金融工程

证券研究报告
2017 年 11 月 16 日

协方差矩阵的常用估计和评价方法

作者

吴先兴 分析师
SAC 执业证书编号：S1110516120001
wuxianxing@tfzq.com
18616029821

张欣慰 分析师
SAC 执业证书编号：S1110517010003
zhangxinwei@tfzq.com

韩谨阳 联系人
hanjinyang@tfzq.com

相关报告

1 《金融工程：因子正交全攻略——理论、框架与实践》2017-10-30
2 《金融工程：基于动态风险控制的组合优化模型》2017-09-21
3 《金融工程：专题报告-MHKQ 因子择时模型在 A 股中的应用》2017-08-15
4 《金融工程：专题报告-利用组合优化构建投资组合》2017-08-14
5 《金融工程：专题报告-半衰 IC 加权在多因子选股中的应用》2017-07-22
6 《金融工程：专题报告-反转现象的选择性交易策略》2017-05-31

协方差矩阵的估计在量化投资中有广泛的应用，许多量化策略都会使用到协方差矩阵（或其逆矩阵）。例如多因子选股中，最大化复合因子 IC_IR 的加权方式会用到因子 IC 序列协方差矩阵的逆；在组合优化中，需要使用股票收益率序列的协方差矩阵来估计组合未来的波动率。工欲善其事，必先利其器，如何准确估计协方差矩阵是一个很重要的议题。本文梳理了常用的协方差估计方法，然后在一套新的评价体系下比较了不同协方差的估计效果。

常见的改进股票收益率协方差估计的方法有因子模型、压缩估计、随机矩阵理论模型等。

因子模型

可以通过给协方差矩阵加以一定的结构，从而减少数据的维数，降低估计误差。这种结构可以来源于因子模型，如单一指数模型（市场模型）、多因子模型（行业因子、宏观因子、基本面因子、统计因子）。

压缩估计

为了避免因子模型中的因子选择问题，可以将样本协方差矩阵与其他结构化模型进行加权，以此来设定结构。

随机矩阵理论模型

可以根据随机矩阵理论分离样本协方差矩阵中的信息与噪声，通过调整相关系数矩阵的特征根来降低协方差矩阵的估计误差。

固定相关系数模型

固定相关系数模型假设股票之间的相关系数是相同的，并以所有股票间的样本相关系数的平均值作为固定相关系数的估计。

其他模型

其他协方差矩阵的估计方法还有固定相关系数模型、时变模型、合成聚类模型等。

协方差矩阵估计量的评价指标

本文采用了一种基于协方差矩阵特征分解的评价指标，通过比较与两个协方差矩阵相关的组合风险能够达到的最大差别，来比较不同的协方差矩阵。该评价指标具有明确的经济意义及统计意义。

应用与实证

实证结果表明，常用的 Ledoit & Wolf 压缩估计及包含风格及行业因子的多因子模型的表现较好，以多因子模型为压缩目标的压缩估计量表现最好。然而，使用多因子模型作为压缩目标既没有回避因子选取的问题，也没有简化运算，无法体现出压缩估计的优势。基于此，在实务中，本文较为推荐以固定相关系数或者对角阵为压缩目标的 Ledoit & Wolf 压缩估计，以及包含风格及行业因子的多因子模型。

风险提示：市场环境变动风险，模型失效风险。

内容目录

1. 样本协方差 4
1.1. 样本协方差存在的问题 4
1.2. 协方差估计的改进思路 5
2. 因子模型 6
2.1. 市场模型 6
2.1.1. 无调整 Beta 模型 6
2.1.2. Blume 调整 Beta 模型 6
2.1.3. Vasicek 调整 Beta 模型 6
2.2. 多因子模型 7
2.2.1. 基本形式与估计方法 7
2.2.2. 时间序列回归 7
2.2.3. 横截面回归 8
2.2.4. 主成分模型 9
3. 压缩估计 9
3.1. Ledoit & Wolf 压缩估计 10
3.2. 估计量组合 11
3.3. 小结 11
4. 随机矩阵理论模型 12
5. 其他方法 12
5.1. 固定相关系数模型 12
5.2. 时变模型 13
5.3. 合成聚类模型 13
6. 协方差矩阵估计量的评价指标 13
6.1. 统计类指标 13
6.2. 经济类指标 14
6.3. 基于特征距离的协方差矩阵估计量评价方法 14
6.3.1. 特征距离的定义与含义 14
6.3.2. 特征距离的计算 15
7. 应用与实证 16
7.1. 协方差估计量列表 16
7.2. 实证结果分析 17
8. 协方差矩阵估计的一些补充说明 18
8.1. 含有缺失值的样本协方差 18
8.2. N/T比值对协方差估计量相对表现的影响 19
9. 参考文献 19
风险提示 20

图表目录

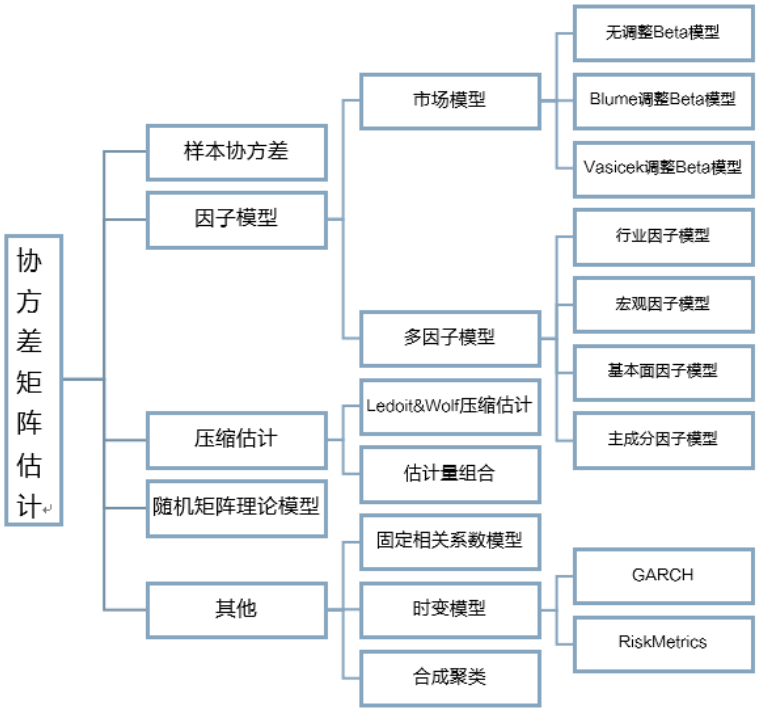
图 1：协方差矩阵估计方法梳理 4
图 2：压缩估计量几何图示 11
图 3：协方差估计量与下期实际协方差矩阵的特征距离 17
图 4：协方差矩阵估计量的成对比较 18

表 1：协方差矩阵估计量说明 16
表 2：风格因子列表 16
表 3：协方差估计量与下期实际协方差之间特征距离的均值 17

协方差矩阵的估计在量化投资中有广泛的应用，许多量化策略都会使用到协方差矩阵（或其逆矩阵）。例如多因子选股中，最大化复合因子 *IC_IR* 的加权方式会用到因子 *IC* 序列协方差矩阵的逆；在组合优化中，需要使用股票收益率序列的协方差矩阵来估计组合未来的波动率。样本协方差是一种简单常用的估计方法。然而，由于可取样本的限制，资产的数量通常会大于样本数量，而在这种情况下，样本协方差矩阵是不可逆的，并且存在较大的估计误差。工欲善其事，必先利其器，如何准确估计协方差矩阵是一个很重要的议题。本文梳理了常用的协方差估计方法，然后在一套新的评价体系下比较了不同协方差的估计效果，供投资者参考。

本文首先梳理了常见的协方差矩阵估计方法（如图 1 所示），包括其假设、原理、计算方法、优势与缺陷等。

图 1：协方差矩阵估计方法梳理



资料来源：天风证券研究所

其次，本文为协方差估计量的比较确定了一个相对合理的标准。常用的评价标准如均方根误差、最小方差组合等都存在一定的缺陷，不能较好地反应协方差矩阵所蕴含的关于组合风险的信息，因此本文采用了一种基于组合风险的度量方法（特征距离），来比较各种协方差矩阵估计量的好坏。

