

ETF 一二级市场 T+0 交易数据分析挖掘

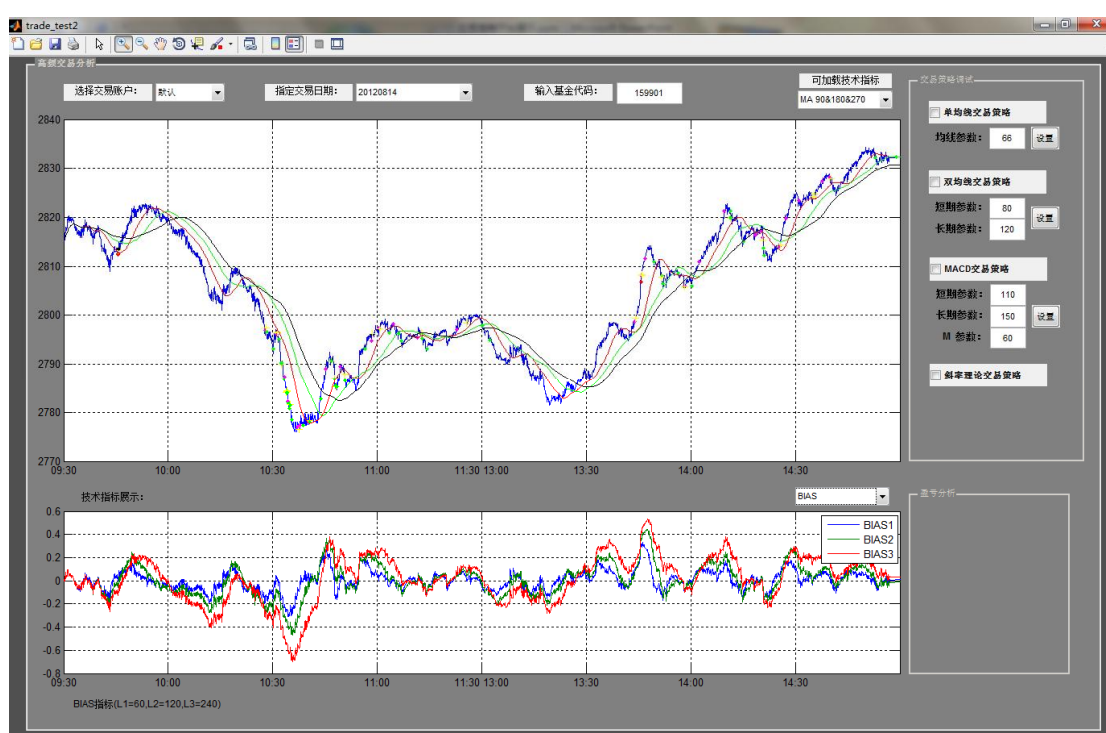
山东大学齐鲁证券金融研究院

项目负责人：杨维强

项目组成员：王海琳、郑骅、田英良

摘要

我们在去年基础之上做了进一步工作，编制了 ETF 策略展现软件，把交易数据与每日分笔数据进行整合，可以方便对交易记录进行观察、猜测、修改参数、对比、验证，并对客户的策略进行深入分析，编了相应的程序化交易策略。



简介

投资者在金融市场里的目的只有一个：赢利，通过对市场的走向判断进行买卖以希望实现盈利。但是由于每个人都有自己的判断，从而形成自己的交易策略。著名的投资者有很多，像索罗斯、巴菲特、西蒙斯、约翰·保尔森等，各有自己的绝招，策略不一而足。

常见的交易方式有手工交易和程序化交易，手段有消息面、基本面、技术分析、形态分析等。手工交易通常是凭借形态、指标、消息进行综合判断等入场出场，有能力的个人和机构可以用程序来完成量化交易。由于指标多种多样，参数多种设定，理论上每个人都可以做

出无穷多个策略，但想让策略能够持续盈利则是非常困难的事情。有一些客户在市场中获得了远远超出其他人和大盘的收益，现在考虑一个反问题，是否可以通过分析这些客户的交易记录和其他可能的信息来挖掘出他们的交易策略。客户怎么思考是黑匣子，如果客户采用的交易思想简单、指标固定，坚决执行策略，特别是程序化交易，则挖掘出他的策略相对容易，而如果客户手工交易，由于各种原因，交易思想不一定贯彻到底，锚定目标不断变化，造成进出场的条件不断发生变化，同时客户也在不断改进自己的策略，参数也在不断发生变化，反演是极其困难的，而如果客户是一个团队在并肩作战，那么复杂度更是大大增加。这有点像不同的液体和溶剂混合在一起，是否能分辨出有哪些成分。一般而言全部分析出是很困难的事情，但常用的手段就那一些，可以尝试分析出一部分。

我们已知某些客户是通过ETF一二级市场进行T+0交易。可能采用的工具和方法有：通过ETF套利软件交易；通过ETF分时线看形态手工交易；搭配指标信号进行手工或者程序化交易；有更为复杂的程序化交易系统。我们需要不断缩小范围，挖掘出客户的策略。

客户的交易记录数据非常复杂，不仅仅是购买—申购/赎回—卖出的轮回这么有规律，存在很多例外，例如卖出一篮子股票绝大部分都是深100ETF成分股，但是其中夹杂了12只成分股之外的股票；或者在正常的交易轮回中夹杂着一两只股票的交易，对这些交易的判定有很大难度。

L430														66	
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	
1	OFERTIME	REPORTTIME	STKCODE	BSFLAGNAME	ORDERPRICE	ORDERQTY	MATCHQTY	CANCELQTY	ORDERFR019	MATCHAMT	ORDERDATE				
424	11292787	11292787	159916	1	1.829	90400	90400	0	165364.33	165341.6	20120227				
425	11292787	11292787	159916	4		90400	90400	0	18231.2		20120227	股票不好判断下面这	股票		
426	13003311	13003311	423	2	43.35	1500	500	1000	0	21675	20120227	238只股票			
427	13003320	13003320	2024	2	9.36	11000	11000	0	0	102960	20120227	有12只在ETF100之外			
428	13003328	13003328	425	2	16.59	3000	3000	0	0	49770	20120227				
429	13003330	13003332	680	2	10.86	1500	1500	0	0	16290	20120227	22 深赤湾A			
430	13003336	13003336	2056	2	19.58	500	500	0	0	9790	20120227	66 长城电脑			
431	13003350	13003350	527	2	14.37	4500	600	3900	0	8622	20120227	89 深圳机场			
432	13003355	13003355	24	2	20.39	2000	2000	0	0	40780	20120227	488 晨鸣纸业			
433	13003357	13003357	877	2	25.26	500	500	0	0	12630	20120227	717 韶钢松山			
434	13003391	13003391	39	2	15.71	2000	2000	0	0	31420	20120227	725 京东方A			
435	13003400	13003400	758	2	22.19	1000	1000	0	0	22190	20120227	883 湖北能源			
436	13003404	13003404	927	2	7.11	1000	1000	0	0	7110	20120227	900			
437	13003411	13003412	762	2	26.47	500	500	0	0	13235	20120227	932			
438	13003414	13003414	60	2	10.74	3500	3500	0	0	37590	20120227	951			
439	13003416	13003418	930	2	6.81	2000	2000	0	0	13620	20120227	2110			
440	13003426	13003426	61	2	12.2	1500	1500	0	0	18300	20120227	2493			
441	13003428	13003428	933	2	12.17	2500	2500	0	0	30450	20120227				
442	13003438	13003439	63	2	17.72	4500	600	3900	0	10632	20120227				
443	13003450	13003451	780	2	12.99	1000	1000	0	0	12990	20120227				
444	13003453	13003455	69	2	7.62	6500	6500	0	0	49530	20120227				
445	13003458	13003458	2155	2	27.09	1000	1000	0	0	27090	20120227				
446	13003473	13003473	581	2	34.05	1000	1000	0	0	34050	20120227				
447	13003477	13003478	338	2	36.4	2000	200	1800	0	7280	20120227				
448	13003486	13003487	792	2	38.74	1500	1500	0	0	58110	20120227				
449	13004878	13004879	2008	2	8.19	2000	2000	0	0	16380	20120227				

