**Summary**

**Keywords:**

# Introduction

## Motivation

21 世纪以来，中国金融市场蓬勃发展。越来越多的普通人成为了初级投资者，进入金融市场寻求机会；同时金融机构也在不断地寻找有效的分析方法和工具，优化在可控投资风险范围内的投资回报，其中对价格序列的预测是一个研究重点。

此外，每日股指价格时间序列包含各种有效信息（开盘价格和收盘价、最高价格和最低价格等），它具有很强的时间间隔特征。而重估的时间序列只反映了每日股指价格的平均值，并可能遭受波动性信息的损失（Sun et al.，2018a）。因此，值得对高频区间股指价格预测进行进一步研究。

## Related Works

### 高频股指价格预测

根据现有文献，高频股指价格预测的研究主要可分为两类，关于高频股票指数价格预测的研究可分为两大类：统计模型和人工智能模型。

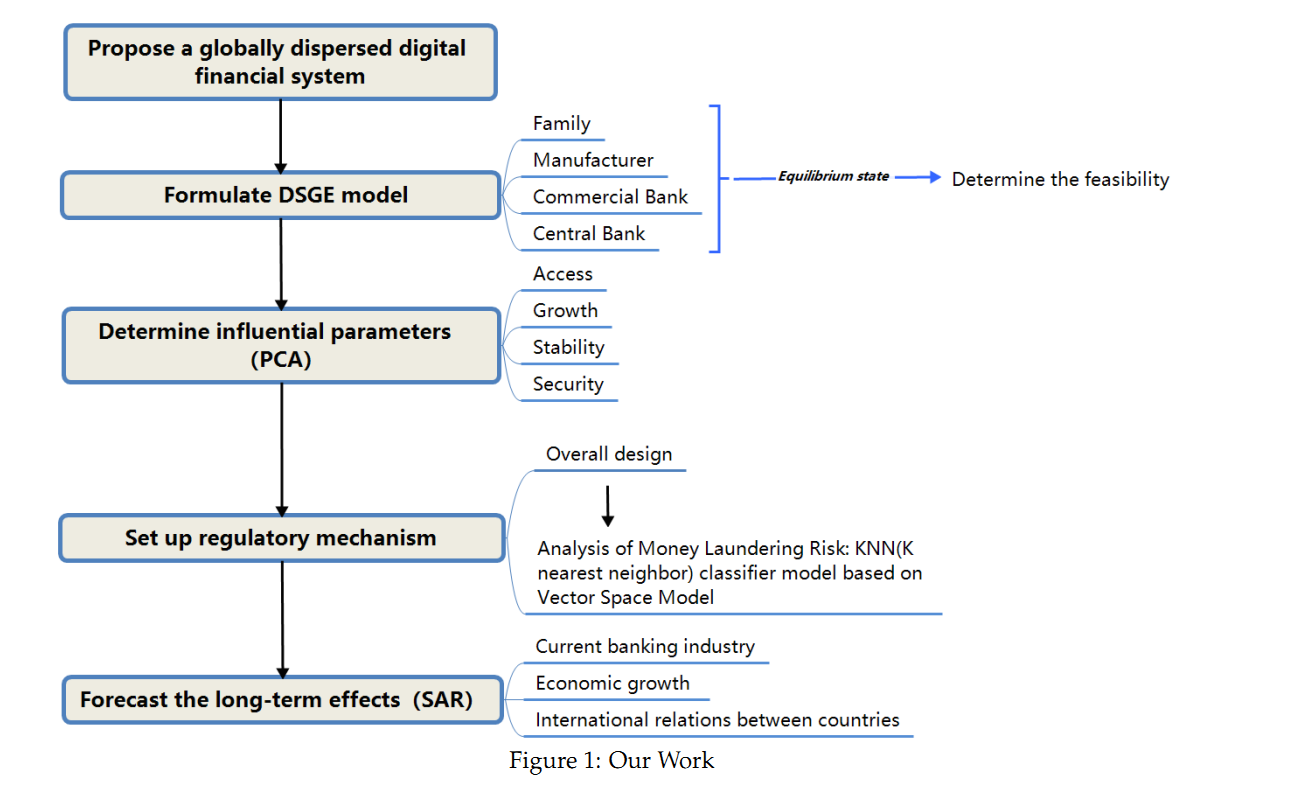
统计模型包括指数平滑、自回归积分移动平均（ARIMA）和广义自回归条件异方差（GARCH）波动率等技术。这些模型旨在捕捉时间序列中的线性模式。在预测高频金融时间序列时，统计模型有一些局限性。其一是它们旨在捕获数据中的线性关系和模式，而这些关系和模式可能并不总是存在于高频金融时间序列中。 其二，统计模型可能无法处理高频数据的复杂性和非平稳性。例如， 李等人。 (2020) 发现传统的自回归综合移动平均 (ARIMA) 模型对高频金融时间序列的预测存在较大偏差。

