(g,θ_g)的函数，寻找最由贝叶斯网的过程可以分成两步：第一步寻找最优的依赖结构g^*，第二部优化参数θ_g^*，即

$$l^*(g|D)=\sup_{\theta_g} l(g,\theta_g|D)$$

$$l^*(g^*|D)=\max_{g} l^*(g|D)$$

此即最优参数对数似然函数的处置方法。

在选定模型的得分函数后，接下来就是要找出得分最高的网络结构。实际上，穷举法是最为直观的方法，即将所有的网络结构都进行考虑，计算出评分最高的结构。然而，穷举法在实际中并不可行，因为即使随机变量的个数不多，其排列组合形成的网络结构数目也是巨大的。实际上，设$f(n)$是由n个节点组成的有向无圈图个数，Robinson（1977）已经证明了

$$f(1)=1$$

$$f(n)=\sum_{i=1}^{n}(-1)^{i+1}\frac{n!}{(n-i)!i!}f(n-i),n>1$$

且由于行程贝叶斯网络还需要对每个节点进行命名，由于有多重命名方式，以这么多变量为节点的贝叶斯网络数目会更多。因此，穷举法很难实现。

因此，大部分的模型计算过程都是利用算法进行搜索，最为传统的两种方法是 K2（Cooper and Herskovits,1992）算法和爬山法。K2 算法的我出发点是一个包含所有节点但却没有变的无边图，在搜索过程中，K2 算法按照顺序考察每个变量，并确定其父节点，然后添加相应的边。需要注意的是，K2 算法需要输入变量顺序和父节点个数的上界，因此仍然有一定的主观性。爬山法的目的则是找出得分最高的模型，从一个初始的无边模型出发，利用搜索算子对当前模型进行局部的修改，得到一系列的候选模型，然后计算所有候选模型的评分，并将最优候选模型与当前模型比较，若修正后的模型得分更高，则继续搜索，否则停止搜索。其中，搜索算子主要就是加边、减边和转边三种情况。与 K2 方法不同的时，爬山法可以适用于各种得分函数，而 K2 方法主要是应用于 CH 得分函数。

2.3 因果关系与贝叶斯网络

2.1 部分对贝叶斯网络的基本概念以及如何确定进行了简单的讨论，从参数学习

和结构学习两个方面给出了贝叶斯网络的基本构造和形成过程。然而，在应用贝叶斯网络时，很多时候可以事先利用因果关系来确定贝叶斯网的结构。考虑一个简单地例子：肺炎（T）、肺癌（L）、支气管炎（B）、出访（A）、抽烟（S）、呼吸困难（D）以及X光胸透结果（X）。对于这些熟悉的变量，我们事先知道其相互之间一定的因果关系：抽烟会引起肺癌和支气管炎，出访可能感染肺炎，无论是肺炎、肺癌还是支气管炎都可能导致呼吸困难，另外如果有肺炎或者肺癌，X光胸透的结果可能是阳性。把这些因果关系结合起来，就可以得到如下的网络结构。

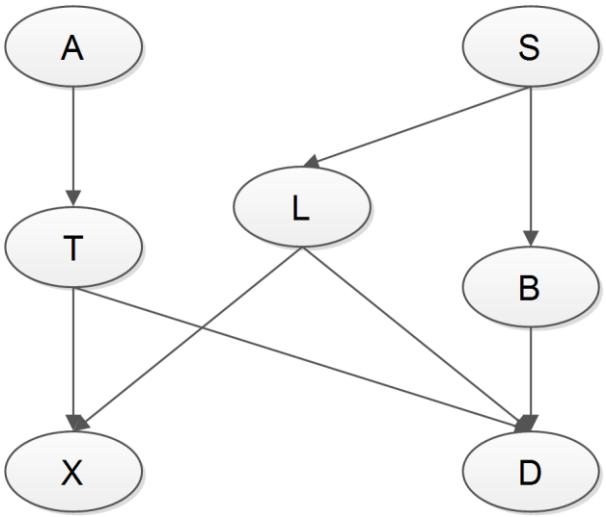

图5 因果关系示例

在利用因果关系建立起来的模型中，可以通过变量与变量之间的边看到明确的因果关系，而非简单的概率以来关系。然而，再利用因果关系建立贝叶斯网络结构的弊端也十分明确。首先，因果关系本就是贝叶斯网络需要探究的一个问题，对不确定的变量之间进行依赖关系的探究是贝叶斯网络的一个重要任务之一，先验的因果关系可能没有更强的说服力，在贝叶斯网络的语义下，因果关系的先验给定实际上是在对变量之间的条件独立性给出假设；其次，因果关系并没有一个能被广泛接受的严格定义，比如经济学中讨论的计量模型，就一直致力于探寻变量之间的因果关系，但即使能够得到统计上的显著性，仍然不能断言变量之间存在因果关系，还需要经济学理论作为基础的判断。

第四章　上市公募基金绩效评价

3.1　指标、数据和算法说明

3.1.1 文献总结

关于基金评价的文献相对也比较多，国内外对此类问题的研究主要集中在两个方面。首先，是对基金评价的指标展开的讨论。Markowitz(1952)首次引入了“均值-方差”模型，采用计量经济学的方法度量证券投资组合的收益与风险的关系，该模型在出事前提假设条件下成功的度量了投资组合的风险，为构建有效投资组合奠定了重要的理论基础。此后，Sharpe、Treynor 和 Jensen 分别在该理论的基础上发展出了 Sharpe 指数、Treynor 指数和 Jensen 指数。Treynor(1965)首次提出了风险调整收益绩效评价的概念，他认为科学有效的构建投资组合完全可以消除单一投资所造成的费系统性风险。这也是第一次出现了不仅以收益率同时加入风险指标的评价方式。此后，Sharpe(1966)指出由于基金经理的管理能力有所不同，因此不同的基金经理对于基金运营过程中可以消除风险的程度也不同，绝大部分基金根本无法完全消除非系统性风险，于是采用标准差衡量，得到了 Sharpe 指数。而 Jensen(1968)的理论主要是在 CAPM 模型的基础上集中讨论了资本市场现，检验超额收益水平，不同于前面两个指标，Jensen 指数是一种绝对数指标，因此该方法也无法用来衡量不同类型基金的绩效水平。此后，Sortino 和 Meer(1991)、Desai(1997)以及 Fama 和 French(1992)分别引进了波动率、数据包络分析和三因素与五因素模型来进行基金业绩的评价。其中，Famade 和 French 的三因素和五因素模型影响深远。除此之外，Goodwind(1998)的信息比以及 Franco Modigliani 和 Leah Modigliani(1997)的风险调整绩效指数也各自都有其侧重点。传统文献中的方法到目前为止，仍然是衡量证券投资基金业绩最为有效的方法，然而随着金融创新的不断发展，特别是金融危机爆发以来，投资组合的风险控制问题受到基金管理者和评价机构越来越多的重视。此类文献也突出了其贡献。Acharya 和 Pederson(2002)认为资产流动性依据时间变化而变化，因此提出了基于流动性的资本资产定价模型，并用此来测度流动性风险。Mcneil 和 Frey(2000)则应用 GARCH 模型来评估资产收益率，采用极值利率计算模型残差项的 VAR。除此之外，投资者偏好（Harve 和 Sliddique, 2000）、系统协峰度（Hwang 和 Satchell, 1999）以及效用推导框架（Alexandros Kostakis, 2008）等也被加入风险的测度。

