# 红墙泰和因子方向实习生笔试题目

**本次笔试题目共包含2道开放性题目，请您于收到题目后的24h内完成作答并提交，本笔试题目不存在标准答案，请诚信应试，严禁使用AI进行作答。**

一、目前量化投资市场现有因子库存量庞大，很多因子已经被发现，如何能够在现有因子库的基础上发现新的因子？请提供你的解决思路，包括但不限于（1、是否有某个方向研究较少有挖掘潜力，请详细说明这个领域的因子设计思路 2、现有一批新挖掘因子与因子库相关性较高，能否设计一种方法对其中的增量信息进行提取，降低相关性 3、如何对最原始数据进行处理得到一系列新的因子素材数据，在这批素材数据上挖掘因子能够产生较好的表现）

1. **提供1个较少有挖掘潜力的方向：**

宏观经济领域

像财务指标、量价因子等大多数集中在企业或更微观的层面，但宏观经济变量因子往往较少被深入探讨。尤其是近两年国内贾盾、胡悦等学者在国内外顶级期刊发表有关宏观经济公告前的溢价效应。这种溢价效应并非主要有内幕交易引起的，更多公布时间不定形成的不确定性补偿，并且这种补偿随着公布时间的临近会渐渐降低。

因此，可以利用宏观经济数据（如M2货币增长率、GDP增长率、失业率、消费者信心指数等）设计因子。①公告预期溢价因子：这种因子衡量的是公告发布前几天或几周内，相关市场资产的价格变动，反映了市场对公告内容的预期。例如，在GDP或失业率等重要数据发布前，某些股票或资产可能会出现上涨趋势，反映市场对该公告的积极预期；②公告前的不确定性补偿因子：计算宏观数据发布前的隐含波动率（VIX指数），反映市场对公告内容不确定性的反应，并根据公告发布的临近性调整。

**2. 新挖掘因子与现有因子相关性较高，如何设计方法提取增量信息？**

通过主成分分析 PCA或者正交化减少因子之间的相关性，从而提取新的增量信息。PCA可以得到一组无关的主成分，最大程度地保留数据的方差，同时降低原始因子之间的相关性；正交化使因子之间的相关性最小化，从而提取更加独立的因子。

或者使用现有因子的加权平均来构造一个新的复合因子。这种复合因子可以减少因子之间的冗余，同时保留不同因子提供的不同信息。具体地，可以利用岭回归或LASSO回归来选择和加权现有因子。

**3. 如何对最原始数据进行处理得到一系列新的因子素材数据？**

首先，在特征提取阶段。对于部分数值型数据，需要先进行对数计算等；对于类别型数据，常转换成独热编码（也称为one-hot编码，将每个类别转换为一个二进制特征）、标签编码（将每个类别赋予一个整数标签）、目标编码（根据目标变量的平均值对类别进行编码）等；对于文本数据，将文本转换为词频向量、或考虑词频和逆文档频率的词向量、词嵌入（如等。

其次，在特征选择阶段。目标是选择出对预测目标变量有最大贡献的特征，同时也能够降低模型的复杂性。在模型训练过程中进行特征选择，例如L1正则化、决策树等模型都有特征选择的功能。

第三，在特征构造阶段。常见的比如，平方、对数等基本数学变换，均值、标准差等统计量计算，滑动窗口操作，移动平均线、MACD、布林带等技术指标。

第四，特征缩放，特征缩放是将特征的值缩放到同一范围内。使用标准化（将特征缩放到均值为0，标准差为1的范围内）或归一化（将特征缩放到0-1的范围内）。

二、请举例说明你在进行因子研究时的整个思维过程与工作模式（例如，从看到一篇研报，产生想法，设计因子形式，代码落地，数据输出，回测检验，根据反馈调整因子设计），请给出具体的例子，并重点说明该因子的设计思路以及个人如何根据反馈进行调优的过程的细节。

首先，看到有关日内不平衡指标的研报时，思考这个因子的理论基础是什么是否预测收益逻辑合理，有什么不足。比如，日内不平衡指标常用主买与主卖的成交单差异衡量，实际上反应的是一种主买后的上升趋势，与未来收益正相关，但是这势必影响当前的涨跌幅状况，而当前涨跌幅又与未来收益呈现反转效应。另一方面，该指标未考虑量的问题。

其次，在因子构建方面，先以常用的方式构建出来，然后考虑因子构建的频率15min/30min/1d，以及选用全部/非孤立的成交单构建，计算完成之后将因子降频到月度层面，并以市值行业去中心化，最后标准化得到最终的因子。有时高频数据操作需要追求高效，根据运行时间修改成更向量化的操作。

1. 观察因子的月度IC、ICIR值以及多空对冲组合表现。根据因子值选取前10%的股票进行回测，得到累计资产净值以及年均收益、年均夏普比率、年均最大回撤等指标。
2. 修改因子参数。例如，将因子的计算窗口缩短、并且只采用非孤立成交单样本，将得到的因子与涨跌幅因子正交化、以及与市场风格因子正交化，检验对未来收益是否仍然具有预测能力。