下面首先对各种协方差估计方法进行介绍。为了便于表示，本文统一以矩阵X（$N \times T$）表示N个变量的T个样本（即N个资产T期的收益率序列）。

1. 样本协方差

1.1. 样本协方差存在的问题

在收益率服从正态分布的假设下，样本协方差是无偏的极大似然估计量，即给定数据下最可能的参数，也就是说“完全让数据说话”。在估计参数时，如果样本数量足够大，那么样本协方差具有良好的性质。而在小样本下，使用该估计量可能会出现过拟合。

样本协方差矩阵为

$$S = \frac{1}{T} X \left(I - \frac{1}{T} \mathbf{1}\mathbf{1}' \right) X'$$

其中$\mathbf{1}$为元素全为 1 的列向量（$N \times 1$），I为单位矩阵（$N \times N$）。从样本协方差的计算公式可以看出，样本协方差S的秩最多等于矩阵$\mathrm{I} - \mathbf{1}\mathbf{1}'/T$的秩，即$T-1$。因此，当矩阵的维数$N$超过$T-1$时，样本协方差矩阵是不满秩的，也是不可逆的。

此外，在样本协方差矩阵中，需要估计$N(N+1)/2$个元素，而总样本量为$N \times T$。在实际应用中，由于收益序列的非平稳性，通常不会取较长的时间区间；而待求解的股票集合往往很大。因而，当股票数量的数量级与样本数量相当甚至更大时，样本数量的缺少给样本协方差带来较大的估计误差。

1.2. 协方差估计的改进思路

从以上分析可以看到，由于样本数量的限制，样本协方差矩阵具有不可逆以及估计误差较大两个主要的缺陷。

面对这个问题，一个很直接的解决方法是增加样本容量。这两个问题都与样本数量少于股票数量有关，因此可以考虑通过使用较高频或者较长时间区间的数据。然而，这种方法在实际投资中的作用是有限的。

首先，使用较高频的数据可能会带一些与市场微观结构相关的问题；此外，由于组合的调整频率通常较低，使用高频数据获得的协方差矩阵并不适宜作为组合管理的输入。

其次，通过增加时间区间而获得更多的样本也可能存在一些问题。较长时间之前的样本包含的信息可能具有一定的滞后性，当股票收益率序列非平稳时，使用滞后信息估计协方差矩阵会造成较大的偏差。

因而，单纯地试图通过增加样本容量来改善样本协方差是不够的，要想获得更合适的协方差估计，应当从估计方法的改善上入手。

应对估计误差常见的方法是在在估计过程中使用一定的结构化模型，但是这在减小估计误差的同时，也会引入设定偏差。与其他估计量一样，协方差矩阵的估计也不可避免地包含误差，然而估计误差与设定偏差是有区别的。估计误差通常出现在样本数量与待估参数相比数量不够大时，而设定偏差指对模型结构进行假设时存在的偏差。在估计误差与设定偏差之间存在一定的权衡，而我们希望在减少协方差矩阵估计误差的同时，不要引入太多的设定偏差。

在实务中常见的改进股票收益率协方差估计的方法有因子模型、压缩估计、随机矩阵理论模型等。

（1） 因子模型

前文提到，可以通过给协方差矩阵加以一定的结构，从而减少数据的维数，降低估计误差。这种结构可以来源于因子模型，如单一指数模型（市场模型）、多因子模型（行业因子、宏观因子、基本面因子、统计因子）。然而，因子模型的缺陷在于，关于模型中应当包含几个因子、包含哪些因子并没有统一的标准。因而，并不能提前知晓在特定环境下应该使用什么模型，这就使得因子模型的设置往往具有一定的“艺术性”。

（2） 压缩估计

为了避免因子模型中的因子选择问题，可以将样本协方差矩阵与其他结构化模型进行加权，以此来设定结构。如 *Ledoit & Wolf*（2003）压缩估计，估计量组合等。

（3） 随机矩阵理论模型

除了在估计中引入结构化，也可以根据随机矩阵理论来分离样本协方差矩阵中的信息与噪声，即通过调整相关系数矩阵的特征根来降低协方差矩阵的估计误差。

下面就详细介绍这几类模型的估计方法。

2. 因子模型

因子模型认为有一些公共的因子驱动了资产的收益率，从而使得资产的收益率之间产生联系。因子模型能够减少协方差矩阵估计的维度。常见的有单因子模型（市场模型）和多因子模型。

2.1. 市场模型

2.1.1. 无调整 Beta 模型

市场模型认为股票收益率仅受到市场收益率的影响：

$$x_{it} = \alpha_i + \beta_i x_{mt} + \varepsilon_{it}$$

其中x_{mt}为市场收益率，残差收益率ε_{it}与市场收益率不相关，并且不同股票的残差收益率不相关。因此，在该模型下，收益率协方差矩阵可以表示为：

$$\Sigma = \beta \Omega_m \beta' + E$$

其中，β（$N \times 1$）为股票 *Beta* 向量，Ω_m为市场收益率的方差，E（$N \times N$）是对角线元素为特质波动率的对角阵。股票收益率对市场收益率进行时间序列回归即可估计得到该股票的β_i。

可以看到，使用市场模型估计协方差矩阵时，最关键的是对 *Beta* 的估计。可以直接将股票收益率对市场收益率的回归系数作为 *Beta*，即无调整 *Beta* 模型；也可以进一步对 *Beta* 的估计值进行调整，如 *Blume*（1971）调整和 *Vasicek*（1973）贝叶斯调整。

2.1.2. Blume 调整 Beta 模型

Blume（1971）调整将历史上相邻两期的 *Beta* 之间的关系外推，从而形成对当期 *Beta* 的估计。具体地，设当前时间为t期，将个股$t-1$期无调整的 *Beta*（β_{t-1}）对$t-2$期无调整的 *Beta*（β_{t-2}）进行回归，即

$$\beta_{t-1,i} = \alpha + b\beta_{t-2,i} + \varepsilon_i$$

其中i为样本中的股票i。

假设连续两期 *Beta* 之间的关系是稳定的，则可根据$t-1$期无调整的 *Beta* 估计t期的 *Beta*，即将β_{t-1}作为自变量代入上式，得到 *Blume* 调整 *Beta*，即

$$\beta_{Blume,i} = \alpha + b\beta_{t-1,i}$$

可以看到，*Blume*（1971）调整 *Beta* 相当于对个股 *Beta* 值与 1 加权平均，该权重根据历史 *Beta* 进行估计，很多机构会直接确定一个固定的权重。这样的调整方法利用了 *Beta* 系数均值回归的性质。

2.1.3. Vasicek 调整 Beta 模型

Blume（1971）对所有股票的 *Beta* 都进行了相同的调整，而 *Vasicek*（1973）根据个股 *Beta* 估计的不确定性对个股 *Beta* 进行调整。如果一只股票 *Beta* 的估计误差较大，那么其有较大的可能会偏离均值较远，因此应对其进行较大的调整。具体地，记$t-1$期个股 *Beta* 的横截面均值为$\bar{\beta}$，则t期股票i的 *Vasicek*（1973）调整 *Beta* 为横截面均值与股票i在$t-1$期 Beta 的加权平均，即

$$\beta_{Vasicek,i} = \frac{\sigma_{\beta_i}^2}{\sigma_{\beta}^2 + \sigma_{\beta_i}^2}\bar{\beta} + \frac{\sigma_{\beta}^2}{\sigma_{\beta}^2 + \sigma_{\beta_i}^2}\beta_i$$

其中，σ_{β}^2为$t-1$期个股 *Beta* 的横截面方差。$\sigma_{\beta_i}^2$为$t-1$期使用时间序列回归估计股票i的 *Beta* 的标准差平方，该值衡量了个股 *Beta* 估计的不确定性。可以看到，当 *Beta* 估计的不

确定性即$\sigma_{\beta_i}^2$越高时，*Vasicek*（1973）调整 *Beta* 在 *Beta* 横截面均值上的权重越大，而在历史 *Beta* 上的权重越小，即向均值调整的幅度越高。

2.2. 多因子模型

因子模型都包含一定的结构化假设，因子越少，结构越强。市场模型是建立在资本资产定价模型（*Capital Asset Pricing Model, CAPM*）的基础上，模型中仅包含市场因子，这种较强的假设可能引入较大的模型设定偏差。因此，在使用因子模型估计协方差矩阵时，实际上同时面临着估计误差和设定偏差两种误差。在结构化模型中增加因子可能减少模型的设定偏差，而根据套利定价模型（*Arbitrage Pricing Theory, APT*）建立的多因子模型就是一个选择。

