# 基于稀疏高维多元统计回归模型的行业轮动策略

## 背景介绍

### ■ 指数共同基金

有效市场假说，认为市场交易者获取公开信息，并对信息处理形成交易观点，实施交易策略，从而促使市场有效性发生变化。它告诉我们，市场是群体智慧的结晶。交易者将他们的信息纳入市场价格，但在这个过程中，最初吸引他们的获利机会很快消失了。在有效市场理论的基础上，主动型投资策略理论在金融学发展中形成了套利定价理论 APT、资本资产定价模型 CAPM、马克维茨最优资产组合等理论基石，衍生了 Fama-French 三因子模型等，它和期权定价模型共同成为金融经济学中量化革命的一部分。

| 假说类型 | 主要观点 | 价格预测 | 推论 |
|---|---|---|---|
| 无效市场 | 市场价格无法反映证券价格信息，包括“成交价、成交量，卖空金额、融资金额”等 | 历史成交信息对未来的预测有效 | 技术分析有效 |
| 弱有效市场 | 市场价格充分反映证券价格信息， | 基本面分析对股票未来的预测有效 | 技术分析失效，基本面分析有效 |
| 半有效市场 | 市场价格只反映所有的公开信息 | 非公开信息决定超额收益 | 只能从隐含市场信息中获得 |
| 强有效市场 | 市场价格反映所有信息，包括内幕消息 | 价格随机游走 | 未来是不确定的，不需要做预测 |

市场越有效率，市场价格的变动越随机。最有效率的、最全面的市场中价格变动是完全随机的，而且不可预知，这不是自然选择的结果，是市场的直接参与者试图从其所掌握的信息中获益的结果。换言之，有效市场假说告诉我们，可以投资被动的、低成本和充分分散风险的共同基金，平衡投资组合的回报和风险，跟随市场而不是通过公开信息挑选股票来打败市场。毫不夸张地说，有效市场假说带来了指数共同基金业务的繁荣。

| 基金名称 | 基金代码 | 基金成立日 | 基金规模 |
|---|---|---|---|
| 汇添富中证主要消费 ETF | 159928.SZ | 2013-08-23 | 15.26 |
| 汇添富中证医药卫生ETF | 159929.SZ | 2013-08-23 | 1.23 |
| 汇添富中证能源 ETF | 159930.SZ | 2013-08-23 | 0.15 |
| 汇添富中证金融地产 ETF | 159931.SZ | 2013-08-23 | 0.24 |
| 广发中证全指信息技术 ETF | 159939.SZ | 2015-01-08 | 3.45 |
| 广发中证全指原材料 ETF | 159944.SZ | 2015-01-08 | 0.18 |

指数 ETF 作为被动型基金类型，有股票型（规模型、行业型、主题型等）、债券型（国

债、城投）、商品型（黄金）等风格。以投资 A 股的股票型 ETF 行业指数为例子，我们选取锁定中证行业指数的 6 个 ETF 基金，下面的表格给出了在 2019 年 4 月，这些行业指数 ETF 基金的重仓股比例：

| 汇添富中证主要消费ETF | | 汇添富中证医药卫生ETF | | 汇添富中证能源ETF | | 汇添富中证金融地产ETF | | 广发中证全指信息技术ETF | |
|---|---|---|---|---|---|---|---|---|---|
| 股票名称 | 净值比 | 股票名称 | 净值比 | 股票名称 | 净值比 | 股票名称 | 净值比 | 股票名称 | 净值比 |
| 贵州茅台 | 15.33 | 恒瑞医药 | 9.57 | 中国石油 | 15.47 | 中国平安 | 15.24 | 海康威视 | 5.04 |
| 伊利股份 | 15.23 | 云南白药 | 4.23 | 中国石化 | 13.75 | 招商银行 | 6.53 | 京东方A | 3.30 |
| 五粮液 | 11.52 | 美年健康 | 3.06 | 中国神华 | 12.70 | 兴业银行 | 4.67 | 东方财富 | 2.32 |
| 洋河股份 | 6.66 | 康美药业 | 3.05 | 陕西煤业 | 10.62 | 交通银行 | 3.99 | 科大讯飞 | 1.91 |
| 海天味业 | 6.50 | 片仔癀 | 2.86 | 广汇能源 | 5.42 | 民生银行 | 3.57 | 立讯精密 | 1.84 |
| 永辉超市 | 3.51 | 爱尔眼科 | 2.74 | 上海石化 | 3.91 | 农业银行 | 3.46 | 三安光电 | 1.47 |
| 泸州老窖 | 3.47 | 长春高新 | 2.64 | 海油工程 | 3.86 | 中信证券 | 3.19 | 大族激光 | 1.38 |
| 双汇发展 | 2.72 | 沃森生物 | 2.59 | 东华能源 | 3.32 | 万科A | 2.91 | 恒生电子 | 1.36 |
| 海大集团 | 2.13 | 复星医药 | 2.59 | 西山煤电 | 3.09 | 浦发银行 | 2.89 | 航天信息 | 1.36 |
| 牧原股份 | 2.10 | 乐普医疗 | 2.44 | 中煤能源 | 3.04 | 工商银行 | 2.86 | 大华股份 | 1.09 |

## ■ 交易策略分类

一般来讲，主动型策略基于“预测收益”，对应着更高频和更大幅度的仓位调整，被动型基于“控制风险”，倾向于长期持有一个较稳定比例的组合。与有效市场理论不同，行为金融学理论则支持技术分析，包括从量价指标中寻找交易机会。我们总结一些交易策略如下：

| 市场有效性 | 主要交易策略类型 | | | 主要模型和方法 | |
|---|---|---|---|---|---|
| | 类型 | 客观方法 | 主观方法 | | |
| 弱/无 | 预测收益 | 动量预测、趋势跟踪和周期预测、高频套利等 | 行为金融学预测，如相对吸引力、主题投资、技术面和交易支撑分析等。 | 基于历史交易信息对未来的预测 | 机器学习和统计建模、时间序列分析 |
| 半有效 | 混合 | 分析师选股，CAPM 组合和多因子模型等。 | 细分行业研究和隐含舆情等、事件驱动、宏观对冲等 | 基于收益分解和驱动因素变化 | 因子收益分解、市场中性策略、期权和波动率对冲等 |
| 市场有效 | 控制风险 | 风险平价策略，指数跟踪策略。 | 指数和行业配置,FoF 和全球配置，债股配置等。 | 波动率控制汇率和流动性风险等 | 组合优化和风险分散模型（VaR）。 |