为了模拟高频金融时间序列中的非线性模式，许多研究转向了**人工智能方法**。 **人工智能模型**则包括人工神经网络、k-最近邻、逻辑回归、朴素贝叶斯、随机森林分类器、支持向量机和极端梯度提升分类器等技术。例如，[[1]](#endnote-1)**Akyildirim** 等人。 (2022) 比较了各种高级预测技术的性能，如人工神经网络、k 最近邻、逻辑回归、朴素贝叶斯、随机森林分类器、支持向量机和极端梯度提升分类器，以根据过去的价格预测股价走势 使用在伊斯坦布尔证券交易所交易的 27 只蓝筹股的高频数据。 他们的发现表明，在选定的方法中，随机森林和支持向量机能够以令人满意的水平捕捉未来价格方向和百分比变化。 Lanbouri 和 Achchab (2020) 提出了一个基于技术指标和长短期记忆 (LSTM) 的模型，使用标准普尔 500 指数盘中交易的高频数据提前一分钟、五分钟和十分钟预测股票价格。另一项研究 使用支持向量机、人工神经网络和隐马尔可夫模型等机器学习技术来预测高频数据中的股价趋势。使用神经网络等人工智能方法进行高频金融时间序列预测的优势之一是它们能够捕捉数据中的非线性关系和模式。 它们还可以处理高频数据的复杂性和非平稳性，并且可以在大量数据上进行训练以提高其准确性。 张等 (2023) 提出了处理原始数据的新建模策略，以提高使用道琼斯 30 成份股高频数据预测中价股的机器学习算法的性能。

上述研究表明，目前高频点值股票指数价格预测的研究主要集中在开发和改进准确预测股票价格的方法。然而，使用点值数据进行高频金融时间序列预测的一个局限性是它可能无法捕捉高频数据的全部复杂性和非平稳性。因此，开发更准确的区间股票指数价格预测模型是很自然的。  
**Table 1. Literature review of 高频股指价格预测**

### Interval-valued time series forecasting

## Our Work



# Notations

The key mathematical notations used in this paper are listed in Table 1.

Table 1: Notations used in this paper

|  |  |  |
| --- | --- | --- |
| **Symbol** | **Description** | **Unit** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