另一方面的主要研究则是对基金绩效影响因素的归纳，此类文献主要集中在 2000 年以后。Chevalier(1999)的文章重点讨论了基金经理的作用，研究了基金经理的能力、知识和努力程度对基金业绩的影响，主要指标包括了基金经理年林、学历和 SAT 成绩，

最终发现高 SAT 成绩的基金经理业绩表现更好。而 Chen 和 Hong(2004)则讨论了基金规模对基金绩效的影响，文章发现基金规模对基金绩效有负的影响，并进一步探讨了原因。文章发现，基金规模主要影响了基金的流动性管理和资产配置效率，一方面，大规模的基金流动性往往波动较大，对净值管理的要求加高；另一方面，大规模的基金由于资产配置的需求，若配置擅长不同类别资产投资的基金经理，那么可能会导致配置能力的提升。这两方面共同影响了基金的业绩水平。Bauer、Koedijk 和 Otten(2005)则讨论了基金投资风格以及基金成立时间对投资业绩的影响，基金投资风格通过基金表现的长期数据进行归纳总结，发现不同投资风格的基金业绩表现有明显差别；而基金成立时间同时也会影响投资业绩，文章认为存在这样一种学习效应，时间长的基金能够快速的跟上市场节奏，因此能够保持良好的业绩。Ferreira 等(2012)则对基金业绩决定因素进行了归纳，主要分为两个方面。其一，是经济发展、市场监督和基金市场发展水平等宏观经济指标；其二，则是基金规模、同类基金规模、基金存续期、基金费率和申购费、资金流动、历史表现以及管理结构等基金结构指标。

表 1 基金绩效影响因素

➢ **基金业绩** ✓ 绝对收益水平：几何&算数平均 ✓ 相对收益水平 ● 基准指标 ● 基金类型 ✓ 基金规模 ✓ 单位净值	➢ **经济因素** ✓ 经济环境 ● 宏观经济 ● 资产轮动 ✓ 市场监管 ✓ 市场规模 ✓ 基金投向
➢ **基金指标** ✓ 同类基金规模、基金存续期、基金费率和申购费、资金流动 ✓ 基金规模 ● 流动性管理 ● 单个经理&经理组 ✓ 基金经理各项指标	➢ **其他指标** ✓ 基金类型 ● 投资标的 ● 市场分类 ✓ 基金经理更换情况 ✓ 基金所属公司：内部操作等

3.1.2 指标选取

文章在上一节已经对基金评价的指标进行了文献综述，根据文献综述的结果，本文选取了几类影响基金指标的因素。首先是基金本身业绩指标，选取成立以来年化收益；其次是基金基本信息，包括基金成立年限、基金类型、管理费率和申购费率等信息；再次是基金经理信息，包括是否为基金经理组、是否更换主要基金经理、管理基金时间、基金经理从业年限、经理组从业平均年限、基金经理任职以来回报、基金经理管理基金数、基金经理性别和基金经理学历等变量；最后是投资组合的相关信息，

包括基金最大回撤、基金资产规模、基金资产净值规模、股票占比、债券占比、存款占比和其他占比等信息。

表 2 基金绩效影响因素

变量分类	变量代码	变量名	变量类型
回报	Return	成立以来年化回报	连续
基本信息	Year	成立年限	连续
	Type	基金类型	离散
	Mfee	管理费率	连续
	Bfee	申购费率	连续
基金经理	PM_team	是否基金经理组	离散
	PM_change	是否更换主要基金经理	离散
	PM_date	管理基金时间	连续
	PM_year	基金经理从业年限	连续
	PMT_year	经理组从业平均年限	连续
	PM_return	基金经理任职以来回报	连续
	PM_sex	基金经理性别	离散
	PM_edu	基金经理学历	离散
组合信息	Drawback	基金最大回撤	连续
	Asset	基金资产规模	连续
	Asset_NAV	基金资产净值规模	连续
	Stock	股票占比	连续
	Bond	债券占比	连续
	Money	存款占比	连续
	Other	其他占比	连续

一个需要说明的问题是，第二部分讨论贝叶斯网络图的结构学习和参数估计时，文章所讨论的模型均为离散形式的模型，前述的各种关于似然函数和分布函数的假设，很多都是基于离散分布。实际上，贝叶斯网除了能够处理离散分布外，还能够处理连续分布问题。然而，本文的数据是混合型的，即包含了一部分离散变量，也包含了一部分连续变量。在现有算法中，这一问题能够得到解决。

3.1.3 数据描述

本文选取了截止 2016 年 12 月 31 日的中国上市公募基金中 3111 支基金的数据。实际上，截止 2016 年 12 月 31 日，中国上市公募基金总数共 3820 支，本文只选取其

中 3111 支基金的原因在于：一方面，本文只选择了四类基金，包括股票基金、债券基金、混合基金和 QDII 基金等四类；另一方面，部分数据缺失较为严重的基金予以剔除。最终，共 3111 支基金进入最终的样本中。数据的描述统计结果如下所示。

表 3 样本描述统计

	平均	中位数	标准差	最小值	最大值	观测数
Return	5.08	5.40	10.28	-43.77	79.18	3111
Year	4.11	2.99	3.14	1.01	15.50	3111
Mfee	1.02	1.00	0.43	0.00	2.50	3111
Bfee	0.21	0.20	0.05	0.05	0.35	3111
PM_date	813.06	647.00	598.95	4.00	4687.00	3111
PM_year	4.02	3.23	2.67	0.01	14.35	3111
PMT_year	3.78	3.22	2.47	0.01	14.35	3111
PM_return	4.18	3.99	11.03	-40.61	65.17	3111
Drawback	-17.83	-7.80	17.99	-66.59	0.00	3111
Asset	14.72	6.44	27.64	0.00	494.92	3111
Asset_NAV	10.83	3.70	25.02	0.00	494.91	3111
Stock	44.18	26.75	40.45	0.00	102.81	3111
Bond	43.21	8.85	47.63	0.00	193.11	3111
Money	10.51	6.35	17.96	0.02	701.82	3111
Other	7.08	1.96	13.09	0.00	107.93	3111

样本中，平均年化回报在 5.08%，而基金成立平均年限在 4.11 年，基金经理管理基金平均年限 2 年不到，但基金经理和基金经理组平均工作年限均超过 3 年；投资组合各类资产平均占比中，股票和债券占比均在 40%以上，而现今占 10%左右，其他投资占比较小。除此之外，各离散变量的含义如下：基金类型（Type）中 1、2、3 和 4 分别对应表股票型基金、债券型基金、混合型基金和 QDII 基金；是否更换主要基金经理

（PM_change）记录为 1 若在基金存续期间存在更换主管基金的基金经理这一情况，否则记为 0；是否为基金经理组（PM_team）记录为 1 若该支基金由多个基金经理进行管理，否则记为 0；基金经理性别（PM_sex）记录为 1 若该支基金第一基金经理为女性，否则记为 0；基金经理学历(PM_edu)中 0、1、2 和 3 分别对应专科、本科、硕士和博士。