20 世纪 90 年代开始，美国科技型对冲基金使用机器学习技术来管理资产，取得了长期稳定的收益业绩。基于高频率交易和预测技术构建策略的基金成功，并不完全否定有效市场理论的作用，但至少质疑它的假设条件。美国文艺复兴公司使用的机器学习模型，D. E. SHAW

资本和世坤投资、TWOSIGMA 等对冲基金管理的资产，甚至完全使用数据驱动挖掘出的投资信号完成交易。这些科技基金实际上不再固定于某种特定的收益风格模式，解释某些策略的成功（收益来源）也变得更为困难。

行业 ETF 策略分为主动和被动型两种，传统的策略如“行业周期”、“经济周期”、“行为金融学”角度进行分析，实际上要求资产的做更高比例的动态调仓，我们归为主动型策略。而从“趋势确立”、“风险平衡”角度构建的策略，通常归为被动型策略。前者，经典的是“宏观对冲基金”，后者是“风险平价策略基金”。

■ **机器学习配置**

实际上，基于机器学习技术（算法技术）构建策略的一个优势是，它可以更灵活得介于两种投资风格之间，以数据驱动的方式，更好地平衡收益和风险。另一个角度就是，传统的主动型策略，很难综合考虑较多数量的投资标的，我们很难使用多个相互孤立的宏观行业预测逻辑，来决定行业配置方案。简单的动量或反转模式也难以获得持续、稳定的超额收益；传统的被动型策略，也面临资产数目较多时候的组合风险模型失效和厚尾风险问题。

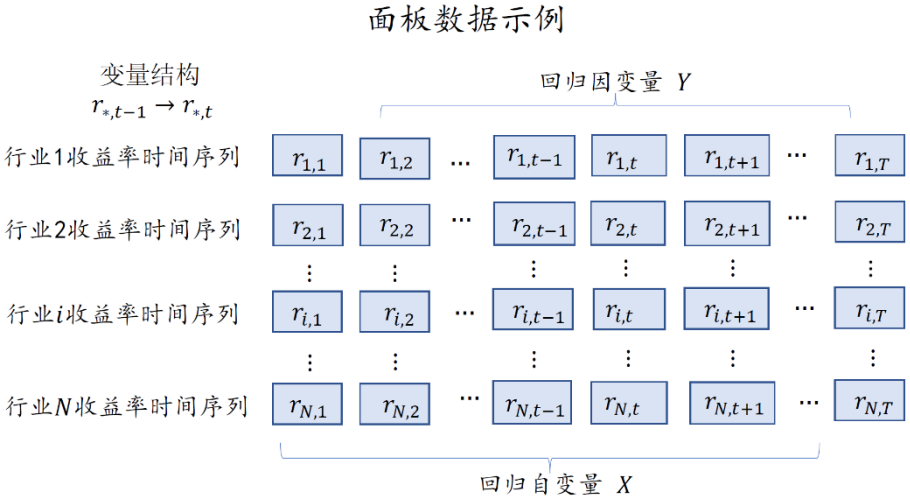
我们给出一种基于高维回归预测模型来挖掘行业轮动规律并指导行业配置的方法，其核心思想是利用所有行业的收益动量因子对每个行业下期收益率做回归，考察行业收益率之间是否存在显著的交叉预测现象，进而构建轮动策略。其优势有以下几点：**交叉预测**：高维模型使多行业间的动量交叉预测成为可能，可以找到行业间的轮动规律；**风险控制**：基于动量因子的收益分解，为配置后的组合配置风险提供了预测基础；**多资产管理**：稀疏学习做动量因子的特征选择，可以管理较多数量的资产，控制交叉板块的连接数量；**可扩展性**：稀疏学习可以限制组别间的选择和组别数量，即每个资产的预测都基于稀疏个数的行业动量因子，自动处理多重共线性和变量选择，即控制预测方差。**可解释性**：利用回归学习框架，找到交叉预测结构的高维机器学习算法，避免了处理多重共线性常用的预处理方法，如 PCA 降维和因子正交化，有助于我们分析和迭代轮动策略。

# 框架介绍

■ **预测框架**

各个行业之间存在许多广泛而紧密的联系，一个行业收益率的变化可能引起其他行业的收益率变化，同时该行业也可能受到其它行业收益率变化的影响。比如钢铁行业，既有可能受益

于下游需求旺盛而上涨，其本身的产能扩张也会影响到上游能源行业。通过回归模型，可以定量刻画这种行业之间的传导规律， 建立起复杂的行业关联关系网络。下面给出只基于一阶动量（每日收益率）预测的示意图：




这是最简单的情形，即我们使用全体行业的当期收益率来预测下期收益率。值得注意的是，预测框架决定了变量间的关系结构。如果我们试图添加“往期收益率”的某些动量性因子作为预测下期的自变量，就默认了这种变量关系结构的存在。事实上，我们并没有将面板数据（Panel Data）作为一个多元时间序列来处理，仅作为多元截面数据（Cross sectional Data）加入到回归模型中进行预测。为了扩充截面模型的特征和维度，我们需要做时间序列的截面化处理，这一部分内容将放在特征工程中介绍。

## ■ 稀疏多目标回归 PSRCE

多目标回归(Multi Response Regression)是经典多元回归模型的推广，我们考虑有$p$个自变量和$q$个因变量的情形，即$x_i = \left(x_i^{(0)}, x_i^{(1)}, \dots, x_i^{(p-1)}\right)^T$和$y_i = \left(y_i^{(0)}, y_i^{(1)}, \dots, y_i^{(q-1)}\right)^T$代表第$i$个观察样本，多元回归由下式给出：

$$y_i = \boldsymbol{B}^T x_i + \epsilon_i \quad for \quad i = 1, \dots, n \tag{1}$$

