清華大學
Tsinghua University

基于机器学习方法的增强型配对交易策略

Enhancing a Pairs Trading strategy with the application of Machine Learning

Simão Moraes Sarmento, Nuno Horta

- 汇报人：刘长　张胜楠

研究背景

配对交易（Pairs Trading）是指八十年代中期华尔街著名投行Morgan Stanley的数量交易员Nunzio Tartaglia成立的一个数量分析团队提出的一种市场中性投资策略，其成员主要是物理学家、数学家、以及计算机学家。

配对交易分为两种类型：一类是基于统计套利的配对交易，一类是基于风险套利的配对交易。

基于统计套利的配对交易策略是一种市场中性策略，配对交易利用配对间的短期错误定价，通过持有相对低估，卖空相对高估，因此其本质上是一个反转投资策略，其核心是学术文献中的股票价格均值回复。

研究背景

配对交易中存在的主要问题

1. 如何选择可获利的配对资产？ （配对选择）

2. 如何避免配对资产价值长时间的发散而导致收益的长期下降？（交易策略）

配对选择—传统方法

方法一：在选中的证券集合中，对所有可能的组合进行穷举。

方法二：先分成几个类别，然后在每个类别中选出合适的配对组合。

法一可能会找到更想不到的组合，但是这些关联可能是站不住脚的。

法二没有办法发现有趣的潜在关系。

如何确定选出配对是否胜任交易?

传统方法使用协整方法进行筛选。

这个方法要求Y_t和X_t是协整的。这样的话，我们可以定义价差序列（spread series）如下：

$S_t = Y_t - \beta X_t$（β是一个协整因子）这样得到的价差序列是平稳的。在这样的情况下，S_t可被认为是均值回归的，意味着每一个价差偏离均值之后都会趋于收敛。

配对选择—创新方法

分为三个要点，**降维，聚类，确定选择配对的标准**

降维

利用PCA方法进行主成分分析。不是直接对股价序列使用PCA，而是使用下面的进行了归一化之后的序列。

$$R_{i,t} = \frac{P_{i,t} - P_{i,t-1}}{P_{i,t-1}}$$

聚类

- 采用无监督学习方法，对证券集进行聚类，然后从每一类中选择配对组合。
- DBSCAN算法固定聚类半径的缺陷，如右图所示。
- OPTICS 算法解决了这个问题。 OPTICS 基于 DBSCAN，引入了一些实现变化的 ε 的方法。 在这个增强的设置中，投资者只需要指定参数 minPts，因为该算法能够为每个集群检测最合适的 ε。

