降维、预测与组合构建—一种“倒向切片回归”方法

分析师：包赞 S1230518090006
baozan@stocke.com.cn TEL：021-80108127

◆**研究背景：**

预测问题是金融研究的核心问题，机构投资者很多时候面临的是美丽的烦恼，就是可用来预测的变量太多，各种风格因子、异像因子、基本面因子，多达上百个。AI 就是在这些超高维数据结构下合理挖掘信息的一种有效方法，引用萨金特的话，AI 本质上就是统计学，所以，作者试图在 JASA 等统计前沿杂志上，寻找适合金融预测的优秀统计方法。本文利用算法主要参考两位优秀华人统计学家的成果，一个是 UCLA 教授李克昭在 JASA 上的 SIR（Sliced Inverse Regression）方法，一个是普林斯顿范剑青教授在 Econometrics 的发表。

◆**SIR 思想：**

常规降维方法是主成分分析法（PCA）或者类似的改进方法，但是此类方法有一个重要的缺陷，就是**因子降维时只考虑了因子的信息，被预测变量的信息完全被忽略，降维后得到的是公共因子，即任何被预测变量都采用相同的主成分因子。而 SIR 方法，降维时考虑被预测变量信息，不同的被预测变量下，降维后的得到的主成份也不同，极大的提高了拟合精度。**利用 SIR 方法会得到“充分预测变量”，然后利用该变量进行进一步的预测研究，预测效果会大幅提升，可以说“SIR 方法就是为预测而生的”。

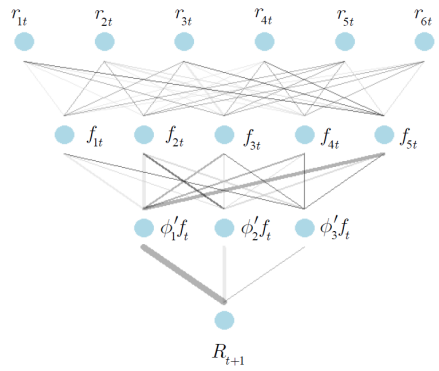
◆**SIR 金融预测运用：**

由于 SIR 方法无需过度研究因子本身，也**无需研究被预测股票与因子之间的关系**，简化了很多的分析工作。**本文利用 SIR 对**$\{(R_{t+1}, f_t)\}$ **进行降维分析，得到充分预测向量。然后利用**$\{(R_{t+1}, \beta' f_t)\}$**做回归分析，得到线性模型系数参数。利用 t+1 期股票收益和 t 期充分预测变量的线性关系，带入 t 期的充分预测变量数值，得到未来股票收益率的预测值。鉴于单一股票预测值本身没有意义，该方法适合对一个股票组合给出相对排序，获得特定样本下股票超低配的建议。对主动权益投资、指数增强策略、对冲策略能提供一些参考意见。**

◆**策略效果：**

见右边图，模型利用过往 60 个交易日的数据，进行主成分分析，取前 25 个主成分作为因子，然后利用 T 期个股收益与滞后一期的 25 个因子做 SIR 降维分析，得到预测变量，利用预测变量与前一期股票收益的线性关系，预测股票未来收益，取收益率排名前 30 的股票作为多头组合，后 30 作为空头组合。组合权重全部利用等权重。模型每隔 30 个交易日调仓，重新获得新的组合。**机构投资者可以利用更有效“因子”来提升预测效果。**

SIR 方法的 AI 表示：



策略效果：

正文目录

1. 前言 3

2. 倒向切片回归介绍 4

2.1. 充分降维思想 4
2.2. SIR 基本思想 6
2.3. SIR 算法与 R 程序 7
2.4. SIR 的 AI 表述与模拟计算 10

3. 倒向切片回归金融运用 11

3.1. SIR 预测性能展示 11
3.2. 金融上因子预测模型 12
3.3. 股票收益预测 SIR 算法 13

4. 应用 SIR 方法构建组合举例 13

图表目录

图 1：中心子空间举例图示 6
图 2：SIR 算法流程直观描述 8
图 3：SIR 预测流程 AI 化展示 10
图 4：SIR 预测向量具有更好预测性质 11
图 5：金融投资下 SIR 预测流程 AI 化展示 12
图 6：AI 化的预测流程 14
图 7：预测组合累积收益情况 14

表 1：2018 年以来多头股票持仓明细 15

附录：

1、参考文献
2、R 代码

1. 前言

预测问题是金融研究的核心问题，机构投资者很多时候面临的是美丽的烦恼，就是可用来预测的变量太多，各种风格因子、异像因子、基本面因子，多达上百个，如何有效利用这些因子构建投资组合获取收益是最重要、最被关心的课题。在 AI 算法日益盛行的今天，各类机器学习、人工智能算法充斥各行各业，引用萨金特的话，AI 本质上就是统计学，所以，作者试图在 JASA 等统计前沿杂志上，寻找适合金融预测的统计方法。本文利用算法主要参考两位优秀华人统计学家的成果，一个是 UCLA 教授李克昭在 JASA 上的 SIR 方法，一个是普林斯顿范剑青教授在 Econometrics 的发表。

当然 AI 算法并不是否定金融学，AI 只是提升预测准确性的工具，范剑青在一次报告中展示预测债券风险溢价的模型，他们利用 131 个宏观经济变量数据，其它数据依据这些数据挖掘获得，最终用了 8 个汇总宏观经济系列指标，用已有的信息来预测。发现有专业知识指导的机器学习可以改进预测的效果。如果只是用简单的算法来预测，大概可以预测 18%，如果用因子选择，可以预测到 24%左右，如果说因子选择的更合适，可以预测到 32%，如果再加入神经网络，最后的结果可以达到 45%左右，这个例子就说 AI 算法是有用的，但是一定要和金融专业知识相结合，尽量去选择更好的因子，这样才有最合适的预测效果。

本文着重向大家介绍“倒向切片回归方法”（Sliced Inverse Regression, SIR），并且举例构建投资组合。由于作者因子库尚未搭建完成，本文用主成份分析对沪深 300 成分股抽取 30 个主成份，作为模型基础数据构建组合，因为文章目标是向投资者介绍 SIR 方法，所以，这样做不影响文章展示效果，反而，“不合适”的因子下，如果组合业绩良好，能更好证实 SIR 方法在金融投资方面应用的有效性。