其中，$\boldsymbol{B} \in \mathbb{R}^{p \times q}$是回归系数，并假设$\epsilon_i \sim N_q(0, \Sigma)$来自$q$维高斯噪音。我们将模型转换为矩阵形式，对于$n$观察样本，则令矩阵$\boldsymbol{X} = (x_0, x_1, .. x_{n-1})^T$，即$\boldsymbol{X} \in \mathbb{R}^{n \times q}$，它的第$i$行为$x_i^T$;同时令$\boldsymbol{Y} = (y_0, y_1, .. y_{n-1})^T$,即$\boldsymbol{Y} \in \mathbb{R}^{n \times p}$，它的第$i$行为$y_i^T$；令$\boldsymbol{E} \in \mathbb{R}^{n \times q}$，它的第$i$行为$\epsilon_i^T$。我们将多变元回归模型表示如下：

$$\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{E} \tag{2}$$

如果$q = 1$，则模型简化为经典回归模型，$\boldsymbol{B}$是一个$p$维回归系数向量。为了简化起见，我们默认$\boldsymbol{Y}$和$\boldsymbol{X}$所有的列向量都做了中心化处理，进而省略截距项。中心化处理的一般方法是，将所有维度移到原点$\boldsymbol{X} - \frac{\boldsymbol{1}}{\boldsymbol{n}}\boldsymbol{1}^{\boldsymbol{T}}\boldsymbol{X}$，并将维度$s$做归一化处理$\frac{x_i^{(s)}}{max\{x^{(s)}\}-min\{x^{(s)}\}}$。公式（2）导出的 Log 似然是优化$(\boldsymbol{B},\boldsymbol{\Sigma})$的目标函数：

$$g(\boldsymbol{B},\boldsymbol{\Sigma}) = tr\left[\frac{1}{n}(\boldsymbol{Y}-\boldsymbol{XB})^T(\boldsymbol{Y}-\boldsymbol{XB})\boldsymbol{\Sigma}^{-1}\right] - log\left|\boldsymbol{\Sigma}^{-1}\right| \tag{3}$$

最大似然的封闭解，称为“原始最小二乘解”，则为$\widehat{\boldsymbol{B}}^{OLS} = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{Y}$。在对$\boldsymbol{\Sigma}^{-1}$无约束条件下，$\boldsymbol{B}$的最优解并不依赖$\boldsymbol{\Sigma}^{-1}$。经验概率，当$\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}$可逆，$n \to \infty$这个估计具有一致性。而当$n$不足，$\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}$不可逆，且维度间存在多重共线性时，则需要进行因子的降维（PCA）和因子的正交化处理。

因子降维方法，提高了预测能力和方差控制，但直接对因变量进行降维处理，进而损失了维度的信息，得到的结果不具有可解释性，难以通过参数估计的得到来得到对投资逻辑（板块轮动）的应证。另外一个重要的因素是，稀疏学习提供的约束项，能确保选择的行业交叉预测的数量和发挥作用的维度仅维持在少数几个，这不仅符合统计机器学习对于模型选择的“奥卡姆剃刀”原则，也更有利于我们寻找行业轮动驱动因素。

综上，我们引入稀疏惩罚多元回归，令$\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$，并最大化下面的目标函数：

$$(\widehat{\boldsymbol{B}},\widehat{\boldsymbol{\Omega}}) = argmin_{\boldsymbol{B},\boldsymbol{\Omega}}\left\{ g(\boldsymbol{B},\boldsymbol{\Omega}) + \lambda_1 \sum_{j' \neq j} |\omega_{j'j}| + \lambda_2 \sum_{j=1}^{p}\sum_{k=1}^{q} |b_{jk}| \right\} \tag{4}$$

其中，$\lambda_1 \geq 0$和$\lambda_2 \geq 0$是超参数，分别是回归系数$\boldsymbol{B}$和逆协方差$\boldsymbol{\Omega}$的$L_1$收缩系数。我们称这个模型为“惩罚稀疏多目标回归 PSRCE”Penalized Sparse Response Regression with Covariance Estimation。在讨论完模型惩罚项的作用后，将会介绍求解它的算法。

## ■ 模型解释

1. 逆协方差Ω的稀疏约束

在 PSRCE 模型中，我们为多目标引入了噪音变元的逆协方差$\Omega$，即$\epsilon \sim N(0, \Omega^{-1})$。$|\Omega|$约束允许$q \geq n$情形的目标函数最优值是有限的，其次它减少了$\Omega$矩阵的参数数量，这在高维情形$q \gg n$有着对参数估计非常重要的作用。

本质上$\epsilon \sim N(0, \Omega^{-1})$是对拟合错误（bias）的先验假设，$\Omega$控制了拟合错误的先验分布密度，一方面$|\Omega|$的大小本身是$\Omega$的模长，$\lambda_1$是收缩系数。$\lambda_1$越大对应$\Omega^{-1}$的较小特征值分量估计越稳定。或者换一个角度，$\Omega$的稀疏性说明$\epsilon$可以从$q$个独立正太分布的稀疏线性组合得到，从而控制了模型拟合错误的可容忍自由度。

在我们的框架中，**机器学习**观点视作$\epsilon$的度量是训练集拟合的错误率（bias rate），进而我们降低自由度即拟合错误率，来寻求预测的可靠性。**统计学**角度$\Omega^{-1}$被视作回归异常点自适应的残差项，通过剥离残差来提高拟合稳定性。在**金融物理**中，有着更为复杂的残差分解模型（噪音先验假设），例如随机波动方程等。在**信号处理领域**，基于自适应滤波、趋势滤波等技术，实际上从时间序列角度直接移除残差影响。

如何找到有效的预测性参数$\boldsymbol{B}$，不同领域的知识都指向一点，残差假设在模型中发挥着非常重要的作用。同时，在本文投资学角度，残差项的协方差$\Omega^{-1}$也被视作多资产的组合风险的一部分，我们将在后文介绍如何计算基于预测模型的组合风险。

2. 多元回归系数$\boldsymbol{B}$的稀疏约束

传统的统计计量，并不允许在不通过实证分析确认单变量的收益贡献下，就加入到模型中，这会增加模型过拟合风险。高维统计利用稀疏结构，允许对高维因变量进行回归建模，稀疏参数确保了发挥作用的变量维度数量大大减少。

具体地，在行业轮动中的作用见第一部分“机器学习配置”的优势，即“交叉预测”、“风险控制”、“多资产管理”、“可扩展性”和“可解释性”。可扩展性内容超出了本文的介绍范围，考虑$\boldsymbol{B}$由$q$个分块组成（$q$个行业因子块），则稀疏惩罚的一个扩展版 Group Lasso 的公式如下：

$$g(\boldsymbol{B}, \boldsymbol{\Omega}) + \lambda_1 \sum_{j' \neq j} |\omega_{j'j}| + \lambda_2 \sum_{g=1}^{q} \left|\left|B_g\right|\right|_2 \quad (5)$$