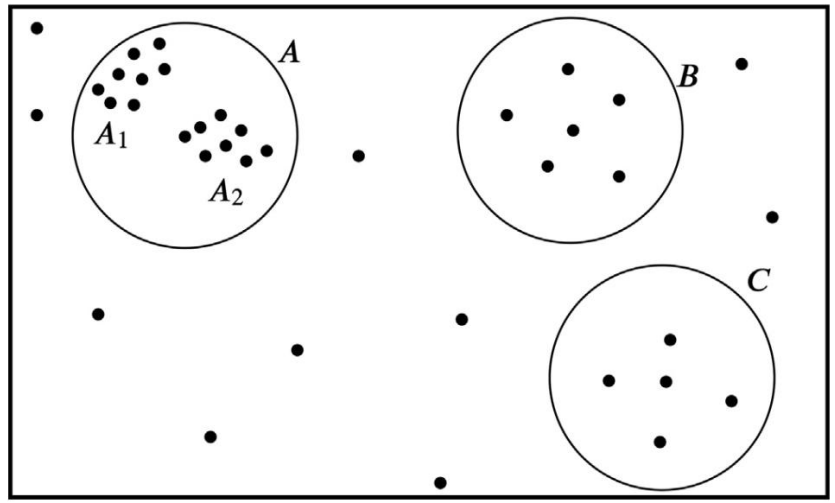


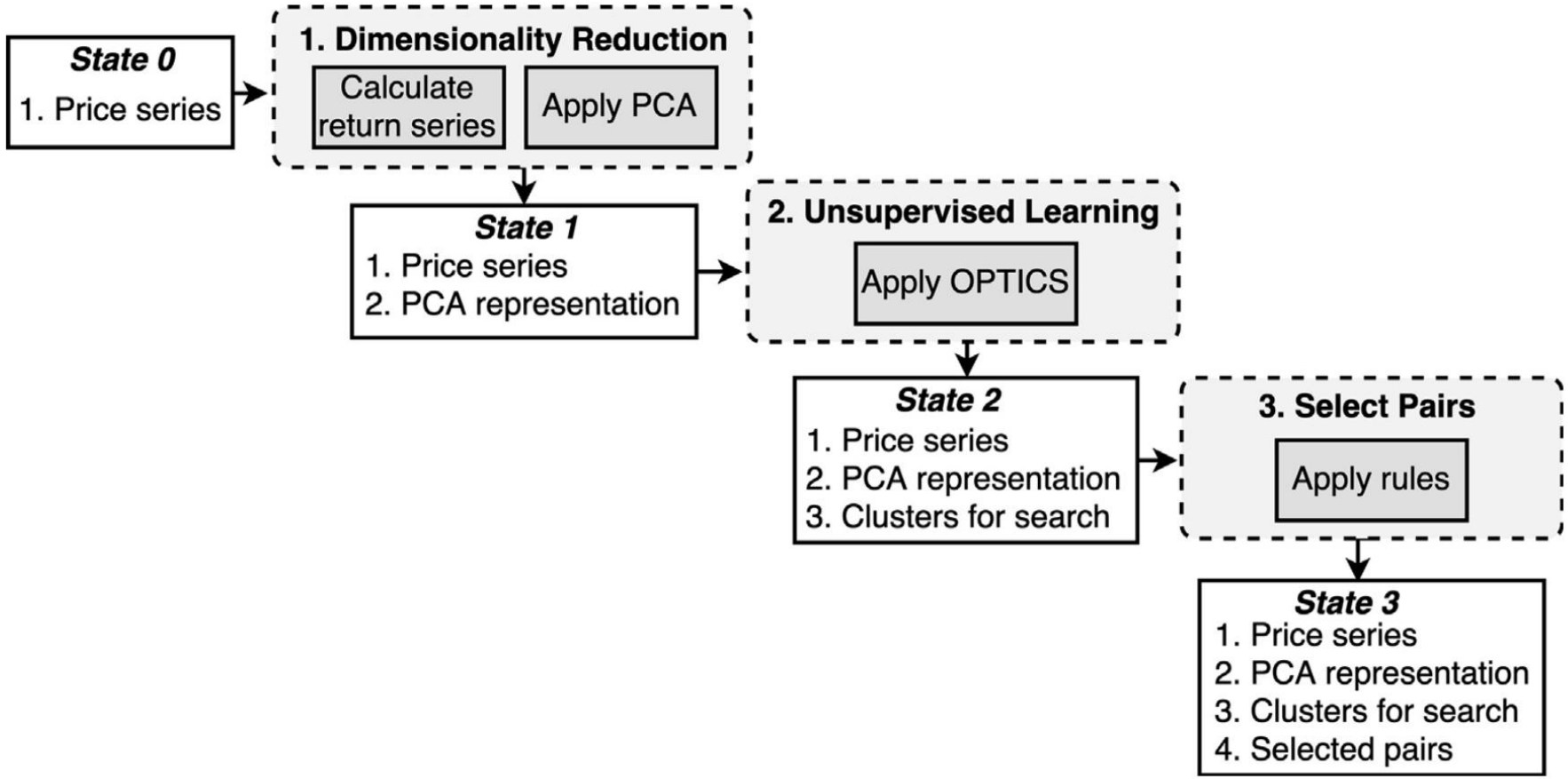
Fig. 1. Clusters with varying density. Adapted from: Ankerst et al. (1999).

配对选择—创新方法

配对的选择标准：

1. 两种资产的历史价格序列必须是协整的。检验方法使用EG两步法（Engle-Granger test）
2. 为了确保价差序列的均值特性，需要进行额外的验证，采用赫斯特指数。如果H的范围在0～0.5之间，则可以说明这个序列是均值回归的。
3. 需要丢弃不合适时间段的平稳配对。因为均值回归本身并不一定产生利润，只有当交易周期和价差序列的半衰期匹配时，配对才具有可以利用的价值。
4. 我们强制要求每一个价差跨过它的均值12次，平均每个月至少执行一次交易。

配对选择—创新方法


State 0
1. Price series
1. Dimensionality Reduction
Calculate return series
Apply PCA
State 1
1. Price series
2. PCA representation
2. Unsupervised Learning
Apply OPTICS
State 2
1. Price series
2. PCA representation
3. Clusters for search
3. Select Pairs
Apply rules
State 3
1. Price series
2. PCA representation
3. Clusters for search
4. Selected pairs

交易模型—传统方法

1. 计算配对组合的价差的平均值和标准误。
2. 定义模型的阈值，触发多头头寸的阈值α_L，触发空头头寸的阈值α_S，退出的阈值α_{exit}
3. 监控差值的变化，也就是$S_t = Y_t - X_t$的变化，当某个阈值被触发时做出相应的操作。
4. 如果α_L被跨过，通过买入Y并且卖出X来做多价差。反之亦然。

这个方法的弊端是，交易决策中不包含有关后续价差方向的信息，因此定义的切入点不一定是最佳的。

Table 1
Threshold-based model parameters.