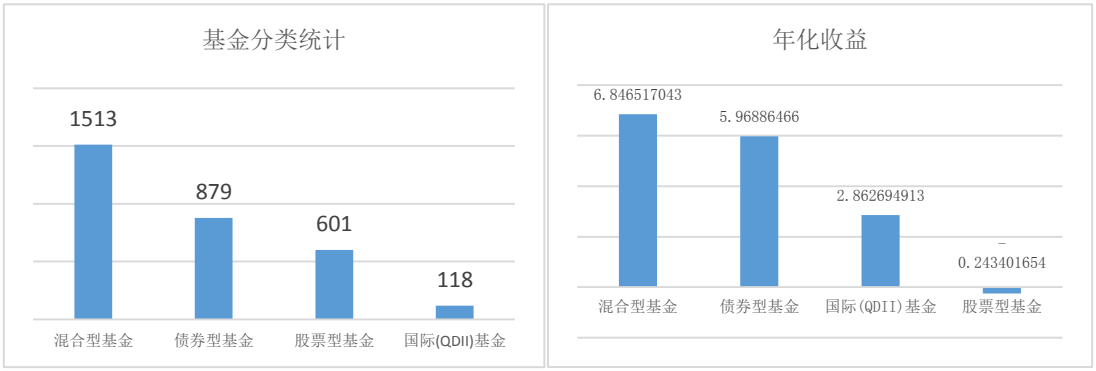

图 6 基金分类统计

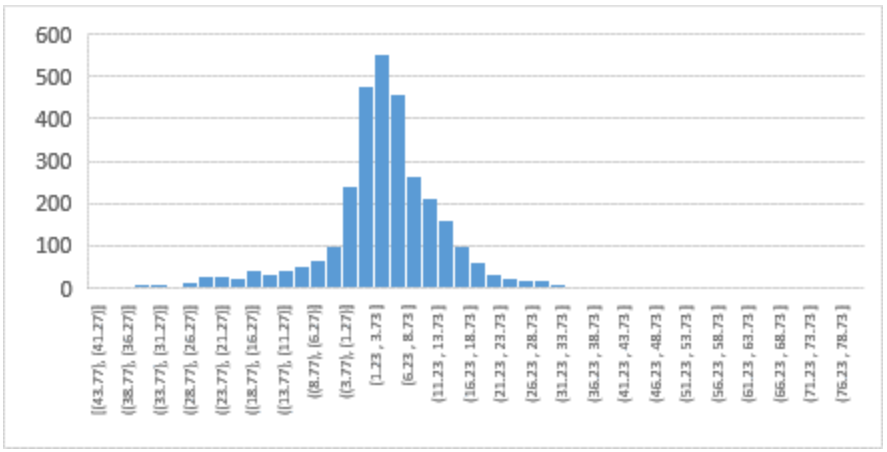

图 7 基金收益分布

从基金分类统计图中可以看到在所有基金中，混合基金占比最大，其次是债券基金；而成立至今年化收益也是混合型基金收益最高，而股票型基金成立至今平均年化收益为负数。从基金收益分布图可以看到，所选基金的收益基本上服从正态分布，收益率最集中的点在 4%-6%之间。在第二部分的结构学习和参数学习过程中，本文的数据还会进行进一步的处理，主要包括连续数据的离散化，主要是为了让贝叶斯网络模型能够更好地拟合的情况而不因算法导致结果的巨大差异。

3.1.4 算法说明

本文使用 R 对贝叶斯网进行学习，R 软件中有多个包可以实现贝叶斯网的创建、学习和推断，详细见下面表格。每个不同的包均有其特点，其中，catnet 和 gRain 侧重于离散数据的分析，而 pcalg 和 gRbase 也能对连续数据进行分析，bnlearn 和 deal

则能对混合数据进行分析。本文的数据不仅包括了离散数据，同时还会进行混合数据的学习，因此本文选取 bnlearn 和 deal。但综合而言，bnlearn 几乎涵盖了贝叶斯网络分析的所有要求，唯一的劣势在于，结构学习的算法 bnlearn 只涵盖了爬山法这一种，对运行效率可能有一定的影响。

表 4 R 中贝叶斯网络学习包总结

计算包	bnlearn	catnet	deal	pcalg	gRbase	gRain
离散数据	Y	Y	Y	Y	Y	Y
连续数据	Y	N	Y	Y	Y	N
混合数据	N	N	Y	N	N	N
基于约束的学习	Y	N	N	Y	N	N
基于得分的学习	Y	Y	Y	N	N	N
混合学习	Y	N	N	N	N	N
结构操作	Y	Y	N	N	Y	N
参数估计	Y	Y	Y	Y	N	N
预测	Y	Y	N	N	N	Y
近似推断	Y	N	N	N	N	Y

除此之外，本文对连续数据也进行了离散化的处理，离散化处理主要应用分位数的方法，本文将所有连续变量经过采取 4 分位数分位四个区段，具体分割如下表所示。具体赋值采用如下办法，在最小值到 25%分位点之间赋值 1，在 25%到 50%分位点之间赋值 2，在 50%到 75%之间赋值 3，在 75%到最大值之间赋值 4。借此方法，可以得到离散化的数据集。前面已经说过，离散化处理是由于考虑到算法的有效性以及结果的可说明性，因此，贝叶斯网络构造主要采用离散化的数据进行处理。

表 5 离散化处理分位点

变量名	Min	0.25	0.5	0.75	Max
Return	-43.77	1.68	5.40	9.85	79.18
Year	1.01	1.75	2.99	5.69	15.50
Mfee	0.00	0.65	1.00	1.50	2.50
Bfee	0.05	0.20	0.20	0.25	0.35
PM_date	4.00	447.00	647.00	987.50	4687.00
PM_year	0.01	1.89	3.23	5.65	14.35
PMT_year	0.01	1.87	3.22	5.10	14.35
PM_return	-40.61	0.07	3.99	8.90	65.17
Drawback	-66.59	-33.79	-7.80	-2.14	0.00
Asset	0.00	2.09	6.44	16.96	494.92
Asset_NAV	0.00	0.91	3.70	11.85	494.91
Stock	0.00	1.82	26.75	87.26	102.81
Bond	0.00	1.60	8.85	89.32	193.11
Money	0.02	2.34	6.35	12.88	701.82

Other	0.00	0.28	1.96	7.19	107.93

下面就利用得到的两个数据集对贝叶斯网络进行结构学习和参数学习，并对学习结果进行解释。值得提出的是，本文采用贝叶斯网络构造金融数据因果关系属于尝试性工作，其结果可能与常识出现一定的偏差。未来的研究方向是，在加入了先验的结构设计基础上，可能出现更加符合直觉的学习结果。

3.2 结构设计和变量初步筛选

在进行完整的贝叶斯结构学习和参数学习之前，文章首先利用混合数据进行了测试，直接利用贝叶斯网络结构学习的方法来对数据进行学习。即不利用任何先验的结构假设，仅仅从数据出发，考察能否得到一个符合经验直觉的结果。实际上，从文献综述出发，理想的贝叶斯网络结构应该是经济环境、基金特征和基金经理三个方面的因素共同决定了基金业绩（本文并未加入经济环境这一变量，主要是因为本文并未采用动态的数据，经济环境数据对所有基金来说截面上并无差异，因此加入经济环境变量并无意义），本文利用贝叶斯网络初步学习也希望得到这样的网络结构。