其中，$B_g$是第$g$个行业的回归参数，它被归为一组。这样 Group Lasso 惩罚的作用是只

有稀疏个组别的参数不为 0，而组别内的变量要么全选中要么全为 0。简洁地理解惩罚项发挥的作用，如果我们让模长为某一个给定值，$L_1$模即绝对值和，它倾向于集中给越少的维度，$L_2$倾向于分散。Group Lasso 提出，是为了解决哑变量编码问题，即一个因子衍生的多个哑变量需要同时被选择或者不被选择，传统的 Lasso 型稀疏约束无法解决这一问题。在我们的问题中，行业衍生指标和价格衍生指标，都可能需要做组别间稀疏选择，这提供了“可扩展性”和“可解释性”的工具基础。

# 特征工程

## ■ 时间序列截面化

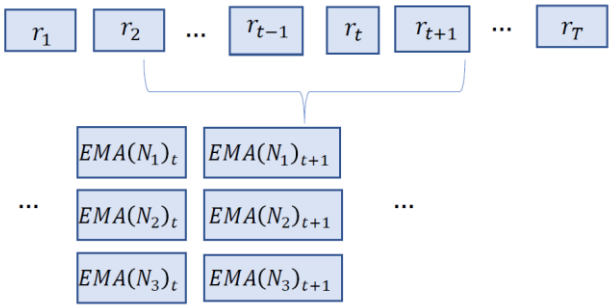
在本文中，我们不再考虑数据的时序依赖关系，也是多目标回归的基本假设。截面化处理，即将历史时间序列（在我们的指数配置场景中，也包括日内价格的变动和交易量等）做特征提取转变为截面型数据。

对于日间策略，考察一天的收益率作为动量是最简单的情形。在这里，我们给出一个动量型技术指标刻画价格变化的趋势和周期的简单分类描述，TicTacTec LLC.提供了技术型指标标准库 Technical Analysis Library 的开源程序库，其它商业版本类似，情况如下：

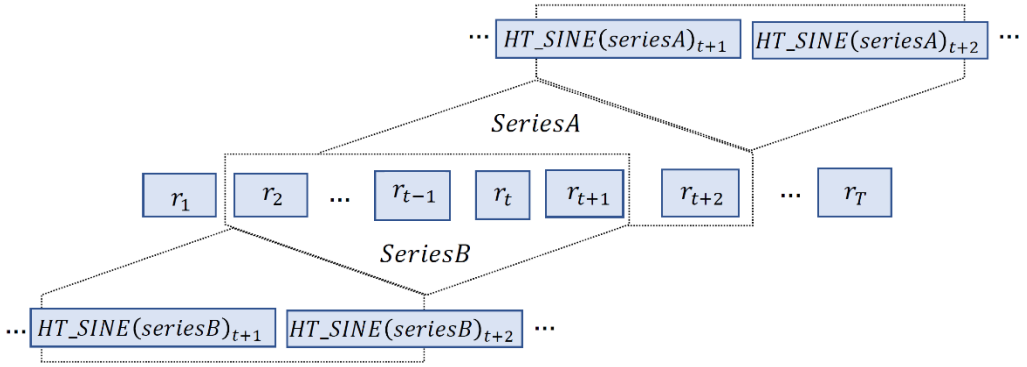
| 指标类型 | 典型指标 | 内容描述 | 特点 |
|---|---|---|---|
| 平滑指标 | BBANDS/DEMA/TEMA/TRIMA/EMA | 包括“指数滑动平均”、“加权移动平均”等，实际上是一种经验性滤波技术得到平滑时间序列 | 向前依赖 |
| 动量指标 | ADX/ADXR/MACD/MOM/ROC/STOCH/RSI | 包括“价格变动”和“价格变动率”以及“平均移动价差”和“相对强弱”，描述平滑指标的一阶变动率 | 向前依赖 |
| 波动指标 | ATR/NATR/TRANGE | 包括“波动幅度均值”和“归一化波幅均值”等，描述变动幅度 | 区间依赖 |
| 周期指标 | HT_SINE | 包括“希尔伯特变换 SINE”等，周期性函数刻画价格变动的循环分量的运动规律，可以探查价格序列的翻转 | 区间依赖 |
| 统计函数 | LINEARREG/BETA/CORREL | 统计函数计算线性拟合的“截距”、“回归系数”和“标准差”和“方差”等 | 区间依赖 |
| 运算函数 | ADD/MAX/SUM/FLOOR/SIN/COS/LOG10 | 最大值，和，三角函数变换和指数变换等 | 单点变换 |
| 流量指标 | AD/ADOSC/OBCV | 交易量流进流出指标 | 单点变换 |
| 模式指标 | ThreeLineStrike/ThreelnsideUpDown/Breakaway/Hammer | 技术性蜡烛图（Bar 线）的技术识别和经验性特征等 | 区间依赖 |

上述技术性因子表，实际上大体囊括了对历史价格分析的指标类型。例如“指数光滑均线 EMA”接受“收盘价”和“事件区间宽度”（$N$）两个参数，计算公式 $EMA(t) = Price(t) *$

$\frac{2}{N+1} + EMA(t-1) * \frac{N-1}{N+1}$。很明显，$N$控制了平滑和滞后性，给定一组$N$可以得到对应的一组$EMA$。下面的图示来说明这个特征变换过程：




$EMA$在$t+1$时刻的截面化处理只依赖时刻$t+1$之前的数据，因此我们将这类技术指标的特性归为“向前依赖”，这类技术指标可以由滑动窗口参数进行扩展。另外一类是“区间依赖”技术性指标，这类指标的特性是根据所给定的区间段进行变换，如下面的示意图：




很明显，HT_SINE 输入一个时间序列片段，并得到周期变换后的序列。周期变换函数利用“快速傅里叶变换”FFT 等算法对时间序列区间内的序列做频谱变换，所以变换后的时间序列利用了整个区间的信息。$SeriesA$和序列$SeriesB$有着时序重叠部分，周期技术指标通常被交易员用来确定翻转（reversal）信号，因而我们将此类技术指标的特性归为“区间依赖”型，在区间内部，$HT_SINE$信号变换依赖整个区间的周期性。类似“希尔伯特周期变换”的周期分析指标，利用线性回归获得趋势斜率，或者计算区间内的波动率等都具有此类特性。