多因子模型试图捕捉除了市场收益外的其他引起股票收益率相互联系的因素，如行业因子、宏观因子、基本面因子、统计因子（主成分因子）等。

2.2.1. 基本形式与估计方法

多因子模型认为股票收益率与多个因子之间存在线性关系，即

$$X = \alpha + \beta f + \mu$$

其中，X为股票收益率向量，β为N只股票在K个因子上的因子暴露矩阵（$N \times K$），f为K个因子的因子收益向量，μ为股票特质收益率向量。

同样地，当股票特质收益率与公共因子不相关时，预期风险可以分解为公共因子解释的风险以及特质风险，即

$$\Sigma = \beta \Omega_f \beta' + E$$

其中，Ω_f为K个因子的因子收益协方差矩阵（$K \times K$），β为股票在K个因子上的因子暴露矩阵（$N \times K$），E为股票的特质波动率对角阵（$N \times N$）。

使用多因子模型估计协方差矩阵需要估计因子暴露以及因子收益。在学术以及实务中，估计因子暴露及因子收益主要有两种方法，一种是 *Fama & French*（1993）的时间序列回归，通过个股收益率序列对因子收益回归估计因子暴露；一种是横截面回归，在每一期通过个股收益率对因子暴露回归，估计因子收益。

在时间序列回归的估计窗口内，因子暴露是固定的，而因子收益是变化的。因此，这样估计得到的因子暴露对于市场的变化反应较慢。而在横截面回归中，因子暴露会根据公司特征的变化而及时变化。对于基本面因子模型，两种方法都具有一定的适用性，时间序列回归方法在学术论文中比较常见，而商业中使用较多的是横截面回归，如 *Barra* 风险模型。对于宏观因子模型，时间序列回归方法更加适用。

虽然多因子模型是一种常用的降维方法，但是关于因子的内涵以及模型中因子的数目尚未形成共识。因此，难以确定应当使用的多因子模型。虽然增加因子的数量能够提高模型样本内的解释力度，但是会降低其对样本外协方差的预测能力。这使得多因子模型的建立与使用具有很大的“艺术性”。

下面具体介绍几种多因子模型的形式以及协方差矩阵的估计方法。

2.2.2. 时间序列回归

下面首先介绍时间序列回归估计行业因子模型的方法。

同一行业的股票收益率更容易受到相同事件的影响，因此会更加相关。行业因子模型认为个股收益率受到市场收益率和所属行业收益率的驱动，而行业收益率与市场收益率及其他行业收益率不相关。为了避免回归模型中的多重共线性，可使用两步法估计行业因子模型。首先，各行业收益对市场收益进行回归，取其残差作为经市场收益调整的行业收益，即

$$R_{gt} = a_g + b_g R_{mt} + Res_{gt}$$

其中，R_{gt}为行业g的收益率（$g = 1,2 \ldots M$，M为行业数量），R_{mt}为市场收益率，Res_{gt}为行业残差收益率，其与市场收益率不相关。

其次，将股票收益率对市场收益率以及行业残差收益率进行回归，即

$$R_{it} = \alpha_i + \beta_i R_{mt} + \gamma_i Res_{gt} + \varepsilon_{it}$$

其中，R_{it}为个股收益率，Res_{gt}为股票i所属行业g的行业残差收益率，ε_{it}为不能被市场收益和行业收益解释的个股残差收益率。回归系数γ_i是股票i对其所属行业的暴露。

因此，协方差矩阵可以通过下式估计：

$$\Sigma = \beta \Omega_m \beta' + \gamma \Omega_{ind} \gamma' + E$$

其中，β是元素为β_i的向量（$N \times 1$），Ω_m为市场收益率方差，γ为股票行业暴露矩阵（$N \times M$），Ω_{ind}为行业因子收益方差对角阵，E为特质波动率对角阵（$N \times N$）。由于假设各行业收益不相关，因此Ω_{ind}中非对角线元素为 0。在行业暴露矩阵γ中，当股票i不属于行业j时，元素γ_{ij}为 0；当股票i属于行业j时，γ_{ij}为第二步回归系数中的γ_i。

可以看到，在时间序列回归中，首先需要确定因子收益。对于行业因子，可以使用行业指数。而对于风格因子等则可借鉴 *Fama & French*（1993）的分组方法。

2.2.3. 横截面回归

横截面回归在每一期使用个股收益率对股票因子暴露进行横截面回归，从而得到因子收益序列。*Barra* 风险模型使用了横截面回归的方法。在本报告中，使用横截面回归估计了两种多因子模型，分别为行业因子模型以及风格行业因子模型，其中风格行业因子模型中包含风格因子与行业因子，而行业因子模型中仅包含行业因子。下面以包含风格因子及行业因子的多因子模型为例，介绍以横截面回归估计因子收益的步骤。此处借鉴了 *Barra* 风险模型的方法，具体可参考 *Barra USE4（The Barra US Equity Model）*、*CNE5(The Barra China Equity Model)*。

（1）因子收益的估计

股票收益率可以表示为市场收益率、行业收益率、风格因子收益率以及特质收益率的线性组合：

$$r_n = f_m + \Sigma_i X_i^I f_i^{\mathrm{I}} + \Sigma_i X_i^S f_i^{\mathrm{S}} + u_n$$

其中 r_n 为股票收益率，f_m为市场收益，f_i^{I}为行业i的因子收益，f_i^{S}为风格i的因子收益，X_i^I、X_i^S分别为股票n对行业i、风格i的因子暴露，u_n为特质收益率。

由于股票收益率存在异方差性，因此以根号市值作为权重，使用加权最小二乘法（*Weighted Least Squares, WLS*）估计以上模型。使用这种加权方法是由于很多研究发现个股的特质风险与股票规模成反比。

此外，由于行业因子暴露使用哑变量表示全部行业，加入截距项 f_m 使得模型中存在共线性。可通过增加以下约束条件使得该模型具有唯一解

$$\Sigma_i w_i f_i^{\mathrm{I}} = 0$$

即市值加权的行业因子平均收益为 0，其中w_i为行业i的市值权重。

在该约束条件下，截距项 f_m 的含义就更加明确了，f_m 实际上代表了全市场市值加权收益率；而行业因子的回归系数代表了行业的超额收益，风格因子的回归系数代表了在控制行业因素的影响后，风格因子的超额收益。

（2）因子收益协方差矩阵的估计

因子收益加权协方差可以直接根据因子收益序列计算得到，即

$$\sigma_{i,j}^2 = \Sigma_t w_t (f_{i,t} - \bar{f}_i)(f_{j,t} - \bar{f}_j)$$

其中，$\sigma_{i,j}$ 为因子i和因子j的半衰加权协方差，w_t为半衰权重，$\bar{f}_i$为因子i半衰加权的因子收益均值。

然而，考虑到因子收益之间的相关系数比因子波动更加稳定，本报告参考 *Barra USE4*、*CNE5* 的方法，先分别估计因子收益的相关系数以及各因子的波动性，然后再据此计算因子协方差矩阵。具体地，因子收益协方差可由因子收益波动率及因子收益相关系数计算得到，即

$$\sigma_{i,j}^2 = \rho_{i,j}\sigma_i\sigma_j$$

其中，σ_i, σ_j分别为因子i, j的因子收益标准差，$\rho_{i,j}$为因子i, j的因子收益相关系数。

这种估计方法的优点在于，可以对因子收益相关系数以及因子波动使用不同的半衰期进行加权计算。由于因子相关系数较为稳定，可以选择较长的半衰期；而因子波动变化较大，可以选择较短的半衰期，从而更加迅速地反应因子风险的变化。当因子相关系数与因子波动的半衰期相同时，通过这样的方法得到的因子收益协方差矩阵与直接使用因子收益计算的加权协方差矩阵是一样的。