数据

分析的首要基础工作是把交易记录和分笔价格数据放到一起对比，把买卖时点和价位标在走势图上，给人进出场的直观形象，然后才能做进一步判断。数据是研究的基础，基础不

牢，则研究的可信度就低。由于前面提到客户交易的方法多种多样，而手工交易随时会发单，因此对数据的要求极高，如果是做全面而精细的分析需要客户交易记录、ETF 分笔数据和 ETF 成分股的分笔数据。但如果遇到数据捉襟见肘的情况则需要把问题做适当的简化，分步骤进行分析。

数据准备

余冠敏和孙华民经理为数据的准备工作付出了辛苦努力，从青岛香港中路营业部提取了一些客户的交易记录和 ETF 的价格数据，这个过程花了两个星期左右。每天收盘后至第二天开盘前只能提取一个客户半年的交易记录，约 50 万条交易记录，这是由于买入卖出 ETF 成分股会生成多条交易记录，例如一轮 180ETF 交易包含约 180 条左右的记录。数据提取的过程也不是一帆风顺，由于不同的数据在不同的表中，第一次提取的数据缺少某些字段，后来又重新提取，通过对多张表进行查询提取工作。

ETF 的价格数据最初余经理提供的是四只最活跃的品种一分钟 K 线数据，这个频率对于分析来说是不够的。后来我们收集到每日 ETF 分笔交易的数据进行分析。

下图是每日分笔数据压缩包，每年有 200 多个压缩包。

☒ 20110711.tar.gz	20110812.tar.gz	20110915.tar.gz	20111026.tar.gz	201
20110712.tar.gz	20110815.tar.gz	20110916.tar.gz	20111027.tar.gz	201
20110713.tar.gz	20110816.tar.gz	20110919.tar.gz	20111028.tar.gz	201
20110714.tar.gz	20110817.tar.gz	20110920.tar.gz	20111031.tar.gz	201
20110715.tar.gz	20110818.tar.gz	20110921.tar.gz	20111101.tar.gz	201
20110718.tar.gz	20110819.tar.gz	20110922.tar.gz	20111102.tar.gz	201
20110719.tar.gz	20110822.tar.gz	20110923.tar.gz	20111103.tar.gz	201
20110720.tar.gz	20110823.tar.gz	20110926.tar.gz	20111104.tar.gz	201
20110721.tar.gz	20110824.tar.gz	20110927.tar.gz	20111107.tar.gz	201
20110722.tar.gz	20110825.tar.gz	20110928.tar.gz	20111108.tar.gz	201
20110725.tar.gz	20110826.tar.gz	20110929.tar.gz	20111109.tar.gz	201
20110726.tar.gz	20110829.tar.gz	20110930.tar.gz	20111110.tar.gz	201
20110727.tar.gz	20110830.tar.gz	20111010.tar.gz	20111111.tar.gz	201
20110728.tar.gz	20110831.tar.gz	20111011.tar.gz	20111114.tar.gz	201
20110729.tar.gz	20110901.tar.gz	20111012.tar.gz	20111115.tar.gz	201
20110801.tar.gz	20110902.tar.gz	20111013.tar.gz	20111116.tar.gz	201
20110802.tar.gz	20110905.tar.gz	20111014.tar.gz	20111117.tar.gz	201
20110803.tar.gz	20110906.tar.gz	20111017.tar.gz	20111118.tar.gz	201
20110804.tar.gz	20110907.tar.gz	20111018.tar.gz	20111121.tar.gz	201
20110805.tar.gz	20110908.tar.gz	20111019.tar.gz	20111122.tar.gz	201
20110808.tar.gz	20110909.tar.gz	20111020.tar.gz	20111123.tar.gz	201
20110809.tar.gz	20110912.tar.gz	20111021.tar.gz	20111124.tar.gz	201
20110810.tar.gz	20110913.tar.gz	20111024.tar.gz	20111125.tar.gz	201
20110811.tar.gz	20110914.tar.gz	20111025.tar.gz	20111128.tar.gz	201

每个压缩包内的数据包含当日所有交易品种的分笔数据，每个文件大小从 10k-500K 不等，这样一个压缩包的大小有几十 M，每年的数据有几十个 G。

名称	大小	压缩后大小	修改时间	模式	用户
SH000001.csv	457 698	457 728	2011-10-11 0...	0rwxrwxrwx	Administ
SH000002.csv	426 227	426 496	2011-10-11 0...	0rwxrwxrwx	Administ
SH000003.csv	365 127	365 568	2011-10-11 0...	0rwxrwxrwx	Administ
SH000004.csv	425 808	425 984	2011-10-11 0...	0rwxrwxrwx	Administ
SH000005.csv	417 381	417 792	2011-10-11 0...	0rwxrwxrwx	Administ
SH000006.csv	415 040	415 232	2011-10-11 0...	0rwxrwxrwx	Administ
SH000007.csv	420 581	420 864	2011-10-11 0...	0rwxrwxrwx	Administ
SH000008.csv	420 830	420 864	2011-10-11 0...	0rwxrwxrwx	Administ
SH000009.csv	423 367	423 424	2011-10-11 0...	0rwxrwxrwx	Administ
SH000010.csv	424 387	424 448	2011-10-11 0...	0rwxrwxrwx	Administ
SH000011.csv	375 550	375 808	2011-10-11 0...	0rwxrwxrwx	Administ
SH000012.csv	18 289	18 432	2011-10-11 0...	0rwxrwxrwx	Administ
SH000013.csv	192 756	193 024	2011-10-11 0...	0rwxrwxrwx	Administ
SH000015.csv	419 872	420 352	2011-10-11 0...	0rwxrwxrwx	Administ

分笔数据示例，数据差不多 3 秒钟一个，包括时间、价格、成交量、成交额、买卖五档。

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Time	Price	Volume	Amount	SP1	SP2	SP3	SP4	SP5	SV1	SV2	SV3	SV4	SV5	BP1
9:25:00	0.551	68769	3789171	0.551	0.552	0.553	0.554	0.555	280704	44027	65632	8680	19044	0.55
9:30:05	0.55	40437	2227243	0.551	0.552	0.553	0.554	0.555	252288	42155	73056	44448	43518	0.55
9:30:08	0.55	1200	66000	0.551	0.552	0.553	0.554	0.555	252288	42155	73056	44448	43518	0.55
9:30:10	0.55	1470	80850	0.551	0.552	0.553	0.554	0.555	252288	42155	73056	44448	43518	0.55
9:30:15	0.551	1692	93061	0.551	0.552	0.553	0.554	0.555	252288	42155	73056	44448	43518	0.55
9:30:23	0.55	694	38170	0.551	0.552	0.553	0.554	0.555	252288	42155	73056	44448	43518	0.55
9:30:26	0.551	300	16530	0.551	0.552	0.553	0.554	0.555	256992	42155	73056	44448	43518	0.55
9:30:28	0.55	10000	550000	0.551	0.552	0.553	0.554	0.555	256992	42155	73056	44448	43518	0.55
9:30:32	0.55	2000	110000	0.551	0.552	0.553	0.554	0.555	256992	47155	73056	44448	43518	0.55
9:30:42	0.55	1050	57750	0.551	0.552	0.553	0.554	0.555	257184	51175	73056	44448	43818	0.55

数据展示

1. 客户交易记录

交易记录数据包括：代码为 053230028571 的客户 2011 年下半年，2012 年全年，2013 年上半年的交易记录。代码为 53230025779 的客户 2012 下半年的交易记录。代码为 053230026242 的客户 2013 年上半年的交易记录。

