基于分位数回归和关联规则挖掘的投资组合选择模型①

房 勇，黄 亮

（中国科学院 数学与系统科学研究院，北京 100190；中国科学院大学 经济与管理学院，北京 100190）

摘 要：传统投资组合选择模型的输入参数一般都是基于历史数据计算得到的点估计值。本文采用分位数回归估计特定因子下的股票收益率的边缘分布，利用关联规则挖掘算法挖掘股票的相关性，然后采用蒙特卡罗模拟方法估计出特定因子下股票收益率的联合分布，进一步根据估计的联合分布构建股票的投资组合选择模型，修正了传统的投资组合选择模型。最后，采用上证 50 指数 18 只成分股的实际数据对模型进行了验证，结果表明，修正的投资组合选择模型优于传统的投资组合选择模型。

关键词：投资组合选择模型；分位数回归；关联规则挖掘

中图分类号：F830 **文献标识码**：A **文章编号**：(2019)01-0038-11

0 引言

Markowitz 提出了经典的投资组合理论，即均值方差模型，他用投资组合的历史收益率的期望来衡量投资组合的未来收益，方差来衡量投资组合的风险，然后根据历史收益率求出给定风险下收益最高的投资组合或给定收益下风险最小的投资组合，得到的投资组合即为最优的投资组合[1]。均值方差模型自提出以来，投资组合选择问题便引起了国内外众多学者的研究。国外学者中，Rockafellar 和 Uryasev 将 CVaR 作为度量风险的方式，构建了基于 CVaR 的投资组合选择模型[2]。Norkin 和 Boyko 从尾部风险出发，将低于给定阈值的概率作为风险度量的方式，构造了基于第一安全准则的投资组合选择模型[3]。Bulmuş 和 Özekici 从投资者效用出发，构建了基于投资者效用的投资组合模型[4]。Czichowsky 构造了连续时间下的投资组合选择模型[5]。国内研究中，更多的是结合我国证券市场的实际情况构造投资组合选择模型。例如：陈志英认为市场状态并不是固定不变的，而是随着宏观经济的变化而变化，因此其构造了不同市场状态下投资组合选择模型[6]。姚海祥等采用下半方差和下半偏差度量风险，用非参数方法估计股票收益率均值，然后构建了投资组合选择模型[7]。柏林等将 Black-Litterman 模型推广到多期框架下分析，并建立了考虑投资者观点的多期 Mean-CVaR 投资组合模型[8]。

然而，传统的投资组合选择模型的参数估计一般是建立在历史数据基础上的，最重要的缺陷是认为股票收益率的未来分布与过去分布是相同的。大量的实证研究表明股票收益率会受到众多因素的影响如公司的盈利能力、财务状况等，因此，需要对股票收益的未来分布予以重新估计。在估计股票收益的未来分布时，一方面是需要估计股票收益的边缘分布，另一个方面是需要估计股票收益的相关性。

股票收益的边缘分布会受到各种因素的影响。Sharpe 等在 Markowitz 研究的工作基础上，提出了现代金融学的核心内容：资本资产定价模型（CAMP 模型）。资本资产定价模型本质上建立的是股票收益率与市场组合收益率的单因子模型[9]。Fama 和 French 对美国股票市场进行实证研究，认为股票的收益率除了会受到市场组合收益率的影响，还会受到公司规模、账面市值比两个因素的影响，由此提出了影响股票收益的三因素模型[10]。Fama 和 French 在其三因素模型的基础上，认为股票收益还受到公司盈利能力和公司投资额的影响，提出了五因素模型[11]。Daniel 和 Moskowitz 认为虽然短期内动量因子有可能失效，但是长

① 基金项目：国家自然科学基金项目（71631008）。

作者简介：房勇（1974—），男，山东聊城人，博士，中国科学院数学与系统科学研究院副研究员，博士生导师，研究方向：金融工程与风险管理、决策科学，E-mail：yfang@amss.ac.cn；黄亮（1990—），男，重庆忠县人，硕士，马上消费金融股份有限公司，研究方向：量化投资，E-Smail：huangliang14@mails.ucas.ac.cn。

期来讲，动量因子对股票收益率的影响较为显著，动量因子本质上刻画的是股票收益率的自相关性[12]。国内研究中，田利辉和王冠英在 Fama-French 三因素模型的基础上，通过实证研究表明，我国股票收益率还受到成交额和换手率的影响[13]。王宜峰等研究了上市公司投资对股票收益的影响，认为中国股票市场存在着显著的投资效应：投资额较高的公司前期的股票收益高于投资额低的公司，在后期则相反[14]。

传统的多因素模型往往采用最小二乘法进行计算，但是最小二乘法的运用是建立在残差服从独立、同分布的正态分布的基础上的。为了更加准确的刻画股票收益率与其影响因素的相关关系，本文采用 Koenker 和 Bassett[15] 提出的分位数回归方法来刻画股票的收益率分布。一方面分位数回归模型对残差的假设较少，另一方面分位数回归模型建立不同分位点的股票收益率与特定因子的回归模型，因此，通过分位数回归模型可以直接得到特定因子下股票收益率的概率分布函数，即可以得到股票收益的边缘分布。

在估计股票的相关性方面，一般采用相关系数来刻画股票的相关性，但是相关系数是建立在线性关系的基础上，无法刻画股票的非线性关系。为了能够刻画股票高度复杂和高度非线性的关系，本文采用 Agrawal 等 [16] 提出的关联规则挖掘算法来挖掘股票的相关性，其核心思想是将股票的收益率离散化，分成若干个区间，然后采用关联规则挖掘算法挖掘出股票收益率的关联规则，用关联规则来刻画股票的相关性。进一步根据蒙特卡罗模拟方法估计出股票收益率的联合分布，最后根据估计的联合分布构建修正的投资组合选择模型。