（3）特质波动率的估计

特质风险为特质收益率的方差，即

$$\sigma_{u,i}^2 = \Sigma_t w_t \left(f_{u,i,t} - \bar{f}_{u,i}\right)^2$$

其中，$\sigma_{u,i}^2$ 为股票i特质收益率的方差，w_t为半衰权重，$f_{u,i,t}$为股票i在t期的特质收益率，$\bar{f}_{u,i}$为股票i特质收益率的半衰加权均值。根据时间序列估计的特质波动率在样本外不一定具有持续性，尤其是当特质波动率特别高或者特别低时，特质风险存在均值回复的可能性，因此需要对特质风险的估计值进行调整。本报告使用贝叶斯收缩（*Bayesian Shrinkage*）的方法对特质风险进行调整，具体可参见 *Barra USE4*、*CNE5*。

2.2.4. 主成分模型

前面提到的因子，不论是市场因子、行业因子，还是风格因子都有明确的含义。然而，用什么因子、用哪些因子并没有一个统一的结论。为了避免因子选取带来的误差，可以考虑通过因子分解的方法，从样本的协方差矩阵中去发现隐含的因子。主成分分析就是一种常见的方法。主成分因子的含义并不明确，但是其优势在于，主成分因子之间相互正交，并且可能找到一些通常难以发现的因子。具体模型如下

$$R_{it} = \alpha_i + \Sigma_{j=1}^{P} \beta_{ji} R_{jt} + \varepsilon_{it}$$

其中R_{jt}为选取的主成分，P为选取主成分的数量。则该模型的协方差矩阵可表示为

$$\Sigma = \beta \Omega_{pc} \beta' + E$$

其中，β为载荷矩阵（$N \times P$），Ω_{pc}是对角线元素为各主成分方差的对角矩阵（$P \times P$）（由于主成分之间相互正交，因此主成分协方差矩阵的非对角线元素为 0），E为特质波动率对角阵（$N \times N$）。

3. 压缩估计

压缩估计的方法来源于 *Stein*（1955）。*Stein* 估计量实际上是贝叶斯估计，即将统计量向先验信息压缩。一种简单的方法就是对估计量和先验估计量以恰当的权重取平均，而给先验估计量的权重就是压缩强度。

从统计的角度来看，在协方差矩阵估计中，样本协方差是完全基于数据的，而先验的协方差矩阵可以来源于主观判断、历史经验或者模型等。样本协方差矩阵是无偏的，但是含有大量的估计误差；而先验的协方差矩阵由于其具有较严格的假设，是有设定偏差的，但是因为待估计的参数较少，其具有较少的估计误差。因而，可以将样本协方差向一个先验的

协方差矩阵压缩，以减少样本协方差的估计误差，从而在设定偏差与估计误差间达到一个最优的平衡。

从金融的角度来看，压缩估计量也可以看作两个极端的折中。样本协方差矩阵可以认为是一个N-因子模型，即每只股票都是一个因子，并且该模型中没有残差；而压缩目标是一个简单的模型。

因而，协方差矩阵的压缩估计量可以表示为样本协方差矩阵与压缩目标的线性组合，即

$$\Sigma(\alpha) = \alpha F + (1-\alpha)S$$

其中，F为压缩目标，S为样本协方差矩阵，α为压缩强度。

3.1. Ledoit & Wolf 压缩估计

压缩估计量的关键在于压缩强度的确定，可以通过一定的损失函数来估计压缩强度。*Ledoit & Wolf*（2003）使用协方差压缩估计量与真实协方差矩阵之间的距离作为损失函数，该距离通过弗罗贝尼乌斯范数（*Frobenius norm*）来衡量，即

$$\mathrm{L}(\alpha) = \| \alpha F + (1-\alpha)S - \Sigma \|^2$$

其中Σ为真实协方差矩阵。

则最优的压缩强度α可通过求解下式的最小值得到：

$$\mathrm{R}(\alpha) = \mathrm{E}\big(\mathrm{L}(\alpha)\big) = \Sigma_{i=1}^{N}\Sigma_{j=1}^{N}\mathrm{E}\big(\alpha f_{ij} + (1-\alpha)s_{ij} - \sigma_{ij}\big)^2$$

其中f_{ij}、s_{ij}、σ_{ij}分别为F、S、Σ的元素。

Ledoit & Wolf（2003）推导了最优压缩强度α的一致估计量，具体过程本文不再赘述。在此，仅对该压缩强度的涵义进行介绍。*Lediot & Wolf*（2003）提出最优压缩强度为（取最大值、最小值是为了保证压缩强度在 0 到 1 之间）：

$$\alpha = \max\{0, \min\{\frac{\hat{\kappa}}{T}, 1\}\}$$

其中，T为样本数量，

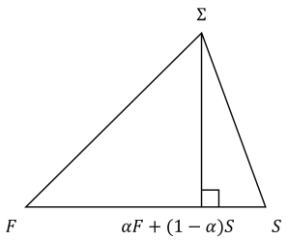
$$\hat{\kappa} = \frac{\hat{\pi} - \hat{\rho}}{\hat{\gamma}}$$

$\hat{\pi}$、$\hat{\rho}$、$\hat{\gamma}$分别为π、ρ、γ的一致估计量。π为样本协方差矩阵元素的渐进方差之和，ρ为样本协方差矩阵元素与压缩目标元素的渐进协方差之和，γ衡量压缩目标的设定偏差。对于$\hat{\pi}$、$\hat{\rho}$、$\hat{\gamma}$这三个变量具体的计算公式可参考 *Lediot & Wolf*（2003）、*Lediot & Wolf*（2004）。

由上式可以看出，压缩目标的权重随着样本协方差矩阵误差的增加而增加（通过$\hat{\pi}$），随着压缩目标设定偏差的增加而减少（通过$\hat{\gamma}$）。参数$\hat{\rho}$反应了样本协方差估计误差与压缩目标估计误差的相关性。贝叶斯估计通常要求先验信息与样本信息相互独立，而在对协方差的压缩估计中，会使用相同的数据。因而 *Lediot & Wolf*（2003）在压缩强度的估计中直接考虑了先验信息与样本信息的相关性。

也可以从几何角度来理解最优压缩强度：压缩估计量是真实协方差矩阵在样本协方差矩阵与压缩目标连线上的正交投影，如图 2 所示。

图 2：压缩估计量几何图示



资料来源：Ledoit & Wolf(2003)，天风证券研究所

关于压缩目标的选取，*Lediot & Wolf*（2003, 2004, 2004）使用了三种压缩目标，包括市场模型、单位矩阵及固定相关系数模型。理论上来讲，压缩目标的选取是不受限制的，因为在其推导过程没有用到关于压缩目标结构的假设，仅需满足压缩目标为有偏估计及其他较弱的假设。

3.2. 估计量组合

与 *Lediot & Wolf* 压缩估计使用损失函数确定压缩强度的方法不同，估计量组合（*Portfolio of estimators*）对样本协方差及其他协方差估计量取等权平均。这是因为当对不同协方差估计量的估计误差知之甚少时，等权加权是一种相对保险的做法。估计量的等权加权避免了 *Ledoit & Wolf* 压缩估计量中求解最优压缩强度的问题，使用起来更加简便。

和 *Ledoit & Wolf* 压缩估计量一样，这种方法也是在估计误差与设定偏差之间取得平衡，认为其他协方差矩阵主要包含设定偏差，因此与含有估计误差的样本协方差平均后，可以提高估计量的整体效果。

如 *Jagannathan and Ma*（2000）使用三种协方差矩阵的等权平均，分别为市场模型、样本协方差以及方差对角阵：

$$\Sigma = \frac{1}{3}F + \frac{1}{3}S + \frac{1}{3}D$$

其中F为市场模型，S为样本协方差，D为对角线元素为样本方差的对角阵，即

$$\mathrm{D} = \begin{bmatrix} s_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & s_{NN} \end{bmatrix}$$

可以看到，这个模型中包含了单因子模型（F）、N因子模型（S）及 0 因子模型（D）。

估计量组合实际认为，根据不同假设建立的协方差估计量会包含不同方向的误差。因而，将不同的估计量结合起来，不同方向的误差相互抵消，能够获得总体表现更好的估计量。按照这个思路，待合成的估计量以及加权方式可以有多种选择。