SIR 方法的名字就能字面理解该统计方法，倒向切片回归中的“倒向”指的是 $E[X \mid Y]$,普通 OLS 回归关心的是 $E[Y \mid X]$，“切片”指的是我们把 Y 切成不同的区间来估计倒向的条件均值，具体算法请见后文。

尽管 SIR 是经典统计方法，但是后文展示也可看出，该方法也是标准的 AI 算法。该方法由 UCLA 教授、著名统计学家李克昭提出。通常情况下，如果用大量的因子，大到因子的个数大于样本时间长度，来预测一个时间序列变量，会采用降维的方法，通过降维抽取能够涵盖因子的主要信息的特征向量，然后利用该向量来进行预测分析。常规方法是主成分分析法（PCA）或者类似的改进方法，但是此类方法有一个重要的缺陷，就是因子降维时只考虑了因子的信息，被预测的信息完全被忽略，降维后得到的是公共因子，即任何被预测变量都采用相同的主成分因子。而 SIR 方法，降维时考虑被预测变量信息，不同的被预测变量下，降维后的得到的主成份也不同，极大的提高了拟合精度，利用 SIR 方法会得到“充分预测变量”，然后利用该变量进行进一步的预测研究，预测效果会大幅提升。

高维统计方法中，降维方法很多，相比主成分分析，SIR 考虑了被预测变量的信息;相比有偏估计的 Lasso，SIR 是无偏估计，精度方面有优势；相比有特定分布假设下的参数方法，SIR 是不需要参数分布假设的非参方法，所以，SIR 方法尤其适合金融建模，之所以，该方法在金融方面的运用不多，主要是由于该方法的数学基础是矩阵论、线性空间和泛函分析，理解起来尤其抽象，影响了该方法的推广。

除了预测方面的运用，在做金融分析时候，经常会利用定价因子或者行业因子，比如 Fama 五因子来计算 R 方、alpha 等指标来进行分析，具体来说，利用 FF-5 计算股票 alpha，然后构建 alpha 动量策略，或者在基金选择方面，利用 R 方来进行基金优选。这些分析都用利用 SIR 方法，由于其“充分预测变量”具有更好的拟合效果，其 alpha 更“纯洁”，R 方也更加客观，否则，回归分析的 R 方会随着被预测变量个数的增多而不断增大。

本文第二部分重点介绍 SIR 方法的理论基础，然后介绍其算法，为了充分服务机构投资者，文中给出 SIR 方法的 R 语言程序。在第三部分，着重介绍该算法在金融中的运用，并且在统计模拟上，证明该方法在预测方面有更好效果。第四部分给出组合构建举例。

2. 倒向切片回归介绍

在信息技术时代，数据的来源和类型多样化，数据的规模越来越大。如果直接处理这样大规模的数据，可能会导致“维数灾难”。把高维数据降低为低维数据，并且使得降低维数的数据能够反映数据样本所表达的信息，这样的降维过程就变得非常有意义。目前，在参数模型下的自变量降维已经有非常成熟的方法，如 Lasso 回归（Tibshirani，1996），平滑截断绝对偏差(SCAD；Fan&Li，2001)等等。然而，在实际问题中，常常缺乏足够的信息支持一个参数模型的设定。因此如何在非参数环境下进行自变量的降维成为一个重要问题，SIR 就是非参数模型下的降维方法。

2.1. 充分降维思想

在回归中的降维问题中，一些传统的降维方法，比如主成分分析、因子分析、偏最小二乘方法，在实际的计算中是非常有用的。但是主成分分析以及因子分析没有考虑响应变量的信息。因此，可能要损失一部分回归的信息。而偏最小二乘方法只是考虑了线性模型的情况。对于更加一般的情况，比如非线性的时候。我们将介绍充分降维的思想。这种想法的一个很重要的特点便是把 X 用一些低维的线性组合 $X\beta$ 替代，但是却不损失条件分布的任何信息。而且，这种降维的方法不假定任何的参数模型。因此，我们认为，这种充分降维的方法对金融分析是非常有用的。为了不损失信息、提升拟合精度，不能把 Y 分开来而只讨论 X。下文，我们介绍充分降维，即在不导致信息损失的条件下降低 X 的维度。

充分降维的思想是在不假定任何参数模型以及不损失条件分布 $F[X \mid Y]$ 中所含有的信息的前提下，通过数据中高维的自变量的一些线性组合（个数较少），以之代替原自变量，而不导致原始回归信息的损失，来达到降维的目的。寻找原自变量的若干个线性组合这是解决高维自变量问题的一个合理方案。考虑因变量为 Y 关于自变量，在回归问题中，我们有模型如：

$$y_t = f(\beta_1' x_t, \ \cdots, \beta_K' x_t, \varepsilon_t)$$

其中：$x_t = (x_{1t}, \ \cdots, x_{pt})$ 是一个 $p \times 1$ 矩阵。

这里非随机常数向量 β 是未知的列向量，K 其中未知但远远小于 X 的维数 p，随机误差 ε 和 X 独立但分布未知，f 是一个未知的函数。

如果说，我们可以找出这样的向量$\beta_1,\beta_2,\cdots,\beta_K$对某个函数$f$成立，那么，我们就将回归问题变成了$Y$对$\beta_1'x_t,\ \ldots,\beta_K'x_t$进行回归拟合。充分降维就是要找到这样的一组向量使得K尽可能小，并且在给定$\beta_1'x_t,\ \ldots,\beta_K'x_t$时，$X$与$Y$独立。也可以假设存在矩阵$\beta$（其维数$K$远远小于自变量X的维数p），$\beta$的列向量由$\beta_1,\beta_2,\cdots,\beta_K$构成，在给定$X\beta$时，Y和X条件独立，即

$$Y \perp\!\!\!\perp X \mid X\beta$$

其中，“$\perp\!\!\!\perp$”表示独立。这个模型等价于：$Y \mid X$与$Y \mid X\beta$有相同的条件分布。也就是说，p维的向量可以被K维的线性组合$X\beta$代替，但是不损失Y关于X的回归的任何信息。这样的β总是存在的，且不唯一。因此，我们实际上是寻找自变量X张成的某个子空间S，满足

$$Y \perp\!\!\!\perp X \mid P_S X$$

这里的P_S表示关于内积的投影算子。满足这个条件的空间我们称为降维子空间。在一些较弱的条件满足时，所有满足这一些条件的空间的交集依然是一个降维空间。这时，我们称这个交集为中心降维子空间（Central dimension reduction subspace）。今后，我们记这个CDR子空间为$S_{Y|X}$。我们通常假定是存在$S_{Y|X}$，并且记$S_{Y|X}$的维数K是Y关于X的回归的结构维数。