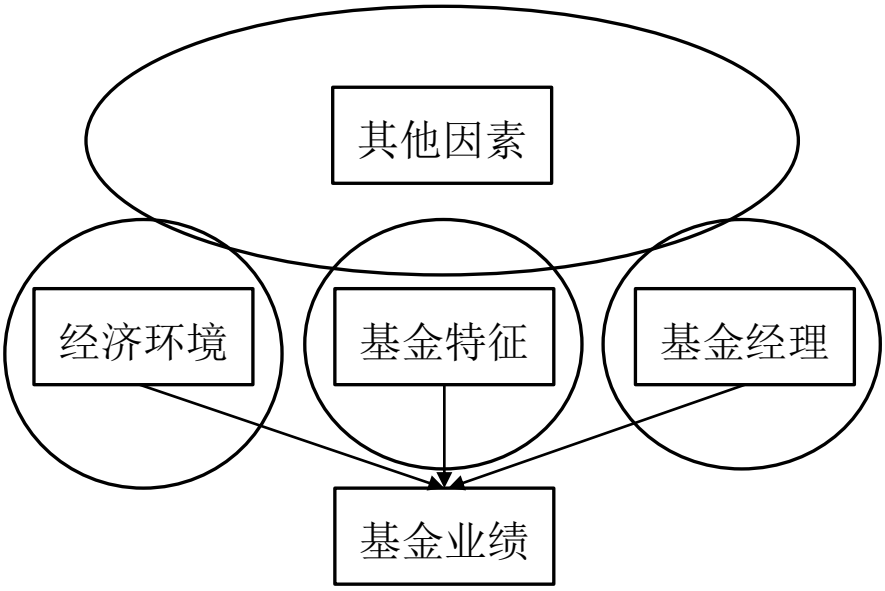

图 8 理想的贝叶斯网络结构

实际上，本文事先对混合数据采取了三种计算方法，得到了三个网络图计算结果。

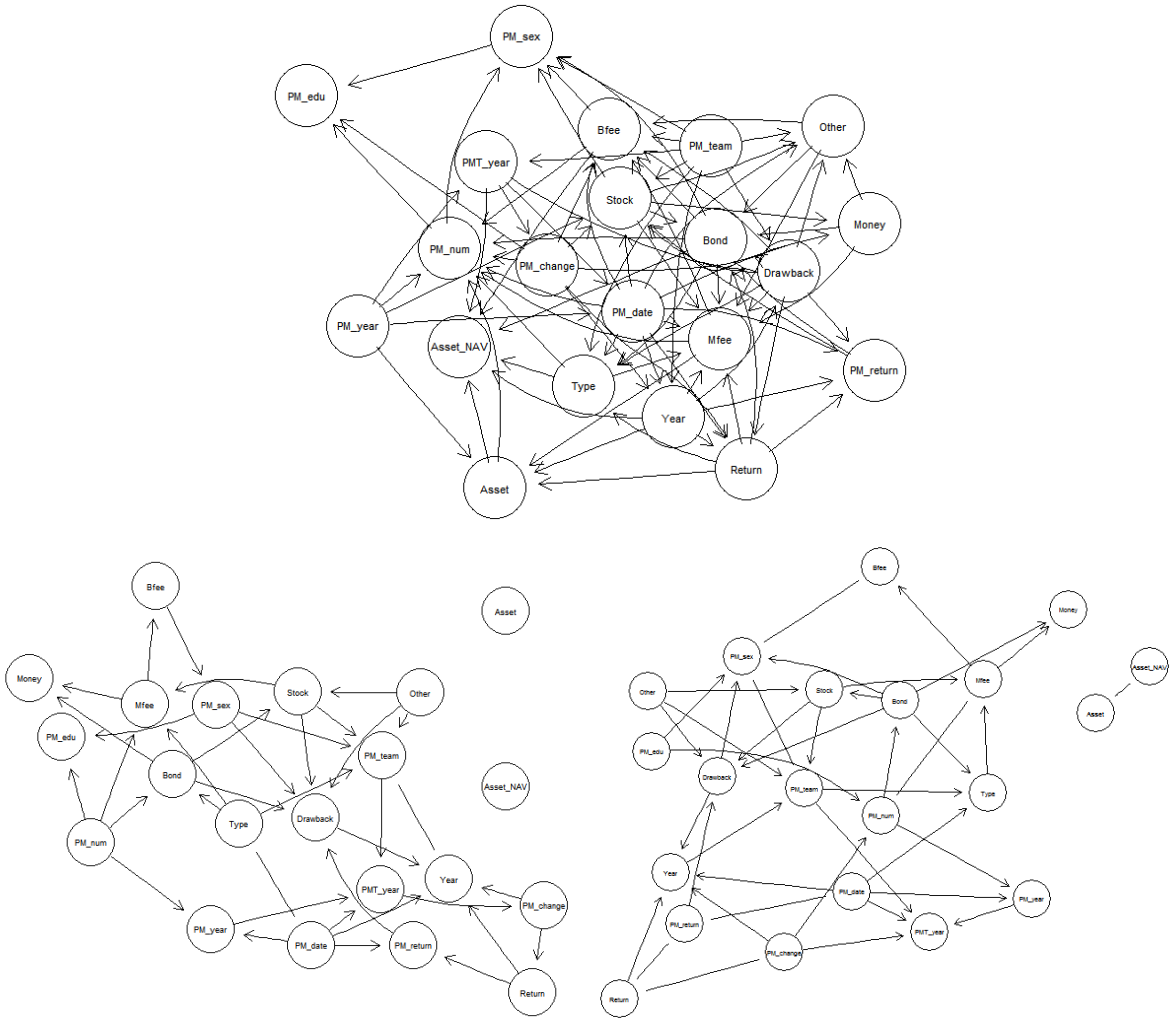

图 9 未加入任何假设的贝叶斯网络学习结果

这三个网络图学习的结果分别来自于三种算法，这三种算法分别来自于基于得分的贝叶斯网络学习和基于约束的贝叶斯网络学习，分别是 HC 算法（爬山法），Grow-Shrink 算法和 Incremental Association 算法。从这三种图中可以得到如下几个信息：其一，不同的算法计算结果差异较大，基于得分的算法得到的网络结构明显比基于得分的网络结构更加复杂；其二，若不对网络结构进行任何的先验设置，得到的网络结构与直觉差异巨大，一个难以解释的问题是，收益率作为最终需要被验证的变量，其与大多数变量之间表现出了条件独立性的同时，还出现了其作为父节点成为其他子节点条件的情况，这与文章需要讨论的问题是不符合的；其三，在基于约束的两种算法之下，得到的网络结构除了部分细微的差异，其他部分的差异相对较小；其四，基金资产规模（Asset）和基金资产净值规模（Asset_NAV）这两个变量在利用基于约束的模型时，出现了与其他所有变量独立的情况，本文认为一个可能的原因是，公募基金在进行投资时，往往会选择流动性相对较好的品种（从资产配置的比例也可以看出），因此，有足够多流动性好的资产可以供公募基金进行配置，规模与收益率和其他

变量之间的相关性相对较弱。基于上述发现，本文认为，需要做出如下三个方面的处理：其一，对网络结构进行简单的先验给定，即限制“成立以来年化回报→其他变量”这一情形，这一改变是为了文章讨论的问题所给定，而不是其他变量依赖关系的先验设定，因此，不会对基于数据贝叶斯网络学习造成较大影响；其二，剔除基金资产规模和基金资产净值规模这两个变量，减少干扰变量的同时提升计算效率；其三，对各种结构学习方式进行比较，主要基于损失函数的预计损失。