例如，可以将压缩估计以及随机矩阵理论（见下文）结合起来。使用包含K个因子的主成分模型估计协方差矩阵P_k，将其作为压缩目标，而其中主成分因子数目K则根据随机矩阵理论来确定。即

$$\Sigma = \alpha P_k + (1-\alpha)S$$

其中，S为样本协方差，最优压缩强度α可根据 *Lediot & Wolf*（2003）的方法来估计。

3.3. 小结

协方差压缩估计量实际上通过一种不同于因子模型的方式来增加结构化。因子模型的建立需要选取因子、计算因子、不断维护与更新，因而通过因子模型估计协方差矩阵的成本是较高的。此外，一些商业机构如 *APT*、*Barra* 开发的风险模型是保密并且需要付费使用的，因此外部的使用者难以知晓其具体的建模过程。而压缩估计量使用简便，能够避免因子模

型存在的一些问题。

4. 随机矩阵理论模型

随机矩阵理论（*Random matrix theory*）模型提供了一种去除样本协方差矩阵噪音的方法。根据随机矩阵理论，当股票数量相对于样本数量较大时，位于一定范围内的协方差矩阵特征根与完全随机的收益序列的协方差矩阵特征根相近。因此，通过对该特定范围内的特征根进行修正，可以提高样本协方差矩阵所包含的信息量。

设X是元素独立同分布的随机矩阵（$N \times T$），其样本相关系数矩阵为C。

根据随机矩阵理论，在$\frac{T}{N} \to Q$时，当N和T趋近于无穷时，矩阵C的特征根分布具有如下密度函数：

$$p(\lambda) = \frac{Q}{2\pi\lambda}\sqrt{(\lambda_{max} - \lambda)(\lambda - \lambda_{min})}，\ \lambda_{min} < \lambda < \lambda_{min}$$

其中，$\lambda_{max} = \left(1 + \sqrt{Q^{-1}}\right)^2$，$\lambda_{min} = \left(1 - \sqrt{Q^{-1}}\right)^2$

也就是说，对于一个完全随机的矩阵，其特征根服从以上分布，并且具有以上阈值。如果把样本协方差矩阵的特征根分布与以上分布相比，两者重合的特征根则认为其为噪声；而样本协方差矩阵特征根中与以上分布不重合的，即大于λ_{max}的特征根，则认为其含有信息。因此，可以通过调整相关系数矩阵的特征根，过滤掉其中的噪声。

具体的调整方法为，对于样本相关系数矩阵的特征根Λ，保留大于λ_{max}的特征根，而将小于λ_{max}的特征根替换为所有小于λ_{max}的特征根的平均值，这样可以保证相关系数矩阵的迹不变。根据特征分解，

$$C = \Sigma_{i=1}^{N}\lambda_i\xi_i\xi_i'$$

其中，$\lambda_1 > \lambda_2 \ldots > \lambda_N$为$C$的特征值，$\{\xi_i\}$为相应的特征向量。假设有$k$个特征值大于$\lambda_{max}$，则一种调整相关系数矩阵的方式为：

$$\bar{C} = \Sigma_{i=1}^{k}\lambda_i\xi_i\xi_i' + \alpha I$$

其中I为对角矩阵（$N \times N$），α为使得$\bar{C}$与C的迹相等的常数，即

$$\alpha = \frac{\lambda_{k+1} + \cdots + \lambda_N}{N}$$

则调整后的协方差矩阵为

$$\Sigma = D^{1/2}\bar{C}D^{1/2}$$

其中D为对角线元素为样本方差的对角矩阵。（需要注意的是，虽然$\bar{C}$为半正定矩阵，但是其并不是一个相关系数矩阵，因为其对角线元素不一定为 1。而Σ为半正定矩阵，是协方差矩阵。）可以看到，在对原始相关系数矩阵进行调整时，$\sum_{i=1+1}^{N}\lambda_i\xi_i\xi_i'$被对角矩阵替代了，因此这种减少误差的方式实质上是将原相关系数矩阵的非对角元素向 0 压缩。

其他运用随机理论模型减少相关系数矩阵误差的方法，具体可参考 *Bai & Shi*（2011）。

5. 其他方法

本节介绍一些其他的协方差估计方法，包括固定相关系数模型、时变模型、合成聚类模型。

5.1. 固定相关系数模型

固定相关系数模型（*Constant correlation model*）假设股票之间的相关系数是相同的，并以所有股票间的样本相关系数的平均值作为固定相关系数的估计。平均相关系数为

$$\bar{\rho} = \frac{2}{N(N-1)} \Sigma_{i=1}^{N-1} \Sigma_{j=i+1}^{N} \rho_{ij}$$

则固定相关系数模型可根据样本方差和平均相关系数得到，其元素为

$$f_{ii} = s_{ii}, \quad f_{ij} = \bar{\rho}\sqrt{s_{ii}s_{jj}} \quad (i \neq j)$$

其中s_{ii}为样本方差。

5.2. 时变模型

由于市场环境是不断变化的，因此收益率的协方差结构也会不断变化。传统的刻画协方差矩阵时变特征的模型如 *GARCH* 类模型等，需要估计的参数比较多，不适用于估计维数较高的股票协方差矩阵。因此，通常协方差模型中较少考虑协方差矩阵的时变性。*RiskMetrics* 的指数加权移动平均值（*Exponentially weighted moving average, EWMA*）模型在估计方差及协方差时，给近期的股票收益率更多的权重，从而更加及时地反应市场的变化。该模型认为越陈旧的信息与未来股票收益率之间的关系越弱。*RiskMetrics* 模型的关键在于衰减系数（*Decay factor*）的选取。衰减系数越小，给与最新信息的权重越大。

然而，当股票数量多于样本数量时，*RiskMetrics* 存在与样本协方差类似的问题，即估计误差以及不可逆。

5.3. 合成聚类模型

合成聚类（*Agglomerative hierarchical clustering*）根据不同元素间的相似性，将数据点中最为相似的两个元素进行组合，并反复迭代，从而生成一个有层次的聚类树。

具体过程如下：

1.每个元素都代表一个类别，根据两个元素时间序列间的相关系数确定元素的相似性，将相似性最高的两个元素合并为一类；

2.计算所有新的类别之间的相似性，将相似性最高的两个类别合并（不同的聚类方法在定义类别之间的距离时有所区别）；

3.重复第 2 步直到所有元素都合并为一个类。

聚类树的各节点与连接其的两个元素之间的相关性有关，因而可以根据聚类树构建一个相关系数矩阵，并由此根据样本方差形成协方差矩阵。

6. 协方差矩阵估计量的评价指标

协方差矩阵估计量的比较既与估计量本身有关，也与评价方法有关。

在学术论文中，常见的协方差估计量评价方法主要有统计类指标以及经济类指标。

6.1. 统计类指标

统计类指标直接比较协方差矩阵在元素上的差别，如均方根误差（*Root mean square error, RMSE*），平均绝对误差（*Mean absolute error, MAE*）等。

例如，协方差估计量与真实协方差矩阵间差距的 *RMSE* 为

$$RMSE = \sqrt{\frac{1}{N^2} \Sigma_{i=1}^{N} \Sigma_{j=1}^{N} \left(s_{ij} - \sigma_{ij}\right)^2}$$

其中s_{ij}、σ_{ij}分别为协方差估计量与真实协方差矩阵的元素。

统计类指标比较了协方差矩阵元素一对一的差距，但是忽略了协方差矩阵的结构所包含的信息。在实务中，使用协方差矩阵主要是为了获得关于风险的度量，而统计类指标并不能

揭示与风险有关的信息。

6.2. 经济类指标

经济类指标考察根据协方差矩阵估计量生成的组合的样本外表现。比如，比较组合在有效前沿上的位置，比较不同风险水平下组合的风险收益特征，比较最小方差组合（*Minimum variance portfolios, MVP*）的实际波动性等。

与统计类指标相比，经济类指标与量化投资中应用协方差矩阵的目标更加一致，能够直接比较协方差矩阵在构建组合中的优劣。但是常见的 *MVP* 检验仅能够考察一个组合的表现，并且其检验与优化模型的设置、约束条件等有关，难以充分反应协方差矩阵蕴含的风险信息。

有研究使用组合优化的方法比较不同协方差矩阵相对于样本协方差的表现，然而，这种比较方法存在以下问题：

（1）样本协方差不可逆

不少文献中用于估计样本协方差的时间序列长度通常小于股票数量，即$Q = N/T$比值大于1。在这种情况下，样本协方差矩阵本身是不满秩的，在求解最小方差组合时，虽然可以通过广义逆矩阵（*Generalized inverse*）来获得不满秩矩阵的逆，然而这样求得的解只是二次规划问题的一个解。而基于该结果比较协方差矩阵的效果，结论还是难以令人信服的。