Parameters	Values
Long Treshold	$\mu_s - 2\sigma_s$
Short Threshold	$\mu_s + 2\sigma_s$
Exit Threshold	μ_s

交易模型—创新方法

$$\Delta_{t+1} = \frac{S^*_{t+1} - S_t}{S_t} \times 100,$$

如何得到预测的价差？ARMA，LSTM， LSTM Encoder-Decoder.

$$x_t = \frac{S_{t+1} - S_t}{S_t} \times 100.$$

正的变化率序列和负的变化率序列x_t分别有分布函数$f^+(x)$和$f^-(x)$

$$P_{t+1}: \begin{cases} \text{if } \Delta_{t+1} \geqslant \alpha_L, & \text{Go long} \\ \text{if } \Delta_{t+1} \leqslant \alpha_S, & \text{Go short}. \\ \text{otherwise}, & \text{Wait} \end{cases}$$

$$\{\alpha_S, \alpha_L\} = \underset{q}{\operatorname{argmax}} R^{\text{val}}(q),$$

$$q \in \left[\left\{Q_{f^-(x)}(0.20), Q_{f^+(x)}(0.80)\right\}\left\{Q_{f^-(x)}(0.10), Q_{f^+(x)}(0.90)\right\}\right]$$

交易模型—创新方法

交易模型框架图

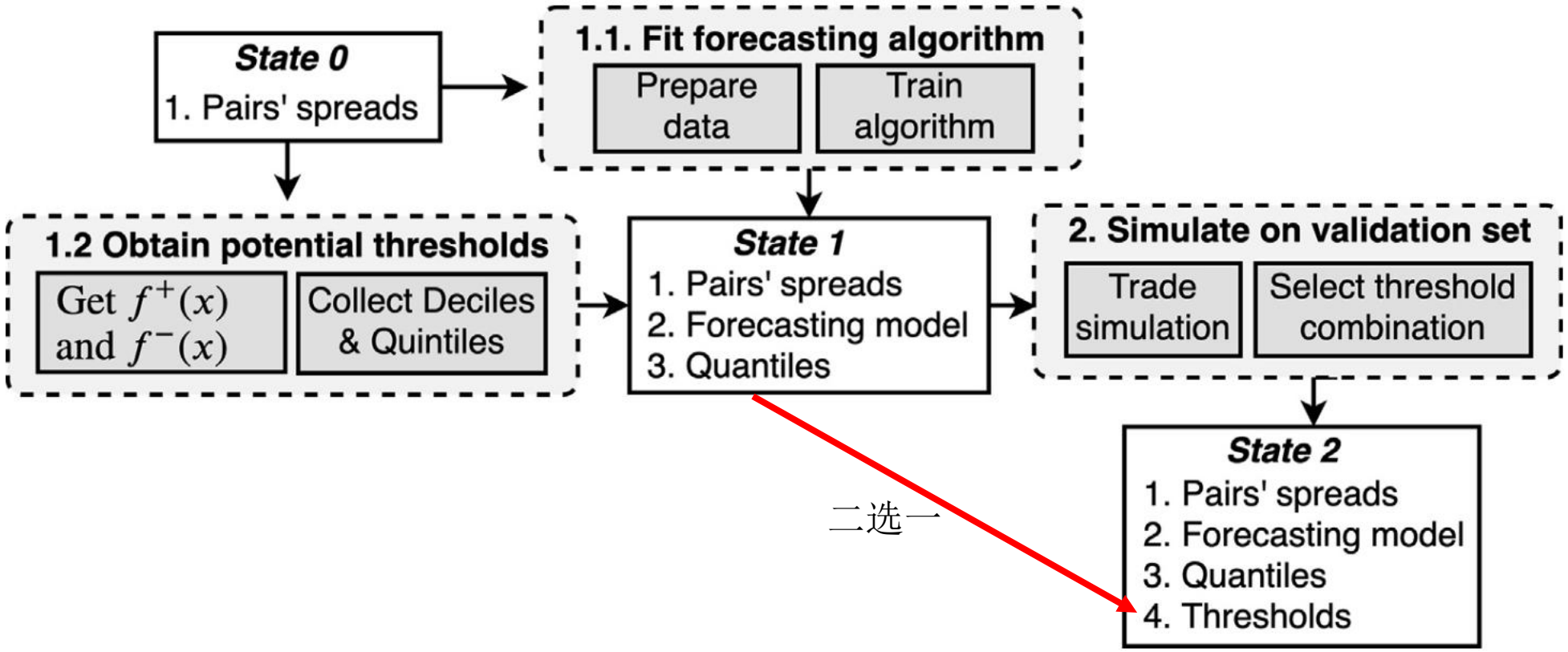


Fig. 3. Proposed model construction diagram.