本部分对应国赛论文中的符号（英文翻译为notation或者symbol）说明部分。

# 典型相关分析：构建17个目标之间的相互联系网络

## 问题分析

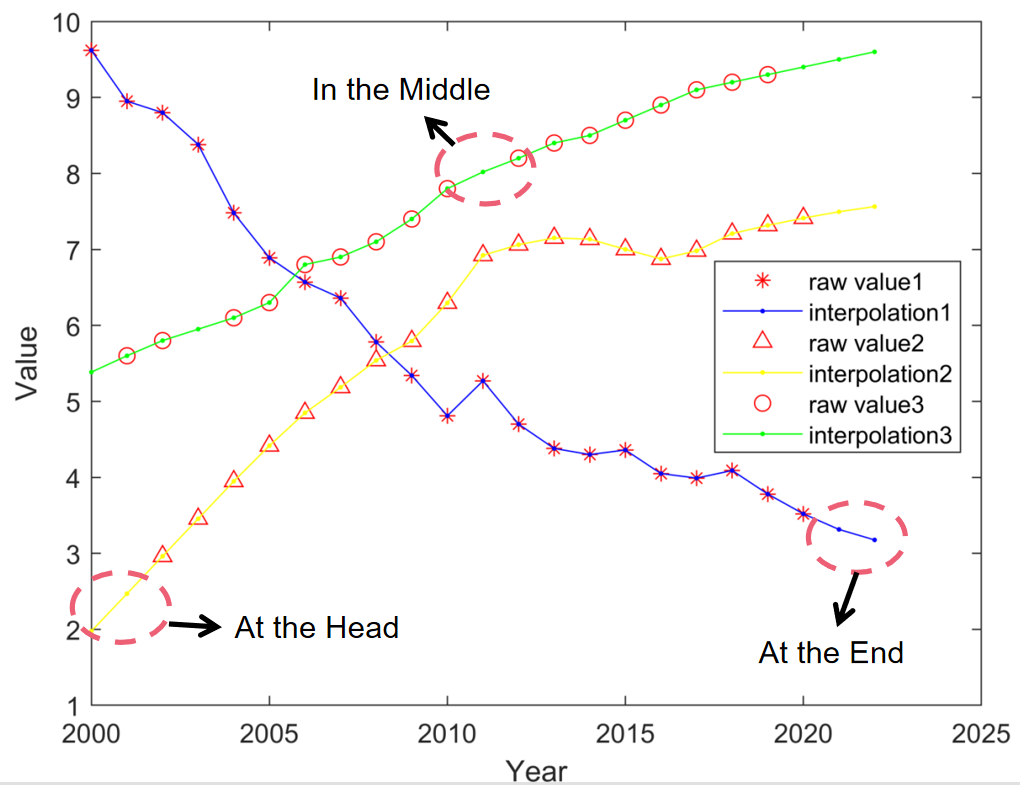
问题一要求构建涵盖17个目标之间相互关系的网络。（蓝色为引入参考文献）斯德哥尔摩的第二页文献提到，对可持续发展目标的基准测试只有在地区的背景下才有意义。经合组织国家、南亚和撒哈拉以南非洲等，它们都有不同的区域背景，制定普遍的全球目标或基准是不合适的。因此，本文分别针对发达国家、中高收入发展国家、低收入发展国家和全球分别构建网络。

每一个网络要求分析17个目标中每两个目标之间的联系（包括强弱以及正负）。每个目标下有多个子指标，若采用简单相关分析的方法，只是孤立的考虑了单个X与单个Y的关系，而没有考虑X、Y变量组内部各变量间的相关。因此，考虑采用研究两组变量之间相关关系的多元统计方法——典型相关分析，识别每两个目标两组变量之间的关系，考虑两组变量的线性组合，并研究它们的典型相关系数p.

每一个国家对应的每一张网络图需要做C 217次典型相关分析，下文以基于澳大利亚的目标4（。。。）和目标11（。。。）的相关关系为例进行详细说明。

## **数据预处理**

经过对数据的查找，我们发现部分年份数据缺失，不足以支撑分析的进行。在此我们通过分段三次埃尔米特插值法对这种缺失数据进行处理。“模拟产生”一些新的但又比较靠谱的值来满足需求，这就是插值的作用。然而，直接使用埃尔米特插值得到的多项式次数较高，也存在**龙格现象**，所以在实际应用中，往往使用**分段三次埃尔米特插值**多项式。  
  
原理：插值多项式要求在插值节点上函数值相等，有的实际问题还要求在节点上的导数值相等，甚至**高阶导数值**也相等，满足这种要求的插值多项式成为埃尔米特插值多项式。

通过Matlab计算，我们得到结果如图所示。value1是德国的sdg4\_tertiary(Tertiary educational attainment(% of population aged 25 to 34)) ； value2是中国的sdg13\_co2gcp(CO₂ emissions from fossil fuel combustion and cement productiont(CO2/capita)) ； value3是Papua New Guinea 的sdg7\_cleanfuel(Population with access to clean fuels and technology for cooking (%))。结果如下图所示：  
 

## 模型建立：（模型流程图）

步骤一：建立原始矩阵；  
根据上述4.2中的数据，考虑到子目标的代表性和信息的完整性，我们选取目标4中的3个子目标为向量X=（X1,X2,X3）’（具体选取哪四个见附录）,选取目标11中的3个子目标为向量Y=（Y1,Y2,Y3）’（具体选哪两个指标见附录）, Z为3+2总体的23次中心化观测数据阵：