如果X的协方差矩阵Σ_X是正定的话，我们标准化为：

$$Z = \hat{\Sigma}_X^{-1/2}(X - E(X))$$

Cook（1998）证明了$S_{Y|X} = \Sigma_X^{-1/2} S_{Y|Z}$。也就是说，基于X以及基于标准化的Z得到的两个空间之间可以自由转换的。因此，我们以后不妨假定X是标准化的随机变量。文献上有很多比较好的办法来估计CDR空间。比如说，倒向切片回归（SIR）、切片平均方差估计（SAVE）、以及等高线回归（SCR）。本文采用最传统的SIR方法。

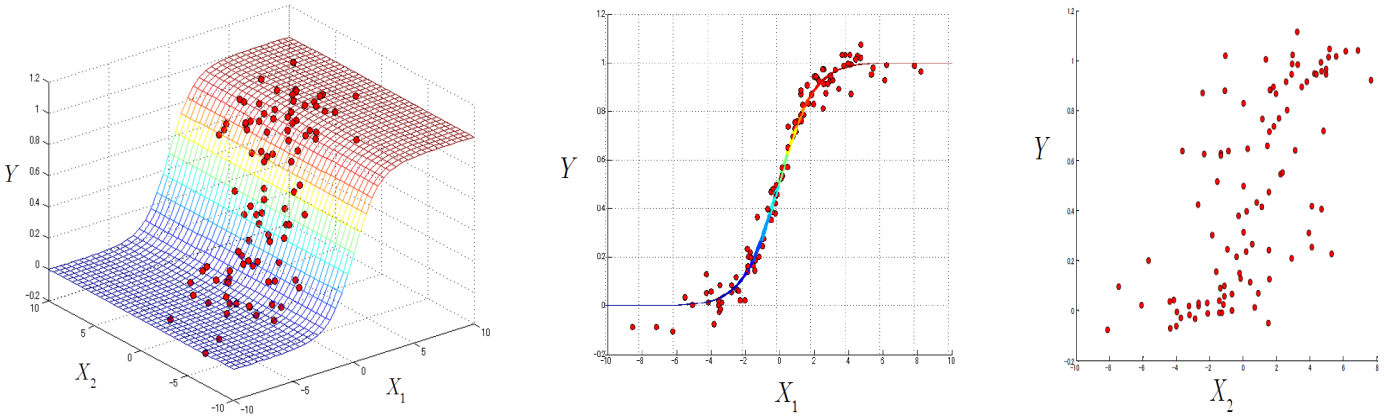
关于中心子空间，我们给出一个直观的举例：

$$Y = \frac{1}{1+e^{-X_1}} + X_2 \ ;\ X_2 \sim N(0, 0.1^2)$$

$$Y = \frac{1}{1+e^{-X_1}} + N(0, 0.1^2)$$

由于$Y \mid X$和$Y \mid X_1$ 分布相同，所以X_1是中心子空间。

图 1：中心子空间举例图示



*数据来源：浙商证券研究所

2.2. SIR 基本思想

切片逆回归是由 Li(1991)提出，这种方法操作简便，且较为稳健可靠，至今仍然被广泛使用。它沿用了主成分的基本思想，可以看作是主成分方法在回归背景下的改进。

记$\Sigma_\eta = \mathrm{cov}(E(x \mid Y))$，$\Sigma_x = \mathrm{cov}(x)$。求满足下式的相对特征向量$\beta_1, \cdots, \beta_k$

$$\Sigma_\eta b_i = \lambda_i \Sigma_x b_i$$

称第 i 个相对特征向量β_i为第 i 个切片逆回归（SIR）方向。

若仅有前 k 个特征值$\lambda_i, i = 1, \cdots, k$显著非零，则取其对应的 k 个相对特征向量所张成的空间$L(\beta_1, \cdots, \beta_k)$为降维空间。

此方法基于的矩阵是$\Sigma_\eta = \mathrm{cov}(E(x \mid Y))$，考虑到了因变量 Y 与自变量 x 之间的关系。

为了确保逆回归函数$E(X \mid Y)$均在子空间 CDR 中，切片逆回归假定了一个关键的条件：

$$\text{线性条件：} \quad E(X \mid X^T B) = P_B X$$

由于实际问题中 B 是未知的，因此我们通常要假定这个线性条件对任意的矩阵 B 都要成立。此时，这个线

性条件等价于假定自变量 X 是来自于椭球对称分布（Eaton，1986）。同时也指出线性条件不是一个很强的条件，因为高维数据的低维投影是渐近正态的。

当线性条件成立的时候，我们有：

$$E(X \mid Y) = E[E(X \mid Y, X^T B) \mid Y] = E[E(X \mid X^T B) \mid Y] = P_B E(X \mid Y)$$

所以，$M = Var[E(X \mid Y)]$**的列向量张成的空间总是属于 CDR 子空间的。**这里的 M 就称为切片逆回归的核矩阵了。此时，矩阵的非零特征值所对应的特征向量就是中心降维子空间的一个估计了。很显然，当 Y 是低维，比如说一维的时候，很多非参数的方法都是可以直接使用的。**其中，Li（1991）提出了“切片”的想法。也就是把响应变量按照大小分成若干“切片”，得到每一个“切片”以内对应的自变量的平均值，这些平均值构成的方差矩阵便是倒向切片回归的核矩阵的一个很好的估计了。**