交易模型—创新方法

理想状况下的交易过程：

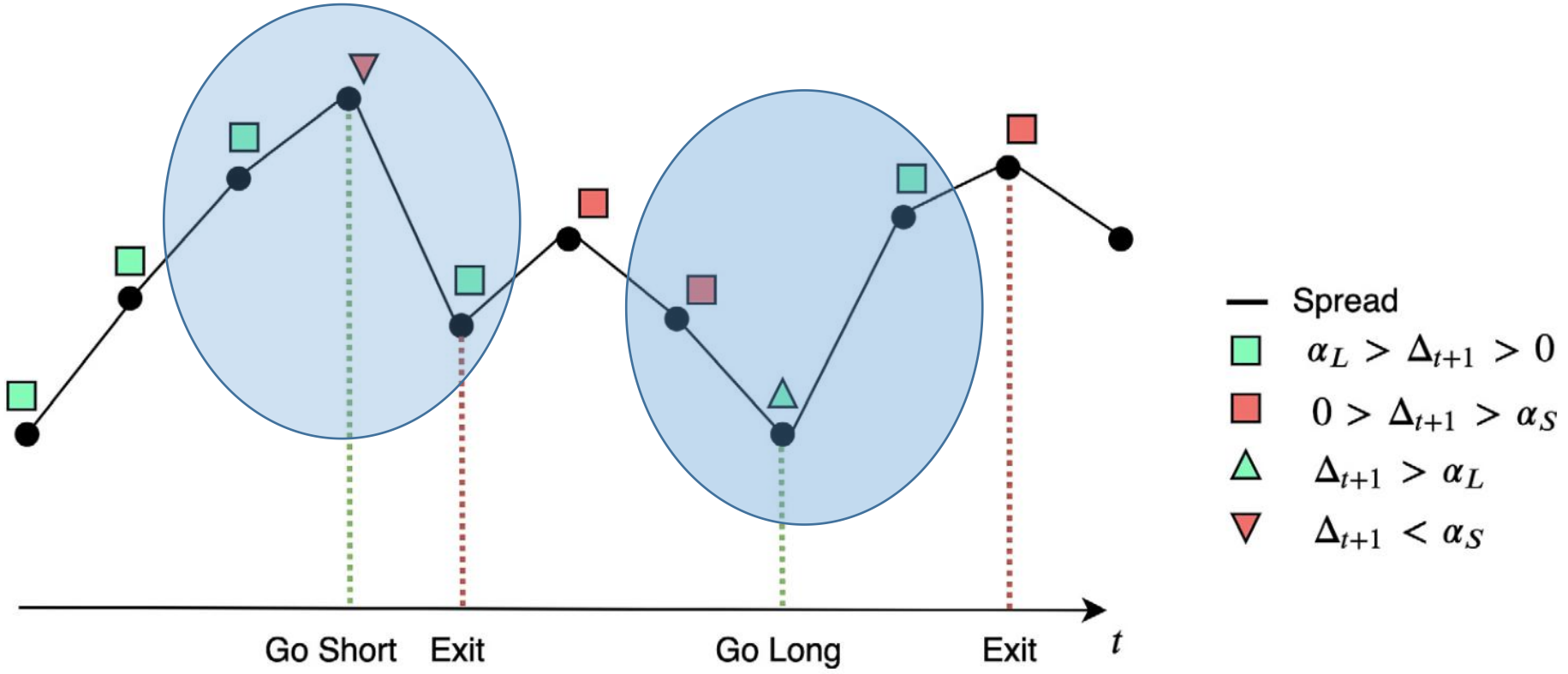


Fig. 4. Example of the proposed forecasting-based strategy.

实验—配对选择

Table 2
Selected pairs using different search methods.

Formation Period	2012–2015	2013–2016	2014–2017
No Clustering			
Number of clusters	1	1	1
Possible combinations	4465	5460	6670
Pairs selected	101	247	150
By Category			
Number of clusters	5	5	5
Possible combinations	2612	3318	4190
Pairs selected	59	51	51
OPTICS			
Number of clusters	9	13	12
Possible combinations	185	140	129
Pairs selected	39	40	18