（2）组合优化的约束条件

组合优化的约束条件如卖空限制、权重上限等也会影响对协方差矩阵的比较。由于较大的方差与协方差会求解出来负权重，即卖空，而卖空限制相当于将较大的方差和协方差向正常值压缩。因此，不同文献设置的优化模型也会影响协方差矩阵的相对表现。

因此，在组合优化中，协方差矩阵对组合优化结果的影响与估计量的选取、约束条件的设置、N/T比值等因素都有关。虽然改善协方差估计量主要是为了较好地控制组合的风险，但是由于影响组合优化结果的因素较为复杂，单纯从组合优化的结果来看并不能筛选出较好的估计量。

从以上分析可以看到，统计类指标与经济类指标都有其各自的局限性。鉴于此，为了更加合理地比较不同协方差估计量的差别，需要一种兼具统计意义与经济意义的评价方法，以相对全面地评估协方差矩阵所包含的风险信息。本文采用了 *Liu & Lan*（2007）提出的一种基于协方差矩阵特征分解的评价指标，通过比较与两个协方差矩阵相关的组合风险所能够达到的最大差别，来比较不同的协方差估计方法。

6.3. 基于特征距离的协方差矩阵估计量评价方法

6.3.1. 特征距离的定义与含义

设组合权重向量为$w(N \times 1)$，股票收益率协方差矩阵为Σ，则在该协方差矩阵下的组合收益率方差即组合风险为

$$\mathrm{w}'\Sigma\mathrm{w}$$

可以定义一个基于组合风险的特征距离$d(V_1, V_2)$，该距离衡量了在协方差矩阵V_1和V_2下，组合风险比值所能达到的最大值的对数，具体公式如下：

$$d(V_1, V_2) = \log\left(\frac{max_x \frac{x'V_1x}{x'V_2x}}{min_y \frac{y'V_1y}{y'V_2y}}\right)$$

其中x、y为组合权重向量（$N \times 1$）。

可见，组合x在协方差矩阵V_1和V_2上的风险差别最大，并且其在协方差矩阵V_1上的风险大于在V_2上的风险；组合y在协方差矩阵V_1和V_2上的风险差别也是最大，且其在协方差矩阵V_2上

的风险大于V_1。因此，$d(V_1,V_2)$衡量了组合在协方差矩阵V_1和V_2上所能达到的最大的风险差别。

与最小方差组合的衡量方法相比，特征距离衡量了两个协方差矩阵在风险角度上存在的最大的差别，而不是仅比较一个特定组合的风险，其经济意义更加鲜明。在下文关于其计算方法的推导中，也可以看出其具有明确的统计意义。

在协方差矩阵V_1和V_2都为正定矩阵、不限制做空的假设下，可以通过较为简便的方法计算两个协方差矩阵的特征距离。

6.3.2. 特征距离的计算

对于任意协方差矩阵V（对称矩阵），都可以进行如下特征分解：

$$\mathrm{V} = \mathrm{T}\Lambda\mathrm{T}'$$

其中Λ是对角矩阵（$N \times N$），其对角线元素是V的特征值λ_i（$i = 1,2 \dots N$），T为特征向量矩阵（$N \times N$），其每一列为相应的特征向量ξ_i（$i = 1,2 \dots N$），并且$\mathrm{T}'\mathrm{T} = \mathrm{I}$，I为单位矩阵，即特征向量正交且模为 1。上式可写作

$$\mathrm{T}'\mathrm{V}\mathrm{T} = \Lambda$$

如果将特征向量看作组合权重，那么特征向量ξ_i构成的组合方差为相应的特征值λ_i。然而，将特征值解释为组合方差有一个问题，这些特征向量的权重和并不为 1（特征向量模为 1，即组合权重平方和为 1）。当然总能通过归一化将特征向量转化一个和为 1 的组合权重，从而将特征值转换为组合的方差。为了简单起见，可以认为特征值是经过调整的组合方差，其组合权重的平方和为 1。

为了计算$d(V_1,V_2)$，定义协方差矩阵

$$\mathrm{V} = \left(\mathrm{T}_1\Lambda_1^{-\frac{1}{2}}\right)' V_2 \left(\mathrm{T}_1\Lambda_1^{-\frac{1}{2}}\right)$$

其中T_1、Λ_1分别为协方差矩阵V_1的特征向量矩阵和特征值矩阵。

可以看到，V为一些组合在V_2下的协方差矩阵，而这些组合在V_1下是相互独立并且方差为 1 的，即

$$\mathrm{V}_1 = \mathrm{T}_1\Lambda_1\mathrm{T}_1'$$

$$\left(\mathrm{T}_1\Lambda_1^{-\frac{1}{2}}\right)' V_1 \left(\mathrm{T}_1\Lambda_1^{-\frac{1}{2}}\right) = \mathrm{I}$$

矩阵V可以称作V_1到V_2的变形矩阵，并且当$V_1 = V_2$时，$V = \mathrm{I}$。

由于需要找到两个协方差矩阵所能生成的组合风险的最大差别，那么这些相互独立并且在V_1下方差为 1 的组合可以作为一个基准，通过衡量它们在V_2下的方差，来衡量V_1与V_2的相对距离。

为了找到这些组合在V_2下的最大方差，可以求解以下问题：

$$\max_X X'VX \ \ s.t. \ \ X'X = 1$$

即求解V的特征向量与特征值。记λ_{max}、λ_{min}分别为矩阵V最大和最小的特征值，则这两个值代表相对于V_1，在协方差矩阵V_2下最极端的两个组合的风险。因此，特征距离可表示为

$$d(V_1,V_2) = \log\left(\frac{\lambda_{max}}{\lambda_{min}}\right)$$

$d(V_1,V_2)$越大，从风险角度来看，两个协方差矩阵之间的差别越大，反之亦然。

7. 应用与实证

7.1. 协方差估计量列表

表 1 列出了本文比较的协方差矩阵估计量。

表 1：协方差矩阵估计量说明

类别	指标名称	指标说明	编号
市场模型	BetaUnadjCov	无调整 Beta 模型	1
	BetaVasicekCov	Vasicek 调整 Beta 模型	2
多因子模型	BetaInduCov	行业因子模型（时间序列估计）	3
	Industry	行业因子模型（横截面回归）	4
	Factors	风格及行业因子模型（横截面回归）	5
	PCACov	主成分模型	6
Ledoit&Wolf 压缩估计	LWConstant	压缩目标为固定相关系数	7
	LWFactors	压缩目标为风格及行业因子模型（横截面回归）	8
	LWIdentity	压缩目标为对角阵	9
	LWIndustry	压缩目标为行业因子模型（横截面回归）	10
	LWMarket	压缩目标为无调整 Beta 模型	11
随机矩阵理论	RandomMatrix	随机矩阵理论模型	12
RiskMetrics	RiskMetrics	RiskMetrics	13
样本协方差	SampleCov	样本协方差矩阵	14
估计量组合	SimpleAverage	样本协方差矩阵与无调整 Beta 模型等权平均	15

资料来源：天风证券研究所

其中，因子模型中的行业使用中信行业分类。具体地，在中信一级行业分类的基础上，采用中信二级行业分类进一步划分银行和非银金融行业，即将银行业分为国有银行和股份制与城商行，将非银行金融分为证券、保险和信托及其他，共计 32 个行业。

风格因子为从规模、技术反转、流动性、波动性、估值、成长、质量等 7 个维度选取的 7 个典型因子，如表 2 所示。

表 2：风格因子列表

类型	因子名称	因子含义	行业和市值中性化
规模	市值对数	总市值对数	否
技术反转	反转	过去 20 个交易日涨跌幅	是
流动性	换手率	过去 20 个交易日日均换手率	是
波动	特异度	Fama 三因子回归后$1-R^2$	是
估值	BP	Book to Price	是
成长	净利润增速	单季度净利润同比增速	是
质量	ROETTM	滚动 ROE	是

资料来源：天风证券研究所

本文使用特征距离来比较各协方差估计方法。在每一期，比较各协方差矩阵估计量与下一期实际协方差矩阵之间的特征距离。该特征距离在推导中假设两个协方差矩阵都为正定矩阵，然而会存在协方差矩阵不满秩的情况，如样本协方差矩阵。因此在实际计算中，进行了如下调整：