3.3 离散模型的学习

3.3.1 离散模型：结构学习

前面已经提到，在 bnlearn 中利用贝叶斯网络进行结构学习时，包括多种思路，每种速录下也有不同的算法。具体而言，bnlearn 中包括的算法有基于得分的计算方法、基于约束的计算方法和混合方法，而每个思路下均有两种算法。而每个算法下学习的贝叶斯网络图都有所不同，因此，需要找到一个标准进行判断。实际上，损失函数可以作为一个较好的标准。损失函数一般是用来估计模型的预测值与真实值的不一致程度，损失函数越小则模型的稳定性越好。一般来说，贝叶斯网络结构学习损失函数包括对数似然损失函数、平方损失函数和分类误差损失函数等。这几类不同的损失函数针对不同的模型，本文采用的离散模型，因此选择了对数似然损失函数。可以发现，得分的排序为 Tabu Search < Hill-Climing < Incremental Association < Grow Shrink < Restricted Maximization。从思路而言，基于得分的计算方法最优。因此，本文选择保留这基于得分的两种算法下得到的贝叶斯网络结构。

表 6 各算法下预计损失

	算法	排序	预计损失
基于得分	Hill-Climbing	2	9.375049
	Tabu Search	1	9.269375
基于约束	Grow-Shrink	5	11.15711
	Incremental Association	3	10.36366
混合方法	Max-Min Hill-Climbing	4	10.97018
	Restricted Maximizaition	6	11.23163

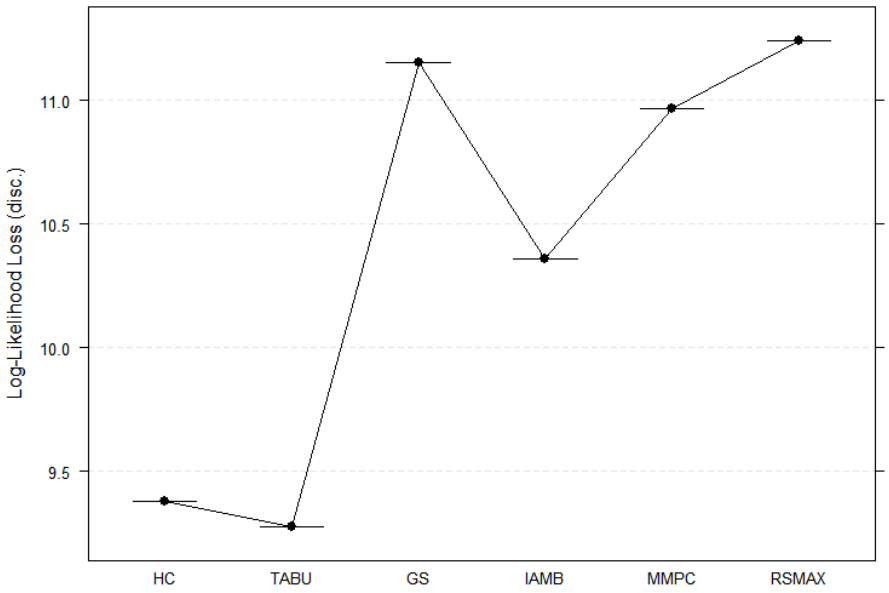

图 10 各算法下预计损失

如下为基于得分的两种算法所得到的贝叶斯网络结构图，可以看到的是，在设置了“黑名单”后，得到的网络图含义较之前的网络图相对更为清晰。成立至今年化回报的父节点仅有两个，一个是管理该基金的基金经理年化回报，一个是管理该基金的基金经理是否发生变动。也就是说，从这种相对稳定的网络结构图中，可以得到这样的独立关系，即

$$(\mathrm{Re}turn \perp all\ other\ \mathrm{var}iables \mid PM\_return, PM\_change)$$

从基金经理年化回报和基金经理是否变动的父节点来看，这两个变量的父节点集合为（股票持仓，基金成立年限，基金经理组平均工作年限），这也是符合直觉的。股票持仓反映了基金经理的资产配置选择，进一步影响收益率；而基金成立年限越长，其更换基金经理的概率率越大，基金经理组工作年限同理。在往上一层的父节点来看，上一个父节点集合的父节点集合为（基金类型，基金经理工作年限）。综合而言，Hill-Climbing 和 Tabu Search 方法得到的依赖关系可以这样表示

Return←(PM_return; PM_change)←(Stock;PMT_year,Year) ←(Mfee;PM_year,Type) ←(PM_date)