切片估计的相合性是统计中的一个重要问题。Hsing and Carroll（1992）以及 Zhu and Ng（1995）证明了切片数 H 从$\sqrt{n}$ 到$n/2$时，渐近正态性以及$\sqrt{n}$ 相合性质总是成立的。这里 n 是样本点的个数。Li（1991）以及 Zhu，OhtakiandLi（2005）的一些模拟例子说明了切片估计得到的切片逆回归的效果对切片数非常不敏感的。很显然，这个发现得到了 Hsing and Carroll（1992）以及 Zhu and Ng（1995）的理论支持。另外，Zhu and Fang（1996）利用核函数来估计切片逆回归的核矩阵。当每个窗宽范围以内包含$n^{1/2}$到$n^{3/4}$个样本点的时候，有$\sqrt{n}$相合性以及渐近正态性的。很显然，从这个角度来说，Li（1991）的切片估计由于计算简单，具有一定的优势的。

2.3. SIR 算法与 R 程序

先简单介绍一下数学推导：

传统的协方差矩阵：

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(X_i - \bar{X}\right)^t \qquad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

分组内协方差矩阵：

$$\widehat{W} = \sum_{j=1}^{h}\frac{n_j}{n}\hat{\Sigma}_j$$

$\hat{\Sigma}_j$ 是第 j 个切片的协方差估计，$n_j = card\left(C_j\right)$ 为第 j 组样本个数。

切片间协方差矩阵：

$$\widehat{B} = \sum_{i=1}^{n}\frac{n_j}{n}(\bar{X}_j - \bar{X})(\bar{X}_j - \bar{X})^t \qquad \bar{X}_j = \frac{1}{n_j}\sum_{X_i \in C_j} X_i$$

依据矩阵运算规则：

$$\hat{\Sigma} = \hat{B} + \hat{W}$$

依据拟合规则，我们需要最小化给定 Y 值后 $b^t X$ 的方差，由于总方差恒定，相当于最大化切片间方差。

$$\hat{b} = \arg\max_b b^t \hat{\Gamma} b$$

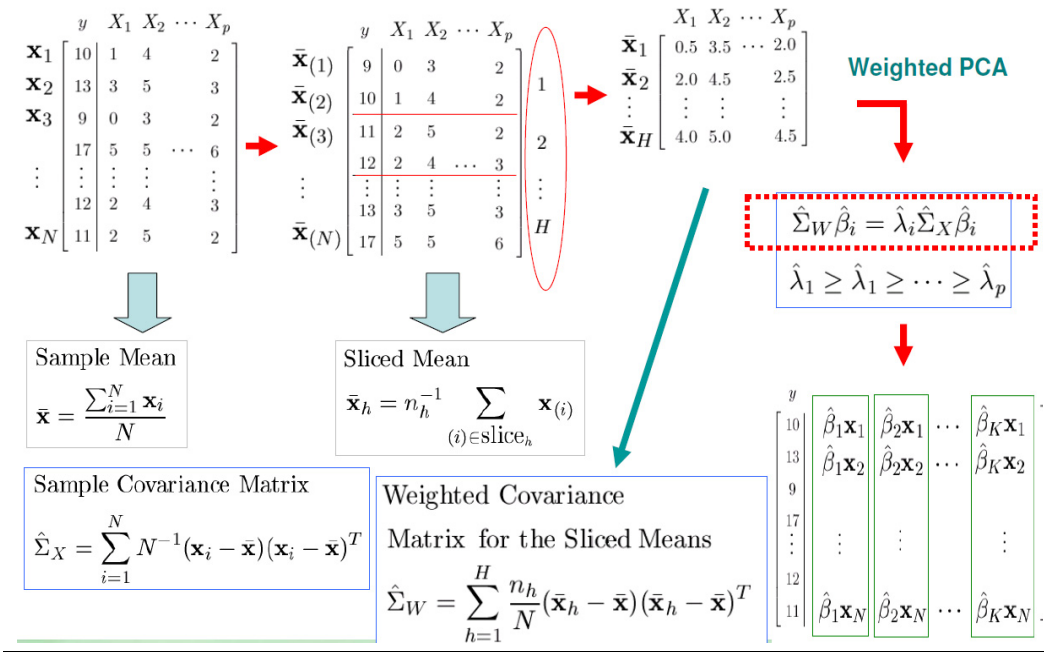
$$b^t \hat{\Sigma} b = 1$$

其中：$\hat{\Gamma} = \sum_{j=1}^{h} \frac{n_j}{n} (\bar{X}_j - \bar{X})(\bar{X}_j - \bar{X})^t$，$\bar{X}_j = \frac{1}{n_j} \sum_{Y_i \in S_j} X_i$

这个优化的解 $\hat{b}$ 就是 $\hat{\Sigma}^{-1} \hat{\Gamma}$ 的特征向量。

尽管 Li 关于 SIR 理论的描述涉及到矩阵论、线性空间理论与泛函分析，实际运用和理解，可以从更简洁直观的方式来理解，下图用直观性形式来描述。

图 2：SIR 算法流程直观描述



*注：红色虚线框内两边同乘 $\hat{\Sigma}_X^{-1}$ 即标准化后的特征向量形式。

为了编程考虑，我们写出 SIR 算法流程，并且给出 SIR 算法的 R 代码。

SIR 算法：

1、 计算样本均值和样本方差，然后标准化：

$$\hat{\mu} = E(X) \quad \hat{\Sigma} = \mathrm{var}(X)$$

$$Z_i = \hat{\Sigma}^{-1/2}(X_i - \hat{\mu}) \qquad i = 1,...,n$$

2、 估计：$E[Z \mid Y \in J_h]$ ，H 为切片个数：

$$E[Z \mid Y \in J_h] = \frac{E[ZI(Y \in J_h)]}{E[I(Y \in J_h)]} \qquad h = 1,...,H$$

3、 估计：$\mathrm{var}[Z \mid g(Y)]$，$g(Y) = \sum_{h=1}^{H} hI(Y \in J_h)$

$$\hat{\Lambda} = \sum_{h=1}^{H} E[I(Y \in J_h)]E[Z \mid I(Y \in J_h)]E[Z^T \mid I(Y \in J_h)]$$