如果我们考察时刻 t 的“区间依赖”特性的指标，则区间窗口的大小和区间起始范围的不同，都可能得到技术指标给出时刻 t 截然不同的数值，因而我们截取截止 t 时刻之前的时间序列进行区间依赖计算，并提取截面特征，例如“收益回归的截距”项、“收益回归的 beta 系数”等都属于此类。

事实上，其它基于无监督学习和标注，以及模式匹配等对时间序列进行状态判断的数值因子都可以作为截面特征工程，融入到当前的框架中。具体地，参考实验部分细节。

# 模型求解

## ■ 数值分析

优化公式（3）中的目标函数可以简写为：

$$(\widehat{\boldsymbol{B}},\widehat{\boldsymbol{\Omega}}) = argmin_{\boldsymbol{B},\boldsymbol{\Omega}}\{\, g(\boldsymbol{B},\boldsymbol{\Omega}) + \lambda_1|\boldsymbol{\Omega}| + \lambda_2|\boldsymbol{B}|\} \tag{6}$$

目标函数（6）中的$\boldsymbol{\Omega}$和$\boldsymbol{B}$的优化并不满足凸函数，但是固定其中一个优化另外一个则满足凸性（Convex）。我们做如下分析：

### 1. 固定参数 $\boldsymbol{B} = \boldsymbol{B_0}$

$$\widehat{\boldsymbol{\Omega}}(\boldsymbol{B}_0) = argmin_{\boldsymbol{\Omega}}\left\{ tr\left(\widehat{\boldsymbol{\Sigma}_R}\boldsymbol{\Omega}\right) - \log(|\boldsymbol{\Omega}|) + \lambda_1 \sum_{j'\neq j}|\omega_{j'j}| \right\} \tag{7}$$

其中，$\widehat{\boldsymbol{\Sigma}_R} = \frac{1}{n}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}_0)^T(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}_0)$，很明显，这是标准的$l1-penalized$协方差估计问题，因而我们可以利用 Graphical Lasso 算法来快速求解（7），我们在后续给出这一算法流程。

### 2. 固定参数 $\boldsymbol{\Omega} = \boldsymbol{\Omega_0}$

$$\widehat{\boldsymbol{B}}(\boldsymbol{\Omega}_0) = argmin_{\boldsymbol{B}}\left\{ tr\left(\frac{1}{n}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B})^T(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B})\boldsymbol{\Omega}_0\right) + \lambda_2 \sum_{j=1}^{p}\sum_{k=1}^{q}|b_{jk}| \right\} \tag{8}$$

如果$\boldsymbol{\Omega}_0$是半正定矩阵，则目标函数（8）的二阶 Hessian 矩阵$2\mathrm{n}^{-1}\Omega_0 \otimes X^T X$是半正定矩阵，显然满足凸性。目标函数（8）是凸函数，即存在为 0 的次梯度（sub-gradient），并记满足条件的全局最优解为$\boldsymbol{B}^{opt}$。用矩阵形式表示：

$$0 = 2n^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{B}^{opt}\boldsymbol{\Omega} - 2n^{-1}\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{\Omega} + \lambda_2\boldsymbol{\Gamma} \tag{9}$$

进而导出$\boldsymbol{B}^{opt}$满足 $\boldsymbol{B}^{opt} = \hat{B}^{OLS} - \lambda_2(2n^{-1}\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}$，其中$\boldsymbol{\Gamma} \equiv \boldsymbol{\Gamma}(\boldsymbol{B}^{opt}) \in \mathbb{R}^{p\times q}$并且其中的元素$\gamma_{ij} = sign\left(b_{ij}^{opt}\right) if\ b_{ij}^{opt} \neq 0\ and\ otherwise.$ 如果忽略残差相关性错误则有$\boldsymbol{\Omega}^{-1} = \boldsymbol{I}$，因此，高相关性假设对$\boldsymbol{B}^{opt}$的收缩作用要远大于低相关性。

### 3. 方向导数

我们令$f(\boldsymbol{B},\boldsymbol{\Omega})$表示（6）中的目标函数，则有下面的方向导数(directional derivatives)：

$$\frac{\partial f^+}{\partial \boldsymbol{B}} = \frac{2}{n}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{B}\boldsymbol{\Omega} - \frac{2}{n}\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{\Omega} + \lambda_2\mathbf{1}(b_{ij} \geq 0) - \lambda_2\mathbf{1}(b_{ij} < 0)$$
$$\frac{\partial f^-}{\partial \boldsymbol{B}} = -\frac{2}{n}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{B}\boldsymbol{\Omega} + \frac{2}{n}\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{\Omega} - \lambda_2\mathbf{1}(b_{ij} > 0) + \lambda_2\mathbf{1}(b_{ij} \leq 0) \quad (10)$$

其中，$\mathbf{1}$是矩阵形式的指示函数，当且仅当条件成立为 1 否则为 0。令$\boldsymbol{S} = \boldsymbol{X}^T\boldsymbol{X}$，$\boldsymbol{H} = \boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{\Omega}$并且$u_{rc} = \sum_{j=1}^{p}\sum_{k=1}^{q} b_{jk}s_{rj}\omega_{kc}$, 对于$\boldsymbol{B}$的某个分量$b_{rc}$，我们有方向导数：

$$\frac{\partial f^+}{\partial b_{rc}} = \frac{2}{n}u_{rc} - \frac{2}{n}h_{rc} + \lambda_2 1(b_{rc} \geq 0) - \lambda_2 1(b_{rc} < 0)$$
$$\frac{\partial f^-}{\partial b_{rc}} = -\frac{2}{n}u_{rc} + \frac{2}{n}h_{rc} - \lambda_2 1(b_{rc} > 0) + \lambda_2 1(b_{rc} \leq 0) \quad (11)$$