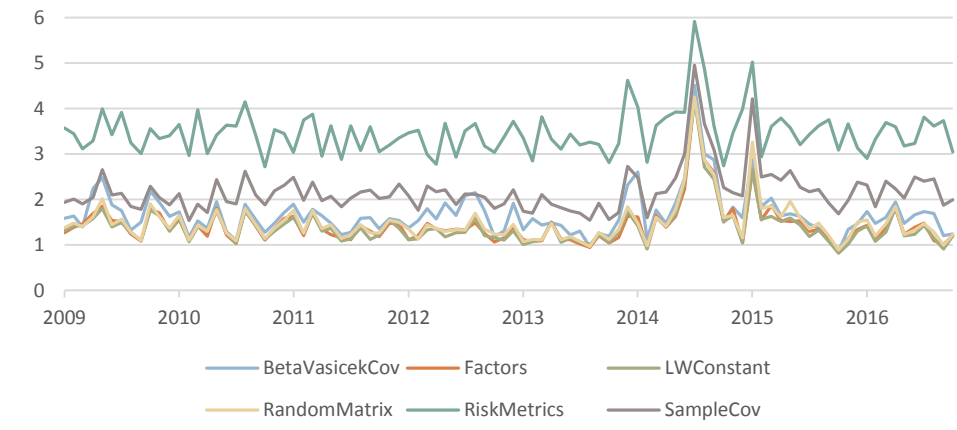
1. 当协方差矩阵不满秩时，使用伪逆矩阵作为矩阵的逆矩阵；
2. 当矩阵V不满秩时，λ_{max}、λ_{min}分别为矩阵V最大和最小的非零特征值，这使得求得的距离值是有限的。

7.2. 实证结果分析

在实际应用中，通常需要估计全市场股票的协方差矩阵，因此本文比较了各种方法估计全市场股票协方差的效果。在每月最后一个交易日，根据过去 250 个交易日的日度数据估计股票收益率协方差矩阵，然后计算协方差矩阵估计量与下月实际收益率协方差矩阵的特征距离，回测区间是 2010 年 1 月-2017 年 10 月。

在每一期，计算各协方差估计量与下期实际协方差矩阵之间的特征距离，如图 3 所示。其中下期实际协方差矩阵为下月日度收益率的样本协方差。为了便于展示，仅选取了部分协方差估计量。

图 3：协方差估计量与下期实际协方差矩阵的特征距离



资料来源：Wind，天风证券研究所

可以看到，除了 *RiskMatrics* 及样本协方差以外，其他协方差矩阵估计量与下期实际协方差之间的特征距离较为相近。

为了更加直观地比较各协方差估计量与下期实际协方差之间的特征距离，表 3 计算了各期特征距离的均值，即

$$E(d(V_{it}, V_{rt}))$$

其中，V_{it}为t期的协方差估计量i，V_{rt}为t期相应的下期实际协方差。

表 3：协方差估计量与下期实际协方差之间特征距离的均值

估计量	平均特征距离	估计量	平均特征距离	估计量	平均特征距离
LWFactors	1.3548	LWMarket	1.4206	Industry	1.5784
LWIndustry	1.3699	Factors	1.4316	BetaUnadjCov	1.6674
LWConstant	1.4056	RandomMatrix	1.4890	BetaVasicekCov	1.6714
LWIdentity	1.4084	BetaInduCov	1.4932	SampleCov	2.1770
SimpleAverage	1.4155	PCACov	1.4960	RiskMatrix	3.4728

资料来源：Wind，天风证券研究所

可以看到，压缩估计、风格行业多因子模型估计量、随机矩阵理论模型相对优于市场模型、主成分模型、样本协方差及 *RiskMatrics* 等。

虽然均值可以比较出来各估计量的相对效果，但是在每一期估计协方差矩阵时，使用了相同的信息，而唯一的差别在于估计方法，因此更好的比较方法是成对地比较任意两种估计量在各期的差别。

因而，本文计算了在每一期，任意两个协方差估计量与下期实际协方差矩阵的特征距离的差值，并对该差值序列进行均值检验。检验的结果以矩阵形式给出，如图 4 所示，其中元素t_{ij}为每期协方差估计量i、j与下期实际协方差矩阵特征距离差值的t值，即：

$$t_{ij} = \frac{E(d(V_{it}, V_{rt}) - d(V_{jt}, V_{rt}))}{std(d(V_{it}, V_{rt}) - d(V_{jt}, v_{rt}))/\sqrt{T}}$$

其中，V_{it}为t期的协方差估计量i，V_{rt}为t期相应的下期实际协方差。

因此，t_{ij}小于（或大于）0，表明估计量i与下期实际协方差矩阵的特征距离小于（或大于）估计量j；并且t_{ij}的绝对值越大，则估计量i与估计量j的差别越显著。

图 4：协方差矩阵估计量的成对比较

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
BetaUnadjCov	1		-1.94	12.20	4.40	12.06	9.36	12.28	13.58	11.46	13.24	12.18	8.20	-46.09	-19.87	12.09
BetaVasicekCov	2	1.94		12.61	4.88	12.41	9.56	12.48	13.79	11.70	13.50	12.35	8.42	-45.53	-19.93	12.28
BetaInduCov	3	-12.20	-12.61		-6.10	6.83	-0.30	8.40	11.77	7.02	11.44	7.67	0.32	-51.71	-31.55	7.86
Industry	4	-4.40	-4.88	6.10		9.95	4.48	9.47	11.83	9.58	12.04	8.42	5.01	-44.82	-28.21	8.65
Factors	5	-12.06	-12.41	-6.83	-9.95		-6.55	3.30	10.76	3.03	9.21	1.33	-5.15	-52.82	-35.18	1.99
PCACov	6	-9.36	-9.56	0.30	-4.48	6.55		12.25	13.93	8.27	13.03	11.36	0.70	-53.26	-32.50	11.38
LWConstant	7	-12.28	-12.48	-8.40	-9.47	-3.30	-12.25		10.03	-0.48	7.71	-4.55	-9.36	-54.28	-34.89	-3.42
LWFactors	8	-13.58	-13.79	-11.77	-11.83	-10.76	-13.93	-10.03		-8.73	-6.13	-10.56	-11.85	-54.30	-34.56	-11.17
LWIdentity	9	-11.46	-11.70	-7.02	-9.58	-3.03	-8.27	0.48	8.73		6.66	-1.68	-9.82	-52.98	-37.32	-1.04
LWIndustry	10	-13.24	-13.50	-11.44	-12.04	-9.21	-13.03	-7.71	6.13	-6.66		-8.61	-11.34	-53.82	-34.95	-8.80
LWMarket	11	-12.18	-12.35	-7.67	-8.42	-1.33	-11.36	4.55	10.56	1.68	8.61		-6.78	-53.65	-33.47	3.16
RandomMatrix	12	-8.20	-8.42	-0.32	-5.01	5.15	-0.70	9.36	11.85	9.82	11.34	6.78		-51.86	-39.38	7.46
RiskMatrix	13	46.09	45.53	51.71	44.82	52.82	53.26	54.28	54.30	52.98	53.82	53.65	51.86		36.22	53.39
SampleCov	14	19.87	19.93	31.55	28.21	35.18	32.50	34.89	34.56	37.32	34.95	33.47	39.38	-36.22		33.34
SimpleAverage	15	-12.09	-12.28	-7.86	-8.65	-1.99	-11.38	3.42	11.17	1.04	8.80	-3.16	-7.46	-53.39	-33.34	

资料来源：Wind，天风证券研究所

在图 4 中，t_{ij}小于 0 的区域标为蓝色，表明纵轴对应的估计量i优于横轴对应的估计量j；t_{ij}大于 0 的区域标为粉色，表明纵轴对应的估计量i差于横轴对应的估计量j。在第i行中，可以看出估计量i与其他估计量的相对表现。

可以发现：

1、整体来看，协方差估计效果从高到低排序为：*Ledoit & Wolf* 压缩估计>估计量组合≈多因子模型>随机矩阵理论模型>市场模型>样本协方差>*RiskMetrics*。