论文安排如下：第 1 部分介绍传统的投资组合选择模型并指出值得改进的地方；第 2 部分介绍分位数回归模型和关联规则挖掘算法；第 3 部分给出基于分位数回归和关联规则挖掘的股票收益率联合分布的估计方法以及修正模型的主要思路；第 4 部分对提出的投资组合选择模型进行数值验证，第 5 部分给出结论与展望。

1 传统的投资组合选择模型

投资组合选择模型最经典的模型是 Markowitz 提出的均值方差模型。均值方差模型用均值来刻画股票的收益，用方差来刻画股票的风险，通常高收益伴随着高风险，因此，总是希望在给定风险下收益最大或者给定收益下风险最小，假设 n 只股票的历史收益率均值为 $\bar{r}_i$，$i=1$，2，$\cdots$，n，协方差矩阵为 $\sum_{i,j}$，$i=1$，2，$\cdots$，n，$j=1$，2，$\cdots$，n。假设每只股票的权重为 w_i，$i=1$，2，$\cdots$，n，则均值方差模型可以表述为

$$\max_w \sum_{i=1}^{n} w_i \bar{r}_i$$

$$\text{s. t.} \sum_{i=1}^{n} \sum_{j=1}^{n} w_j \sum_{i,j} w_i \leqslant \sigma^2$$

$$\sum_{i=1}^{n} w_i = 1$$

该模型表示给定风险时，期望收益最大；给定收益时，风险最小的模型也很容易同理得到；同时，由于各国股票市场制度的不同，也可以对 w_i 添加限制各种条件，如不允许卖空，则添加条件 $w_i \geqslant 0$。

传统的投资组合选择模型，大多是采用历史数据来估计输入参数，即期望收益和协方差矩阵，最重要的缺陷是认为未来分布与过去分布是相同的，忽视了特定因子对股票收益的影响，因此需要估计特定因子下股票收益的联合分布，可以将其分成两个问题：股票边缘分布的估计和股票相关性的估计。

2 分位数回归模型和关联规则挖掘算法

2.1 分位数回归模型

最小二乘法是传统的回归分析中研究最广的方法，但是最小二乘法具有一系列严格的假设：残差服从独立、同分布的正态分布，当违背其中一条假设时，估计的参数是有偏的。因此，这一系列严格的假设在一定程度上限制了最小二乘法的应用。相对于最小二乘法而言，分位数回归的理论假设弱于最小二乘法，

同时，最小二乘法研究的是被解释变量的期望值与解释变量的关系，分位数回归研究的是被解释变量的分位点与解释变量的关系，因此，分位数回归得到的信息更加丰富。

设被解释变量 y，解释变量为 x_i，$i=1$，2，…，m，解释变量与被解释变量的线性回归模型：

$$y = \alpha + \sum_{i=1}^{m} \beta_i x_i + \varepsilon$$

其中 ε 为误差项，并且假设样本数据为 $\{y_t, x_{i,t}\}$，则 τ 分位数回归则是求满足：

$$\min_{\alpha(\tau),\beta(\tau)} \sum_{t=1}^{T} \rho_\tau (y_t - \alpha - \sum_{i=1}^{m} \beta_i x_{i,t})$$

的解 $\alpha(\tau)$ 和 $\beta_i(\tau)$，其中 $\rho_\tau(u) = [\tau - I(u<0)]u$。在线性条件下，给定 x_i 后，y 的 τ 分位数函数为：

$$Q(\tau \mid x) = \alpha(\tau) + \sum_{i=1}^{m} \beta_i(\tau) x_i, \ \tau \in [0, 1]$$

分位数回归自提出后，便受到广泛的研究和应用。例如，Chernozhukov 等研究了分位数回归模型中的内生性问题[17]，Tsai 用分位数回归模型研究了亚洲国家的股票价格和汇率的相关关系[18]。

2.2 关联规则挖掘算法

近年来，随着人类积累的数据量越来越大，大数据时代已经来临，在大数据时代，数据之间的关系呈现出高度复杂和高度非线性的特点，采用传统的相关系数刻画数据的相关性已经无法满足用户的要求，如何从海量数据中挖掘这些高度复杂和高度非线性的关系成为学者研究的重点。关联规则挖掘正是基于此而产生的，它是数据挖掘领域中的一个应用非常广泛的领域，主要用于发现隐藏在大数据中有意义的关系和联系。它反映了多个事件之间的关联和依赖。

关联规则挖掘最早是 1993 年由 Agrawal 等提出的。具体而言，他们提出了关联规则挖掘中经典的算法 Apriori 算法，通过关联规则挖掘算法提取超市购物中顾客购买商品的关联规则，从而为超市制定决策提供依据[16]。

设 $I=\{i_1, i_2, \cdots, i_n\}$ 是 n 个不同项目的集合，D 是针对 I 的事务集，每一个事务 T 包含若干个项目，且每一个事务都有一个唯一的标识 TID。一个常见事务集是超市中每位顾客的购买产品，项目包括超市的所有商品。

