CHAPTER 11

Financial and Economic Data Applications

The use of Python in the financial industry has been increasing rapidly since 2005, led largely by the maturation of libraries (like NumPy and pandas) and the availability of skilled Python programmers. Institutions have found that Python is well-suited both as an interactive analysis environment as well as enabling robust systems to be developed often in a fraction of the time it would have taken in Java or C++. Python is also an ideal glue layer; it is easy to build Python interfaces to legacy libraries built in C or C++.

While the field of financial analysis is broad enough to fill an entire book, I hope to show you how the tools in this book can be applied to a number of specific problems in finance. As with other research and analysis domains, too much programming effort is often spent wrangling data rather than solving the core modeling and research problems. I personally got started building pandas in 2008 while grappling with inadequate data tools.

In these examples, I'll use the term *cross-section* to refer to data at a fixed point in time. For example, the closing prices of all the stocks in the S&P 500 index on a particular date form a cross-section. Cross-sectional data at multiple points in time over multiple data items (for example, prices together with volume) form a *panel*. Panel data can either be represented as a hierarchically-indexed DataFrame or using the three-dimensional Panel pandas object.

Data Munging Topics

Many helpful data munging tools for financial applications are spread across the earlier chapters. Here I'll highlight a number of topics as they relate to this problem domain.

Time Series and Cross-Section Alignment

One of the most time-consuming issues in working with financial data is the so-called *data alignment* problem. Two related time series may have indexes that don't line up perfectly, or two DataFrame objects might have columns or row labels that don't match. Users of MATLAB, R, and other matrix-programming languages often invest significant effort in wrangling data into perfectly aligned forms. In my experience, having to align data by hand (and worse, having to verify that data is aligned) is a far too rigid and tedious way to work. It is also rife with potential for bugs due to combining misaligned data.

pandas take an alternate approach by automatically aligning data in arithmetic operations. In practice, this grants immense freedom and enhances your productivity. As an example, let's consider a couple of DataFrames containing time series of stock prices and volume:

```
In [16]: prices
Out[16]:
              AAPL    JNJ      SPX    XOM
2011-09-06  379.74  64.64  1165.24  71.15
2011-09-07  383.93  65.43  1198.62  73.65
2011-09-08  384.14  64.95  1185.90  72.82
2011-09-09  377.48  63.64  1154.23  71.01
2011-09-12  379.94  63.59  1162.27  71.84
2011-09-13  384.62  63.61  1172.87  71.65
2011-09-14  389.30  63.73  1188.68  72.64

In [17]: volume
Out[17]:
                AAPL       JNJ       XOM
2011-09-06  18173500  15848300  25416300
2011-09-07  12492000  10759700  23108400
2011-09-08  14839800  15551500  22434800
2011-09-09  20171900  17008200  27969100
2011-09-12  16697300  13448200  26205800
```

Suppose you wanted to compute a volume-weighted average price using all available data (and making the simplifying assumption that the volume data is a subset of the price data). Since pandas aligns the data automatically in arithmetic and excludes missing data in functions like `sum`, we can express this concisely as:

```
In [18]: prices * volume
Out[18]:
                  AAPL         JNJ  SPX         XOM
2011-09-06  6901204890  1024434112  NaN  1808369745
2011-09-07  4796053560   704007171  NaN  1701933660
2011-09-08  5700560772  1010069925  NaN  1633702136
2011-09-09  7614488812  1082401848  NaN  1986085791
2011-09-12  6343972162   855171038  NaN  1882624672
2011-09-13         NaN         NaN  NaN         NaN
2011-09-14         NaN         NaN  NaN         NaN

In [19]: vwap = (prices * volume).sum() / volume.sum()
In [20]: vwap            In [21]: vwap.dropna()
Out[20]:                 Out[21]:
AAPL    380.655181       AAPL    380.655181
JNJ      64.394769       JNJ      64.394769
SPX            NaN       XOM      72.024288
XOM      72.024288
```