这是客户交易记录的片段，字段很多，这里只展示了约 1/3 的字段

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
ORIGINNAME	CUSTID	FUNDID	SECID	ORDERSNO	ORDERID	OPERTIME	REPORTTIME	MARKETNAME	STKCODE	STKNAME	PRODCODE	ESFLAGNAME	MONEYNAME	
青岛香港中路营业部	313030030418	53230026242	A217846121	3068	1517000117	9224175	9224175	沪A	601618	中国中冶	证券买入	人民币		
青岛香港中路营业部	313030030418	53230026242	A217846121	3760	1517000120	9230830	9230831	沪A	600316	洪都航空	证券买入	人民币		
青岛香港中路营业部	313030030418	53230026242	A217846121	3818	1517000123	9232723	9232724	沪A	601336	新华保险	证券买入	人民币		
青岛香港中路营业部	313030030418	53230026242	0134577989	99251	Q4006384	11311014	11311014	深A	159902	中小板	证券卖出	人民币		
青岛香港中路营业部	313030030418	53230026242	A217846121	99269	1517005610	11312214	11312214	沪A	510050	50ETF	证券卖出	人民币		
青岛香港中路营业部	313030030418	53230026242	A217846121	99304	1517005611	11314915	11314917	沪A	510110	周期ETF	证券卖出	人民币		
青岛香港中路营业部	313030030418	53230026242	A217846121	149527	1517009463	14303334	14303334	沪A	600873	梅花集团	证券卖出	人民币		
青岛香港中路营业部	313030030418	53230026242	0134577989	154579	Q4010181	14413076	14413076	深A	000826	桑德环境	证券卖出	人民币		
青岛香港中路营业部	313030030418	53230026242	0134577989	4278	Q4000123	9223857	9223857	深A	159908	深F200	证券卖出	人民币		
青岛香港中路营业部	313030030418	53230026242	0134577989	4292	Q4000124	9224555	9224555	深A	159908	深F200	证券卖出	人民币		
青岛香港中路营业部	313030030418	53230026242	0134577989	4304	Q4000126	9224918	9224918	深A	159908	深F200	证券卖出	人民币		
青岛香港中路营业部	313030030418	53230026242	0134577989	4657	Q4000150	9244958	9244958	深A	159903	深成ETF	证券买入	人民币		
青岛香港中路营业部	313030030418	53230026242	A217846121	5819	1517000154	9283402	9283402	沪A	510050	50ETF	证券买入	人民币		
青岛香港中路营业部	313030030418	53230026242	A217846121	6398	1517000154	9300825	9300825	沪A	510050	50ETF	证券买入	人民币		
青岛香港中路营业部	313030030418	53230026242	A217846121	7037	1517000373	9304243	9304244	沪A	510050	50ETF	证券买入	人民币		
青岛香港中路营业部	313030030418	53230026242	A217846121	7102	1517000373	9304742	9304743	沪A	510050	50ETF	证券买入	人民币		
青岛香港中路营业部	313030030418	53230026242	0134577989	7328	Q4000150	9304910	9304910	深A	159903	深成ETF	证券买入	人民币		
青岛香港中路营业部	313030030418	53230026242	0134577989	7365	Q4000256	9305074	9305074	深A	159903	深成ETF	证券买入	人民币		
青岛香港中路营业部	313030030418	53230026242	0134577989	8737	Q4000256	9313795	9313795	深A	159903	深成ETF	证券买入	人民币		
青岛香港中路营业部	313030030418	53230026242	0134577989	8976	Q4000267	9313960	9313960	深A	159903	深成ETF	证券买入	人民币		
青岛香港中路营业部	313030030418	53230026242	0134577989	10162	Q4000281	9325154	9325155	深A	159903	深成ETF	ETF赎回	人民币		
青岛香港中路营业部	313030030418	53230026242	0134577989	10932	Q4000293	9333159	9333159	深A	000001	平安银行	证券卖出	人民币		
青岛香港中路营业部	313030030418	53230026242	0134577989	10933	Q4000294	9333159	9333160	深A	000069	华侨城A	证券卖出	人民币		
青岛香港中路营业部	313030030418	53230026242	0134577989	10934	Q4000295	9333161	9333163	深A	000423	东阿阿胶	证券卖出	人民币		
青岛香港中路营业部	313030030418	53230026242	0134577989	10935	Q4000296	9333161	9333163	深A	000024	招商地产	证券卖出	人民币		
青岛香港中路营业部	313030030418	53230026242	0134577989	10936	Q4000297	9333161	9333163	深A	000157	中联重科	证券卖出	人民币		

2. ETF 价格以及 ETF 指数的分笔数据

每天每支产品的分笔数据为一个 csv 文件，当天所有的数据文件为一个压缩包。

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Time	Price	Volume	Amount	SP1	SP2	SP3	SP4	SP5	SV1	SV2	SV3	SV4	SV5	BP1
9:25:00	0.551	68769	3789171	0.551	0.552	0.553	0.554	0.555	280704	44027	65632	8680	19044	0.55
9:30:05	0.55	40437	2227243	0.551	0.552	0.553	0.554	0.555	252288	42155	73056	44448	43518	0.55
9:30:08	0.55	1200	66000	0.551	0.552	0.553	0.554	0.555	252288	42155	73056	44448	43518	0.55
9:30:10	0.55	1470	80850	0.551	0.552	0.553	0.554	0.555	252288	42155	73056	44448	43518	0.55
9:30:15	0.551	1692	93061	0.551	0.552	0.553	0.554	0.555	252288	42155	73056	44448	43518	0.55
9:30:23	0.55	694	38170	0.551	0.552	0.553	0.554	0.555	252288	42155	73056	44448	43518	0.55
9:30:26	0.551	300	16530	0.551	0.552	0.553	0.554	0.555	256992	42155	73056	44448	43518	0.55
9:30:28	0.55	10000	550000	0.551	0.552	0.553	0.554	0.555	256992	42155	73056	44448	43518	0.55
9:30:32	0.55	2000	110000	0.551	0.552	0.553	0.554	0.555	256992	47155	73056	44448	43518	0.55
9:30:42	0.55	1050	57750	0.551	0.552	0.553	0.554	0.555	257184	51175	73056	44448	43818	0.55

图：深 100ETF（代码 159901）2012 年 8 月 27 号的部分分时数据

3. 可利用数据

由于在分析 ETF 交易策略时需要同时考虑 ETF 客户交易情况以及 ETF 价格及其跟踪指数的分时数据，因此可以进行分析的数据即代码为 053230028571 的客户 2011 年下半年及 2012 年全年的交易记录以及代码为 53230025779 的客户 2012 下半年的交易记录。

数据提取及处理

1. 对客户交易记录数据的处理

原客户交易记录中有 33 列数据，为减少输入程序的数据，提高运算效率，只提取其中 11 列有价值的数用来分析。

A	B	C	D	E	F	G	H	I	J	K
OPERTIME	REPORTTIME	STKCODE	BSFLAGNAME	ORDERPRICE	ORDERQTY	MATCHQTY	CANCELQTY	ORDERFR019	MATCHAMT	ORDERDATE
9295224	9295225	510180	证券买入	0.6430000	1000000	0	1000000	0.0000000	0.0000000	20110701
9295226	9295226	510180	证券买入	0.6430000	1000000	0	1000000	0.0000000	0.0000000	20110701
9295228	9295228	510180	证券买入	0.6430000	1000000	0	1000000	0.0000000	0.0000000	20110701
9302092	9302093	159901	证券买入	0.7800000	1000000	0	1000000	0.0000000	0.0000000	20110701
9304498	9304498	159901	证券买入	0.7800000	1000000	1000000	0	0.0000000	0.0000000	20110701
9313292	9313292	510180	证券买入	0.6430000	1000000	1000000	0	0.0000000	0.0000000	20110701
9313296	9313296	510180	证券买入	0.6430000	1000000	1000000	0	0.0000000	0.0000000	20110701
9313300	9313300	510180	证券买入	0.6430000	1000000	1000000	0	0.0000000	0.0000000	20110701
9314699	9314699	159901	证券买入	0.7770000	1000000	1000000	0	777106.8400000	777000.0000000	20110701
9315188	9315188	159901	证券买入	0.7760000	1000000	0	1000000	0.0000000	0.0000000	20110701
9315867	9315867	159901	证券买入	0.7750000	1000000	0	1000000	0.0000000	0.0000000	20110701
9315868	9315869	159901	证券买入	0.7750000	1000000	0	1000000	0.0000000	0.0000000	20110701
9395754	9395754	159903	证券买入	1.2280000	1000000	0	1000000	0.0000000	0.0000000	20110701
9395756	9395756	159903	证券买入	1.2280000	200000	0	200000	0.0000000	0.0000000	20110701
9404026	9404026	159903	证券买入	1.2280000	1000000	1000000	0	0.0000000	0.0000000	20110701
9404030	9404030	159903	证券买入	1.2280000	200000	200000	0	0.0000000	0.0000000	20110701
9461332	9461333	159903	证券买入	1.2300000	1000000	0	1000000	0.0000000	0.0000000	20110701
9461334	9461335	159903	证券买入	1.2300000	800000	0	800000	0.0000000	0.0000000	20110701
9471638	9471638	159903	证券买入	1.2300000	1000000	1000000	0	0.0000000	0.0000000	20110701
9471643	9471643	159903	证券买入	1.2300000	800000	800000	0	0.0000000	0.0000000	20110701
9572447	9572447	159903	证券买入	1.2260000	300000	0	300000	0.0000000	0.0000000	20110701
9575455	9575455	159903	证券买入	1.2260000	300000	300000	0	0.0000000	0.0000000	20110701
9580589	9580589	159906	证券买入	0.9060000	1000000	0	1000000	0.0000000	0.0000000	20110701
9593001	9593001	159906	证券买入	0.9060000	1000000	1000000	0	0.0000000	0.0000000	20110701