关联规则挖掘就是在事务集 D 中找到形如 $X \rightarrow Y$ 这样的蕴含式，其中 X，$Y \subset I$ 且 $X \cap Y = \varnothing$，表示 X 在发生时，Y 也发生或者发生的概率很大，X 和 Y 分别称为事务集中的项集，若某个项集包含 k 个项，则该项集称为 k 项集。因此，可以根据事务集 D 简单的计算 X 发生时，Y 也发生的频率值大小来检验某条规则是否为关联规则，这里计算的频率称为规则的置信度，表示为 $C(X \rightarrow Y) = P(X \cap Y \mid X) = \frac{|X \cap Y|}{|X|}$。根据置信度来确定关联规则用了一个事实：可以用频率来近似的估计概率。但是频率估计概率是在样本量很大的情况下，当样本量很小时，频率与概率的值有可能差距较大，因此需要对 X 和 Y 同时发生的频率做一个限制，X 和 Y 同时发生的频率被称为规则的支持度，表示为 $S(X \rightarrow Y) = P(X \cap Y) = \frac{|X \cap Y|}{|D|}$。在寻找关联规则过程中，还需确定两个阈值：最小支持度和最小置信度，把支持度大于等于最小支持度的项集称为频繁项集；支持度大于等于最小支持度且置信度大于等于最小置信度的规则称为关联规则。

因此，关联规则挖掘可以分成两个步骤：首先，找出事务集 D 中的所有频繁项集，即找到支持度大于等于最小支持度的项集；然后根据频繁项集产生关联规则，即根据频繁项集找到置信度大于等于最小置信度的规则。

若 I 中有 n 个项目，这些项目生成的项集数量为 $2n$，当 n 很大时，通过计算所有项集的支持度来查找频繁项集的计算量非常巨大。Apriori 算法的思想是首先计算所有 1 项集的支持度，若所有 1 项集的支持度均小于等于最小支持度，则该事务集中无频繁项集；否则，找出 1 项集中支持度大于等于最小支持度的项集，该项集为 1 频繁项集；根据该频繁项集与自身连接产生 2 项集，同 1 项集一样，计算产生的所有 2 项集的支持度，寻找其支持度大于等于最小支持度的 2 项集，若无，则已找到全部频繁项集。若存在，则以此类推，计算 3 项集的频繁项集，4 项集的频繁项集等等，直至无频繁项集产生为止。则可以找出所有的

频繁项集，找出频繁项集后，则根据频繁项集产生关联规则。

3 基于分位数回归和关联规则挖掘的股票收益率联合分布估计

3.1 股票收益率边缘分布估计

传统上常常假设股票的收益率服从一个给定的分布，一般情况下是假设服从正态分布或者其他分布，然后采用极大似然方法估计出该分布的参数，则可以得到股票的收益率分布，然而这种方法存在着很大的局限。首先是忽略因素的影响，事实上股票的收益率分布会受到众多因素的影响，如公司的盈利能力；其次是假设了股票收益率的分布形式，而实际上股票收益率可能并不服从该分布。应用分位阿数回归恰恰可以解决这个问题，一方面是考虑了因素的影响，另一方面是没有假设分布的形式。

假设第 i 只股票的收益率为 r_i，影响因子为 $f_{i,j}$，其中 $i=1, 2, \cdots, n$，$j=1, 2, \cdots, m$，则建立收益率与影响因子的分位数回归模型：

$$r_i(\tau) = \alpha_i(\tau) + \sum_{i=1}^{m} \beta_{i,j}(\tau) f_{i,j} + \varepsilon_i(\tau)$$

估计出参数 $\alpha_i(\tau)$ 和 $\beta_{i,j}(\tau)$ 后，则可以得到给定因子 $f_{i,j}$ 下股票收益率的 τ 分位数条件分布为

$$Q_i(\tau \mid f) = \alpha_i(\tau) + \sum_{j=1}^{m} \beta_{i,j}(\tau) f_{i,j},\ \tau \in [0, 1]$$

因为 $[0, 1]$ 中的实数不可数，不可能估计出所有的分位数条件分布，只能估计出其中的一部分的分位数条件分布。假设给定 $\tau_1, \tau_2, \cdots, \tau_p$，其中满足 $0=\tau_1<\tau_2<\cdots<\tau_{p-1}<\tau_p=1$，则可以估计出给定因子 $f_{i,j}$ 下的 τ_k 分位数条件分布 $Q_i(\tau_k \mid f) = \alpha_i(\tau_k) + \sum_{j=1}^{m} \beta_{i,j}(\tau_k) f_{i,j}$。对于任意的 τ，若 $\exists k$ 使得 $\tau=\tau_k$ 则 τ 分位数条件分布是 $Q_i(\tau_k \mid f)$，若 $\forall k$ 都有 $\tau \neq \tau_k$，则必然存在 k 使 $\tau_k<\tau<\tau_{k+1}$，此时可以采用线性插值或者其他插值公式计算出 τ 分位数条件分布，其中线性插值计算公式为

$$Q_i(\tau \mid f) = \frac{Q_i(\tau_{k+1} \mid f) - Q_i(\tau_k \mid f)}{\tau_{k+1} - \tau_k}(\tau - \tau_k) + Q_i(\tau_k \mid f)$$

也可以采用二次插值、三次样条插值等方法进行求解。因此，通过分位数回归模型则可以求解给定因子下股票的收益率分布，即可以得到个股收益率的边缘分布。

3.2 股票相关性估计

传统上常常根据股票收益率的相关系数来刻画股票的相关性，若两只股票的相关系数为正，则表明这两只股票有正相关关系，其值越大，代表着相关性越大；若这两只股票的相关系数为负，则表明这两只股票有负相关关系，其值越小，代表着相关性越大。但是相关系数是建立在线性关系的基础上的，无法刻画非线性关系。