Return←(PM_return; PM_change)←(Stock;PMT_year,Year) ←(PM_year,Type) ←(PM_date)。

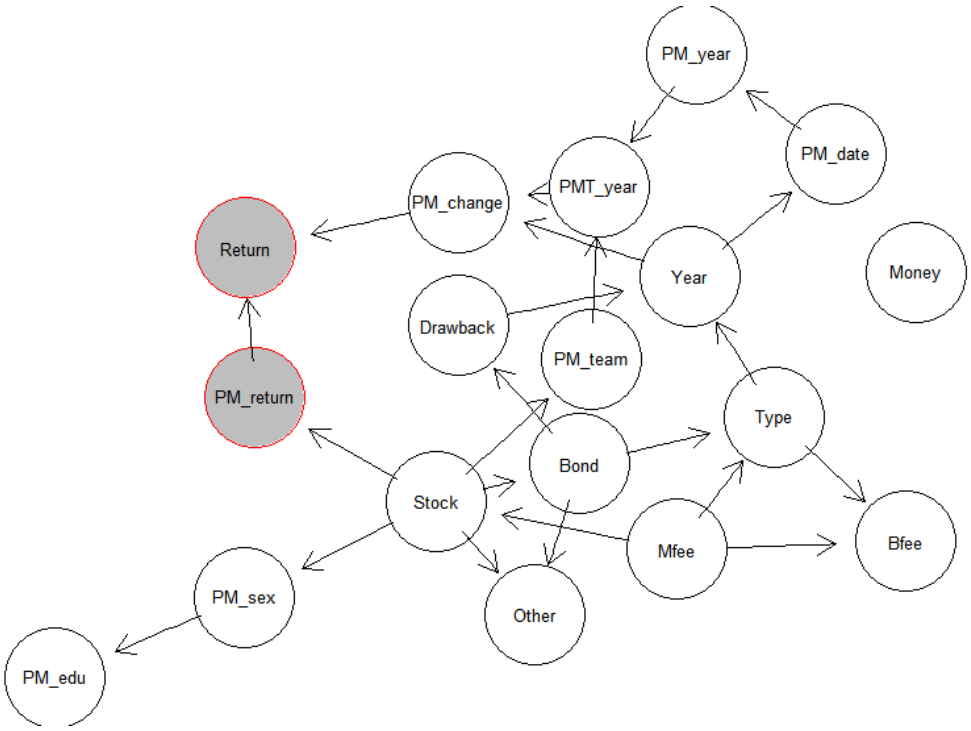

图 11 基于 HC 的贝叶斯网络结构

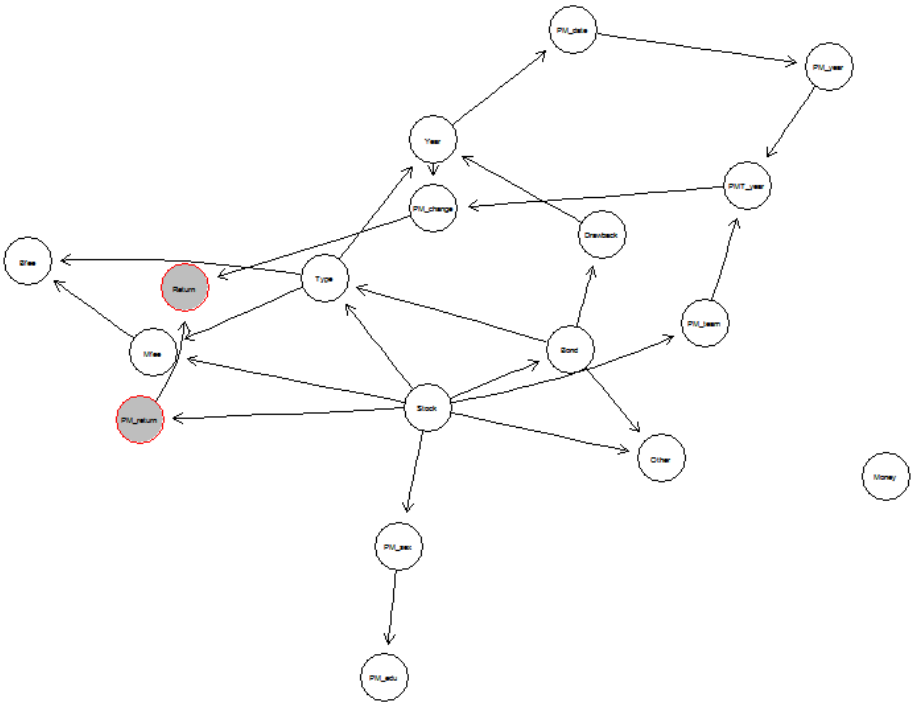

图 12 基于 TABU 的贝叶斯网络结构

3.3.2 离散模型：参数学习

参数学习部分指的是将所有离散变量的条件概率估计出来，由于涉及变量较多，本文只展示“基金成立至今年化回报（Return)”这一变量的条件概率分布，其父节点只包括了“基金经理年化回报（PM_return)”和“是否更换基金经理（PM_change)”。参数学习结果如下所示。需要说明的是，HC 算法和 TABU 算法得到的参数学习结果是相同的。其原因在于，HC 算法和 TABU 算法均为基于极大似然函数下的得分函数算法，其参数估计依照极大似然函数的估计，而 3.2.2 中得到的贝叶斯网络结构明显除了在 Mfee 这一变量处有细小差异外，无其他差异。因此，联合概率分布的分解在 Return 这一随机变量处不会有差别，参数估计自然也不会受到影响。

表 7 HC 和 TABU 参数学习——Return 条件概率分布

PM_change=A				
Return/PM_return	A	B	C	D
A	0.9872	0.0433	0.0000	0.0000
B	0.0128	0.9567	0.5390	0.0000
C	0.0000	0.0000	0.4610	0.2000
D	0.0000	0.0000	0.0000	0.8000
PM_change=B				
Return/PM_return	A	B	C	D
A	0.1919	0.0251	0.0146	0.0000
B	0.7778	0.9287	0.8041	0.6429
C	0.0303	0.0455	0.1813	0.2857
D	0.0000	0.0007	0.0000	0.0714

参数学习的结果是符合逻辑的。整体而言，无论是否更换基金经理，若基金经理过往业绩较好，则该基金获得较高回报的概率较大；而对于过往业绩不好的基金经理，其获得较差回报的概率较大。具体而言：第一，若不更换基金经理，好的基金经理其回报 98%以上的概率会落在最好的区间范围内，而过往业绩表现最差的基金经理其回报 80%的概率仍然会落在最差的区间范围内；第二，不更换基金经理的情况下，除了 A 档历史收益的基金经理有概率使得管理的基金收益落入更低档，其他档历史收益的基金经理均有正的概率提升其管理基金的收益，且不会使得其管理基金落入更低档，且 C 档基金经理实现跨越的可能性最大；第三，更换基金经理后，基金的成立至今收益率分布相较于不变动基金经理变得更为广泛且概率大小发生反转；第四，更换基金经历后，历史收益为 A 档的基金经理大概率（77%以上）下降至 B 档；第五，更换基金经理后，历史收益为 C 档的基金经理以 80%的概率提升其业绩至 B 档，且有概率提升至 A 当，而历史收益为 D 档的基金经理则以 64%的概率跳升至 B 当，以 28%的概率升至 C 档。也即是说，在判断基金的投资价值时，可以参考这样一个逻辑：1、若该基金未更换主

要基金经理，则根据该基金经理历史业绩来判断是否购买该基金；2、若该基金确定更换主要基金经理且新的基金经理历史业绩较好，则警惕其历史业绩下滑的可能性；3、若该基金确定更换主要基金经理且新的基金经理历史业绩较差，则考虑该基金收益上升的空间。

3.3.3 模型推理：各变量对收益率的影响

前面两部分已经学习了一个完整的基金评价变量贝叶斯网络结构和分布体系，本部分主要是想讨论利用这一体系“回答问题”，即在给定各个变量的新信息后，计算收益率可能的范围，即已知某些变量的状态，讨论收益率状态的分布。本文利用的是联合数算法来执行推理，即是把贝叶斯网变换成一个联合树（junction tree）来执行再用原贝叶斯网的局部分布参数去计算联合树中复合节点的参数集合。下表给出了收益率两个父节点“PM_change”和“PM_return”对收益率分布的推理，并给出了另外两个更高节点“PMT_year”和“Stock”对收益率分布的推理。

表 8 HC 和 TABU 参数学习——Return 条件概率分布

	基金收益率			
基金经理历史收益	A	B	C	D
>75%	0.4722	0.5082	0.0196	0.0000
50%-75%	0.0322	0.9395	0.0279	0.0004
25%-50%	0.0094	0.7092	0.2814	0.0000
<25%	0.0000	0.4040	0.2539	0.3421
更换基金经理				
否	0.0863	0.8386	0.0703	0.0048
是	0.0332	0.8974	0.0685	0.0010
基金经理组工作年限				
>75%	0.0482	0.8817	0.0680	0.0021
50%-75%	0.0573	0.8706	0.0694	0.0027
25%-50%	0.0677	0.8550	0.0741	0.0033
<25%	0.0766	0.8395	0.0801	0.0038
配置中股票占比				
>75%	0.0352	0.9228	0.0400	0.0021
50%-75%	0.0525	0.8560	0.0911	0.0003
25%-50%	0.0739	0.8290	0.0940	0.0031
<25%	0.0719	0.8255	0.0998	0.0028
注：以上百分比均为分位数				

基金经理历史收益对基金收益率推理概率分布的影响较为明显，在不同的基金历史收益率水平下基金的收益率分布表现差异较大。具体而言，基金经理历史收益 75%分为以上时，基金收益落入 A 档的概率最高，为 47.22%，而基金经理历史收益在其他分位时，基金收益落入 B 档的概率最高，且该概率逐步下降，落入 C 档和 D 档的概率逐渐提高。股票占比作为基金经理历史收益的父节点，其对收益率分布的推理概率在四个分位下表现类似，均呈现出 B 档概率远高于起他档，C 档次之，而 A 档和 D 档更低的情况。也即是说，通过观察股票占比很难推断出选择何种基金。