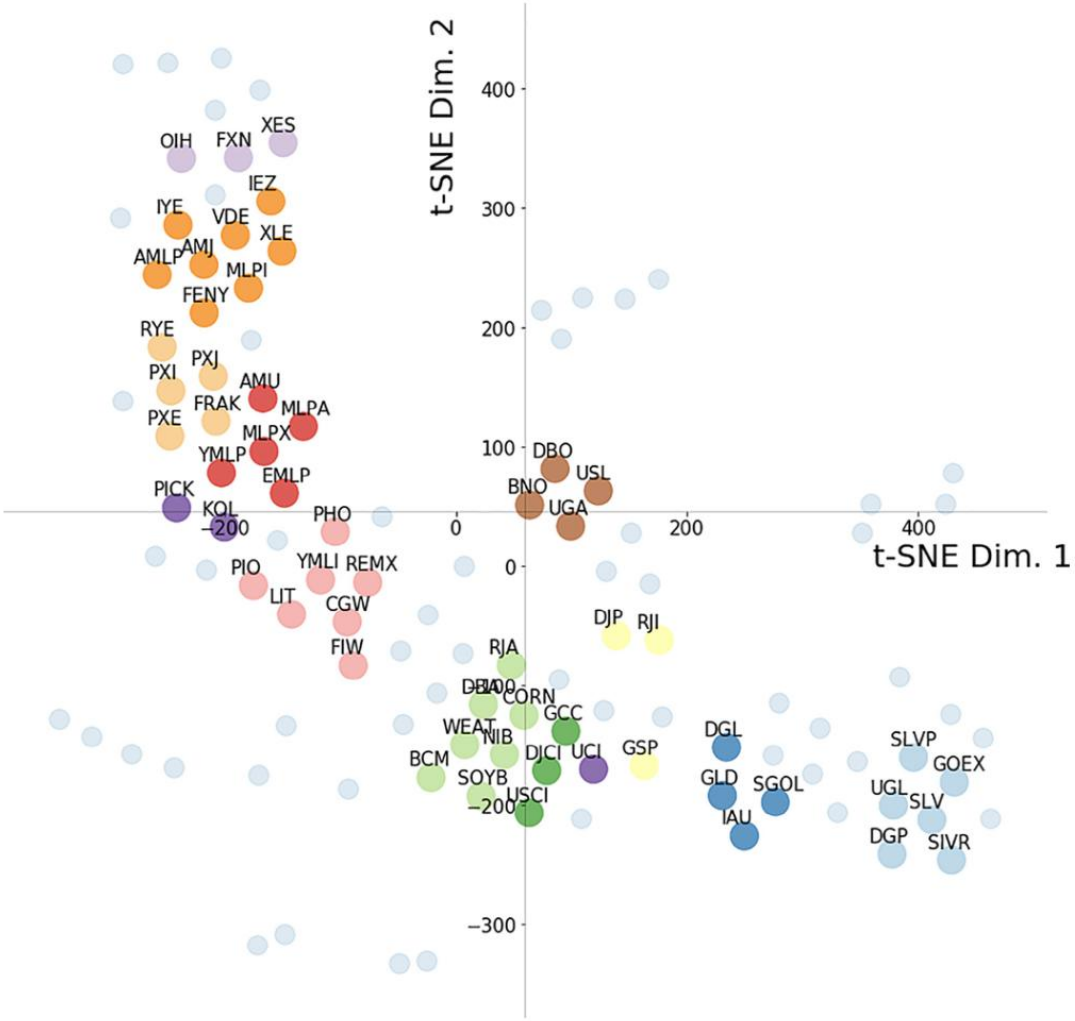
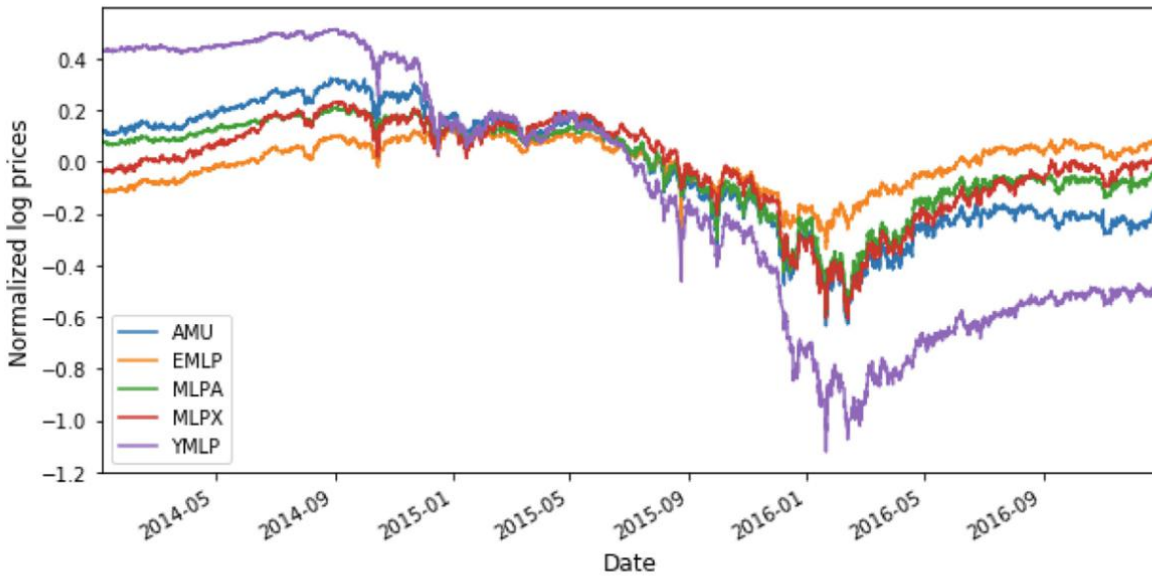
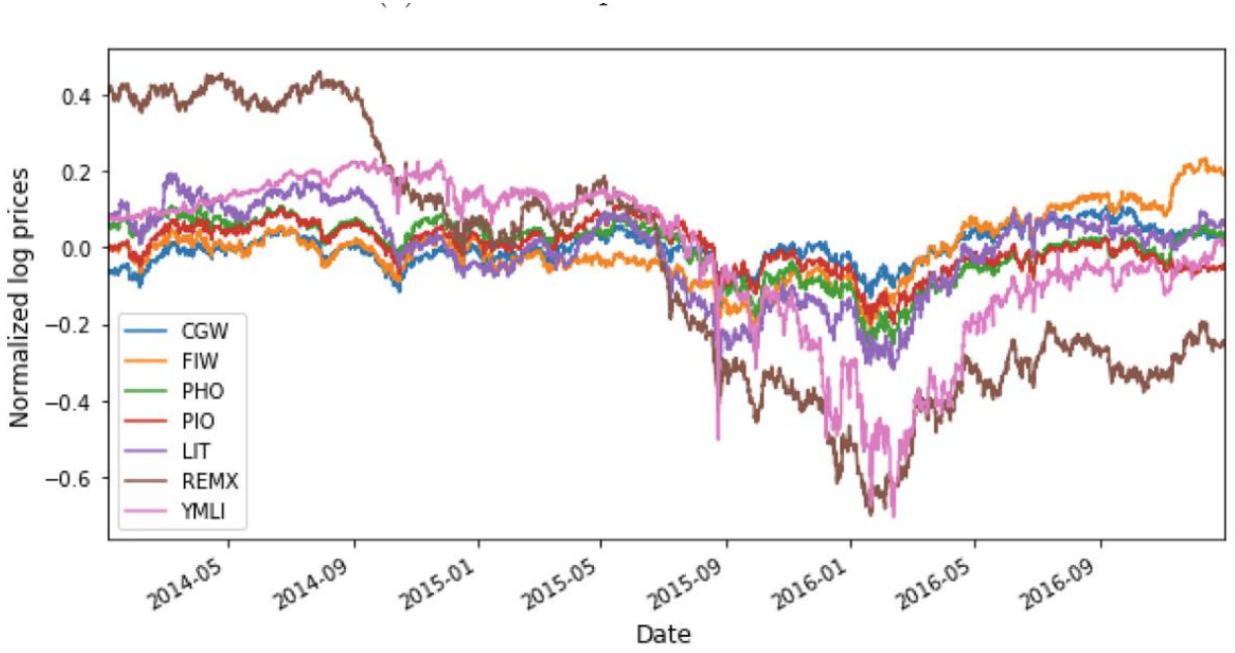