随着股票市场的不断发展，股票之间的关系呈现出高度复杂和高度非线性的特点。为了更加准确的刻画股票的相关关系，挖掘其非线性特征，本文采用关联规则挖掘算法来挖掘股票收益的相关性，用关联规则来刻画股票的相关性。股票收益的关联规则可以表达成：股票 k_1 的收益率为 r_1，股票 k_2 收益率为 r_2，$\cdots$，股票 k_p 的收益率为 r_p，则可以推出股票 m_1 的收益率为 R_1，股票 m_2 的收益率为 R_2，$\cdots$，股票 m_q 的收益率为 R_q。因此，若知道关联规则条件里面的股票的收益率，则可以得到结果里面的几只股票的收益率。因此，股票的关联规则挖掘得到的结果可以描述为若某几只股票的收益率为某个值或在某个区间时，则可以得到另外几只股票的收益率为相应的值或相应的区间。同时，股票的收益率是在整个实数区间上连续变动的，而关联规则挖掘的项集的数量往往是有限的，因此需要将股票收益离散化，即将股票收益划分为若干个子区间。设股票的收益率在区间 $[a, b]$ 上波动，将其划分为 m 个区间 $[a_0, a_1]$，$[a_1, a_2]$，$\cdots$，$[a_{m-1}, a_m]$，其中 $a_0=a$，$b_0=b$，记为 $i_1, i_2, \cdots, i_m$ 并且假设有 n 只股票，则项的集合可以表达为 $\{(1, i_1), \cdots, (1, i_m), (2, i_1), \cdots, (2, i_m), \cdots, (n, i_1), \cdots, (n, i_m)\}$，其中 (j, i_k) 表示第 j 只股票的收益率在区间 i_k 里。然后调用关联规则挖掘算法挖掘股票的关联规则，得到的关联规则即可描述股票的相关性。在实际的股票市场中最重要的就是股票的涨或者跌，因此，可以将股票收益划分成两

个区间，即股票收益大于 0 和股票收益小于等于 0。

3.3 股票联合分布估计

采用分位数回归估计出股票的边缘分布后，采用关联规则挖掘算法挖掘出股票的相关性，还有一个重要的工作就是估计股票收益的联合分布，股票的联合分布无法直接计算得到，但是可以采用蒙特卡罗方法估计出股票的联合分布。其思想是：随机模拟一只股票的收益率，然后在关联规则库中找到是否有根据已知股票收益为条件产生的关联规则，若存在，则模拟关联规则结果的股票的收益率，反之，若没有，则继续随机模拟股票的收益率，直至所有的股票均模拟完为止，在模拟前，需要对股票进行编号，设 n 只股票的编号分别为 1，2，…，n。

股票收益的联合分布估计算法表示如下。

1）给定 τ_1，τ_2，…，τ_p，其中满足 $0=\tau_1<\tau_2<\cdots<\tau_{p-1}<\tau_p=1$，则可以估计出给定因子 f 下 n 只股票的 τ_k 分位数条件分布 $Q_i(\tau_k \mid f)=\alpha_i(\tau_k)+\sum_{j=1}^{m}\beta_{i,j}(\tau_k)f_{i,j}$，$i=1,2,\cdots,n$，$k=1,2,\cdots,p$，其中 $Q_i(\tau_k \mid f)$ 表示第 i 只股票在因子 f 下的收益率的 τ_k 条件分位数；

2）利用关联规则挖掘出股票收益的关联规则库 D；

3）设集合 $A=\varnothing$，$B=\{1,2,\cdots,n\}$，$C=\varnothing$，$R=[0,0,\cdots,0]$，A 表示已模拟的股票，B 表示未模拟的股票，C 表示已模拟股票形成的项集，R 表示模拟股票的收益率；

4）产生 [0，1] 之间的均匀分布的随机数 u；

5）必然 $\exists j$ 得 $u\in(j-1,j]$，若 $j\in B$，则根据股票收益的边缘分布根据蒙特卡罗方法模拟第 j 只股票的收益率 r，令 $A=A\cup\{j\}$，$B=B/\{j\}$，$R(j)=r$，根据收益率区间，则必然存在区间 $[a_k,a_{k+1}]$ 满足条件：$r\in[a_k,a_{k+1}]$，$C=C\cup\{(j,[a_k,a_{k+1}])$，转 6）；若 $j\notin B$，则转 4）；

6）在关联规则库 D 中查找是否有前件为 C_1 的关联规则，其中 $C_1\subset C$，若存在，假设关联规则的后件为 $C_2=\{(j_1,[a_{k_1},a_{k_1+1}]),\cdots,(j_m,[a_{k_m},a_{k_m+1}])$，若 $\{j_1,j_2,\cdots,j_m\}\cap A=\varnothing$，转 4）；若 $\{j_1,j_2,\cdots,j_m\}\cap A\neq\varnothing$，则 $A=\{j_1,j_2,\cdots,j_m\}\cup A$，$B=B/(\{j_1,j_2,\cdots,j_m\}\cap A)$，$C=C\cup C_2$，$D=D\setminus\{C_1\rightarrow C_2\}$，则模拟 $\{j_1,j_2,\cdots,j_m\}\cap A$ 股票的收益率 $r_{j_{k_1}},\cdots,r_{j_{k_L}}$，令 $R(j_{k_q})=r_{j_{k_q}}$，$q=1,2,..,L$，重复 6）；

7）若 $B=\varnothing$ 时，则算法结束。

重复上述算法 N 次，则模拟得到 n 只股票的 N 个模拟收益率。模拟的收益率可以作为股票联合分布的估计。进而可以根据估计的联合分布构建投资组合选择模型，修正传统的投资组合选择模型。

4 数值验证

本文选取 2007 年至 2015 年上海证券交易所的 18 只股票的周度数据用于模型验证，这 18 只股票的分别是：600000（浦发银行）、600010（包钢股份）、600015（华夏银行）、600016（民生银行）、600028（中国石化）、600030（中信证券）、600036（招商银行）、600050（中国联通）、600104（上汽集团）、600111（包钢稀土）、600518（康美药业）、600519（贵州茅台）、600585（海螺水泥）、600637（东方明珠）、600795（国电电力）、600837（海通证券）、600887（伊利股份）、600893（中航动力），这 18 只股票均是上证 50 的成分股。