|  |  |
| --- | --- |
|  | () |

步骤二：对原始数据进行标准化变换并进行相关系数矩阵；  
我们首先对2个目标的子目标的数据进行标准化处理，然后计算两样本间的相关系数矩阵R，并将R分为

|  |  |
| --- | --- |
|  | () |

其中，R11、R22分别为目标4和目标11内的相关系数阵，R12和R21为目标4和目标11间的相关系数矩阵。

步骤三：求典型相关系数及典型变量；  
首先求的特征根，特征向量；的特征根，特征向量，则有

|  |  |
| --- | --- |
|  | () |

则随机变量*X* （目标4）和随机变量*Y* （目标11）的典型相关系数为 λ ，典

型变量为

|  |  |
| --- | --- |
|  | () |

步骤四：检验各典型相关系数的显著性。

对典型相关系数进行显著性检验。在作两组变量X和Y的典型相关分析之前，首先应检验两组变量是否相关；如果不相关，即，则讨论的两组变量的典型相关就毫无意义。

## 模型求解与可视化

附件中包含2000年至2022年的3个目标四和3个目标十一的子目标共23年6个指标的原始数据。如下表所示。

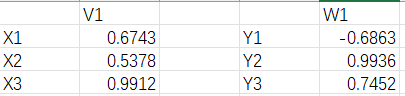
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 表1:选取的6个子目标 | | | | | |
| X1 | X2 | X3 | X4 | Y1 | Y2 |
| ne ona t | u5 mort | tb | hiv | intuse | a rtides |

利用SAS软件对目标4和目标11两个变量进行了典型相关分析，结果如下：

第一对和第二对变量的典型相关系数分别是**0.787644 ,0.398143和0.134586**。第一对的典型相关系数大于0.5，由此可见这对典型变量的解释能力比较强。但要确定典型变量相关性的显著程度，需要进行典型相关系数的显著性检验。通过SAS检验结果显示，P值分别是0.0005,0.2718和0.5646，在0.05的显著性水平下，第一对典型变量之间相关关系显著。

标准化后的典型变量的系数建立典型相关模型以及典型变量与变量的相关关系如下：

|  |  |
| --- | --- |
|  | () |

(还没来得及画表嘞)

典型变量每个分量面前的系数代表着重要程度，从(5)式可以看出，对于目标四而言，最重要的影响因素为子目标三，且X3与V1的相关系数明显大于X1、X2与V1的相关系数，可见V1主要代表了受高等教育程度；从(5)式可以看出，对于目标十一而言，最重要的影响因素为子目标二，且Y2与W1的相关系数明显大于Y1、Y3与V1的相关系数的绝对值，可见W1主要代表了(Access to improved water source, piped)。该对典型变量解释配对变量组方差的累计比例分别为70.25%和87.31%。我们采用了同样的方法计算出9个代表国家17个指标两两间的典型相关系数，如果典型相关系数不止一对通过了显著性检验，我们则取典型相关系数最大的那对典型相关变量。

图：变量之间的对应关系  
  
依照同样的步骤，我们基于国家或地区发展不同的现实分别选取了3个发达国家（A组：美国、德国、澳大利亚）、3个中高收入发展中国家（B组：中国、马来西亚、巴西）和3个低收入发展中国家（C组：Mozambique、Papua New Guinea、Sierra Leone）的数据，分别求解每一个国家的17个指标中每两个指标之间的典型相关系数，总共得到了九个17×17的邻接矩阵。  
  
