F.I.T 套利模型

套利交易，作为期货市场规避风险功能的重要交易方式之一，在国际上广泛被期货基金、投资机构和投机者所应用。据统计，在国外成熟的商品期货市场中，套利交易约占总交易量的 40%以上。虽然我国期货市场经过多年发展逐步完善，参与期货市场的机构和个人逐渐增多，但套利交易并未受到投资者的真正重视和充分挖掘，使得我国期货市场存在着一定的不规律性，价格发现的功能尚未完全发挥。同时，在我国期货市场上，对套利进行数学建模的方法还相对较少。

统计套利交易策略是一种被欧美大型对冲基金公司、投行等机构投资者广泛采用的投资交易策略。其最早源于国际知名投资银行 Morgan Stanley。在上个世纪九十年代早期由其旗下一个研究小组集金融学专家、统计学专家、计算机专家等研发而成。Morgan Stanley 认为统计套利是一种基于模型的投资过程，其目的是通过对相对价格偏离其理论或数量模型预测价格的资产组合构建多头和空头组合而获利。该套利技术基于定价理论、统计决策、模式识别、时间序列分析、计量经济学和现代计算方法等知识，是一种宽学科的综合投资技术。发展至今，已被华尔街的各大投行和基金公司广泛采用，甚至一些投资技术精湛的个人也在投资过程中运用该技术获得丰厚利润。

人工智能 (Artificial Intelligence)，英文缩写为 AI，它是研究、开发用于模拟、延伸和扩展人的智能的理论、方法、技术及应用系统的一门新的技术科学。人工智能在期货交易上的应用主要是通过数据挖掘、神经网络、遗传算法、模糊逻辑第五章期货价差套利策略等具有类人类大脑传递的智能系统来对期货交易进行指导分析，甚至做出决策。人工智能套利就是人工智能在套利方面的应用，这种智能系统的应用能够对交易的模式进行智能识别、智能预测，甚至通过结果进行再学习，以适应不断变化的市场环境。

人工神经网络 (Artificial Neural Networks，ANN)，一种模仿动物神经网络行为特征，进行分布式并行信息处理的算法数学模型。这种网络依靠系统的复杂程度，通过调整内部大量节点之间相互连接的关系，从而达到处理信息的目的。人工神经网络具有自学习和自适应的能力，可以通过预先提供的一批相互对应的输入一输出数据，分析掌握两者之间潜在的规律，最终根据这些规律，用新的输入数据来推算输出结果，这种学习分析的过程被称为“训练”。

程序化套利交易是程序化交易在套利方面的应用。对于大多数个人交易者，无论是业余的还是专业的，都是在存在情绪风险的情况下进行交易。对于某些投资者来说，私欲、贪婪、企盼会降低知觉的敏锐，对于另一些投资者来讲，偏见、恐惧会动摇信心。当这些投资者进入金融市场后，人类的情绪往往使得投资者变得不稳，缺乏自律。这样程序化交易的理念就出现了，而且特别要提出的是这一理念主要是应用于以期货来代表的衍生品市场。程序化交易是克服人类情感的手段。一个以计算机为基础的交易系统仅仅根据事先编好的规则来发出指令。这一系统确定进入市场和离开市场的时机、目标利润以及止损的价位。一个系统的二进制编码是没有感情的，因此系统不存在对交易的偏见、感觉、担忧。所以系统化交易的执行是客观、自律、量化以及科学的。与最初凭借个人临时判断的交易相比，程序化交易具有如下优点：

（1）跟随长期系统化策略。

（2）严格的执行纪律。

（3）资金使用量根据系统合理调整。

（4）多市场分散化交易。

（5）持续关注全球市场机会。

（6）系统的预期收益/损失是对投资者有利的。

尽管程序化交易的优缺点还有争论，但是继部分优秀的个人投资者之后，机构投资者开始大量采用程式化交易的理念，从精细的小规模交易公司、做市商到大规模的对冲基金、投资银行，程序化交易正在成为发展的方向。