4.1 股票边缘分布估计

本文主要采用周度数据验证模型的有效性，由于需要估计下周的股票收益分布（图 1），因此，选取的因素主要是期初因子。本文采用的因子为：上一周的上证指数收益率、上一周的该股票 K 线上影线、上一周该股票 K 线下影线、上一周该股票市净率、上一周该股票收益，并将上述 5 个因子用 X_1，X_2，X_3，X_4，X_5 表示。本文采用滚动预测模型，每次用该股票前面 250 周的数据作为训练样本，建立影响因子与股票收益的分位数回归模型，然后根据回归参数和本周因子大小，预测下周的股票收益分布。

本文以估计浦发银行 2011 年 12 月 5 号至 12 月 9 号这周的股票收益分布为例，则用该周前 250 周的数据作为训练样本，建立分位数回归模型，估计参数为表 1 中所示。

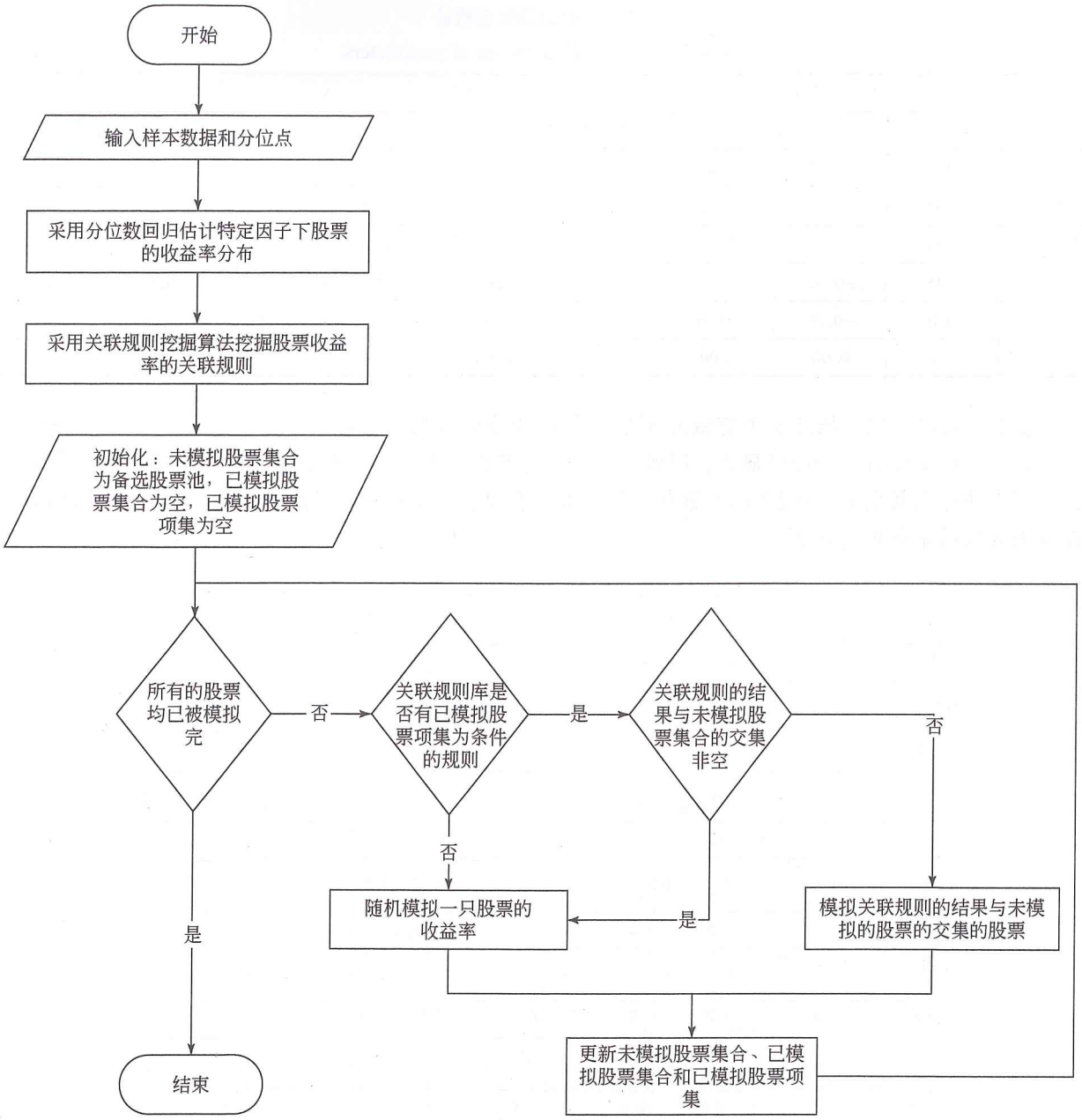

图 1 股票收益联合分布估计流程图

Fig. 1 The flow chart of joint distribution estimation of stock returns

表 1 浦发银行的分位数回归参数

Table 1 The quantile regression parameters of Pudong Development Bank

分位点	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
截距项	-0.01	-0.01	-0.01	0.00	0.00	0.00	0.01	0.02	0.04
X_1	0.10	-0.01	0.08	0.03	-0.02	0.02	-0.07	-0.18	-0.07
X_2	-0.11	-0.05	-0.06	-0.04	0.06	0.01	0.05	0.10	0.14
X_3	0.01	0.18	0.19	0.22	0.24	0.35	0.40	0.37	0.48
X_4	-0.01	-0.01	-0.01	-0.01	0.00	0.00	0.00	0.00	0.00
X_5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