然后，我们分别对每组数据进行算数平均，得到三个邻接矩阵，分别是：  
。受篇幅限制，具体数据在附录中展示。

发达国家，中高收入发展中国家和低收入发展中国家的17个可持续发展目标之间的相互联系网络如下图所示：

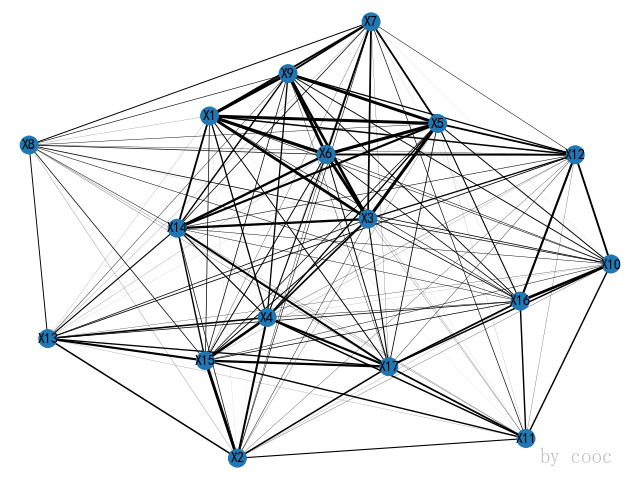
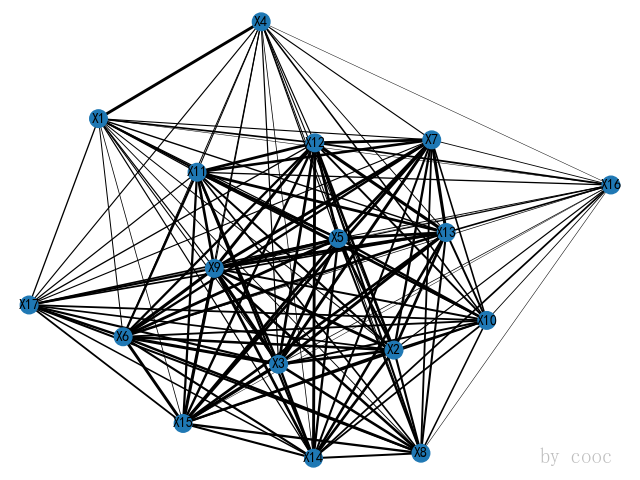


图1 发达国家目标网络图 图2 中高收入发展中国家目标网络图

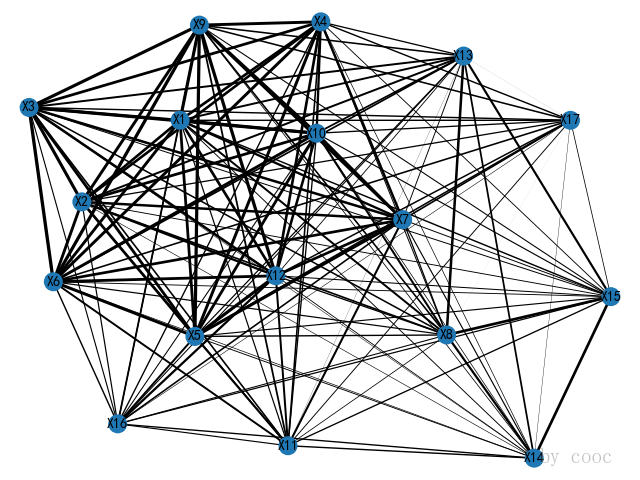
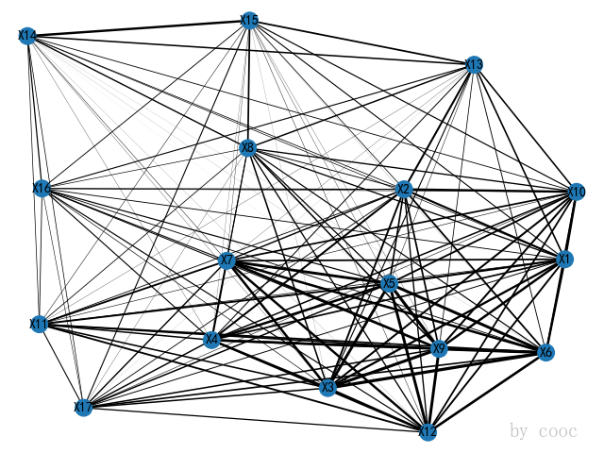
 

图3 低收入发展中国家目标网络图 图4 全球可持续发展目标网络图

这里后面最好换一种形式展示，图中共有17个节点分别对应17个可持续发展目标，节点之间连线代表两个节点相关。其中，连线的粗细程度对应两个节点联系的强弱程度。显然，基于不同国家和地区的发展状况和目标执行情况不同，网络图不同节点联系的强弱程度也不同。另外，三组的各个节点的优先级也不同，这一点在下面将详细阐述。

|  |  |
| --- | --- |
| 这里插入公式 | () |

|  |  |
| --- | --- |
| 这里插入公式 | () |

|  |  |
| --- | --- |
| 这里插入公式 | () |

# 评价不同地区的目标优先级和预测未来

## 问题分析

优先级是一个定性的概念，17个目标各自的优先级需要通过指标来量化。本模型首先计算网络中每一个节点的特征向量中心性，再根据联合国和世界银行构件的全球可持续发展目标数据库数据统计每个目标的完成度。之后，基于不同国家和地区发展情况不同，凭借经验与常识设定调整参数，其目的是突出对上述提到的三类国家各自的最紧迫的制约因素，以体现三类国家对不同目标完成的迫切程度。即不同发展情况的国家和地区应当优先考虑那些通过重新部署现有能力可以提供更大的成功空间的目标（高中心度）和完成度差、迫切度高的目标。这三个指标以相同权重计入优先度得分，最终17个目标的优先度可按最终得分排序（做张图）。