本文简述了一种基于神经网络的统计套利方法，并使得该套利方法能全自动的实现数据抽样，检验，自动交易及参数优化一系列过程。

统计套利是基于两个或多个相关性较高的资产组合，如果在未来继续保持这种相关性，那么一旦两者之间出现了背离的走势，则这种背离在未来将会得到纠正，从而产生套利的机会。对统计套利的实践来讲，如果两者间出现背离，那么应该买进表现相对较差的资产而卖出表现相对较好的资产；当未来两者间的背离得到纠正时，反向平仓。其原理背后的基本概念还是均值回复（即资产价格将会回复到它的均值，或者说资产价格序列平稳）。若价差序列是平稳的，那么可以

构造统计套利的交易策略，即判断价差是否偏离了长期均值。

最常见的统计套利的定义是 HJTW 在 2004 年给出的：统计套利是一种零初始成本，自融资的无限期投资策略，其累计交易收益的折现值 v(t)满足：

（1）$\mathrm{v}(t_0) = 0$

（2）$\lim_{t\to\infty} E^p\,[v(t)] > 0$

（3）$\lim_{t\to\infty} P(v(t) < 0) = 0$

（4）如果对$\forall < \infty$，$\mathrm{P}[(\mathrm{v}(\mathrm{t}) < 0)] > 0$，则有$\lim_{t\to\infty} \frac{Var(v(t))}{t} = 0$

则称该交易策略为一个统计套利机会。若记 V(t)为 t 时刻的累计交易收益，则 v(t)=V(t)e^{-rt}为 V(t)的贴现（r 为无风险利率）。由定义我们知道，统计套利需要满足四个条件：自融资交易策略的初始投入为 0；经无风险利率折现后的价值的极限值为正；损失概率趋于 0；如果损失概率在一定时间内不是 0，则经时间平均的方差极限为 0，或者说随着时间的推移，收益 v(t)的方差的增加速度能够被 t 控制住。

第四个条件当且仅当损失概率大于 0 时成立，这表明统计套利是有风险的，即存在着损失的可能，统计套利在一段时间内并不是无风险方法，而恰恰是这一点使得我们能够区分套利与统计套利的概念。若存在某一时刻 T，对 t≥T，有 P(v(t) < 0) = 0，就是标准形式的套利。

由于直接判断价差的点位具有很大的主观性，我们引入统计上的均值、标准差和神经网络来指导交易，均值反映的是价差的集中程度，标准差反映的是价差的离散程度，根据统计学中的分布知识，价差一般落在均值加减若干倍的标准差之间，当这个参数确定后，套利区间就出来了，为了获得更大的收益和控制套利的风险，我们引入人工神经网络对价差进行预测，最后根据预测值对策略进行优化。

在交易市场中有两种典型的投资策略：趋势追踪(Trend Following)和均值回复(Mean Reversion)。趋势追踪策略的特点是抓住大幅上升趋势，避免大幅下跌趋势，其胜率并不高，但收益率较高；而均值回复策略则是一种反趋势策略，认为一波大幅上涨后容易出现下跌，而一波大幅下跌后容易出现上涨。其特点是捕捉小的机会，其胜率较高，但收益率有限。套利交易中的价差往往表现为小机

会和高胜率等特点，所以均值回复比较适合应用在套利交易中。

移动平均线 MA 是指用统计分析的方法，将一定时期内的证券价格（指数）加以平均，并把不同时间的平均值连接起来，形成一根 MA，用以观察证券价格变动趋势的一种技术指标。在这里，我们采用指数平滑移动平均线（EMA），其算法为：