采用置信区间法对参数进行显著性检验，即估计 95% 可能性的置信区间，若 0 不在该区间内，则该变量是显著的，否则，则不是显著的，估计的置信区间为表 2 中所示。

表 2　参数的显著性检验

Table 2　The significance test of parameters

项目	0.20 分位点			0.50 分位点			0.80 分位点		
	估计值	置信区间		估计值	置信区间		估计值	置信区间	
截距项	-0.01	-0.02	0.00	0.00	-0.01	0.01	0.02	0.00	0.04
X_1	-0.01	-0.18	0.19	-0.02	-0.23	0.30	-0.18	-0.39	0.20
X_2	-0.05	-0.15	0.04	0.06	-0.22	0.22	0.10	-0.13	0.27
X_3	0.18	-0.02	0.33	0.24	0.03	0.53	0.37	0.15	0.71
X_4	-0.01	-0.01	-0.01	0.00	-0.01	0.00	0.00	0.00	0.01
X_5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

从表 2 中可以看到，虽然 5 个变量并不是在所有的分位点都是显著的，但是有的变量在某分位点显著，而其他变量又在另一分位点显著，因此，为了估计参数的方便，将 5 个变量均纳入模型中。同样的方法，也可以估计出其余 17 只股票的参数估计值，并进行显著性检验。估计出参数后，则可以得到 18 只股票在该周的收益率分布（图 2）。

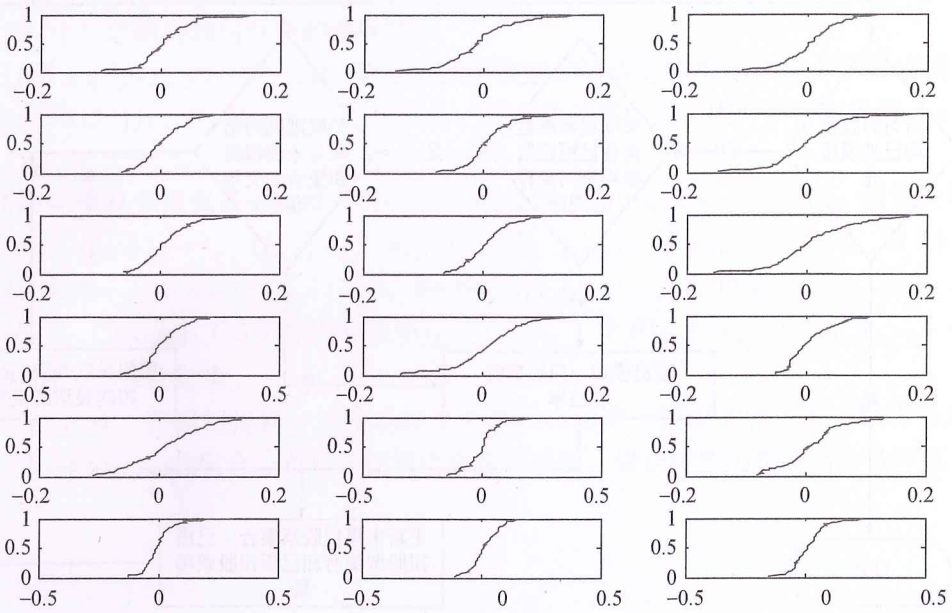

图 2　18 只股票的周度概率分布图

Fig. 2　The weekly probability distribution of 18 stocks

估计出股票在该周的收益率分布后，则可以计算出各股票在该周的收益率均值、方差、分位数、峰度、偏度等统计量。

4.2　股票相关性刻画

在挖掘股票关联规则前，首先对股票进行编号，这里为了简单起见，将股票收益划分成两个区间：涨（收益大于等于 0）或者跌（收益小于 0），并假设 1，2，…，18 分别表示这 18 只股票上涨，19，20，…，36 表示这 18 只股票下跌，并假设最小支持度为 0.1，最小置信度为 0.85，则得到的关联规则数量如表 3。由于关联规则较多，表 4 中仅列出 4 项集的部分关联规则以及相应的支持度和置信度。

表 3　股票关联规则数量

Table 3　The number of stock association rules

类别	2 项集	3 项集	4 项集	5 项集	6 项集	7 项集	8 项集	9 项集
关联规则	2	155	2466	12281	19438	11388	2654	187

表 4 股票关联规则

Table 4 The stock association rules

条件 1	条件 2	条件 3	结果	置信度	支持度
2	3	4	1	0.90	0.19
1	2	3	4	0.94	0.19
2	3	5	1	0.90	0.17
2	3	6	1	0.87	0.16
2	3	7	1	0.92	0.18
1	2	3	7	0.89	0.18
2	3	8	1	0.89	0.16
2	3	10	1	0.86	0.14
2	3	12	1	0.88	0.14
2	3	13	1	0.87	0.17

从第 1 条规则中可知，若第 2 只股票涨、第 3 只股票涨、第 4 只股票涨，则第 1 只股票涨的概率为 90%，该条规则描述了 1，2，3，4 只股票之间的关系。因此，整个关联规则库描述了这 18 只股票之间的关联关系。

4.3 股票联合分布估计

本节依旧以 2011 年 12 月 5 号至 12 月 9 号这周为例，采用分位数回归方法估计出股票的边缘分布，基于关联规则刻画股票的相关性，然后采用蒙特卡洛模拟方法估计出该周股票收益的联合分布，可以得到股票收益率的描述性统计结果（表 5）。