由于 L1 惩罚回归的目标函数（8）是不光滑的凸函数，从上面的式子可以看出导数存在跳变，但具有对称性。

**4. 最优值分析**

首先考虑非惩罚目标函数的导数$\frac{\partial f}{\partial b_{rc}} = \frac{2}{n}u_{rc} - \frac{2}{n}h_{rc}$，若单变元$b_{rc}$初始迭代为$b_{rc}^0$, 则满足最小化的解$\widehat{b_{rc}^*}$满足：

$$\left(\sum_{j=1}^{p}\sum_{k=1}^{q} b_{jk}^0 s_{rj}\omega_{kc} - b_{rc}^0 s_{rr}\omega_{cc} + \widehat{b_{rc}^*}s_{rr}\omega_{cc}\right) - h_{rc} = 0$$
$$\widehat{b_{rc}^*}s_{rr}\omega_{cc} - b_{rc}^0 s_{rr}\omega_{cc} + u_{rc} - h_{rc} = 0 \quad (12)$$

导出$\widehat{b_{rc}^*} = b_{rc}^0 + \frac{h_{rc}-u_{rc}}{s_{rr}\omega_{rr}}$。如果$\widehat{b_{rc}^*} > 0$，则考虑惩罚项的最优值点$\widehat{b_{rc}}$，由目标函数的凸性$0 < \widehat{b_{rc}} < \widehat{b_{rc}^*}$（仅实直线负方向是下降方向，即系数收缩，且目标函数上升速率和 L1 模收缩速率有零界点，即负方向导数为 0）则一定满足：

$$\widehat{b_{rc}^*}s_{rr}\omega_{cc} - b_{rc}^0 s_{rr}\omega_{cc} + u_{rc} - h_{rc} + \frac{n}{2}\lambda_2 = 0$$
$$\widehat{b_{rc}} = max\left(0, \widehat{b_{rc}^*} - \frac{0.5n\lambda_2}{s_{rr}\omega_{rr}}\right) \quad (13)$$

同样地，如果$\widehat{b_{rc}^*} < 0$则有正方向导数为 0：

$$\widehat{b_{rc}} = min\left(0, \widehat{b_{rc}^*} + \frac{0.5n\lambda_2}{s_{rr}\omega_{rr}}\right) \quad (14)$$

$\widehat{b_{rc}^*} = 0$则$\widehat{b_{rc}} = 0$，综合上述情况，可得：

$$\widehat{b_{rc}} = sign\left(b_{rc}^{0} + \frac{h_{rc} - u_{rc}}{s_{rr}\omega_{rr}}\right)\left(\left|b_{rc}^{0} + \frac{h_{rc} - u_{rc}}{s_{rr}\omega_{rr}}\right| - \frac{0.5n\lambda_2}{s_{rr}\omega_{rr}}\right)_{+} \quad (15)$$

■ **迭代算法**

**算法 1：固定 Ω迭代$\boldsymbol{B}$**

给定$\boldsymbol{\Omega}$和起始值$\widehat{\boldsymbol{B}}^{(0)}$, 令$\boldsymbol{S} = \boldsymbol{X}^T\boldsymbol{X}$，$\boldsymbol{H} = \boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{\Omega}$，有下面的两步迭代：

Step1：设置$\widehat{\boldsymbol{B}}^{(m-1)} \to \widehat{\boldsymbol{B}}^{(m)}$, 遍历所有的$\widehat{\boldsymbol{B}}^{(m)}$元素, 对每个$(r,c)$利用公式（15）来更新：

$$\widehat{b_{rc}}^{(m)} = sign\left(\widehat{b_{rc}}^{(m)} + \frac{h_{rc} - u_{rc}}{s_{rr}\omega_{rr}}\right)\left(\left|\widehat{b_{rc}}^{(m)} + \frac{h_{rc} - u_{rc}}{s_{rr}\omega_{rr}}\right| - \frac{0.5n\lambda_2}{s_{rr}\omega_{rr}}\right)_{+}$$

$$u_{rc} = \sum_{j=1}^{p}\sum_{k=1}^{q}\widehat{b_{jk}}^{(m)}s_{rj}\omega_{kc}$$

Step2：如果$\sum_{j=1}^{p}\sum_{k=1}^{q}\left|\widehat{b_{rc}}^{(m)} - \widehat{b_{rc}}^{(m-1)}\right| < \epsilon\sum_{j=1}^{p}\sum_{k=1}^{q}\left|\widehat{b_{rc}}^{(Ridge)}\right|$则停止，否则 Step1；

其中$\widehat{b_{rc}}^{(Ridge)}$是 Ridge 回归估计$(\boldsymbol{X}^T\boldsymbol{X} + \lambda_2\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$用来测试收敛条件。上面的迭代步骤确保收敛，由 trace 的凸性和惩罚项的性质保证，可以参考 Friedman2007 等。$\epsilon$可以取$10^{-4}$，它决定迭代次数，而每次迭代的浮点运算复杂度为$O(vpq)$, 其中$v$是每次迭代的非零元平均数量，最坏情况是$O(p^2q^2)$。

**算法 2：块坐标下降**

基于算法 1，给定超参数$\lambda_1$和$\lambda_2$，初始化$\widehat{\boldsymbol{B}}^{(0)} = 0$并且$\widehat{\Omega}^{(0)} = \widehat{\Omega}(\hat{B}^{(0)})$。循环坐标下降（Coordinate Descent）迭代的步骤：

Step1：基于$\widehat{\Omega}^{(m)}$固定, 利用算法 1 求解公式（8）, 得到$\hat{B}\left(\widehat{\Omega}^{(m)}\right)$, 令$\widehat{\boldsymbol{B}}^{(m+1)} = \hat{B}\left(\widehat{\Omega}^{(m)}\right)$;

Step2：基于$\widehat{\boldsymbol{B}}^{(m+1)}$固定，利用 graphical lasso 算法求解公式（7），得到$\widehat{\Omega}^{(m+1)}$;

Step3：如果$\sum_{j=1}^{p}\sum_{k=1}^{q}\left|\widehat{b_{rc}}^{(m)} - \widehat{b_{rc}}^{(m-1)}\right| < \epsilon\sum_{j=1}^{p}\sum_{k=1}^{q}\left|\widehat{b_{rc}}^{(Ridge)}\right|$则停止，否则 Step1；

上述步骤 Step1 和 Step2 确保了每次目标函数迭代都会下降。为了加速算法的收敛，可以利用如下近似方法。

**算法 3：近似算法**

给定超参数$\lambda_1$和$\lambda_2$，迭代步骤如下：

Step1：运行$q$个独立 Lasso 回归，通过交叉验证找到最优的惩罚系数$\widehat{\lambda_0}$，并确定回归系数$\widehat{\boldsymbol{B}}_{\widehat{\lambda_0}}^{lasso}$；