$$\mathrm{EMA(n)}_t = \frac{2}{n+1} \times (g_t - EMA(n)_{t-1}) + EMA(n)_{t-1}$$

其中 n 的取值为交易观察值的数据，一般 n 取 20

价差标准差的计算公式为：

$$\sigma = \sqrt{\sum_{i=t-n}^{t} (g_i - m)^2}$$

RBF 神经网络即径向基函数(Radial Basis Function)神经网络，是一种三层前馈式神经网络，由输入层、隐层和输出层构成，其结构如图 1 所示。

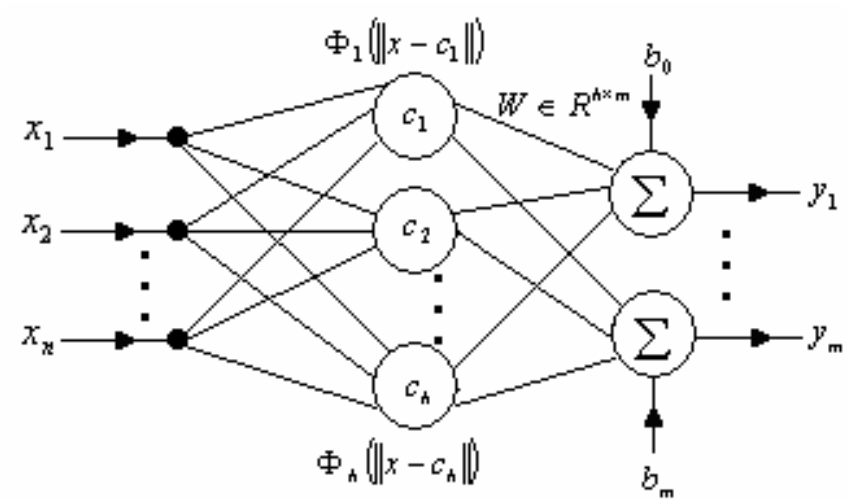

图 1　RBF 网络结构

图 1 为 n-h-m 结构的 RBF 网，即网络具有 n 个输入，h 个隐节点，m 个输出。其中 $\mathrm{x} = (x_1, x_2, \cdots, x_n)^T \in R^n$ 为网络输入矢量，$\mathrm{W} \in R^{h\times m}$ 为输出权矩阵，$b_0, \cdots, b_m$ 为输出单元偏移，$\mathrm{y} = [y_1, \cdots, y_m]^T$ 为网络输出，$\Phi_i(*)$ 为第 i 个隐节点的激活函数。图中输出层节点中 Σ 表示输出层神经元采用线性激活函数，当然输出神经元也可以采用其它非线性激活函数，如 Sigmoid 函数。

我们知道，多层感知器（包括 BP 网）的隐节点基函数采用线性函数，激活函数则采用 Sigmoid 函数或硬极限函数。与多层感知器不同，RBF 网的最显著的特点是隐节点的基函数采用距离函数（如欧氏距离），并使用径向基函数（如 Gaussian 函数）作为激活函数。径向基函数关于 n 维空间的一个中心点具有径向对称性，而且神经元的输入离该中心点越远，神经元的激活程度就越低。隐节点的这一特性常被称为“局部特性”。因此 RBF 网的每个隐节点都具有一个数据中心，如图 1 中 c_i 就是网络中第 i 个隐节点的数据中心值，‖∗‖则表示欧氏范数。

径向基函数$\Phi_i(*)$有多种形式，一般取高斯（Gaussian）函数。

$$\Phi_i(t) = e^{-\frac{t^2}{\delta_i^2}}$$

式中的δ_i称为该基函数的扩展函数（Spread）或宽度。显然，δ_i越小，径向基函数的宽度就越小，基函数就越具有选择性。

于是图 1 中，RBF 网的第 k 个输出可表示为：

$$y_k = \sum_{i=1}^{h} w_i \Phi_i(\|x - c_i\|)$$

下面以两输入单输出函数逼近为例，简要介绍 RBF 网的工作机理。与输出节点相连的隐层第 i 个隐节点的所有参数可用三元组(c_i, δ_i, w_i)表示。由于每个隐层神经元都对输入产生响应，且响应特性呈径向对称（即是一个个同心圆），于是 RBF 网的作用机理可用图 2 表示。图 2 表示输入区域中中有 6 个神经元，每个神经元都对输入 x 产生一个响应$\Phi_i(\|x - c_i\|)$，而神经网络的输出则是所有这些响应的加权和。

