# 第3讲 资产定价和机器学习2

南京大学金融与保险学系

杨 念

2025年3月3日

# 降维

- 基于特征选择的方法: Lasso,…
  - Chinco, Alexander, Adam D Clark-Joseph, and Mao Ye. 2019. “Sparse Signals in the Cross-Section of Returns.” *Journal of Finance* 74 (1): 449–92.
  - Gu, Shihao, Bryan T Kelly, and Dacheng Xiu. 2020b. “Empirical Asset Pricing via Machine Learning.” *Review of Financial Studies* 33 (5): 2223–73.
- 线性降维方法
  - **主成分分析，Principal component analysis, PCA**
    - ✓ Lettau, Pelger. 2020. Estimating Latent Asset-Pricing Factors. *Journal of Econometrics,* 218: 1-31.
    - ✓ Pelger. 2019. Large-Dimensional Factor Modeling Based on High-Frequency Observations. *Journal of Econometrics* 208: 23-42.
  - **偏最小二乘，Partial least squares, PLS**
    - ✓ Kelly, Pruitt. 2013. Market expectations in the cross-section of present values. *Journal of Finance* 68: 1721-1756.
  - **IPCA, Instrumental Principal Component Analysis**
    - ✓ Kelly, Pruitt, Su. 2019. Characteristics are covariances A unified model of risk and return. *Journal of Financial Economics* 134: 501-524
  - ICA, LDA…
- 非线性降维方法
  - **自编码器(AutoEncoder)**
    - ✓ Gu, Kelly, Xiu. 2021. Autoencoder asset pricing models. *Journal of Econometrics* 222: 429-450
  - 基于核函数的方法(Kernel PCA)
  - …
- …

# 主成分分析(Principal component analysis, PCA)

- 主成分分析是一种**无监督降维**方法，它通过正交变换提取数据中方差最大的不相关线性组合（主成分），以保留最重要的信息并减少维度。
- 资产$i$在时间$t$的超额收益 $r_{i,t}$

$$r_{i,t} = \alpha_i + \beta_{i1} f_{t,1} + \beta_{i2} f_{t,2} + \cdots + \beta_{iK} f_{t,K} + \epsilon_{t,i} \qquad (*)$$

- 主成分分析帮助从数据提取最重要的因子$f_{t,k}$
  - 每一个资产都是一个特征，总共有$\boldsymbol{N}$个特征
  - 每一个特征沿时间序列方向都有$\boldsymbol{T}$个观察值
- 假设有$\boldsymbol{N}$ 个资产的历史收益率矩阵$\boldsymbol{R}$

1. 通过PCA，从$\boldsymbol{R}$的$\boldsymbol{N}$个特征中提取潜在的因子$\boldsymbol{F_K}$ ($K \times T$维矩阵)
   a) 计算协方差矩阵
   b) 进行特征值分解(EVD, SVD)
   c) 选取主成分
   d) 提取因子时间序列：$\boldsymbol{F_K}$ ($K \times T$维矩阵)
2. 估计beta系数

$$R = \begin{bmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,T} \\ r_{2,1} & r_{2,2} & \cdots & r_{2,T} \\ \vdots & \vdots & \ddots & \vdots \\ r_{N,1} & r_{N,2} & \cdots & r_{N,T} \end{bmatrix}$$

# PCA在资产定价中的作用

- 纯数据驱动
- 自动提取**收益方差最大**的因子，提高模型的解释力
- **无监督降维**
- 通过 PCA 提取的因子，可以构造因子投资策略(投资或规避对某个风险因子暴露过高的资产)

- 问题：PCA会选择对收益无关但方差较大的因子，预测效果可能一般
- 一个可能的改进：PLS，关注因子与资产收益之间的关系

# 偏最小二乘法(Partial least squares, PLS)

- PLS 是一种**监督降维**方法
  - 用于当自变量和因变量可能存在多重共线性时，从数据中提取最相关的成分。
  - 相较于 PCA，PLS 主要关注 **因子（自变量）与资产收益（因变量）之间的关系**，而不是仅仅解释因子本身的方差
- 仅基于资产收益数据来提取最重要的风险因子，PLS 可以帮助找到一组**最佳的隐含因子**，并用于资产定价
  - 每一个资产都是一个特征，总共有$N$个特征
  - 每一个特征沿时间序列方向都有$T$个观察值
- 预测未来$M$期的资产收益率矩阵$Y$

$$Y = \begin{bmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,T} \\ y_{2,1} & y_{2,2} & \cdots & y_{2,T} \\ \vdots & \vdots & \ddots & \vdots \\ y_{M,1} & y_{M,2} & \cdots & y_{M,T} \end{bmatrix}$$

- PLS通过最大化R和Y之间的协方差来提取因子并构建预测模型

# Instrumental Principal Component Analysis (IPCA)

- Kelly, Pruitt, Su. 2019. Characteristics are covariances A unified model of risk and return. *Journal of Financial Economics* 134: 501-524
- PCA: **因子载荷是静态**的，预测信息完全来自成分因子
- IPCA
  - 模型特征
    - ✓ 这个系统由T个时期内的N个资产组成。
    - ✓ $f_{t+1}$：隐含因子，K*1。
    - ✓ $\boldsymbol{\beta_{i,t}}$：**动态因子载荷**，1*K。
    - ✓ $z_{i,t}$：资产特征的向量，L*1。
    - ✓ $\Gamma_\beta$：L*K，将L维的特征降成K维。

$$r_{i,t+1} = \alpha_{i,t} + \beta_{i,t} f_{t+1} + \epsilon_{i,t+1},$$

$$\alpha_{i,t} = z'_{i,t}\Gamma_\alpha + \nu_{\alpha,i,t}, \quad \beta_{i,t} = z'_{i,t}\Gamma_\beta + \nu_{\beta,i,t}.$$