Step2：基于$\widehat{\boldsymbol{B}}_{\widehat{\lambda_0}}^{lasso}$固定，利用 graphical lasso 算法求解公式（7），得到$\widehat{\boldsymbol{\Omega}}\left(\widehat{\boldsymbol{B}}_{\widehat{\lambda_0}}^{lasso}\right)$；

Step3：基于$\widehat{\boldsymbol{\Omega}}\left(\widehat{\boldsymbol{B}}_{\widehat{\lambda_0}}^{lasso}\right)$固定，利用算法 1 得到$\hat{B} = \hat{B}(\widehat{\boldsymbol{\Omega}})$；

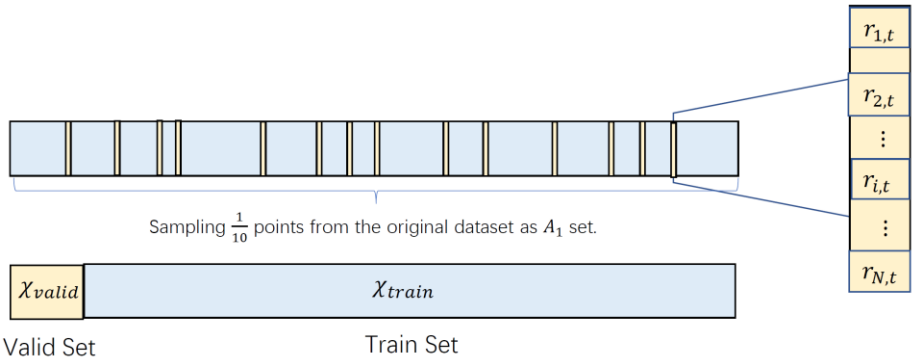
# 训练方法

## ■ 观点回顾

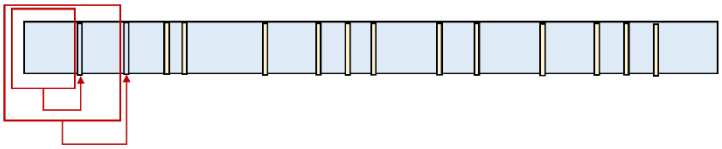
统计学习观点，将最小二乘回归的错误分为偏差 bias 和方差 variance 两个部分，训练集上的优化解决的是如何降低 bias，而 bias 降低必然意味着需要更大的模型容量（参数规模或者模型空间更大）和更精确的优化，这会导致预测 variance 变大（预测和推广能力变差），在机器学习领域，很多简单的方法如早期停止法（Early Stop）等，就是为了避免在训练集上做过度拟合（Overfit）。正则化方法利用结构化经验风险最小化 Structural Empirical Risk Minimization 来降低训练集的拟合程度（VC 维增长），以提高模型的推广能力（泛化界）。在我们的情形中，系数多元回归最终有效特征数量决定了模型的复杂度，这些都由超参数来控制。

## ■ 数据抽样

利用$K = 10$折交叉验证来模拟模型预测误差期望值，我们令对输入的 $T$ 长度的资产收益序列$x_1, \ldots, x_T$，随机抽取 10%的数据样本，并留下 $0.9T$数量的样本。留下的样本作为训练集（Train Dataset），抽取的样本作为验证集（Validation Dataset）。

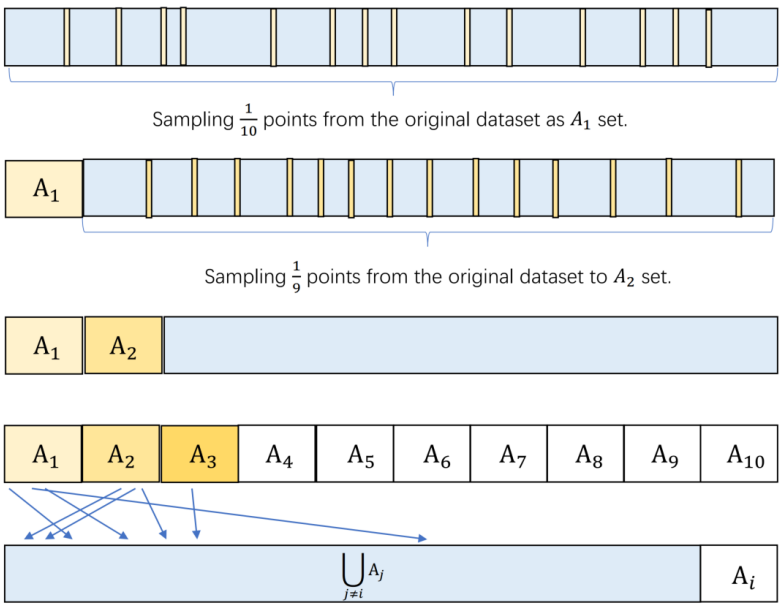


我们建立的预测关系是利用“当期观察的历史价格序列”对“未来区间收益或者下期收益”进行回归预测，为了确保训练集中不出现验证集数据的信息，缺失的时间戳则以缺失值处理，以“维度均值”或者“历史抽样替换”进行填充。例如，对于时间序列$x_1,\dots,x_{20}$，我们得到$A_1=\{x_5,x_{13}\}$，则$x_5$的位置由$x_1,\dots,x_4$以$\frac{1}{4}$概率抽取出一个样本替代，$x_{13}$的位置由$x_1,\dots,x_{12}$以$\frac{1}{12}$概率抽取一个样本替代，以此类推。这种方法我们称为“时间序列移动自助法”（Moving Bootstrap time series）。



如上例，我们约定$\chi_{valid}$为一组序号集合即$\{5,13\}$，$\chi_{train}$为$\{1,\dots 20\}-\chi_{valid}$。原始时间序列中的顺序信息，将被用于构建特征和预测关系，当我们构建滚动预测模型（行业轮动策略时），$\chi_{train}$位置作为训练集（Train Set）预测目标，$\chi_{valid}$将作为验证集（Validation Set）的预测目标。构建时间序列截面信息所使用的是$\chi_{valid}$被填充后的序列。

基于自助法重建序列的策略被广泛用于时间序列建模中，以检验模型训练的参数稳定性。在我们的实例中$K=10$折交叉验证意味着有$\frac{1}{10}$的样本做了随机替换，因而在交叉验证实验中可以多次随机重建获得超参数的评价指标均值。N 折交叉验证（N-fold Cross-Validation）它将数据集划分为 N 份，每次做留 1（leave-one-out）训练，下面给出示意图（N=10）：