图：提取 11 列有价值数据后客户部分交易数据

2. 对 ETF 价格及指数的提取与处理

现有数据为 2011 年及 2012 年每日所有股票，ETF 基金，指数的分时交易数据。可以用把所有成分股数据根据权重求和，算出指数，但是指数每半年会进行调整，特殊情况也可能进行临时调整，如果详细算出成分股综合之后的效果将是巨大的工作量。而其效果上等同于直接使用指数数据。例如上证 180 指数（部分）成分股如下图

上证180指数成分股列表

[基本信息](#) [编制方法](#) [编制细则](#) [成份股列表](#) [成交概况](#) [最新走势](#) [历史模拟走势](#)

→ 成份股列表

上证180指数依据样本稳定性和动态跟踪相结合的原则，每半年调整一次成份股，每次调整比例一般不超过10%。特殊情况时也可能对样本进行临时调整。查看详情，[请点击这里](#)。

2014-04-01发布		
浦发银行 (600000)	首创股份 (600008)	上海机场 (600009)
包钢股份 (600010)	华能国际 (600011)	华夏银行 (600015)
民生银行 (600016)	上港集团 (600018)	宝钢股份 (600019)
华电国际 (600027)	中国石化 (600028)	南方航空 (600029)
中信证券 (600030)	三一重工 (600031)	招商银行 (600036)
保利地产 (600048)	中国联通 (600050)	五矿发展 (600058)
海信电器 (600060)	南京高科 (600064)	宇通客车 (600066)
冠城大通 (600067)	同仁堂 (600085)	特变电工 (600089)
太古城 (600094)	同方股份 (600100)	上汽集团 (600104)
亚盛集团 (600108)	国金证券 (600109)	包钢稀土 (600111)
中国卫星 (600118)	兰花科创 (600123)	兴发集团 (600141)
金发科技 (600143)	中国船舶 (600150)	永泰能源 (600157)
中体产业 (600158)	巨化股份 (600160)	香江控股 (600162)
福田汽车 (600166)	上海建工 (600170)	雅戈尔 (600177)
奈州矿业 (600188)	复星医药 (600196)	金种子酒 (600199)
新湖中宝 (600208)	浙江医药 (600216)	海南航空 (600221)

可以直接使用上证 180 指数，代码为 000010。

ETF 成分股每半年调整一次，同时每日公布申购赎回列表，申赎列表如下格式


```

[ETF180]
Fundid1=510181
CreationRedemptionUnit=3000000
MaxCashRatio=0.50000
Publish=1
CreationRedemption=1
Recordnum=180
EstimateCashComponent=-15931.47
TradingDay=20120228
PreTradingDay=20120227
CashComponent=-15885.47
NAUperCU=1712272.53
NAU=0.571
TAGTAG
600000|浦发银行| 5300|1|0.10000|
600005|武钢股份| 1900|1|0.10000|
600009|上海机场| 500|1|0.10000|
600010|宝钢股份| 1200|1|0.10000|
600015|华夏银行| 1600|1|0.10000|
600016|民生银行| 10800|1|0.10000|
600019|宝钢股份| 2500|1|0.10000|
600028|中国石化| 2000|1|0.10000|
600029|南方航空| 1000|1|0.10000|
600030|中信证券| 3300|1|0.10000|
600031|三一重工| 1400|1|0.10000|
600036|招商银行| 5900|1|0.10000|
600048|保利地产| 1700|1|0.10000|
600050|中国联通| 4000|1|0.10000|
600058|五矿发展| 200|1|0.10000|
600064|南京高科| 200|1|0.10000|
600068|葛洲坝| 1000|1|0.10000|
600085|同仁堂| 300|1|0.10000|

```

再以深证 100 为例，看某日价格数据，由于深 100ETF[159901]变动最小为 1 厘，看上去锯齿状很明显，而深证 100 指数[399330]数量级为 10^3 ，数据按比例缩小后可以与 ETF 数据非常接近，而中间的变化可以小于 1 厘，这就造成 ETF 与指数之间实际上总存在价差，只是大小的问题。



每轮交易的记录提取及算法

为了分析交易策略，首先需要提取出完成每次套利交易所涉及到的交易记录。由于数据数量巨大，因此需编写程序来判断每一条交易记录是属于哪一次套利交易，并将属于不同次交易的记录归类。

通过分析数据，可以得出绝大多数情况交易员在完成一整次套利交易后才会进行下一次套利，因此出现在一次套利交易中穿插另外一次套利交易的记录情况极少，可以不予考虑，在以下的算法中只考虑大多数情况下的套利交易。套利交易分为两种，一种是先买入股票，再申购，最后卖出 ETF，另一种是先买入 ETF，再赎回，最后卖出股票。由于每条记录中包含此条记录是对哪只股票或 ETF 进行操作，进行何种操作，以及进行操作的时刻，因此只需

要通过程序得到以下几个记录即可对每次操作有清晰的认识：1. 每次交易的开始点的记录及结束点的记录 2. 申购（或赎回）开始及结束的记录

算法：

1. 判断出每笔交易 ETF 申购或赎回开始和结束的记录

- (1) 判断第一条记录是否为申购（或赎回），若是则(2)，若不是，则判断下一条记录是否为申购（或赎回）
- (2) 判断此（1）中记录的上一条是否也为对相同 ETF 的申购（或赎回），若不是，则此申购赎回记录为此次交易申购（或赎回）的开始点（记为 a）。若是，则继续考虑其上一条记录，直到找到开始点。
- (3) 考虑开始记录的下一条记录，若此记录为对于相同 ETF 的申购（或赎回），则再考虑下一条记录（记为 c），以此向下，直到找到一条记录不是对同支 ETF 的申购（或赎回）（记为 x），则 x 的上一条记录为 ETF 申购（或赎回）的结束点（记为 x-1）。

2. 判断出每笔交易的开始点的记录及结束点的记录

分为两种情况：交易为申购还是赎回

申购：

- (1) 由 1 知 ETF 申购的开始点记录 a，则先于 a 的且最邻近 a 的多条记录应为股票买入，因此从记录 a 向上寻找不是买入股票的记录，遇到的第一个不是股票买入的记录（记为 y），y 的下一条记录为此次 ETF 交易的起点。
- (2) 由 1 知 ETF 申购的结束点记录为 x-1，则晚于 x-1 且最邻近 x-1 的多条记录应为卖出 ETF，因此从记录 x-1 向下寻找非卖出 ETF 的记录，遇到的第一个不是卖出 ETF 的记录（记为 z），z 的上一条记录为交易结束点记录。

赎回：

- (1) 由 1 知 ETF 赎回的开始点记录 a，则先于 a 的且最邻近 a 的多条记录应为 ETF 买入，因此从记录 a 向上寻找不是买入 ETF 的记录，遇到的第一个不是 ETF 买入的记录（记为 p），p 的下一条记录为此次 ETF 交易的起点。
- (2) 由 1 知 ETF 赎回的结束点记录为 x-1，则迟于 x-1 且最邻近 x-1 的多条记录应为卖出股票，因此从记录 x-1 向下寻找非卖出股票的记录，遇到的第一个不是卖出股票的记录（记为 q），q 的上一条记录为交易结束点记录。