## 模型建立

# The name of model 3

和上一个部分类似的写法。

注意：大多数美赛优秀论文都是对每个问题或者每个模型作为一个一级标题，就像我们上面的这种布局；也有一部分论文建立一个大的一级标题，取名为“Models and Solutions”，然后在这个大的标题下设计每个问题或者模型对应的二级标题，这一种排版布局在国赛中用的更多。

# Sensitivity Analysis

在国赛论文写作视频中（<https://www.bilibili.com/video/BV1Na411w7c2/>），我介绍过怎么写模型的分析和检验这个部分：

模型的分析 ：在建模比赛中模型分析主要有两种，一个是灵敏度(性)分析，另一个是误差分析。灵敏度分析是研究与分析一个系统（或模型）的状态或输出变化对系统参数或周围条件变化的敏感程度的方法。其通用的步骤是：控制其他参数不变的情况下，改变模型中某个重要参数的值，然后观察模型的结果的变化情况。误差分析是指分析模型中的误差来源，或者估算模型中存在的误差，一般用于预测问题或者数值计算类问题。

模型的检验：模型检验可以分为两种，一种是使用模型之前应该进行的检验，例如层次分析法中一致性检验，灰色预测中的准指数规律的检验，这部分内容应该放在模型的建立部分；另一种是使用了模型后对模型的结果进行检验，数模中最常见的是稳定性检验，实际上这里的稳定性检验和前面的灵敏度分析非常类似，等会大家看到例子就明白了。

在美赛的写作中，写的最多的就是灵敏度分析（Sensitivity Analysis），因此这里我们的标题就直接取得是灵敏度分析；如果你既要写灵敏度分析，又要写误差分析（Error Analysis），那么你可以把标题改成： Sensitivity Analysis and Error Analysis

# Model Evaluation and Further Discussion

注：本部分的标题需要根据你的内容进行调整，例如：如果你没有写进一步讨论的话，就直接把标题写成模型的评价。（优缺点一定要写）

## Strengths

这里写论文或者模型的优点

## Weaknesses

这里写缺点：缺点写的个数一般要比优点少

## Further Discussion

进行进一步的讨论，这里可以写模型的改进和拓展：

模型的改进：主要是针对模型中缺点有哪些可以改进的地方；

模型的拓展：将原题的要求进行扩展，进一步讨论模型的实用性和可行性。

# Conclusion

结论部分，这个部分在国赛论文很少见到，但在美赛中出现的频率很高。

这个部分可以是论文中心思想的重申、研究结果或主要观点的归纳，也可以是某些启示性的解释或考虑。

有些论文把“Model Evaluation and Further Discussion”的内容放到了结论部分，这也是可以的，大家可以灵活调整。

# References

参考文献：所有引用他人或公开资料(包括网上资料)的成果必须按照科技论文的规范列出参考文献，并在正文引用处予以标注。

一般新起一页列出参考文献，如果上一个部分的下面有很多空白，那么就不用新起一页了。

美赛中不要出现中文，如果引用中文文献请翻译过来。

# Appendices

|  |
| --- |
| Appendix 1 |
| Introduce: 这里放上附录1的介绍 |
|  |

|  |
| --- |
| Appendix 2 |
| Introduce: 这里放上附录2的介绍 |
|  |

本部分是附录部分，美赛对于附录不是特别看重，今年还限制了论文的页数（从第二页开始编号，不能超过25页）。

一般新起一页列出附录。

在不超过页数限制的条件下，附录中可以包括下面内容：

* 你们写的代码；
* 某一问题的详细证明或求解过程；
* 自己在网上找到的数据；
* 比较大的流程图；
* 较繁杂的图表或计算结果。

1. <https://link.springer.com/article/10.1007/s10479-021-04464-8> Forecasting high-frequency stock returns: a comparison of alternative methods （2022）

   Erdinc Akyildirim, Aurelio F. Bariviera, Duc Khuong Nguyen & Ahmet Sensoy [↑](#endnote-ref-1)