4、 $\upsilon_1,...,\upsilon_r$ 表示 $\hat{\Lambda}$ 前 r 个特征向量，$\hat{\beta}_k = \hat{\Sigma}^{-1/2}\upsilon_k$，$k = 1,...,r$，充分预测变量为：

$$\hat{\beta}_k^T(X_1 - \hat{\mu}),\ \ldots\ , \hat{\beta}_k^T(X_n - \hat{\mu})$$

R 程序函数如下：

R 程序：

1、 矩阵幂函数：

```
mat_power = function(a, alpha){
        a = round((a + t(a))/2,7); tmp = eigen(a)
        return(tmp$vectors%*%diag((tmp$values)^alpha)%*%t(tmp$vectors))
        }
```

2、 $g(Y)$ 函数

```
discretize=function(y,h){
        n=length(y);m=floor(n/h)
        y=y+.00001*mean(y)*rnorm(n)
        yord = y[order(y)]
        divpt=numeric();for(i in 1:(h-1)) divpt = c(divpt,yord[i*m+1])
        y1=rep(0,n);y1[y<divpt[1]]=1;y1[y>=divpt[h-1]]=h
        for(i in 2:(h-1)) y1[(y>=divpt[i-1])&(y<divpt[i])]=i
        return(y1)
        }
```

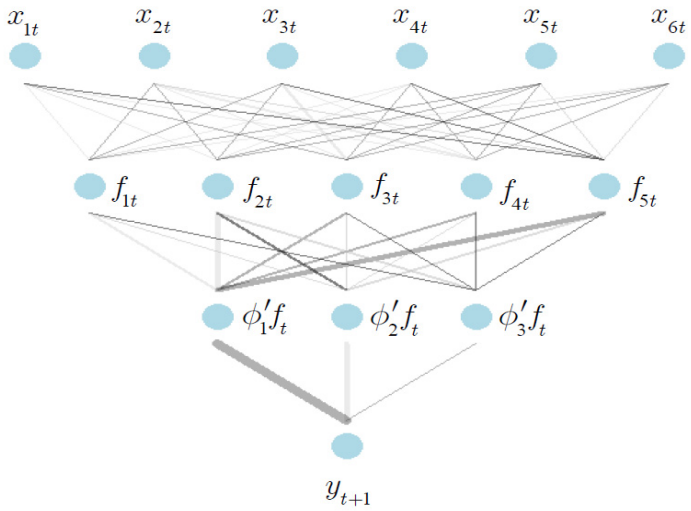
3、 充分预测变量函数：

```
sir=function(x,y,h,r){
                    p=ncol(x);n=nrow(x)
                    signrt=mat_power(var(x),-1/2)
                    xc=t(t(x)-apply(x,2,mean))
                    xst=xc%*%signrt
                    ydis=discretize(y,h)
                    yless=ydis;ylabel=numeric()
                    for(i in 1:n) {if(var(yless)!=0) {ylabel=
                    c(ylabel,yless[1]);yless=yless[yless!=yless[1]]}}
                    ylabel=c(ylabel,yless[1])
                    prob=numeric();exy=numeric()
                    for(i in 1:h) prob=c(prob,length(ydis[ydis==ylabel[i]])/n)
                    for(i in 1:h) exy=rbind(exy,apply(xst[ydis==ylabel[i],],2,mean))
                    sirmat=t(exy)%*%diag(prob)%*%exy
                    return(signrt%*%eigen(sirmat)$vectors[,1:r])
                    }
```

2.4. SIR 的 AI 表述与模拟计算

利用 SIR 的充分预测向量做预测可以理解为人工智能的四层次深度学习结构，从金融上来说， x_t 表示原始的经济信息，f_t 表示依据原始信息构建的指标因子，第三层为利用 SIR 得到的充分预测向量，最终利用这些向量，得出预测值。

图 3：SIR 预测流程 AI 化展示



*数据来源：浙商证券研究所

为了效率考虑，实际计算也可运用 dr 函数包，下面利用 dr 包，来进行 SIR 计算举例。

$$y = 2x_1 + 3x_2 + 6x_3 + 9x_4^3 + x_5^2 + \varepsilon$$

其中 x_i，$i = 1,2,3,4,5$ 为解释变量，y 为被解释变量，ε 为误差项。

通过运用 dr()函数，找到 β_i，$i = 1,\cdots,k$；再运用 lm()函数进行线性拟合。

模拟计算，假设这些变量都服从 1-2 的均匀分布，残差服从均值为 0，标准差为 0.1 的正太分布，通过计算，得到投射向量：beta=c(-0.756,-0.128,-0.327,-3.277,-0.118)。进一步按照下式计算，得到充分预测向量：

```
x_1 <- beta[1,1]*x1 + beta[2,1]*x2 + beta[3,1]*x3 + beta[4,1]*x4 + beta[5,1]*x5
```

利用该向量，对 y 变量进行回归分析，充分预测向量 p 值显著，R 方达到 97.4%，从这看出，SIR 方法做到了在不损耗信息的条件下，达到降维的效果。

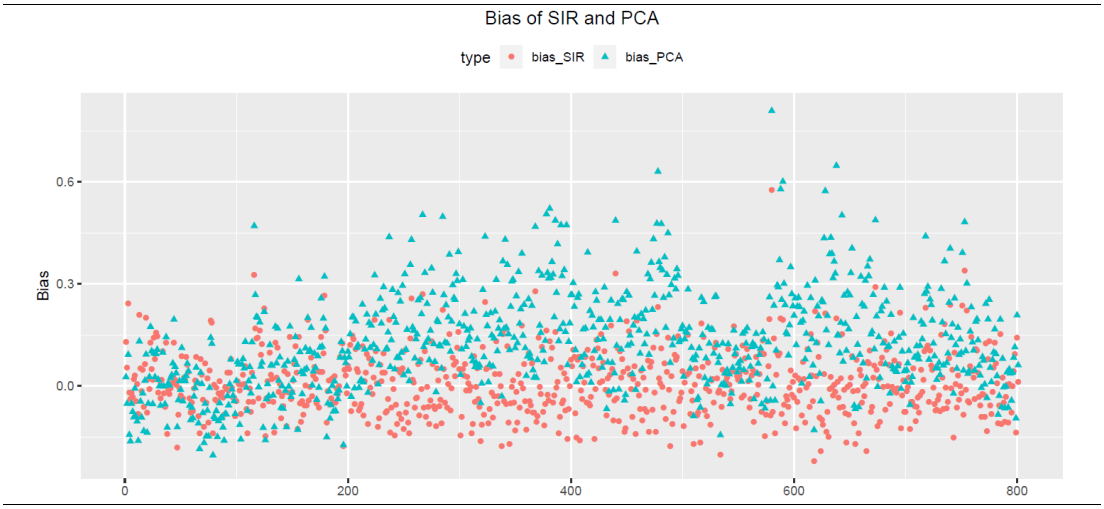
3. 倒向切片回归金融运用

3.1. SIR 预测性能展示

既然本文强烈推介 SIR 在金融上的运用，不仅仅要在数据结构上与解决方案上证明适合金融运用，更要证明其良好的预测性质。这一小结，我们构建一个“粗糙”模拟模型，用来展示 SIR 构建的预测向量具有更好的预测性质。