举例说明：

1	ORDERID	ORDERTIME	REPORTTIME	MARKETNAME	STKCODE	STKNAME	PRODCODE	ESFLAGNAME
2	Q4021097	14410115	14410116	深A	000839	中信国安		证券卖出
3	Q4021098	14410202	14410203	深A	002001	新 和 成		证券卖出
4	Q4021099	14410318	14410319	深A	002069	獐 子 岛		证券卖出
5	Q4021624	14501177	14501177	深A	159901	深100ETF		证券买入
6	Q4021632	14510304	14510304	深A	159901	深100ETF		证券买入
7	Q4021633	14510306	14510308	深A	159901	深100ETF		证券买入
8	Q4021636	14511830	14511838	深A	159901	深100ETF		ETF赎回
9	Q4021638	14512485	14512495	深A	159901	深100ETF		ETF赎回
10	Q4021864	14553998	14553998	深A	000001	深发展A		证券卖出
11	Q4021865	14553999	14554000	深A	000422	湖北宜化		证券卖出
12	Q4021866	14554001	14554002	深A	000651	格力电器		证券卖出
13	Q4021869	14554007	14554007	深A	002008	大族激光		证券卖出
14	Q4021872	14554015	14554016	深A	000858	五 粮 液		证券卖出
15	Q4021874	14554020	14554020	深A	002024	苏宁电器		证券卖出
16	Q4021875	14554022	14554022	深A	000425	徐工机械		证券卖出
17	Q4021878	14554032	14554033	深A	000012	南 玻A		证券卖出
18	Q4021892	14554058	14554060	深A	000878	云南铜业		证券卖出
19	Q4021895	14554064	14554065	深A	000538	云南白药		证券卖出
20	Q4021896	14554066	14554067	深A	000728	国元证券		证券卖出
21	Q4021900	14554073	14554074	深A	000539	粤电力A		证券卖出
22	Q4021904	14554081	14554081	深A	002106	莱宝高科		证券卖出
23	Q4021907	14554086	14554087	深A	000917	电广传媒		证券卖出
24	Q4021911	14554095	14554095	深A	000762	西藏矿业		证券卖出
25	Q4021916	14554106	14554107	深A	000768	西飞国际		证券卖出
26	Q4021918	14554111	14554111	深A	000060	中金岭南		证券卖出
27	Q4021924	14554123	14554123	深A	002146	荣盛发展		证券卖出
28	Q4021926	14554128	14554128	深A	000780	平庄能源		证券卖出
29	Q4021929	14554136	14554137	深A	002155	辰州矿业		证券卖出
30	Q4021930	14554137	14554138	深A	000568	泸州老窖		证券卖出
31	Q4021932	14554141	14554141	深A	000960	锡业股份		证券卖出
32	Q4021935	14554147	14554147	深A	000581	威孚高科		证券卖出
33	Q4021939	14554156	14554157	深A	002233	塔牌集团		证券卖出
34	Q4021945	14554169	14554169	深A	000623	吉林敖东		证券卖出
35	Q4021947	14554172	14554174	深A	000970	中科三环		证券卖出
200	Q4022854	14565966	14565967	深A	000069	华侨城A		证券卖出
201	Q4022855	14565973	14565973	深A	002007	华兰生物		证券卖出
202	Q4022856	14565977	14565980	深A	000100	TCL 集团		证券卖出
203	Q4022857	14565984	14565985	深A	000157	中联重科		证券卖出
204	Q4022858	14565992	14565993	深A	000401	冀东水泥		证券卖出
205	Q4022859	14570000	14570000	深A	000402	金 融 街		证券卖出
206	151700004	9213842	9213843	沪A	601988	中国银行		证券买入
207	151700004	9215687	9215688	沪A	600816	安信信托		证券买入
208	151700004	9221793	9221793	沪A	600816	安信信托		证券买入

上图为客户交易数据的截图，从第 35 行到 208 行都为证券卖出，此处省略中间部分股票。

对于以上例子，运用算法

1. 判断出每笔交易 ETF 申购或赎回开始和结束的记录

- (1) 从第二行开始判断，是否为申购或赎回操作。发现第 8 行为 ETF 赎回。
- (2) 从第 8 行向上找，发现第 7 行不是 ETF 赎回操作，则说明第 8 行为 ETF 赎回的开始点记录。
- (3) 从第 8 行向下找，发现第 9 行仍为 ETF 赎回操作，则考虑第 10 行，第 10 行不为 ETF 赎回操作，因此第 9 行为 ETF 赎回的结束记录。

2. 判断出每笔交易的开始点的记录及结束点的记录

此例为赎回的情况

- (1) 由 1 知第 8 行为 ETF 赎回的开始点，则考虑其上方记录，第 5、6、7 行都为证券买入，第 4 行不为证券买入，说明第 5 行为整次交易的开始点。

(2) 由 2 知第 9 行为 ETF 赎回的结束记录，考虑其后的记录，直到第 205（包括第 205 行都为证券卖出），第 206 不为证券卖出，因此 205 行为此次交易的结束记录。

ETF 申购或赎回	交易开始记录	申购或赎回开始记录	申购或赎回结束记录	交易结束记录	ETF 代码
4	13	15	15	225	510300
4	233	233	233	439	510300
4	443	444	444	444	159901
4	445	446	446	446	159901
4	649	650	650	650	159903
4	694	695	695	735	159903
4	753	754	754	754	159903
4	796	796	796	997	510300
3	998	1294	1294	1294	510300
4	1296	1298	1298	1330	159901
4	1360	1360	1360	1452	159901
4	1505	1507	1507	1710	510300
4	1711	1712	1712	1932	510300
4	1934	1936	1936	2141	510300
4	2143	2144	2144	2144	159901
4	2146	2147	2147	2347	159901
4	2348	2351	2351	2557	510300
4	2558	2559	2559	2661	159901
4	2662	2663	2663	2768	159901
4	2773	2774	2774	2874	159901
3	2875	3072	3072	3072	510300
3	3073	3270	3270	3272	510300
3	3274	3468	3469	3469	159901
4	3470	3470	3470	3470	159901
4	3471	3472	3472	3596	159901
3	3658	3755	3755	3762	159901
3	3763	3859	3859	3860	159902
3	3880	4077	4077	4077	510300
3	4078	4275	4275	4282	510300

图：20120702-0706 客户交易记录经上述算法处理后每次交易的提取

注：1.第一列：ETF 申购为 3，ETF 赎回为 4
2.第二列至第五列的数字表示记录所在原数据的行数

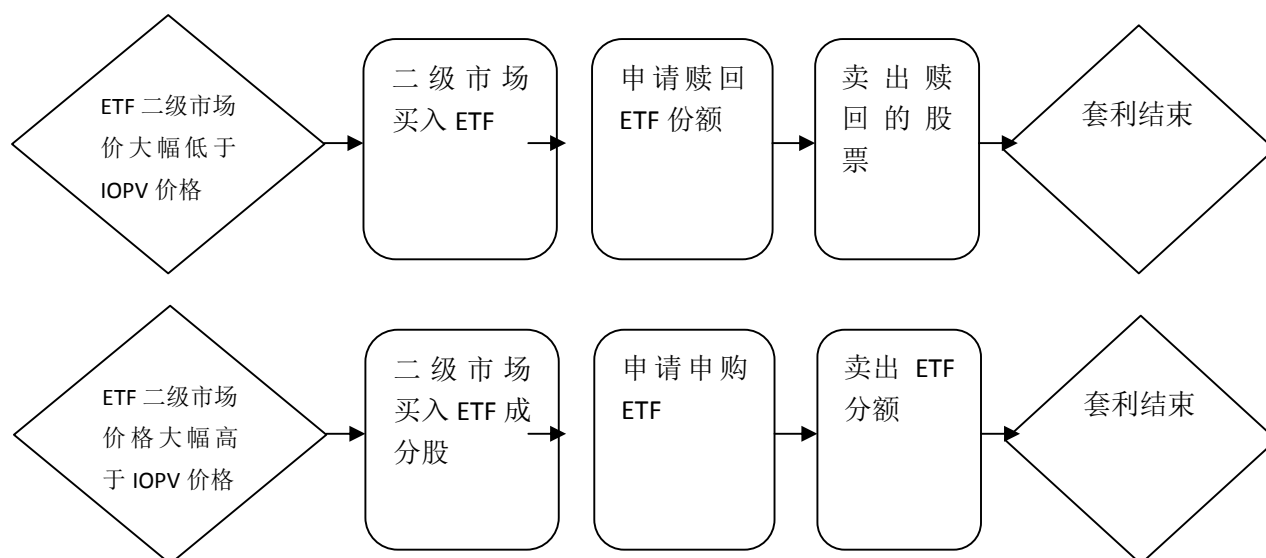
ETF 介绍

ETF 作为一种创新证券品种，能帮助投资者便捷实现指数化投资，并且管理和交易费用低廉，因而自 04 年底第一只 ETF 成立以来，广受市场欢迎，成交日趋活跃。2013 年，ETF 数量已经超越 50 支，市场中存在着各式各样 ETF 交易策略：