  - **动态因子载荷，动态风险补偿，时变的beta提升了模型对于收益率的预测能力**
  - 将可观测的特征纳入模型中
  - L可以很大，K很小，实现了降维
  - 美股数据表明，IPCA样本内外表现均由于传统因子模型
  - 与其他模型相比，**IPCA具有更高水平的样本外夏普比率**

## Unrestricted model （$\Gamma_\alpha \neq 0$）

$$r_{i,t+1} = z'_{i,t}\Gamma_\alpha + z'_{i,t}\Gamma_\beta f_{t+1} + \epsilon^*_{i,t+1}. \qquad (11)$$

一阶条件

$$f_{t+1} = \left(\Gamma'_\beta Z'_t Z_t \Gamma_\beta\right)^{-1} \Gamma'_\beta Z'_t (r_{t+1} - Z_t \Gamma_\alpha), \quad \forall t. \qquad (12)$$

and

$$\mathrm{vec}(\hat{\Gamma}'_\beta) = \left(\sum_{t=1}^{T-1} Z'_t Z_t \otimes \hat{f}_{t+1}\hat{f}'_{t+1}\right)^{-1} \left(\sum_{t=1}^{T-1} \left[Z_t \otimes \hat{f}'_{t+1}\right]' r_{t+1}\right). \qquad (7)$$

求解方法

- **Alternating Least Squares（ALS 交替最小二乘法）**

额外要求

- 为了使结果更具有经济意义，文章提出额外要求：$\Gamma'_\alpha \Gamma_\beta = 0_{1\times K}$
- 方法：先通过上述一阶条件回归出一组$\Gamma_\alpha$ 、$\Gamma_\beta$，然后用$\Gamma_\beta$对$\Gamma_\alpha$做回归，得到的残差定义为$\widehat{\Gamma_\alpha}$
- 这样做的目的是把因子载荷无法解释的正交残差分配给截距。

## Restricted model $(\Gamma_\alpha = 0)$

$$r_{i,t+1} = z'_{i,t}\Gamma_\beta f_{t+1} + \epsilon^*_{i,t+1} \tag{4}$$

where $\epsilon^*_{i,t+1} = \epsilon_{i,t+1} + \nu_{\alpha,i,t} + \nu_{\beta,i,t} f_{t+1}$ is a composite error.[14]

向量形式

$$r_{t+1} = Z_t \Gamma_\beta f_{t+1} + \epsilon^*_{t+1}$$

目标函数

$$\min_{\Gamma_\beta, F} \sum_{t=1}^{T-1} \left(r_{t+1} - Z_t \Gamma_\beta f_{t+1}\right)' \left(r_{t+1} - Z_t \Gamma_\beta f_{t+1}\right). \tag{5}$$

一阶条件

$$\hat{f}_{t+1} = \left(\hat{\Gamma}'_\beta Z'_t Z_t \hat{\Gamma}_\beta\right)^{-1} \hat{\Gamma}'_\beta Z'_t r_{t+1}, \quad \forall t \tag{6}$$

and

$$\mathrm{vec}(\hat{\Gamma}'_\beta) = \left(\sum_{t=1}^{T-1} Z'_t Z_t \otimes \hat{f}_{t+1}\hat{f}'_{t+1}\right)^{-1} \left(\sum_{t=1}^{T-1} \left[Z_t \otimes \hat{f}'_{t+1}\right]' r_{t+1}\right). \tag{7}$$

求解方法

- Alternating Least Squares（ALS 交替最小二乘法）

# 非线性

- 自编码模型
  - Gu, Kelly, Xiu. 2021. Autoencoder asset pricing models. *Journal of Econometrics* 222: 429-450
  - 非线性的因子和载荷估计
- 神经网络，万能逼近定理
- 循环神经网络，RNN模型
  - 对序列数据中的模式进行建模，可以视为对应计量经济学里面的时间序列
  - Deep Sequence Modeling: Development and Applications in Asset Pricing" by Cong, Tang, Wang, and Zhang (2020)
  - 梯度消失或者梯度爆炸 → LSTM
- 长短期记忆网络，LSTM模型

# RNN模型与资产定价

- 循环神经网络（RNNs）可以被视为传统计量经济学中的时间序列分析的类似方法，因为二者都旨在对序列数据中的模式进行建模。
- RNNs **更具灵活性**，因为它们以数据驱动的方式揭示这些模式，并且采用**高度非线性**的方式，通常涉及多个隐藏状态。
- 总结而言，RNNs 具有以下特点
  a. 能够处理变长序列，使其对不同类型的序列数据具有较强的适应性
  b. 能够捕捉长期依赖性，但普通 RNN 可能会面临梯度消失的问题，使其难以学习长距离依赖关系；
  c. 能够保留序列中元素的顺序信息，这对于需要顺序上下文的任务至关重要；
  d. 在整个序列中共享参数，使模型能够在不同部分的序列上有效泛化。
- 梯度消失或者梯度爆炸 → LSTM