Since `SPX` wasn't found in `volume`, you can choose to explicitly discard that at any point. Should you wish to align by hand, you can use DataFrame's `align` method, which returns a tuple of reindexed versions of the two objects:

```
In [22]: prices.align(volume, join='inner')
Out[22]:
(              AAPL    JNJ    XOM
2011-09-06  379.74  64.64  71.15
2011-09-07  383.93  65.43  73.65
2011-09-08  384.14  64.95  72.82
2011-09-09  377.48  63.64  71.01
2011-09-12  379.94  63.59  71.84,
                AAPL       JNJ       XOM
2011-09-06  18173500  15848300  25416300
2011-09-07  12492000  10759700  23108400
2011-09-08  14839800  15551500  22434800
2011-09-09  20171900  17008200  27969100
2011-09-12  16697300  13448200  26205800)
```

Another indispensable feature is constructing a DataFrame from a collection of potentially differently indexed Series:

```
In [23]: s1 = Series(range(3), index=['a', 'b', 'c'])

In [24]: s2 = Series(range(4), index=['d', 'b', 'c', 'e'])

In [25]: s3 = Series(range(3), index=['f', 'a', 'c'])

In [26]: DataFrame({'one': s1, 'two': s2, 'three': s3})
Out[26]:
   one  three  two
a    0      1  NaN
b    1    NaN    1
c    2      2    2
d  NaN    NaN    0
e  NaN    NaN    3
f  NaN      0  NaN
```

As you have seen earlier, you can of course specify explicitly the index of the result, discarding the rest of the data:

```
In [27]: DataFrame({'one': s1, 'two': s2, 'three': s3}, index=list('face'))
Out[27]:
   one  three  two
f  NaN      0  NaN
a    0      1  NaN
c    2      2    2
e  NaN    NaN    3
```

Operations with Time Series of Different Frequencies

Economic time series are often of annual, quarterly, monthly, daily, or some other more specialized frequency. Some are completely irregular; for example, earnings revisions for a stock may arrive at any time. The two main tools for frequency conversion and realignment are the `resample` and `reindex` methods. `resample` converts data to a fixed frequency while `reindex` conforms data to a new index. Both support optional interpolation (such as forward filling) logic.

Let's consider a small weekly time series:

```
In [28]: ts1 = Series(np.random.randn(3),
   ....:              index=pd.date_range('2012-6-13', periods=3, freq='W-WED'))

In [29]: ts1
Out[29]:
2012-06-13   -1.124801
2012-06-20    0.469004
2012-06-27   -0.117439
Freq: W-WED
```

If you resample this to business daily (Monday-Friday) frequency, you get holes on the days where there is no data:

```
In [30]: ts1.resample('B')
Out[30]:
2012-06-13   -1.124801
2012-06-14         NaN
2012-06-15         NaN
2012-06-18         NaN
2012-06-19         NaN
2012-06-20    0.469004
2012-06-21         NaN
2012-06-22         NaN
2012-06-25         NaN
2012-06-26         NaN
2012-06-27   -0.117439
Freq: B
```

Of course, using `'ffill'` as the `fill_method` forward fills values in those gaps. This is a common practice with lower frequency data as you compute a time series of values on each timestamp having the latest valid or "*as of*" value:

```
In [31]: ts1.resample('B', fill_method='ffill')
Out[31]:
2012-06-13   -1.124801
2012-06-14   -1.124801
2012-06-15   -1.124801
2012-06-18   -1.124801
2012-06-19   -1.124801
2012-06-20    0.469004
2012-06-21    0.469004
2012-06-22    0.469004
2012-06-25    0.469004
2012-06-26    0.469004
```

```
2012-06-27   -0.117439
Freq: B
```

In practice, upsampling lower frequency data to a higher, regular frequency is a fine solution, but in the more general irregular time series case it may be a poor fit. Consider an irregularly sampled time series from the same general time period:

```
In [32]: dates = pd.DatetimeIndex(['2012-6-12', '2012-6-17', '2012-6-18',
   ....:                           '2012-6-21', '2012-6-22', '2012-6-29'])

In [33]: ts2 = Series(np.random.randn(6), index=dates)

In [34]: ts2
Out[34]:
2012-06-12   -0.449429
2012-06-17    0.459648
2012-06-18   -0.172531
2012-06-21    0.835938
2012-06-22   -0.594779
2012-06-29    0.027197
```