Fig. 8. Application of t-SNE to the clusters generated by OPTICS.

实验—配对选择



(a) Normalized prices in Cluster 1.



(b) Normalized prices in Cluster 2.

我们选择两个集群并表示出 ETF 的价格系列。

图 9（a） 说明了一个识别出的 ETF 不仅属于同一类别，而且属于同一部类，即美国股票：MLPs。

图 9（b） 展示了 OPTICS 聚类能力超出了在同一细分市场中选择 ETF 的范围，因为我们可以观察到来自不同类别的 ETF，例如农业（CGW、FIW、PHO 和 PIO）、工业金属（LIT 和 REMX）和能源（YMLI）。表示出的价格序列之间存在可见的关系，即使它们不都属于同一类别。

实验—配对选择

Table 3
Trading performance for each pairs search technique.

Test Period	2015			2016			2017			AVG.
Test Portfolio	1	2	3	1	2	3	1	2	3	–
No Clustering										
SR	3.53	4.12	3.32	3.96	4.51	3.56	4.08	4.05	1.11	***3.58***
ROI (%)	10.4%	12.4%	17.4%	24.8%	26.3%	26.0%	11.9%	12.4%	11.5%	***17.0%***
MDD (%)	1.42%	0.97%	2.59%	2.05%	1.98%	2.65%	1.33%	1.38%	9.28%	***2.63%***
Total pairs	101	77	10	247	223	10	150	141	10	*108*
Profitable pairs (%)	70%	80%	70%	86%	86%	90%	69%	70%	90%	*79%*
Total trades	229	173	17	411	361	15	212	195	14	*181*
Profitable trades	180	147	15	369	329	15	172	162	12	*156*
Unprofitable trades	49	26	2	42	32	0	40	33	2	*25*
By Category										
SR	1.56	2.39	3.75	3.48	3.82	3.09	2.17	2.14	0.89	***2.59***
ROI (%)	5.52%	9.38%	17.8%	13.6%	13.9%	20.4%	7.86%	8.42%	8.31%	***11.7%***
MDD (%)	1.77%	1.82%	2.09%	2.06%	2.26%	4.56%	2.47%	2.67%	8.91%	***3.18%***
Total pairs	59	40	10	51	44	10	51	47	10	*36*
Profitable pairs (%)	64%	85%	90%	86%	86%	90%	65%	64%	90%	*80%*
Total trades	154	108	39	107	83	20	64	54	13	*71*
Profitable trades	112	89	36	92	73	19	49	43	12	*58*
Unprofitable trades	42	19	3	15	10	1	15	11	1	*13*
OPTICS										
SR	4.05	3.84	5.08	4.72	4.79	3.80	2.75	2.83	2.27	***3.79***
ROI (%)	12.5%	13.5%	23.5%	10.5%	11.9%	15.2%	7.36%	8.38%	9.98%	***12.5%***
MDD (%)	1.37%	1.66%	1.30%	0.80%	0.83%	1.46%	1.21%	1.35%	2.35%	***1.37%***
Total pairs	39	34	10	40	35	10	18	16	10	*24*
Profitable pairs (%)	82%	82%	100%	80%	83%	90%	78%	81%	100%	86%
Total trades	161	147	68	87	78	30	24	22	17	*70*
Profitable trades	140	128	67	72	66	27	21	20	17	*62*
Unprofitable trades	21	19	1	15	12	3	3	2	0	*8*