ETF 一二级市场价差套利

ETF 套利针对的是 ETF 一级市场申赎价格和二级市场交易价格的差异。当价差足够大时（能够覆盖交易成本），通过两个市场的申赎和买卖就能够实现无风险套利。根据价差的高低可以分为溢价套利和折价套利。

交易流程为



ETF 套利流程

交易是要付出成本的，包括：

- ETF 申购/赎回费用、经手费、过户费、证管费、证券结算金
- 印花税
- ETF 二级市场交易佣金、股票交易佣金
- 市场冲击成本

根据海通证券《量化策略研究》（2013）的研究，“目前，第一部分的交易成本大概在 3bp（1bp = 0.01%）左右，第二部分的印花税成本为 10 bp，只在折价套利卖出股票时缴纳，第三部分的佣金费率的高低则有赖客户同券商的谈判能力，我们假设 ETF 和股票交易的平均佣金费率为 5bp。这三部分成本的总和通常称为固定交易成本，按我们的费率估算数值，溢价套利的固定交易成本为 13bp，折价套利的固定交易成本为 23bp。”市场冲击成本则依赖于 ETF、股票的市场成交量和参与套利交易的资金量大小。由此可以看出，不含冲击成本

的交易成本应当在 13bp 到 23bp 之间。

延时套利

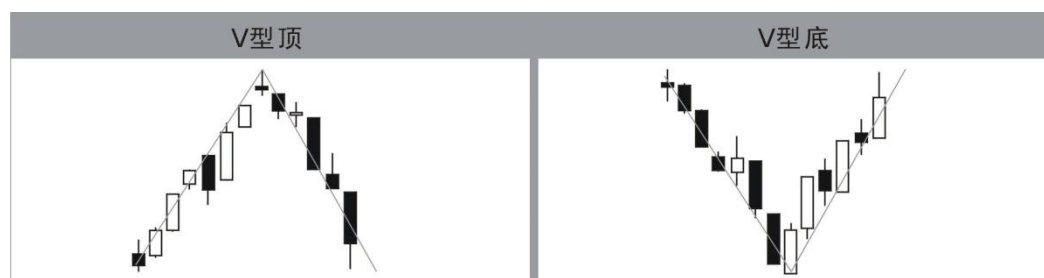
因为 50ETF 的申购赎回，需要 100 万单位作为最低限额，这样大单的买卖使得冲击成本变成 ETF 交易的一个主要成本。而通过算法交易的方法，进行延时下单交易，成为一个降低冲击成本的重要方法。也就是说，不是让大单一次买入或卖出，而是以一个较小的时间段来承受，从而起到减小冲击成本的作用。虽然下单时间长了，存在价格涨跌的不确定风险，但如果一笔成交，势必导致较大的冲击成本。关于冲击成本的影响，从长江证券《程序化交易》（P13, 2008）的研究中可以看出，它的存在可使原本达到 10.21% 的交易总收益率化为 -7.71%。所以延时套利其实是用不确定性来换取确定的冲击成本损失，也可以认为是快速超短线投机。