表 5 模拟收益率的描述性统计量

Table 5 The descriptive statistics of simulated rate of return

项目	股票代码					
	600000	600010	600015	600016	600028	600030
平均值	0.007	0.013	0.019	0.006	0.011	0.019
0.25 分位点	-0.021	-0.020	0.001	-0.019	-0.011	-0.012
中位数	0.002	0.011	0.017	0.003	0.007	0.020
0.75 分位点	0.030	0.050	0.042	0.029	0.032	0.051
标准差	0.044	0.059	0.045	0.035	0.036	0.057
项目	股票代码					
	600036	600050	600104	600111	600518	600519
平均值	0.012	0.016	0.034	0.034	0.023	0.008
0.25 分位点	-0.017	-0.005	0.001	-0.027	-0.002	-0.020
中位数	0.002	0.014	0.028	0.032	0.027	0.001
0.75 分位点	0.032	0.037	0.066	0.087	0.055	0.032
标准差	0.042	0.040	0.070	0.084	0.056	0.041
项目	股票代码					
	600585	600637	600795	600837	600887	600893
平均值	0.027	0.008	0.015	0.018	0.033	0.017
0.25 分位点	-0.001	-0.039	-0.012	-0.020	0.001	-0.034
中位数	0.028	0.009	0.017	0.010	0.027	0.014
0.75 分位点	0.070	0.032	0.037	0.043	0.070	0.052
标准差	0.061	0.090	0.052	0.066	0.059	0.072

4.4 基于估计的联合分布的投资组合选择模型

4.4.1 模型表现

本节将根据估计的联合分布修正传统的均值方差模型，并与传统的投资组合选择模型的结果进行对比。股票收益联合分布估计采用前面 250 周的数据估计得到，传统的均值方差模型也采用前面 250 周的数据构建。考虑到中国股票市场的实际情况，这里假设不允许卖空，股票按照收盘价进行买卖，交易成本为千分之一。均值方差的参数 $\sigma=0.03$。

从图 3 中可以看到，基于估计的联合分布修正的均值方差模型显著优于传统的投资组合选择模型。尤其是在 2015 年股灾时，其收益显著跑赢了传统的均值方差模型。

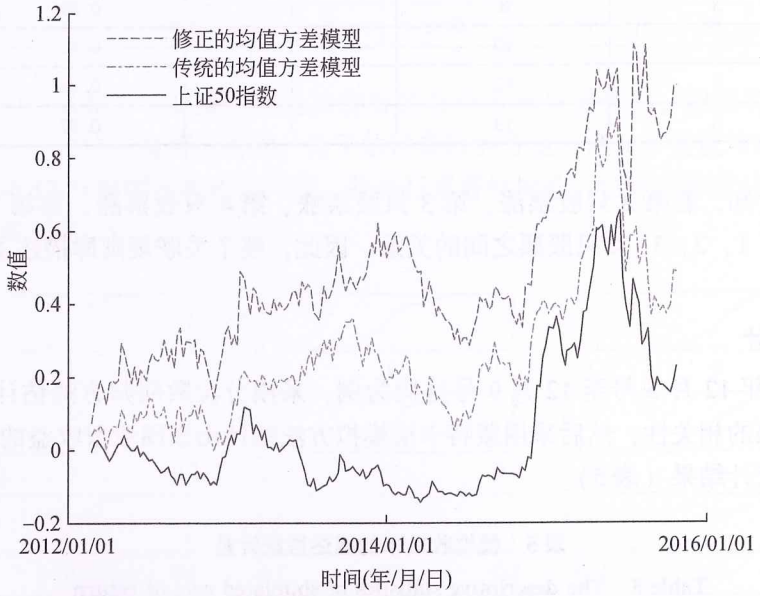

图 3　投资组合模型累积收益对比图

Fig. 3　Comparison of cumulative returns in portfolio selection models

4.4.2 参数的敏感性分析

选择 $\sigma=0.0300$，0.0400，0.0500，0.0600，对比修正的均值方差模型和传统的均值方差模型。其结果如表 6。

表 6　修正均值方差模型和传统均值方差模型对比

Table 6　Comparison between modified mean variance model and traditional mean variance model

项目	σ			
	0.0300	0.0400	0.0500	0.0600
修正均值方差——年化收益	0.2456	0.2734	0.2451	0.2475
传统均值方差——年化收益	0.0837	0.1330	0.1345	0.0991
修正均值方差——夏普比率	0.7125	0.6293	0.5249	0.5056
传统均值方差——夏普比率	0.3337	0.4027	0.3281	0.2095
修正均值方差——最大回撤	0.1497	0.2070	0.2527	0.2790
传统均值方差——最大回撤	0.2546	0.2852	0.3255	0.3866

通过表 6 可以看到，基于估计的联合分布修正的均值方差模型优于传统的均值方差模型，其年化收益高于传统的均值方差模型，夏普比率高于传统的均值方差模型，最大回撤低于传统的均值方差模型。在参数为 0.0300 时，修正的均值方差模型相比传统的均值方差模型，其年化收益提高了 193.43%，夏普比率提高了 113.52%，最大回撤下降了 41.20%。

同时可以发现，随着参数的变大，最大回撤逐渐变大，投资组合的风险逐渐变大，但无论是修正的均

值方差模型还是传统的均值方差模型，其年化收益均是先升后降，这是因为当参数提高到一定程度时，投资组合的权重集中在少数股票上，投资组合的风险分散能力有所下降，因此，其年化收益不增反降。对于夏普比率这个角度而言，修正的均值方差模型呈现出逐渐下降的趋势，但对于传统的均值方差模型，则呈现出先升后降的趋势。