实验—交易模型

5个配对的差值序列（9年的形成期）

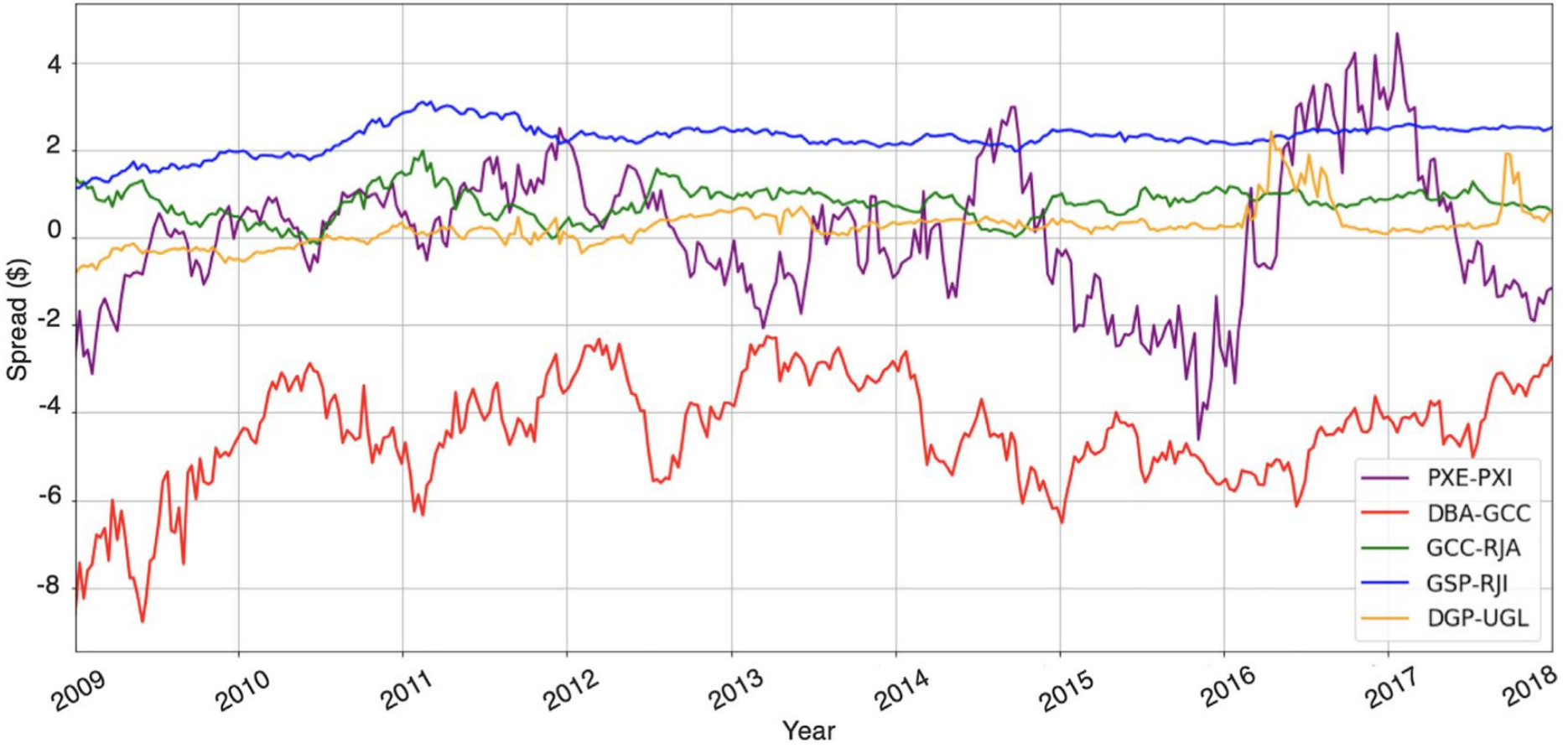


Fig. 10. Pairs identified in Jan 2009-Dec 2017.

实验—交易模型

调参过程，各个模型预测效果对比，选出均方误差最小的模型用于交易策略的实验

Table 4
Forecasting results comparison.

Model	Parameters	Time-step	Validation			Test		
			MSE (E-03)	RMSE (E-02)	MAE (E-02)	MSE (E-03)	RMSE (E-02)	MAE (E-02)
Naive	$Y_{t+1} = Y_t$	(t + 1)	1.87	3.69	1.50	2.60	3.89	1.68
Naive	$Y_{t+2} = Y_t$	(t + 2)	3.34	4.94	2.24	4.47	5.14	2.61
ARMA	$p : 5, q : 2$	(t + 1)	1.511	3.006	1.781	2.271	3.343	1.967
ARMA	$p : 8, q : 3$	**(t + 1)**	**1.508**	**3.004**	**1.780**	**2.264**	**3.339**	**1.964**
RMA	$p : 12, q : 4$	(t + 1)	1.509	3.004	1.780	2.264	3.338	1.964
LSTM	$i_n : 12, h_l : 1, h_n : 10$	(t + 1)	2.73	3.73	2.65	4.99	4.74	3.63
LSTM	$i_n : 24, h_l : 1, h_n : 50$	**(t + 1)**	**1.69**	**3.28**	**2.04**	**3.35**	**4.30**	**3.08**
LSTM	$i_n : 24, h_l : 1, h_n : 60$	(t + 1)	1.91	3.43	2.03	3.54	4.61	3.36
LSTM	$i_n : 12, e_n : 30, d_n : 30$	(t + 1)	2.03	3.60	2.18	5.72	5.75	4.31
Enc.-Dec.		(t + 2)	2.43	3.94	2.51	8.45	6.91	5.52
LSTM	$i_n : 24, e_n : 15, d_n : 15$	**(t + 1)**	**1.71**	**3.32**	**2.13**	**4.31**	**5.21**	**3.94**
Enc.-Dec.		**(t + 2)**	**2.05**	**3.60**	**2.45**	**9.03**	**7.50**	**5.91**
LSTM	$i_n : 24, e_n : 30, d_n : 30$	(t + 1)	1.96	3.56	2.21	6.06	5.95	4.50
Enc.-Dec.		(t + 2)	2.42	3.92	2.51	8.73	7.04	5.57