If you wanted to add the "as of" values in `ts1` (forward filling) to `ts2`. One option would be to resample both to a regular frequency then add, but if you want to maintain the date index in `ts2`, using `reindex` is a more precise solution:

```
In [35]: ts1.reindex(ts2.index, method='ffill')
Out[35]:
2012-06-12         NaN
2012-06-17   -1.124801
2012-06-18   -1.124801
2012-06-21    0.469004
2012-06-22    0.469004
2012-06-29   -0.117439

In [36]: ts2 + ts1.reindex(ts2.index, method='ffill')
Out[36]:
2012-06-12         NaN
2012-06-17   -0.665153
2012-06-18   -1.297332
2012-06-21    1.304942
2012-06-22   -0.125775
2012-06-29   -0.090242
```

Using periods instead of timestamps

Periods (representing time spans) provide an alternate means of working with different frequency time series, especially financial or economic series with annual or quarterly frequency having a particular reporting convention. For example, a company might announce its quarterly earnings with fiscal year ending in June, thus having `Q-JUN` frequency. Consider a pair of macroeconomic time series related to GDP and inflation:

```
In [37]: gdp = Series([1.78, 1.94, 2.08, 2.01, 2.15, 2.31, 2.46],
   ....:              index=pd.period_range('1984Q2', periods=7, freq='Q-SEP'))
```

```
In [38]: infl = Series([0.025, 0.045, 0.037, 0.04],
   ....:               index=pd.period_range('1982', periods=4, freq='A-DEC'))

In [39]: gdp          In [40]: infl
Out[39]:              Out[40]:
1984Q2    1.78        1982    0.025
1984Q3    1.94        1983    0.045
1984Q4    2.08        1984    0.037
1985Q1    2.01        1985    0.040
1985Q2    2.15        Freq: A-DEC
1985Q3    2.31
1985Q4    2.46
Freq: Q-SEP
```

Unlike time series with timestamps, operations between different-frequency time series indexed by periods are not possible without explicit conversions. In this case, if we know that `infl` values were observed at the end of each year, we can then convert to `Q-SEP` to get the right periods in that frequency:

```
In [41]: infl_q = infl.asfreq('Q-SEP', how='end')

In [42]: infl_q
Out[42]:
1983Q1    0.025
1984Q1    0.045
1985Q1    0.037
1986Q1    0.040
Freq: Q-SEP
```

That time series can then be reindexed with forward-filling to match `gdp`:

```
In [43]: infl_q.reindex(gdp.index, method='ffill')
Out[43]:
1984Q2    0.045
1984Q3    0.045
1984Q4    0.045
1985Q1    0.037
1985Q2    0.037
1985Q3    0.037
1985Q4    0.037
Freq: Q-SEP
```

Time of Day and "as of" Data Selection

Suppose you have a long time series containing intraday market data and you want to extract the prices at a particular time of day on each day of the data. What if the data are irregular such that observations do not fall exactly on the desired time? In practice this task can make for error-prone data munging if you are not careful. Here is an example for illustration purposes:

```
# Make an intraday date range and time series
In [44]: rng = pd.date_range('2012-06-01 09:30', '2012-06-01 15:59', freq='T')

# Make a 5-day series of 9:30-15:59 values
```

```
In [45]: rng = rng.append([rng + pd.offsets.BDay(i) for i in range(1, 4)])

In [46]: ts = Series(np.arange(len(rng), dtype=float), index=rng)

In [47]: ts
Out[47]:
2012-06-01 09:30:00    0
2012-06-01 09:31:00    1
2012-06-01 09:32:00    2
2012-06-01 09:33:00    3
...
2012-06-06 15:56:00    1556
2012-06-06 15:57:00    1557
2012-06-06 15:58:00    1558
2012-06-06 15:59:00    1559
Length: 1560
```

Indexing with a Python `datetime.time` object will extract values at those times:

```
In [48]: from datetime import time

In [49]: ts[time(10, 0)]
Out[49]:
2012-06-01 10:00:00      30
2012-06-04 10:00:00     420
2012-06-05 10:00:00     810
2012-06-06 10:00:00    1200
```