由于每个神经元具有局部特性，最终整个 RBF 网最终也呈现“局部映射”特性，即 RBF 网是一种局部相应神经网络。这意味着如果神经网络有较大的输出，必定激活了一个或多个隐节点。

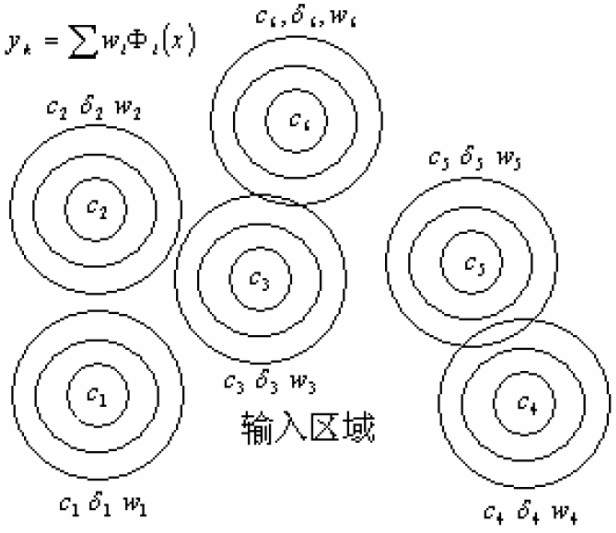

图 2　RBF 网的工作原理

RBF 神经网络是一种性能良好的前向网络。它具有最佳逼近性能，在结构上具有输出——权值线性关系，训练方法快速易行，不存在局部最优问题。该网络的学习算法有很多种，其中最经典的 RBF 网学习算法是由 Moody 与 Darken 提出聚类方法。其思路是先用无监督学习（用 k - means 算法对样本输入进行聚类）方法确定 RBF 网中 h 个隐节点的数据中心，并根据各数据中心之间的距离确定隐节点的扩展常数，然后用有监督学习（梯度法）训练各隐节点的输出权值。

假设 k 为迭代次数，第 k 次叠代时的聚类中心为$c_1(k)$，$c_2(k)$，…，$c_h(k)$，相应的聚类域为$w_1(k)$，$w_2(k)$，…，$w_h(k)$。k - means 聚类算法确定 RBF 网数据中心c_i和扩展常数δ_i的步骤如下：

(1) 算法初始化：选择 h 个不同的初始聚类中心，并令 k = 1。初始聚类中心的方法很多，比如说从样本输入中随机选取，或者选择前 h 个样本输入，但这 h 个初始数据中心必须取不同值。

(2) 计算所有样本输入与聚类中心的距离$\|X_j - c_i(\mathrm{k})\|$，i=1, 2, …, h；j=1, 2, …, N。

(3) 对样本输入X_j，按最小距离原则对其进行分类：即当$\mathrm{i}(X_j) = \min_i \|X_j - c_i(\mathrm{k})\|$, i=1, 2, …, h 时时，$X_j$即被归为第 i 类，即$X_j \in w_i(k)$。

(4) 重新计算各类的新的聚类中心：$c_i(k+1)=\frac{1}{N_i}\sum_{x\in w_i(k)}x$，i=1,2,⋯,h，其中$N_i$为第 i 个聚类域$w_i(k)$中包含的样本数。

(5) 如果 $c_i(k+1)\neq c_i(k)$，转(2)，否则聚类结束，转(6)。

(6) 根据各聚类中心之间的距离确定各隐节点的扩展常数。隐节点的扩展常数取$\delta_i=\kappa d_i$，其中d_i为第 i 个数据中心与其他最近的数据中心之间的距离，即$d_i=\min_{j\neq i}\|c_j-c_i(\mathrm{k})\|$，$\kappa$称重叠系数，一般取$\kappa=1$。