实验—交易模型

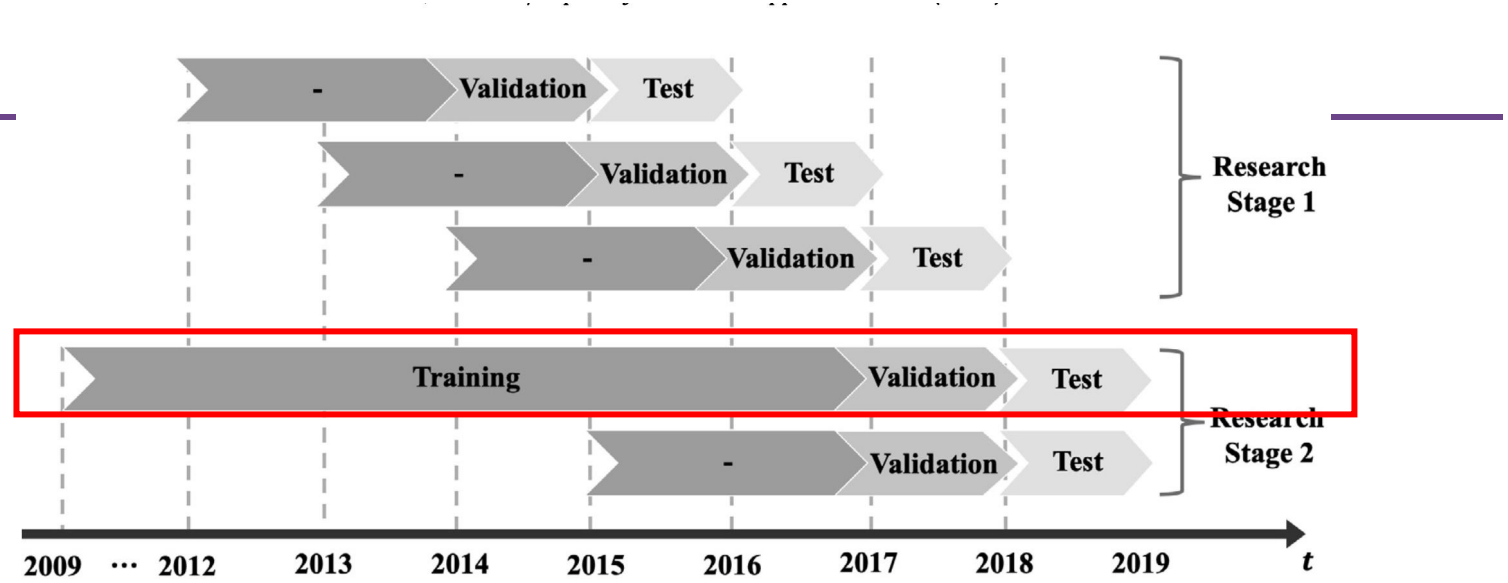


Fig. 6. Trading periods.

Table 5
Trading results comparison using a 9-year-long formation period.

Trading Model	Standard	ARMA based Model	LSTM based Model	LSTM Encoder Decoder based Model
Parameters	*see Table 1	$\alpha_S = Q_{f^-(0.20)}$ $\alpha_L = Q_{f^+(0.80)}$	$\alpha_S = Q_{f^-(0.10)}$ $\alpha_L = Q_{f^+(0.90)}$	$\alpha_S = Q_{f^-(0.10)}$ $\alpha_L = Q_{f^+(0.90)}$
SR	1.85	1.22	0.50	0.98
ROI	6.27%	5.57%	2.93%	4.17%
MDD	1.43%	0.73%	0.47%	1.19%
Days of portfolio decline	**87**	**11**	**2**	**21**
Trades (Positive–Negative)	149 (89–60)	34 (22–12)	8 (6–2)	17 (14–3)
Profitable pairs	3	3	2	2

总结

我们探讨了如何通过集成机器学习来增强配对交易。

首先，我们提出了一种基于 PCA 和 OPTICS 算法的应用来搜索对的新方法。 建议的方法比标准方法获得了更好的风险调整回报：在部门内搜索，或考虑所有可能的配对组合。

其次，我们引入了一个基于预测的配对交易模型，旨在减少与不合时宜的市场头寸和长期分歧配对相关的下降期。 我们证明了所提出的模型能够将平均下降期缩短 75% 以上，尽管这是在所研究的条件下以盈利能力下降为代价的。