“粗糙”模型是生成十变量随机矩阵和依据此线性构成的 y，然后分别利用 SIR 和 PCA 方法，构建预测向量，对 T+1 期进行预测，然后考察预测值与真实值的偏离。此处我们选择“相对误差=（预测值-真实值）/真实值”来比较二者与真实值之间的差距。

图 4：SIR 预测向量具有更好预测性质



*数据来源：建模过程数据

在该图中，红色圆点代表切片逆回归 SIR 降维回归法的相对误差序列，蓝色三角点代表主成分分析 PCA 降维回归法的相对误差序列。由该图可以看出，SIR 的偏差分布大致集中在 0 附近，在（-0.3,0.3）范围内波动且分布呈现均值为零的正态分布。而 PCA 的偏差分布较为离散，有部分偏差点甚至超出了 0.6。由此可见，PCA 的误差较大，预测性能不如 SIR 降维后的预测向量。

3.2. 金融上因子预测模型

考虑以下因子模型，r_{t+1} 是希望被预测的股票在未来的收益率：

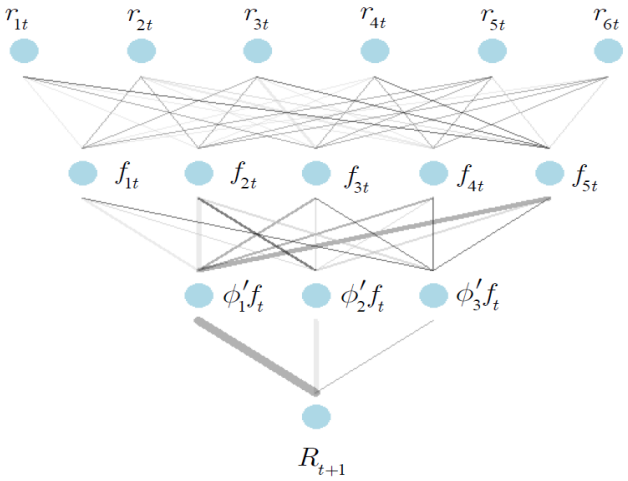
$$r_{t+1} = h(\phi_1' f_t, \cdots, \phi_L' f_t, \varepsilon_{t+1}) \quad (3\text{-}1)$$

$$r_{it} = b_i' f_t + u_{it}, 1 \le i \le p, 1 \le t \le T \quad (3\text{-}2)$$

考虑以下因子模型，r_{it} 是 t 时刻第 i 个股票收益率，b 是 K X 1 维因子载荷向量，$f_t = (f_{1t}, \cdots, f_{Kt})'$ 是 K X 1 维因子，u_{it} 是误差项（异质成分）。为了便于记号，我们记 $r_t = (r_{1t}, \cdots, r_{pt})'$，$B_t = (b_1, \cdots, b_p)'$，$u_t = (u_{1t}, \cdots, u_{pt})'$。3.1 式中 h(·)是未知函数，$\varepsilon_{t+1}$ 是与 f_t 和 u_{it} 独立的随机误差。$\phi_1, \cdots, \phi_L$ 的线性组合是 K 维标准正交向量。显然，这个模型同样适用于横截面回归。由于没有完备因子库，我们下文组合构建举例采用公式 3.2 主成分分析法来构建因子，当然，很多机构投资者已经有因子数据库，可以忽略 3.2 式。

股票收益的充分预测也可用深度学习结构表现，由四层线性或者非线性过程进行降维。充分预测和深度学习的关联见图 5。

图 5：金融投资下 SIR 预测流程 AI 化展示



*数据来源：浙商证券研究所

3.3. 股票收益预测 SIR 算法

SIR 金融运用算法：

1、得到被估计因子 $\{\hat{f}_t\}_{t=1,\cdots,T}$ ；

2、对 $\{\hat{f}_t\}_{t=1,\cdots,T}$ 进行标准化；

3、构建 $\hat{\Sigma}_{f|y}$ ：

$$\hat{\Sigma}_{f|y} = \frac{1}{H}\sum_{h=1}^{H} E(f_t \mid y_{t+1} \in I_h)E(f_t' \mid y_{t+1} \in I_h)$$

4、从 $\hat{\Sigma}_{f|y}$ 的 L 个最大特征向量得到 $\hat{\psi}_1,\cdots,\hat{\psi}_L$ ；

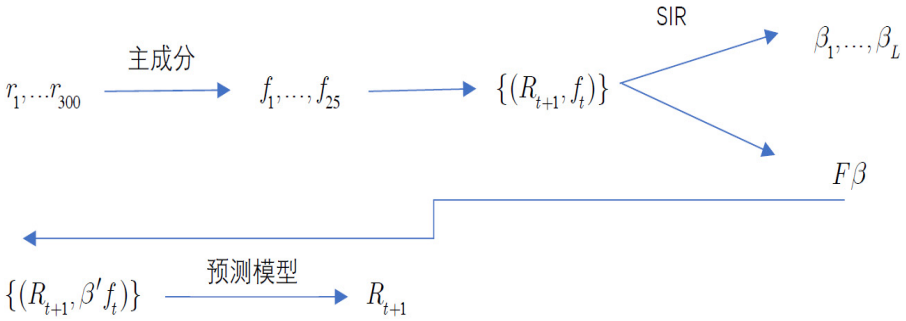
5、构建预测指数 $\hat{\psi}_1'\hat{f}_t,\cdots,\hat{\psi}_L'\hat{f}_t$ 并预测 R_{t+1} ；