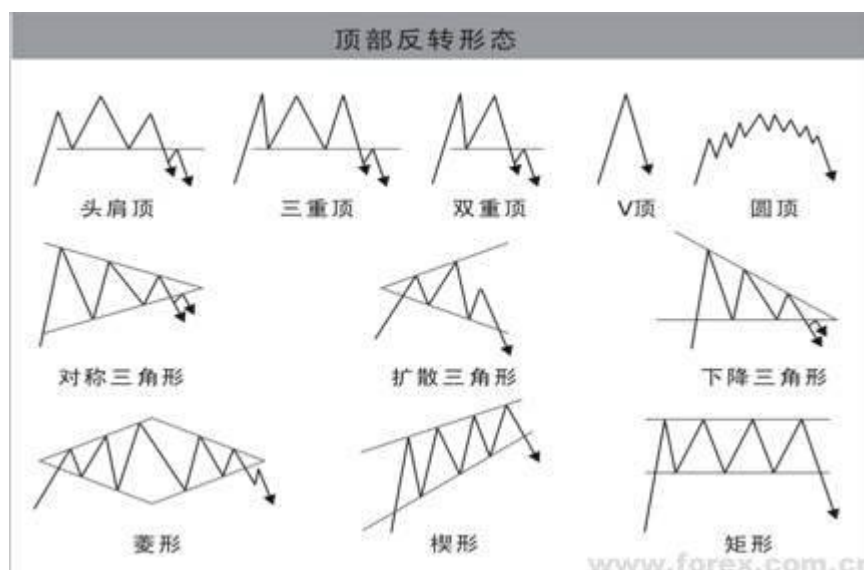
T+0 趋势交易

根据 ETF 形态或指标分析，确定入场出场点，从而达到获利的目的。其时间跨度比延时套利要长。在入场的时候可以利用价差选择买入股票还是买入 ETF。

具体策略

1. 形态识别。在上涨趋势的回调处入场，抓 V 型底等。

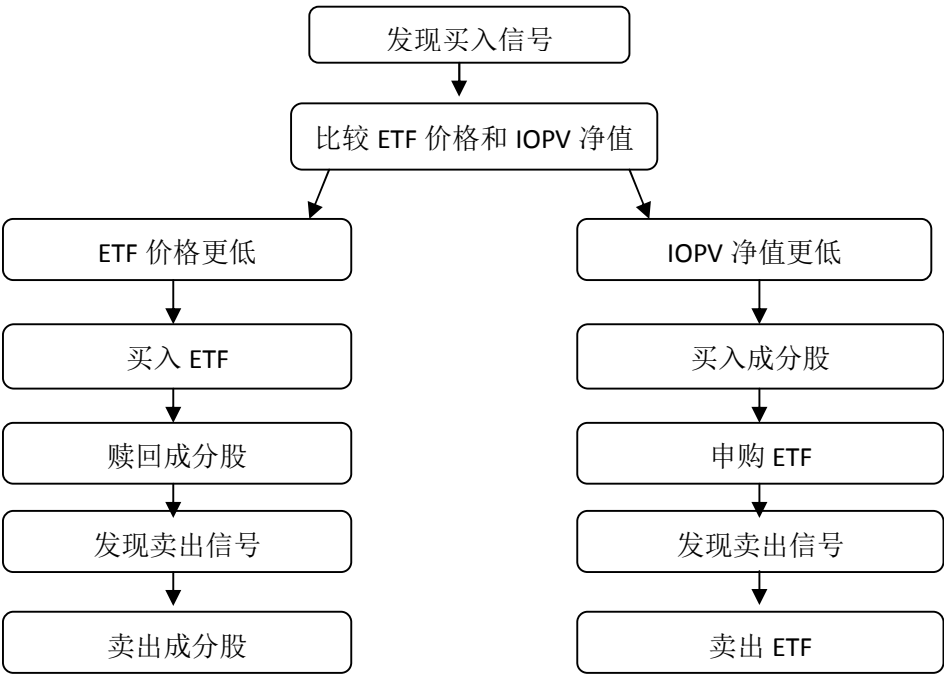




2. 高低点比较策略。如果指数运行刚出现了一个低点，而当前低点的指数高于前一个低点的指数，同时处于空仓状态，则认为出现了一个买点，这时买入指数，转变为持仓；如果指数运行出现一个高点，而当前高点的指数低于前一个高点的指数，同时处于持仓状态，则认为出现了一个卖点，这时卖出指数，转变为空仓。
3. 行情突破策略。在行情出现虽然当前高点要高于前一个高点，但是如果往下突破了上一次低点，那么在突破时卖出；如果当前低点虽然低于上次低点，但如果行情继续往上突破了上个高点，那么在突破时买入。
4. 行情滤波策略。如果布林带很窄，即认为行情进入了整理形态，暂时停止交易。同时继续观察，如果变宽超过该参数，则恢复到交易状态，同时如果其继续变小，小于某参数，则可能展开趋势，则应用布林极限等组合策略判断趋势展开的真伪，如果为真，则进入交易状态，如果判断为假，则维持原状态。
5. 对极点的识别。先画出指数的高频数据的轨迹，然后画出此轨迹的布林线，加减指数高频数据 2 倍标准差的轨迹，就得到上轨线和下轨线。如果指数在上次是出现高点后，往上首次突破上轨线，则认为此时是一个局部低点。如果指数在上次是出现低点后，往下首次突破下轨线，则认为此时是一个局部高点。

交易流程

程序通过分析当前的价格走势情况，给出买入和卖出的信号，然后通过判断比较 ETF 价格和 IOPV 净值来选择套利的类型。关于套利的方面，我们容易理解：由于 ETF 和其成分股在同一时间本应价格趋同，一旦二者不同，就可买入价格相对低的，申购或是赎回，再卖出相对高的，达到盈利。而这里需要注意的是，由于 ETF 不能卖空，我们只能在 ETF 价格上涨时做买入交易，至于是买入 ETF 还是其成分股，则由 ETF 价格和 IOPV 净值的大小决定。



T+0 趋势交易流程图

从这个意义上来说，我们可以知道，整个一次交易的开始与结束，主要与买卖信号有关，而如何得到这些买卖信号，则成了金融分析师的主要研究方向。

增强指数型策略

该策略要求投资者长期持有 ETF（或股票组合），当价格高于 IOPV 一定幅度时，将 ETF 卖出，并买入股票组合申购 ETF；当价格低于 IOPV 一定幅度时，卖出股票组合，买入 ETF 并

赎回得到股票组合。可见，与一般的套利流程相比，该策略的起点和终点发生了变化，并由“先买后卖”变成了“先卖后买”。该策略更适合长期看好 A 股的投资者。由于一直处于持仓状态，该策略有较高的系统性风险。

事件套利

当成份股停牌、涨（跌）停时，可利用 ETF 进行事件套利，套取看涨的股票或减持看空的股票。当某成份股停牌，投资者预期将公布利好消息，或某成份股已封涨停，投资者预期该股票在次交易日会继续大涨时，可买进 ETF，再将其赎回，并抛售其他成份股，从而套取以其他方式无法买入的股票；当某成份股停牌，投资者预期将公布利空消息，或某成份股已封跌停，投资者预期该股票在次交易日会继续大跌时，且投资者已持有该成份股，可买进其他成份股，并以此申购 ETF，再将卖出，从而顺利减持看空的股票（以其他方式无法抛售）。事件套利是一种方向性套利，当成份股后市的走势与投资者预期的一致时，可获利，否则，可能面临亏损。当然，投资者预期成份股的盈利（或规避的损失）与（有利的）价差要能够覆盖交易成本。因此，事件套利一般要求标的成份股的权重较高。但是，如果基金管理人将停牌的成分股设置为必须现金替代的证券，事件套利将无法进行。

策略对比表

		主要获利点	出现的次数	成本	成份股短期停牌的影响	成份股长期停牌的影响	风险
价差套利	折价套利	IOPV-P	较少	较高	成份股复牌后大涨，增厚套利收益，反之侵蚀套利收益	获得成份股只能在次交易日以溢价套利交易流程卖出，成本倍增	小
	溢价套利	P-IOPV	较少	较高	成份股复牌后大跌，增厚套利收益，反之侵蚀套利收益	长期占用资金(20个交易日以上)	小
T+0 趋势交易		短期涨幅	较多	较高	同上（视交易流程而异）	同上（视交易流程而异）	较大
增强指数型策略		价差及指数涨幅	较少	较高	同上（视交易流程而异）	同上（视交易流程而异）	大
事件套利		成份股的涨跌幅	较少	高	可利用停牌成份股获利	可利用停牌成份股获利	较大

从表中来看，而趋势交易机会较多，其他策略的机会较少。

另外，利用 ETF 做日内 T+0 交易有三个缺陷：

1. 一般只能做多头交易，除非是增强指数型策略；
 2. 交易成本高。卖出赎回的成分股时要支付股票交易佣金和印花税；
 3. 执行难度大。ETF 跟踪指数的成分股数量较多时，同时卖出大量股票的交易执行难度大，交易完成时间的不同会造成其对指数的跟踪误差；
- 可以利用融资融券克服这几个问题。

融资融券实现 T+0

用融资融券的交易方式，可以很便捷的实现 ETF 日内 T+0 交易，并避免上述三个问题。对于空头方向的 T+0 交易，可以先融券卖出相应 ETF，需要平仓的时候给交易系统下达“买券换券”指令，通过信用账户买入 ETF 并归还原先的融券，了结交易。这种日内交易方式不受交易次数的限制，而且由于是日内就完成了借券换券，投资者也不需要承担融资融券的利息，整个过程仅需支付 ETF 的交易佣金。对于多头方向的 T+0 交易，上证和深证的 ETF 由于两个交易所交易系统设置上的不同，操作方式上有较大差异。首先都是融资买入 ETF，在需要平仓时融券卖出同等数量的 ETF 份额，锁定收益。之后，上证交易的 ETF 可以马上选择“直接还券”，即用之前买入的 ETF 来偿还之后从券商融出的 ETF 份额，了结交易，不用支付融资利息。而深证交易的 ETF 则不行，必须等到下一个交易日开盘才能“直接还券”，因而投资者需承担期间每天万分之二点五左右的利息费用（融资融券的利息按自然日计算），而且在此期间投资者的资金一直被占用，不能反复操作。

数据分析

由于数据的限制，无法算出 ETF 与 IOPV 的价差，对于套利的判断依据不够，暂时的思路是把一篮子股票的交易看作是对 ETF 的直接交易，这样只看 ETF 数据，简化了原来问题，并推断交易方式。

对 50ETF2013 年 5 月 2 日至 2013 年 11 月 12 的数据进行计算，我们可以算出每分钟 ETF 价格变化的绝对值的均值为约为 0.001（0.001003154）元，即 10bp 左右。如果没有价差，每轮交易想要获利，即使没有冲击成本，买卖间隔平均要大于 1 分。

交易时间分析

通过分析 2013 年五月份（交易记录和 ETF 数据的重叠区间），所有的 ETF 和股票交易，我们可以得到一笔交易的时间间隔以及 ETF 申购赎回与 ETF 买卖的时间间隔的数据。处理后，我们将交易时间间隔的数据做成如下表所示。

表 1：2013 年 5 月交易用时统计

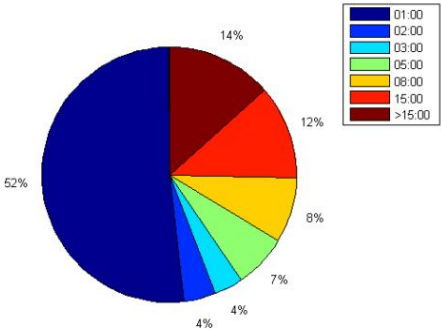
交易用时	次数	比例
<1min	73	50%
<2min	96	65%
<3min	109	74%
<4min	129	81%
<5min	129	88%
>5min	18	12%
总计	147	

可以看出 1 分钟之内的完成交易的次数占 50%，如果价差很小则很难赢利，延时套利在其中应占不小比例，而随着交易时间跨度的增长，投机性越来越强，超过 5 分钟跨度的交易占到 12%。