5 结论与展望

本文首先采用分位数回归方法估计特定因子下的股票收益的边缘分布，采用关联规则挖掘算法挖掘股票之间的相关性，然后采用蒙特卡罗模拟方法估计出股票收益的联合分布，进一步根据估计的联合分布构建了修正的投资组合选择模型。相对于传统的投资组合选择模型，一方面考虑了特定因子对股票收益分布的影响，同时没有假设股票收益分布的具体分布形式，另一方面根据关联规则挖掘算法挖掘股票的相关性，能够得到股票的非线性关系。

同时，本文结合实际的股票交易数据验证了修正的投资组合选择模型的有效性。年化收益、夏普比率和最大回撤等指标表明修正的投资组合选择模型显著优于传统的投资组合选择模型。通过实证研究发现，对于风险偏好型投资者，其投资权重主要集中在少部分股票，投资组合的风险分散能力有所下降，因此，未来可以通过寻找影响股票收益的有效因子来提高风险偏好型投资者的风险分散能力。

参考文献：

[1] Markowitz H. Portfolio selection [J]. Journal of Finance, 1952, 7 (1): 77-91.

[2] Rockafellar T, Uryasev S. Optimization of conditional value at risk [J]. Journal of Risk, 2000, 2: 21-41.

[3] Norkin V I, Boyko S V. Safety-first portfolio selection [J]. Cybernetics and Systems Analysis, 2012, 48 (2): 180-191.

[4] Bulmus T, Özekici S. Portfolio selection with hyperexponential utility functions [J]. OR spectrum, 2014, 36 (1): 73-93.

[5] Czichowsky C. Time-consistent mean-variance portfolio selection in discrete and continuous time [J]. Finance and Stochastics, 2013, 17 (2): 227-271.

[6] 陈志英. 状态变化和学习行为下的最优资产组合选择 [J]. 管理科学, 2013, 26 (2): 81-89.
Chen Z. Optimal portfolio choice under regime-switching and learning behaviors [J]. Journal of ManagementScience, 2013, 26 (2): 81-89. (in Chinese)

[7] 姚海祥, 姜灵敏, 马庆华. 不允许买空时的均值-下方风险投资组合选择——基于非参数估计方法 [J]. 数理统计与管理, 2015, (6): 1077-1086.
Yao H X, Jiang L M, Ma Q H. Mean-Downside risk protfolio selection without short selling: based on nonparametric estimation methodology [J]. Journal of Applied Statistics and Management, 2015, (6): 1077-1086. (in Chinese)

[8] 柏林, 赵大萍, 房勇, 等. 基于投资者观点的多阶段投资组合选择模型 [J]. 系统工程理论与实践, 2017, 37 (8): 2024-2032.
BoL, Zhao D P, Fang Y, et al. Multi-period portfolio selection model based on investor's views [J]. Systems Engineering-Theory&Practice, 2017, 37 (8): 2024-2032. (in Chinese)

[9] Sharpe W F. Capital asset prices: a theory of market equilibrium under conditions of risk [J]. Journal of Finance, 1964, 19 (3):425-442.

[10] Fama E F, French K R. Common risk factors in the returns on stocks and bonds [J]. Journal of Financial Economies, 1993, 33 (1): 3-56.

[11] Fama E F, French K R. Dissecting anomalies with a five-factor model [J]. Review of Financial Studies, 2016, 29 (1): 69-103.

[12] Daniel K, Moskowitz T J. Momentum crashes [J]. Journal of Financial Economics, 2016, 122 (2): 221-247.

[13] 田利辉, 王冠英. 我国股票定价五因素模型：交易量如何影响股票收益率? [J]. 南开经济究, 2014, (2): 54-76.
Tian L H, Wang G Y. Asset pricing model of the chinese stock market: how trading volumes influence the returns? [J]. Nankai Economic Studies , 2014, (2): 54-76. (in Chinese)

[14] 王宜峰, 王燕鸣, 吴国兵. 公司投资对股票收益的影响研究 [J]. 管理评论, 2015, 27 (1): 103-113.
Wang Y F, Wang Y M, Wu G B. Empirical study on the impact of company investment on stock returns [J]. Management Review, 2015, 27 (1): 103-113. (in Chinese)

[15] Koenker R, Bassett G. Regression quantiles [J]. Econometrica, 1978, 46 (1): 33-50.
[16] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large data bases [C]. Proceedings of the ACM SIG -MOD Conference on Management of data, 1993, 22 (2): 207-216.
[17] Chernozhukov V, Fernández- Val I, Kowalski A. Quantile regression with censoring and endogeneity [J]. Journal of Econometrics, 2015, 186 (1): 201-221.
[18] Tsai I. The relationship between stock price index and exchange rate in Asian markets: a quantile regression approach [J]. Journal of International Financial Markets Institutions & Money, 2012, 22 (3): 609-621.

Portfolio Selection Model Based on Quantile Regression and Association Rule Mining

Fang Yong[1], *Huang Liang*[2]

1. Academy of mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China;
2. School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, China

Abstract: The traditional portfolio selection models are based on historical data and assume that the future distribution of stock returns is the same as the past. This paper uses the quantile regression method to estimate the marginal distribution of the stock returns under specific factors, and uses the association rule mining algorithm to find the relationship among the stocks, and then uses Monte Carlo simulation method to estimate the joint distribution of the stock returns. Furthermore, we propose a portfolio selection model based on the estimated joint distribution of stock returns to revise the traditional portfolio selection model. A numerical example is given to illustrate the behavior of the proposed model through the real data that is the 18 constituent stocks of SSE 50. It can be concluded that the revised portfolio selection model is better than tradition portfolio selection model.

Key words: Portfolio Selection Model; Quantile Regression; Association Rule Mining