4. 应用 SIR 方法构建组合举例

本文目标是向机构投资者推荐 SIR 方法，投资者可以利用该方法进行降维，然后用充分预测变量作进一步的预测分析。由于该方法无需过度研究因子本身，也**无需研究被预测股票与因子之间的关系**，简化了很多的分析工作。

本文利用 SIR 对 $\{(R_{t+1}, f_t)\}$ 进行降维分析，得到充分预测向量。受限于因子库、数据流量问题，本文利用沪深 300 成分股收益率矩阵，通过主成分分析法得到公共因子，然后利用前 25 个因子，进行 SIR 分析，得到充分预测变量，然后利用 $\{(R_{t+1}, \beta' f_t)\}$ 做回归分析，得到 t+1 期股票收益和 t 期充分预测变量的线性关系，最后带入 t 期的充分预测变量数值，得到未来股票收益率的预测值。

图 6：AI 化的预测流程



*数据来源：浙商证券研究所

由于沪深 300 指数定期调整，本文为了简化，没有考虑指数调整。模型利用过往 60 个交易日的数据，进行主成分分析，取前 25 个主成分作为因子，然后**利用 T 期个股收益与滞后一期的 25 个因子做 SIR 降维分析，得到预测变量，利用预测变量与前一期股票收益的线性关系，预测股票未来收益，取收益率排名前 30 的股票作为多头组合，后 30 作为空头组合。组合权重全部利用等权重。模型每隔 30 个交易日调仓，重新获得新的组合。**

机构投资这实际运用过程中，可以是用自己的因子数据库，这样会比本文单纯从 PCA 分析得到的因子更有效，预测性能更好。本文只用了简单线性回归来作为线性外推预测模型，实际运用中可以探索更有效的预测模型来提高预测精度。由于该方法能够给出未来股票收益的预测值，其预测值的本身并无意义，但是通过给出样本股票组合的相对排序，可以获得特定样本下股票超低配的建议。对主动权益投资、指数增强策略、对冲策略能提供一些参考意见。

图 7：预测组合累积收益情况

*数据来源：建模过程数据

我们在下表展示今年以来，每一期多头组合的股票明细。

表 1：2018 年以来多头股票持仓明细

2018.02.08		2018.03.29		2018.05.16		2018.06.28		2018.08.09		2018.09.20	
002001.SZ	新和成	600588.SH	用友网络	600874.SH	创业环保	600271.SH	航天信息	600569.SH	安阳钢铁	000063.SZ	中兴通讯
000933.SZ	神火股份	600859.SH	王府井	000002.SZ	万科 A	000959.SZ	首钢股份	600426.SH	华鲁恒升	600196.SH	复星医药
601699.SH	潞安环能	600176.SH	中国巨石	600598.SH	北大荒	600598.SH	北大荒	600068.SH	葛洲坝	002122.SZ	*ST 天马
000425.SZ	徐工机械	600037.SH	歌华有线	000568.SZ	泸州老窖	000917.SZ	电广传媒	000338.SZ	潍柴动力	000792.SZ	盐湖股份
000488.SZ	晨鸣纸业	600030.SH	中信证券	600737.SH	中粮糖业	600380.SH	健康元	600596.SH	新安股份	002024.SZ	苏宁易购
601111.SH	中国国航	600208.SH	新湖中宝	600183.SH	生益科技	600655.SH	豫园股份	600125.SH	铁龙物流	000060.SZ	中金岭南
000983.SZ	西山煤电	600426.SH	华鲁恒升	000858.SZ	五粮液	600183.SH	生益科技	000968.SZ	蓝焰控股	000783.SZ	长江证券
601588.SH	北辰实业	600600.SH	青岛啤酒	600048.SH	保利地产	600132.SH	重庆啤酒	000488.SZ	晨鸣纸业	000725.SZ	京东方 A
600456.SH	宝钛股份	600635.SH	大众公用	600216.SH	浙江医药	000401.SZ	冀东水泥	600741.SH	华域汽车	002038.SZ	双鹭药业
600208.SH	新湖中宝	000729.SZ	燕京啤酒	000983.SZ	西山煤电	000680.SZ	山推股份	600104.SH	上汽集团	601169.SH	北京银行
600547.SH	山东黄金	600895.SH	张江高科	600779.SH	水井坊	002024.SZ	苏宁易购	600004.SH	白云机场	600428.SH	中远海特
000069.SZ	华侨城 A	600299.SH	安迪苏	601628.SH	中国人寿	000999.SZ	华润三九	600037.SH	歌华有线	600299.SH	安迪苏
000825.SZ	太钢不锈	600015.SH	华夏银行	601088.SH	中国神华	600383.SH	金地集团	600066.SH	宇通客车	600050.SH	中国联通
601766.SH	中国中车	000895.SZ	双汇发展	000543.SZ	皖能电力	000488.SZ	晨鸣纸业	600643.SH	爱建集团	600718.SH	东软集团
000422.SZ	*ST 宜化	002024.SZ	苏宁易购	600690.SH	青岛海尔	600596.SH	新安股份	600270.SH	外运发展	000488.SZ	晨鸣纸业
600029.SH	南方航空	601601.SH	中国太保	000768.SZ	中航飞机	000968.SZ	蓝焰控股	000729.SZ	燕京啤酒	600741.SH	华域汽车
600595.SH	中孚实业	600585.SH	海螺水泥	600600.SH	青岛啤酒	600718.SH	东软集团	600028.SH	中国石化	000897.SZ	津滨发展
601899.SH	紫金矿业	600028.SH	中国石化	600362.SH	江西铜业	600109.SH	国金证券	601857.SH	中国石油	600809.SH	山西汾酒
600153.SH	建发股份	600718.SH	东软集团	601991.SH	大唐发电	002155.SZ	湖南黄金	000876.SZ	新希望	600426.SH	华鲁恒升
600638.SH	新黄浦	600643.SH	爱建集团	600812.SH	华北制药	600516.SH	方大炭素	600000.SH	浦发银行	600220.SH	江苏阳光
600031.SH	三一重工	600519.SH	贵州茅台	002001.SZ	新和成	600639.SH	浦东金桥	600839.SH	四川长虹	600096.SH	云天化
600352.SH	浙江龙盛	600048.SH	保利地产	600528.SH	中铁工业	600050.SH	中国联通	600663.SH	陆家嘴	600183.SH	生益科技
002155.SZ	湖南黄金	600528.SH	中铁工业	600027.SH	华电国际	600307.SH	酒钢宏兴	600132.SH	重庆啤酒	600068.SH	葛洲坝
000968.SZ	蓝焰控股	000858.SZ	五粮液	600820.SH	隧道股份	600027.SH	华电国际	600362.SH	江西铜业	000651.SZ	格力电器
600066.SH	宇通客车	601186.SH	中国铁建	601601.SH	中国太保	601699.SH	潞安环能	600816.SH	安信信托	000927.SZ	一汽夏利
601898.SH	中煤能源	600782.SH	新钢股份	601998.SH	中信银行	600117.SH	西宁特钢	601169.SH	北京银行	600100.SH	同方股份
000059.SZ	华锦股份	600383.SH	金地集团	000807.SZ	云铝股份	000002.SZ	万科 A	600123.SH	兰花科创	600276.SH	恒瑞医药
600096.SH	云天化	600029.SH	南方航空	000652.SZ	泰达股份	600547.SH	山东黄金	002244.SZ	滨江集团	601186.SH	中国铁建
000792.SZ	盐湖股份	600881.SH	亚泰集团	600019.SH	宝钢股份	600585.SH	海螺水泥	601919.SH	中远海控	600804.SH	鹏博士
600028.SH	中国石化	600016.SH	民生银行	601766.SH	中国中车	600779.SH	水井坊	600782.SH	新钢股份	601398.SH	工商银行