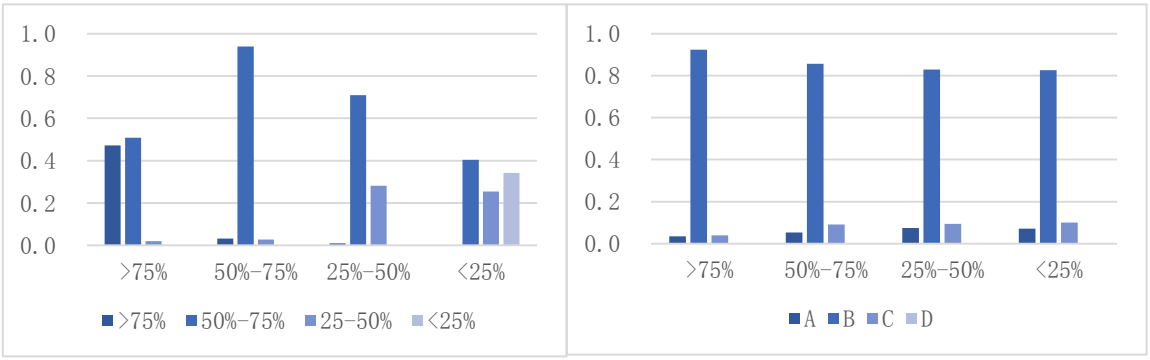

图 13 基金经理历史收益（左）和股票占比（右）对基金收益的推理条件概率分布

是否更换基金经理和其父节点股票占比对基金收益的推理条件概率分布在不同成都下也表现一直。具体而言，无论是否更换基金经理，基金收益的条件概率都是 B 档最高，不更换基金经理落入 A 档概率较更换基金经理落入 B 档概率高。股票占比对基金收益率的推理条件概率分布同样表现出 B 档最高的特征，另一个特征是，股票占比越低，落入 A 档的概率越高。

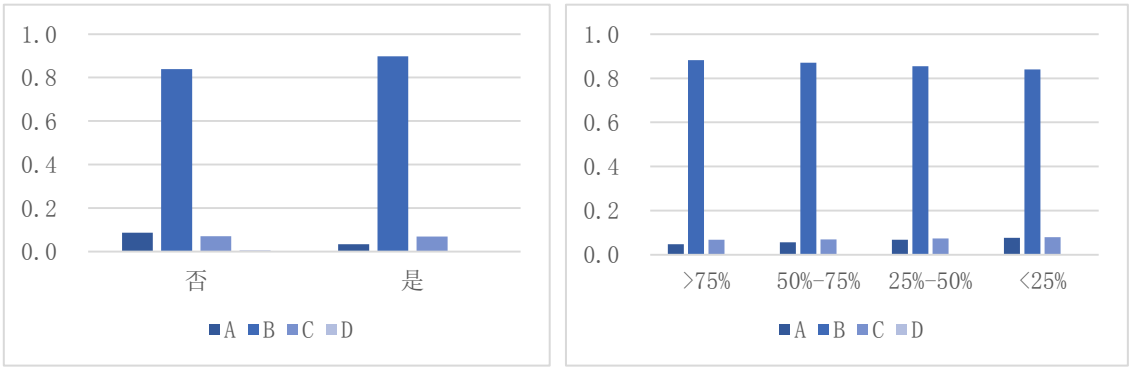

图 14 基金经理历史收益（左）和股票占比（右）对基金收益的推理条件概率分布

3.3.4 实证评述

本部分的实证讨论遵循了如此的过程：首先，根据文献和研究问题初步确定需要进入贝叶斯网络结构的所有可能变量；其次，根据模型需要对数据进行离散化且同时将其设置成虚拟变量；再次，根据初探模型对变量进行进一步的筛选，并确定设置一

定额黑名单以保证研究目的能够在贝叶斯网络图中得到体现；进一步，对贝叶斯网络结构进行学习并根据损失函数选择稳定结构的算法，得到贝叶斯网络结构；最后，对参数进行估计。这一套完整的贝叶斯网络评价结构能够在很多金融数据中得到应用，在本文的应用中也对基金投资起到了一定的指导意义。然而，值得说明的是，由于文章是尝试性的使用此类方法并应用于金融数据评价中，相应的数据处理过程应在更加完善的算法下得到加强，且学习结果中有部分依赖关系难以被直观解释。若要更为精确的研究此类问题，需要对模型加入更为严格的先验设计，且找到相应的文献支持。

3.4 与 LASSO 结果比较

实际上，在探讨收益率的影响因素时，为了减小因为缺少重要自变量而出现的模型偏差，本文在前述讨论时总结了过往文献所有可能影响基金收益率的因素。然而，无论是贝叶斯网络，还是计量方法们都是希望能够找到对因变量具有最强解释力的自变量集合，以提高模型的解释性和预测精度。LASSO 算法实际上是一种能够对变量进行选择且精简模型的估计方法，文章希望通过比较 LASSO 方法和概率图方法得到的结论，来进一步探讨概率图模型的有效性和可靠性。

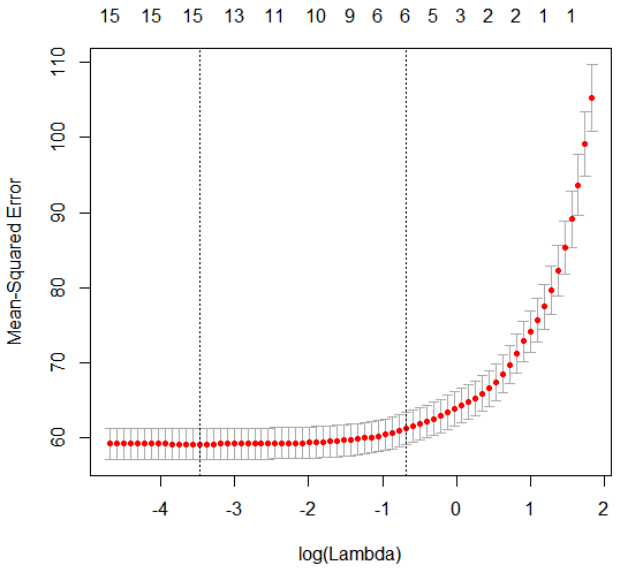

图 15 交叉验证确定模型

表 9 LASSO 选择系数

变量	系数	节点
Year	0.54157	2

Mfee	0.30808	4
PM_change	0.95091	1
PM_return	0.49161	1
Drawback	0.0212	3
Bond	0.00744	4

由上图可知，交叉验证得到模型最佳 lambda 值为 0.5061308。进一步地，由上表可知，由最佳模型选择的变量为(Year,Mfee,PM_change，PM_return,Drawback,Bond)。其中，(PM_change，PM_return)为 Return 父节点，Year 与 Return 距离为 2 步，Drawback 与 Return 距离为 3 步，（Mfee，Bond）与 Return 距离为 4 步。从 LASSO 算法选出变量来看，保留了 PM_return 和 PM_change 这两个变量，与概率图模型结果一致。这进一步说明了，概率图模型在给定了一定的先验设计后，其得到的结果也能从其他计量模型中得到一定的验证。这使得我们有理由相信，在给定完整的先验设计后，概率图模型的结果将会是可靠的。