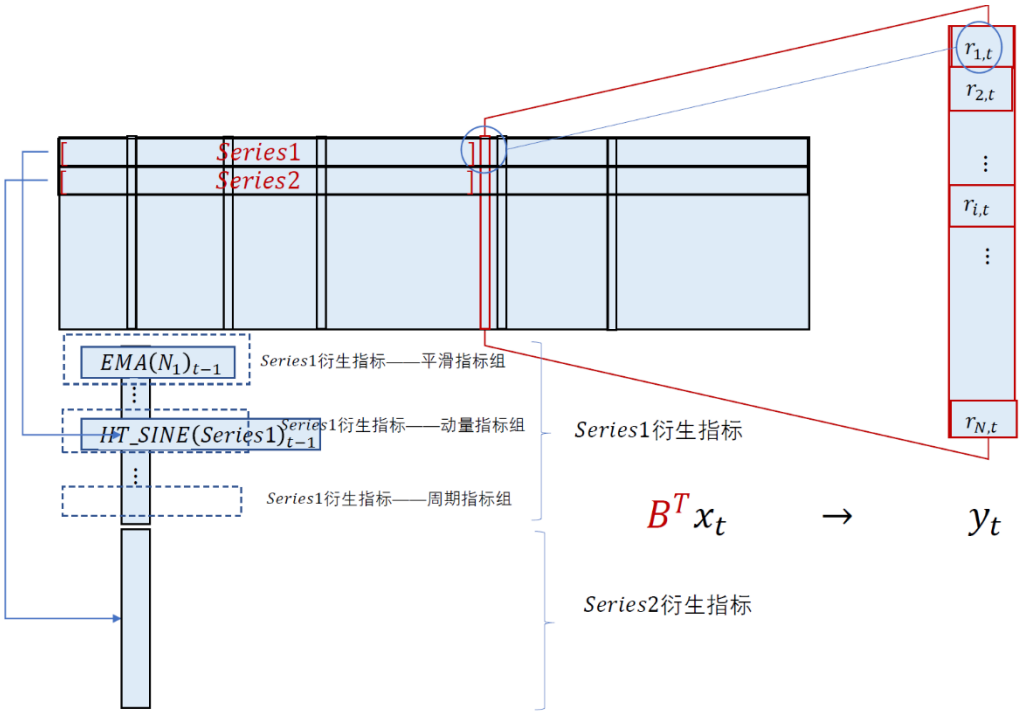



通过平均抽样，将数据集划分为$\{A_1,\dots A_{10}\}$，样本互不交叉$\cap_{j\neq i}A_i=\phi$。每次保留$\cup_{j\neq i}A_j$需

要样本作为$\chi_{train}$，$A_i$作为$\chi_{valid}$。

■ **样本构建**

基于特征工程这一部分的描述，我们给出如下的序列样本构建的示意图：




每种技术指标和截面特征提取的算法，对 N 个资产序列$Series1$、$Series2$构建了 N 个截面化因子，它们堆砌起来进而扩展为一个高维截面向量。值得注意的是，截面化处理函数对历史序列的窗口长度是有要求的，因而利用“滑动窗口”来构建预测样本$x_t \rightarrow y_t$，会截掉前置窗口的截面化特征（去首）。另外，在上面的示意图中$y_t$为一个时间戳上多资产收益向量，通常情况下，“按周轮动的行业间交叉预测”则需要合并 5 个收益时间戳$\left(r_{1,t} = log\frac{P_{1,t}}{P_{1,t-1}} \approx \frac{P_{1,t}}{P_{1,t-1}} - 1\right)$，即$\sum_{i=0}^{4} r_{1,t+i}$，样本构建过程也会截掉后 4 个时间戳对应的样本（去尾）。综上，虽然$K$次实验的的$\chi_{train}$和$\chi_{valid}$的数量一致，但由于我们不确定$\chi_{valid}$会落入“去首、去尾”的区间，因而有效样本数量每次是不一致的。

■ **评价指标**

定义$K$次交叉验证第$i$次实验以$\bigcup_{j\neq i} A_j$序号样本作为$\chi_{train}$，$A_i$作为$\chi_{valid}$，基于上述“时间序列移动自助法”填充后的序列构建样本，并基于$\chi_{train}$得到$\boldsymbol{X}^{(-i)}$，$\boldsymbol{Y}^{(-i)}$用于训练，运行算法-2 或者算法-3 得到参数$\boldsymbol{B}_{\lambda_1,\lambda_2}^{(-i)}$，再基于$\chi_{valid}$得到$\boldsymbol{X}^{(i)}$，$\boldsymbol{Y}^{(i)}$用于验证，有如下的最小二

乘错误：

$$\ell(\lambda_1,\lambda_2)_{train}^{(k)} = \frac{1}{|\chi^{(-k)}|}\left|\left|\boldsymbol{Y}^{(-k)} - \boldsymbol{X}^{(-k)}\boldsymbol{B}_{\lambda_1,\lambda_2}^{(-k)}\right|\right|_F^2$$
$$\ell(\lambda_1,\lambda_2)_{valid}^{(k)} = \frac{1}{|\chi^{(k)}|}\left|\left|\boldsymbol{Y}^{(k)} - \boldsymbol{X}^{(k)}\boldsymbol{B}_{\lambda_1,\lambda_2}^{(-k)}\right|\right|_F^2 \quad (16)$$

我们考虑$K$次交叉验证，以及每次对超参数$(\lambda_1,\lambda_2)$进行$N$次“时间序列移动自助法”重复实验，则得到如下“平均样本验证误差”（Averaging Validation Error Per Sample）：

$$loss(\lambda_1,\lambda_2) = \frac{1}{NK}\sum_{k=1}^{K}\sum_{i=1}^{N}\ell(\lambda_1,\lambda_2)^{(k,i)} \quad (17)$$

给定交叉验证次数$K$和自助法重复次数$N$下，记录$(k,i)$次实验的评价指标$\ell(\lambda_1,\lambda_2)_{train}^{(k,i)}$和$\ell(\lambda_1,\lambda_2)_{train}^{(k,i)}$，它有助于帮助我们分析训练拟合和预测偏差。

## ■ 超参数学习

未完成，待续