Under the hood, this uses an instance method `at_time` (available on individual time series and DataFrame objects alike):

```
In [50]: ts.at_time(time(10, 0))
Out[50]:
2012-06-01 10:00:00      30
2012-06-04 10:00:00     420
2012-06-05 10:00:00     810
2012-06-06 10:00:00    1200
```

You can select values between two times using the related `between_time` method:

```
In [51]: ts.between_time(time(10, 0), time(10, 1))
Out[51]:
2012-06-01 10:00:00      30
2012-06-01 10:01:00      31
2012-06-04 10:00:00     420
2012-06-04 10:01:00     421
2012-06-05 10:00:00     810
2012-06-05 10:01:00     811
2012-06-06 10:00:00    1200
2012-06-06 10:01:00    1201
```

As mentioned above, it might be the case that no data actually fall exactly at a time like 10 AM, but you might want to know the last known value at 10 AM:

```
# Set most of the time series randomly to NA
In [53]: indexer = np.sort(np.random.permutation(len(ts))[700:])
In [54]: irr_ts = ts.copy()

In [55]: irr_ts[indexer] = np.nan

In [56]: irr_ts['2012-06-01 09:50':'2012-06-01 10:00']
Out[56]:
2012-06-01 09:50:00    20
2012-06-01 09:51:00   NaN
2012-06-01 09:52:00    22
2012-06-01 09:53:00    23
2012-06-01 09:54:00   NaN
2012-06-01 09:55:00    25
2012-06-01 09:56:00   NaN
2012-06-01 09:57:00   NaN
2012-06-01 09:58:00   NaN
2012-06-01 09:59:00   NaN
2012-06-01 10:00:00   NaN
```

By passing an array of timestamps to the `asof` method, you will obtain an array of the last valid (non-NA) values at or before each timestamp. So we construct a date range at 10 AM for each day and pass that to `asof`:

```
In [57]: selection = pd.date_range('2012-06-01 10:00', periods=4, freq='B')

In [58]: irr_ts.asof(selection)
Out[58]:
2012-06-01 10:00:00      25
2012-06-04 10:00:00     420
2012-06-05 10:00:00     810
2012-06-06 10:00:00    1197
Freq: B
```

Splicing Together Data Sources

In Chapter 7, I described a number of strategies for merging together two related data sets. In a financial or economic context, there are a few widely occurring use cases:

- Switching from one data source (a time series or collection of time series) to another at a specific point in time
- "Patching" missing values in a time series at the beginning, middle, or end using another time series
- Completely replacing the data for a subset of symbols (countries, asset tickers, and so on)

In the first case, switching from one set of time series to another at a specific instant, it is a matter of splicing together two TimeSeries or DataFrame objects using `pandas.concat`:

```
In [59]: data1 = DataFrame(np.ones((6, 3), dtype=float),
   ....:                   columns=['a', 'b', 'c'],
   ....:                   index=pd.date_range('6/12/2012', periods=6))
```

```
In [60]: data2 = DataFrame(np.ones((6, 3), dtype=float) * 2,
   ....:                   columns=['a', 'b', 'c'],
   ....:                   index=pd.date_range('6/13/2012', periods=6))

In [61]: spliced = pd.concat([data1.ix[:'2012-06-14'], data2.ix['2012-06-15':]])

In [62]: spliced
Out[62]:
            a  b  c
2012-06-12  1  1  1
2012-06-13  1  1  1
2012-06-14  1  1  1
2012-06-15  2  2  2
2012-06-16  2  2  2
2012-06-17  2  2  2
2012-06-18  2  2  2
```

Suppose in a similar example that `data1` was missing a time series present in `data2`:

```
In [63]: data2 = DataFrame(np.ones((6, 4), dtype=float) * 2,
   ....:                   columns=['a', 'b', 'c', 'd'],
   ....:                   index=pd.date_range('6/13/2012', periods=6))

In [64]: spliced = pd.concat([data1.ix[:'2012-06-14'], data2.ix['2012-06-15':]])

In [65]: spliced
Out[65]:
            a  b  c   d
2012-06-12  1  1  1 NaN
2012-06-13  1  1  1 NaN
2012-06-14  1  1  1 NaN
2012-06-15  2  2  2   2
2012-06-16  2  2  2   2
2012-06-17  2  2  2   2
2012-06-18  2  2  2   2
```