在对 2012 年上半年的数据进行统计我们得到如下表格，

表 2：2012 年上半年交易用时统计图 2012 年上半年交易用时比例图

交易用时	比例
<1min	54.8%
<2min	59.05%
<3min	62.97%
<5min	70.1%
<8min	78.87%
<15min	91.4%
>15min	100%



统计分析

通过统计，我们得到 2012 年以下几只 ETF 交易次数以及各自所占比例，具体如下表，

ETF 名称	ETF	交易比例
易方达深证 100ETF	159901	67.3%
华夏中小板 ETF	159902	3.7%
华安上证 180ETF	510180	3.3%
华泰柏瑞沪深 300ETF	510300	2.0%
南方深成 ETF	159903	0.7%
国联安上证商品 ETF	510170	0.7%
华夏上证 50ETF	510050	0.5%
广发中小板 300ETF	159907	0.4%
国泰上证 180 金融 ETF	510230	0.2%
鹏华深证民营	159911	0.1%
建信深证 60ET	159916	0.1%
博时深证 200ETF	159908	0.1%

易方达深圳 100ETF[159901]在所有交易中占相当大的比例。从易方达基金管理公司得到的数据分析表明，深证 100 指数的收益率及波动性（以年化标准差衡量）综合来看好于大部分其它主要指数，如下表所示：

指数	上证 180	上证 50	上证 A 指	深证 100	深证 A 指	深证成份 A 指
日均收益率	-0.05%	-0.04%	-0.05%	-0.03%	-0.05%	-0.03%
日均收益率标准差	1.32%	1.31%	1.35%	1.41%	1.45%	1.40%
年化标准差	20.38%	20.26%	20.85%	21.78%	22.50%	21.66%

数据来源：天相投资分析系统，统计区间为 2004~2005 年。

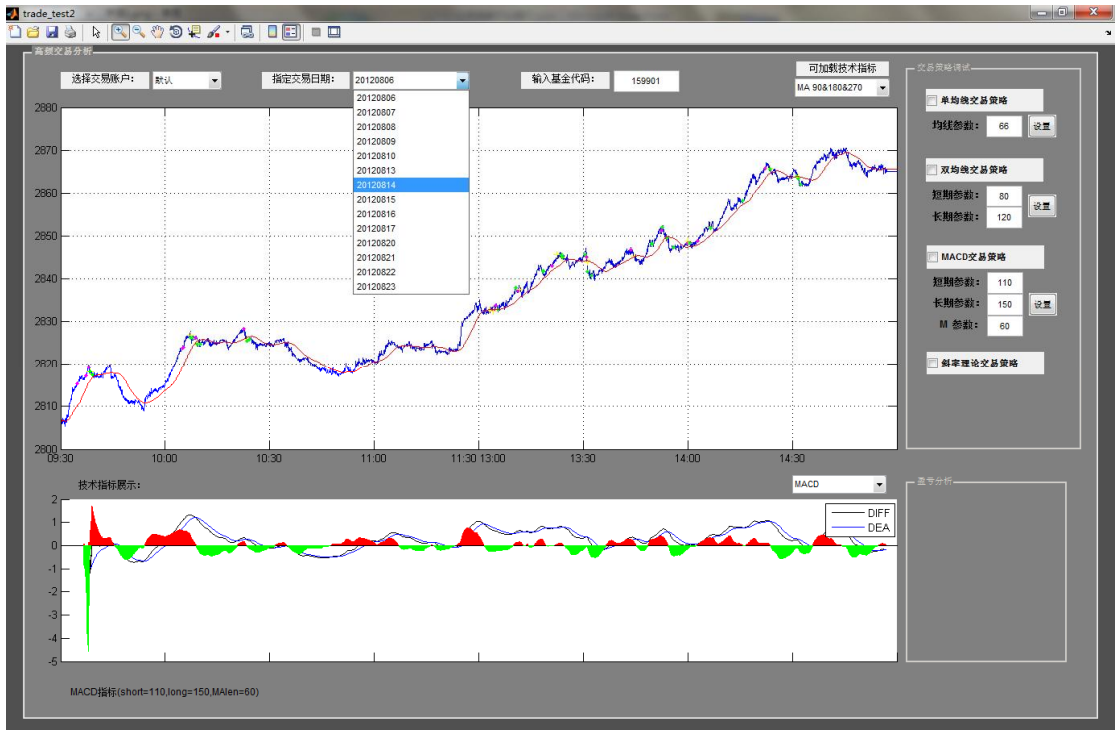
深证 100 指数更适合于短线波段操作，结合 ETF 上市交易、买卖便捷、费率低廉的特点，易方达深证 100ETF 为偏好频繁操作的短线交易者提供了良好的交易品种。

软件展示

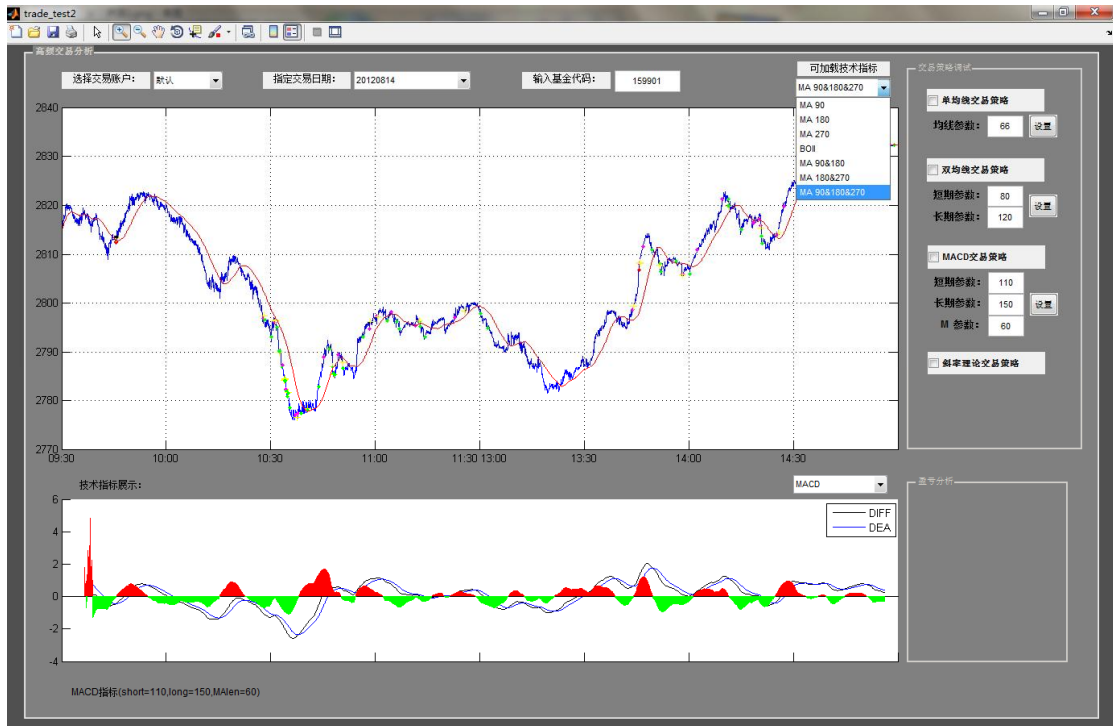
基于软件的分析使得我们可以拥有和交易者相近的视角，从而能够更加容易地判断其交易时所用的策略。我们利用近似的方法，将指数转化为 ETF 净值，与 ETF 价格图画在一起，

并标记上 ETF 交易时间和买卖方向等数据。由此可以形象地看到，每笔交易发生时的净值与价格之差等情况，从而更好的分析其使用的交易策略。

选择某交易日，整合分笔数据和交易数据



选择程序化交易策略



标出理论买卖点和实际买卖点



参考文献

- 【1】 朱剑涛、杨勇。量化策略研究：ETF 相关投资交易策略。海通证券，2013.1。
- 【2】 范辛亭、谭卓。程序化交易：指数日内交易策略研究----利用 ETF 的申购和赎回。长江证券，2008-11-18。
- 【3】 蒋瑛琨等。沪深 300ETF 套利策略分析。国泰君安证券销售交易总部，2010.2
- 【4】 邱小平。ETF 套利策略研究。浙商证券，2009.8