第四章　结论及展望

4.1 结论

本文试图利用概率图模型这一方法来评价中国上市公募基金的绩效，文章讨论了概率图模型的主要理论和应用，在利用中国上市公募基金数据进行讨论时，文章首先采用了不给定先验设计的概率图模型进行估计，得到不符合直觉的贝叶斯网络图；在给定了一定的先验设计后，本文得到了相对较为明确的网络图结构，并得到了有意思的条件概率分布。然而，尽管“收益率”的两个父节点实际上满足直觉且能够提供一定的知识，其他节点的信息却存在明显的偏误。一个直接的例子是“Stock”到“PM_SEX”的父子节点关系，当然，存在这样一种解释，债券型基金的基金经理中女性的比例要高于股票型基金。尽管如此，父子节点的关系却反向了，按照这一逻辑，“PM_SEX”这一变量应当是父节点而非子节点。

因此，从本文的例子来看，利用概率图模型来探讨金融经济变量的网络关系，的确存在不小的问题。可能的原因在于，一方面，变量找寻不全，尽管文章已经探讨了众多影响基金收益的变量，但仍然可能存在共同影响基金收益的变量，如心理因素等这些难以刻画的变量无法寻找；其二，尽管数据能够反映一定的变量关系，但先验的设计十分重要，对于已经存在的明确条件独立关系，应当能够事先给定。而问题的难点在于，先验的设计需要对网络系统有一定的知识，而概率图模型便是用来寻找这种知识。因此，这是一个矛盾。可能解决这一矛盾的方法在于，将变量关系分为两类，一类是共同知识，另一类是关系尚不明确的，对于所有的共同知识事先给定，而另一类不明确的管理利用概率图模型来进行探索，并动态的调整网络结构。也即是说，概率图模型希望达到的目标在金融数据中的应用仍然需要人类经验的总结，单纯从数据出发，获得的网络结构甚至不能战胜常识。

4.2 研究展望

本文的研究存在以下几个方面仍然可以改进：其一，文章采用的数据是静态数据，由于市场环境、政策环境等均是时间相关的变量，故未加入实证部分，若采用动态方法，本文得到的结论可能更加有意义。其二，前面已经提到，金融类问题的先验设计部分十分重要，本文仅仅给出了黑名单“收益率->其他变量”，这只是最为直接的结构设计，本文并未给出其他先验设计，在进一步研究此类问题时，可以给出更为精确的结构设计和分类。最后，本文采用的学习方法和算法均采用已有数据包进行的拓展，且只采用了离散数据进行探讨，进一步可以拓展至连续模型，并采用更为高效的算法。

参考文献

Abramson, B., & Finizza, A. (1995). Probabilistic forecasts from probabilistic models: a case study in the oil market. International Journal of forecasting, 11(1), 63-72.

Acharya, V. V., & Pedersen, L. H. (2005). Asset pricing with liquidity risk. *Journal of financial Economics*, *77*(2), 375-410.

Babalos, V., Caporale, G. M., Kostakis, A., & Philippas, N. (2008). Testing for persistence in mutual fund performance and the ex-post verification problem: evidence from the Greek market. The European Journal of Finance, 14(8), 735-753.

Bauer, R., Koedijk, K., & Otten, R. (2005). International evidence on ethical mutual fund performance and investment style. *Journal of Banking & Finance*, *29*(7), 1751-1767.

Chen, J., Hong, H., Huang, M., & Kubik, J. D. (2004). Does fund size erode mutual fund performance? The role of liquidity and organization. *The American Economic Review*, *94*(5), 1276-1302.

Chevalier, J., & Ellison, G. (1999). Career concerns of mutual fund managers. *The Quarterly Journal of Economics*, *114*(2), 389-432.

Cooper, G. F., & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine learning*, *9*(4), 309-347.

Demirer, R., Mau, R. R., & Shenoy, C. (2006). Bayesian networks: a decision tool to improve portfolio risk analysis. *Journal of applied finance*, *16*(2), 106.

Ferreira, M. A., Keswani, A., Miguel, A. F., & Ramos, S. B. (2012). The determinants of mutual fund performance: A cross-country study. *Review of Finance*, rfs013.

Gemela, J. (2001). Financial analysis using Bayesian networks. *Applied Stochastic Models in Business and Industry*, *17*(1), 57-67.

Kemmerer, B., Mishra, S., & SHENOY, P. P. (2002, August). BAYESIAN CAUSAL MAPS AS DECISION AIDS IN VENTURE CAPITAL DECISION MAKING: METHODS AND APPLICATIONS. In Academy of Management Proceedings (Vol. 2002, No. 1, pp. C1-C6). Academy of Management.

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Nadkarni, S., & Shenoy, P. P. (2001). A Bayesian network approach to making inferences in causal maps. *European Journal of Operational Research*, *128*(3), 479-498.

Pearl, J. (1986). Fusion, propagation, and structuring in belief networks. *Artificial intelligence*, *29*(3), 241-288.

Robinson, R. W. (1977). Counting unlabeled acyclic digraphs. In *Combinatorial mathematics V* (pp. 28-43). Springer Berlin Heidelberg.

Sanford, A. D., & Moosa, I. A. (2012). A Bayesian network structure for operational risk modelling in structured finance operations. *Journal of the Operational Research Society*, *63*(4), 431-444.

致谢

本论文是在胡大源老师的悉心指导下完成的。从论文选题开始，胡老师十分耐心的引导我选择具有挑战性且有实际价值的题目，到论文写作过程中，胡老师从参考文献和论文结构等各个方面入手指导我进行写作，最后的结稿胡老师也提供了十分有建设性的修改意见。总而言之，本文的完成离不开胡老师的指导，我十分感谢胡老师在此期间对我的关心和帮助，我从胡老师这里真正感受到了学者风范和导师的关怀。除此之外，我还要感谢胡老师在研究生这几年期间对我学习和生活的指导。从二年级确定导师开始，每次跟胡老师的交流都十分深入，话题不仅仅局限于学业，关于工作习惯和为人处世的道理，胡老师也对我们谆谆教诲，我受益匪浅。可以说，我很多不良工作习惯和处事习惯都在胡老师这里得到了纠正。我在此真心给胡老师说一声谢谢！

我也要感谢我的父母，从高中开始，我的父亲母亲对我学业和生活上都给予了充分的信任，除非必要的提醒，其余并不多加干涉。这也让我很早学会了独立决策，也免去了很多来自父母的压力和困扰。拿起容易放下难，我在此十分感谢父亲母亲对我的信任，也谢谢二位对我选择的支持。

我还要感谢国家发展研究院和光华管理学院的许多任课老师，在学校期间，我修读了国家发展研究院和光华管理学院的许多课程，在这里感受到了大师的魅力，并进一步提升了自身的经济学素养。也十分感谢研究生办公室的各位行政老师，能够基于我们在学业、生活和职业上诸多帮助。

我还要感谢我的研究生同学和北大的其他同学们，我在生活中得到了很多同学的帮衬，三年生活的点点滴滴难以忘记，最终只能化作一声谢谢，希望未来无论在学界、业界或者政界，大家永远是朋友。

最后，我要特别感谢中国人民大学的韩松老师，韩老师是我的本科班主任，从本科开始我的很多决策和判断都是与韩老师商量后完成的，从入学、保研论文、本科毕业到硕士工作、毕业，韩老师都给予我很多支持和帮助，谢谢！

漫漫人生路，我在校的生活即将结束，梦想就在远处，我会一直努力。

再次感谢你们！

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