一旦各隐节点的数据中心和扩展常数确定了，输出权矢量$\mathrm{w}=[w_1,w_2,\ldots,w_h]^T$就可以用有监督学习方法（如梯度法）训练得到，但更简洁的方法是使用最小二乘方法(LMS)直接计算。假定当输入为X_i，i = 1,2,⋯,N 时，第 j 个隐节点的输出为：

$$h_{ij}=\Phi_J(\|X_i-c_j\|)$$

则隐层输出阵为

$$\widehat{H}=[h_{ij}]$$

则$\widehat{H}\in R^{N\times h}$。如果 RBF 网的当前权值为$\mathrm{w}=[w_1,w_2,\ldots,w_h]^T$（待定），则对所有样本，网络输出矢量为

$$\hat{y}=\widehat{H}w$$

令$\varepsilon=\|\mathrm{y}-\hat{y}\|$为逼近误差，则如果给定了教师信号$\mathrm{y}=[y_1,y_2,\cdots,y_N]^T$并确定了$\widehat{H}$，便可通过最小化下式求出网络的输出权值 w：

$$\varepsilon=\|\mathrm{y}-\hat{y}\|=\|\mathrm{y}-\widehat{H}\mathrm{w}\|$$

通常 w 可用最小二乘法（LMS）求得：

$$\mathrm{w}=\widehat{H}^{+}y$$

其中，$\widehat{H}^{+}$为$\widehat{H}$的伪逆：

$$\widehat{H}^{+}=(\widehat{H}^T\widehat{H})^{-1}\widehat{H}^T$$

表 1 变量说明

变量名称	变量意义	变量解析
g	价差	
m	价差均值	采取指数加权平均的方法计算
σ	价差标准差	
fm	价差预测值	使用神经网络对价差进行预测
$m+k_1\sigma$	不考虑预测值的空单界线	$3<k_1<=6$
$m+k_2\sigma$	考虑预测值的空单界线	$2<k_2<=3$
$m+k_3\sigma$	空单平仓界线	$0<k_3<=2$
$m-k_4\sigma$	多单平仓界线	$0<k_4<=2$
$m-k_5\sigma$	考虑预测值的多单界线	$2<k_5<=3$
$m-k_6\sigma$	不考虑预测值的多单界线	$3<k_6<=6$

空单策略

- 如果 $g_t>=m+k_1\sigma$，不考虑预测值直接建仓（一手）
- 如果 $m+k_2\sigma<=g_t<m+k_1\sigma$，且 $fm_{t+1}>=m+k_1\sigma$，此时价差偏离正常值，但预测值显示价差进一步扩大的可能性比较大，此时不进行任何操作
- 如果 $m+k_2\sigma<=g_t<m+k_1\sigma$，且 $m+k_3\sigma<=fm_{t+1}<m+k_1\sigma$，此时价差偏离正常值，但预测值显示价差未出现明显的回归趋势和进一步扩大的可能，此时建仓（一手）
- 如果 $m+k_2\sigma<=g_t<m+k_1\sigma$，且 $fm_{t+1}<m+k_3\sigma$，此时价差偏离正常值，但预测值显示价差出现回归的可能性比较大，此时可以大胆地建仓（两手）
- 如果 $g_{t+1}<=m+k_3\sigma$，则平仓

多单策略

- 如果 $g_t<=m-k_6\sigma$，不考虑预测值直接建仓（一手）
- 如果 $m-k_6\sigma<g_t<=m-k_5\sigma$，且 $fm_{t+1}<=m-k_6\sigma$，此时价差偏离正常值，但预测值显示价差进一步缩小的可能性比较大，此时不进行任何操作

- 如果 $m-k_6\sigma<g_t<=m-k_5\sigma$，且 $m-k_6\sigma<fm_{t+1}<=m-k_4\sigma$，此时价差偏离正常值，但预测值显示价差未出现明显的回归趋势和进一步缩小的可能，此时建仓（一手）
- 如果 $m-k_6\sigma<g_t<=m-k_5\sigma$，且 $fm_{t+1}>m-k_4\sigma$，此时价差偏离正常值，但预测值显示价差出现回归的可能性比较大，此时可以大胆地建仓（两手）
- 如果 $g_{t+1}>= m-k_4\sigma$，则平仓

该交易策略的原则是：步步为营，大赢，小亏。