资料来源：浙商证券研究所

附录：

1、参考文献：

[1] Li, K.-C. (1991), Sliced inverse regression for dimension reduction (with discussion), Journal of the American Statistical Association 86(414), 316-327.

[2] Cook, R.D. (2007). Fisher lecture : Dimension reduction in regression. Statistical Science, 22(1), 1–26.

[3] Zhong, W., Zeng, P., Ma, P., Liu, J.S. and Zhu, Y. (2005). RSIR : Regularized Sliced Inverse Regression for motif discovery. Bioinformatics, 21(22).

[4] Fan, Jianqing, Xue, Lingzhou, Yao, Jiawei, 2017. Sufficient forecasting using factor models, Journal of Econometrics201(2), 292-306.

[5] Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2000), The generalized dynamic-factor model: identification and estimation, The Review of Economics and Statistics 82(4), 540-554.

[6] Lam, C., Yao, Q. et al. (2012), Factor modeling for high-dimensional time series: inference for the number of factors, The Annals of Statistics 40(2), 694-726.

[7] Ludvigson, S. and Ng, S. (2007), The empirical risk return relation: A factor analysis approach, Journal of Financial Economics 83(1), 171-222.

[8] Schott, J. R. (1994), Determining the dimensionality in sliced inverse regression, Journal of the American Statistical Association 89(3), 141-148.

2、R 代码：

程序问题请联系：baozan@stocke.com.cn;

```
#------------------
mat_power = function(a, alpha){
                    a = round((a + t(a))/2,7); tmp = eigen(a)
                    return(tmp$vectors%*%diag((tmp$values)^alpha)%*%t(tmp$vectors))
}


discretize=function(y,h){
                    n=length(y);m=floor(n/h)
                    y=y+.00001*mean(y)*rnorm(n)
                    yord = y[order(y)]
```

```
                divpt=numeric();for(i in 1:(h-1)) divpt = c(divpt,yord[i*m+1])
                y1=rep(0,n);y1[y<divpt[1]]=1;y1[y>=divpt[h-1]]=h
                for(i in 2:(h-1)) y1[(y>=divpt[i-1])&(y<divpt[i])]=i
                return(y1)
}


sir=function(x,y,h,r){
                 p=ncol(x);n=nrow(x)
                 signrt=mat_power(var(x),-1/2)
                 xc=t(t(x)-apply(x,2,mean))
                 xst=xc%*%signrt
                 ydis=discretize(y,h)
                 yless=ydis;ylabel=numeric()
                 for(i in 1:n) {if(var(yless)!=0) {ylabel=
                 c(ylabel,yless[1]);yless=yless[yless!=yless[1]]}}
                 ylabel=c(ylabel,yless[1])
                 prob=numeric();exy=numeric()
                 for(i in 1:h) prob=c(prob,length(ydis[ydis==ylabel[i]])/n)
                 for(i in 1:h) exy=rbind(exy,apply(xst[ydis==ylabel[i],],2,mean))
                 sirmat=t(exy)%*%diag(prob)%*%exy
                 return(signrt%*%eigen(sirmat)$vectors[,1:r])
}
```

股票投资评级说明

以报告日后的 6 个月内，证券相对于沪深 300 指数的涨跌幅为标准，定义如下：
1、买入 ：相对于沪深 300 指数表现 ＋20％以上；
2、增持 ：相对于沪深 300 指数表现 ＋10％～＋20％；
3、中性 ：相对于沪深 300 指数表现－10％～＋10％之间波动；
4、减持 ：相对于沪深 300 指数表现－10％以下。

行业的投资评级：

以报告日后的 6 个月内，行业指数相对于沪深 300 指数的涨跌幅为标准，定义如下：
1、看好 ：行业指数相对于沪深 300 指数表现＋10%以上；
2、中性 ：行业指数相对于沪深 300 指数表现－10%～＋10%以上；
3、看淡 ：行业指数相对于沪深 300 指数表现－10%以下。
我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重。
建议：投资者买入或者卖出证券的决定取决于个人的实际情况，比如当前的持仓结构以及其他需要考虑的因素。投资者不应仅仅依靠投资评级来推断结论

法律声明及风险提示

浙商证券研究所

上海市浦东南路 1111 号新世纪办公中心 16 层
邮政编码：200120
电话：(8621)80108518
传真：(8621)80106010
浙商证券研究所：http://research.stocke.com.cn