2、在多因子模型中，表现排序为：风格及行业因子模型（横截面回归）>行业因子模型（时间序列估计）>主成分模型>行业因子模型（横截面回归）。

3、在 *Ledoit & Wolf* 压缩估计中，不同压缩目标的表现排序为：风格及行业因子模型（横截面回归）>行业因子模型（横截面回归）>固定相关系数模型≈对角阵>无调整 Beta 模型。

4、随机矩阵理论模型表现差于风格及行业因子模型（横截面回归），与行业因子模型（时间序列估计）及主成分模型接近。

实证结果表明，常用的 *Ledoit & Wolf* 压缩估计及包含风格及行业因子的多因子模型的表现较好。而以多因子模型为压缩目标的压缩估计量表现最好。然而，以多因子模型作为压缩目标与 *Ledoit & Wolf* 压缩估计的初衷并不太相符。使用多因子模型作为压缩目标既没有回避因子选取的问题，也没有简化运算，无法体现出压缩估计的优势。基于此，在实务中，本文较为推荐以固定相关系数或者对角阵为压缩目标的 *Ledoit & Wolf* 压缩估计，以及包含风格及行业因子的多因子模型。

8. 协方差矩阵估计的一些补充说明

8.1. 含有缺失值的样本协方差

由于股票停牌，股票收益率会出现缺失值，即样本矩阵X中存在缺失值，这给样本协方差矩阵的计算带来一些问题。

首先，直接将包含空值的观测删掉是不合理的。

其次，如果用一对一对的变量来计算协方差，较为合理的方法是只保留两者都非缺失的观测，然后计算协方差，即

$$\frac{1}{q}\Sigma_{i=1}^{q}(x_i-\bar{x})(y_i-\bar{y})$$

其中q为保留的观测数量。因此，在计算协方差矩阵中各元素时，分母中的观测数量q可能不完全相同。在这种情况下，得到的样本协方差可能会不是半正定的，即可能会出现负的特征根。这样估计得到的样本协方差矩阵虽然可以通过压缩估计、随机矩阵理论模型等方法来调整或者寻找与其最接近的半正定矩阵等方式，使其成为半正定矩阵，但是负特征根对样本协方差的影响是难以估计的。此外，当两个变量没有共同的观测时，会无法估计出协方差。因而，为了避免这个问题，可以选择填补空值，简单的方法是直接用 0 来填充，当然这也不可避免地会带来一些估计的误差。

实际上，有很多方法可以用来估计包含缺失值样本的协方差矩阵，较为常见的有多重填补（*Multiple imputation*）、全息极大似然估计（*Full information maximum likelihood*）。当然，这些计算方法较为复杂，并且估计得到的样本协方差矩阵本身仍包含较多的估计误差。在实际应用中，可以根据需求选择合适的处理方法。

8.2. N/T比值对协方差估计量相对表现的影响

在比较协方差矩阵估计量时，还要注意到N/T比值对协方差矩阵相对表现的影响。例如，当$N/T>1$时，很多文献发现固定相关系数模型比样本协方差矩阵表现好，这可能由于股票数量多于时间序列样本数量，样本协方差包含较多估计偏差，因此简单的模型能够有效减少估计偏差。而在$N/T<1$时，有文献发现固定相关系数模型没有样本协方差矩阵表现好，这是由于此时时间序列样本量较多，过于简单的模型对于降低估计误差的作用不明显，并且会损失掉部分信息。

Pafka and Kondor（2003）使用模拟的方法，比较了不同N/T比值下协方差矩阵的相对表现，并发现：

1. 当$N/T<1$时，不同协方差矩阵生成的最优组合在实际风险以及分散程度上差别不大。
2. 当N/T接近 1 时，组合的表现与模型中是否添加卖空限制条件有关。当允许卖空时，样本协方差矩阵生成的组合表现最差；而当限制卖空时，样本协方差矩阵与其他估计量生成的组合在风险方面非常相近。正如前文所述，卖空限制相当于对样本协方差矩阵进行修正。
3. 当$N/T>1$时，样本协方差估计的组合在风险上表现最差，而不论是否添加卖空限制，其它协方差估计量生成的组合都具有更好的风险特征。

9. 参考文献

Bai J, Shi S. Estimating high dimensional covariance matrices and its applications[J]. Annals of Economics and Finance, 2011, 12(2): 199-215.

Blume M E. On the assessment of risk[J]. The Journal of Finance, 1971, 26(1): 1-10.

Disatnik D, Benninga S. Estimating the covariance matrix for portfolio optimization[J]. 2006.

Fama E F, French K R. Common risk factors in the returns on stocks and bonds[J]. Journal of financial economics, 1993, 33(1): 3-56.

Laloux L, Cizeau P, Potters M, et al. Random matrix theory and financial correlations[J]. International Journal of Theoretical and Applied Finance, 2000, 3(03): 391-397.

Ledoit O, Wolf M. A well-conditioned estimator for large-dimensional covariance

matrices[J]. Journal of multivariate analysis, 2004, 88(2): 365-411.

Ledoit O, Wolf M. Honey, I shrunk the sample covariance matrix[J]. The Journal of Portfolio Management, 2004, 30(4): 110-119.

Ledoit O, Wolf M. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection[J]. Journal of empirical finance, 2003, 10(5): 603-621.

Liu L. Portfolio risk measurement: the estimation of the covariance of stock returns[D]. University of Warwick, 2007.

Ng K W. Can random matrix theory resolve Markowitz optimization enigma?: the impact of" noise" filtered covariance matrix on portfolio selection[D]. Rutgers University-Graduate School-Newark, 2014.

Pafka S, Kondor I. Estimated correlation matrices and portfolio optimization[J]. Physica A: Statistical Mechanics and its Applications, 2004, 343: 623-634.

Pafka S, Kondor I. Noisy covariance matrices and portfolio optimization II[J]. Physica A: Statistical Mechanics and its Applications, 2003, 319: 487-494.

Pantaleo E, Tumminello M, Lillo F, et al. When do improved covariance matrix estimators enhance portfolio optimization? An empirical comparative study of nine estimators[J]. Quantitative Finance, 2011, 11(7): 1067-1080.

Stein C. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution[C]//Proceedings of the Third Berkeley symposium on mathematical statistics and probability. 1956, 1(399): 197-206.

Vasicek O A. A Note on Using Cross-sectional Information in Bayesian Estimation of Security Betas[J]. The Journal of Finance, 1973, 28(5): 1233-1239.

风险提示

市场环境变动风险，模型失效风险。

分析师声明

本报告署名分析师在此声明：我们具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，本报告所表述的所有观点均准确地反映了我们对标的证券和发行人的个人看法。我们所得报酬的任何部分不曾与，不与，也将不会与本报告中的具体投资建议或观点有直接或间接联系。

一般声明

特别声明

在法律许可的情况下，天风证券可能会持有本报告中提及公司所发行的证券并进行交易，也可能为这些公司提供或争取提供投资银行、财务顾问和金融产品等各种金融服务。因此，投资者应当考虑到天风证券及/或其相关人员可能存在影响本报告观点客观性的潜在利益冲突，投资者请勿将本报告视为投资或其他决定的唯一参考依据。

投资评级声明

类别	说明	评级	体系
股票投资评级	自报告日后的 6 个月内，相对同期沪深 300 指数的涨跌幅	买入	预期股价相对收益 20%以上
		增持	预期股价相对收益 10%-20%
		持有	预期股价相对收益-10%-10%
		卖出	预期股价相对收益-10%以下
行业投资评级	自报告日后的 6 个月内，相对同期沪深 300 指数的涨跌幅	强于大市	预期行业指数涨幅 5%以上
		中性	预期行业指数涨幅-5%-5%
		弱于大市	预期行业指数涨幅-5%以下

天风证券研究

北京	武汉	上海	深圳
北京市西城区佟麟阁路 36 号	湖北武汉市武昌区中南路 99 号保利广场 A 座 37 楼	上海市浦东新区兰花路 333 号 333 世纪大厦 20 楼	深圳市福田区益田路 4068 号卓越时代广场 36 楼
邮编：100031	邮编：430071	邮编：201204	邮编：518017
邮箱：research@tfzq.com	电话：(8627)-87618889	电话：(8621)-68815388	电话：(86755)-82566970
	传真：(8627)-87618863	传真：(8621)-68812910	传真：(86755)-23913441
	邮箱：research@tfzq.com	邮箱：research@tfzq.com	邮箱：research@tfzq.com