Using `combine_first`, you can bring in data from before the splice point to extend the history for `'d'` item:

```
In [66]: spliced_filled = spliced.combine_first(data2)

In [67]: spliced_filled
Out[67]:
            a  b  c   d
2012-06-12  1  1  1 NaN
2012-06-13  1  1  1   2
2012-06-14  1  1  1   2
2012-06-15  2  2  2   2
2012-06-16  2  2  2   2
2012-06-17  2  2  2   2
2012-06-18  2  2  2   2
```

Since `data2` does not have any values for `2012-06-12`, no values are filled on that day.

DataFrame has a related method `update` for performing in-place updates. You have to pass `overwrite=False` to make it only fill the holes:

```
In [68]: spliced.update(data2, overwrite=False)

In [69]: spliced
Out[69]:
            a  b  c   d
2012-06-12  1  1  1 NaN
2012-06-13  1  1  1   2
2012-06-14  1  1  1   2
2012-06-15  2  2  2   2
2012-06-16  2  2  2   2
2012-06-17  2  2  2   2
2012-06-18  2  2  2   2
```

To replace the data for a subset of symbols, you can use any of the above techniques, but sometimes it's simpler to just set the columns directly with DataFrame indexing:

```
In [70]: cp_spliced = spliced.copy()

In [71]: cp_spliced[['a', 'c']] = data1[['a', 'c']]

In [72]: cp_spliced
Out[72]:
             a  b   c   d
2012-06-12   1  1   1 NaN
2012-06-13   1  1   1   2
2012-06-14   1  1   1   2
2012-06-15   1  2   1   2
2012-06-16   1  2   1   2
2012-06-17   1  2   1   2
2012-06-18 NaN  2 NaN   2
```

Return Indexes and Cumulative Returns

In a financial context, *returns* usually refer to percent changes in the price of an asset. Let's consider price data for Apple in 2011 and 2012:

```
In [73]: import pandas.io.data as web

In [74]: price = web.get_data_yahoo('AAPL', '2011-01-01')['Adj Close']

In [75]: price[-5:]
Out[75]:
Date
2012-07-23    603.83
2012-07-24    600.92
2012-07-25    574.97
2012-07-26    574.88
2012-07-27    585.16
Name: Adj Close
```

For Apple, which has no dividends, computing the cumulative percent return between two points in time requires computing only the percent change in the price:

```
In [76]: price['2011-10-03'] / price['2011-3-01'] - 1
Out[76]: 0.072399874037388123
```

For other stocks with dividend payouts, computing how much money you make from holding a stock can be more complicated. The adjusted close values used here have been adjusted for splits and dividends, however. In all cases, it's quite common to derive a *return index*, which is a time series indicating the value of a unit investment (one dollar, say). Many assumptions can underlie the return index; for example, some will choose to reinvest profit and others not. In the case of Apple, we can compute a simple return index using `cumprod`:

```
In [77]: returns = price.pct_change()

In [78]: ret_index = (1 + returns).cumprod()

In [79]: ret_index[0] = 1  # Set first value to 1

In [80]: ret_index
Out[80]:
Date
2011-01-03    1.000000
2011-01-04    1.005219
2011-01-05    1.013442
2011-01-06    1.012623
...
2012-07-24    1.823346
2012-07-25    1.744607
2012-07-26    1.744334
2012-07-27    1.775526
Length: 396
```

With a return index in hand, computing cumulative returns at a particular resolution is simple:

```
In [81]: m_returns = ret_index.resample('BM', how='last').pct_change()

In [82]: m_returns['2012']
Out[82]:
Date
2012-01-31    0.127111
2012-02-29    0.188311
2012-03-30    0.105284
2012-04-30   -0.025969
2012-05-31   -0.010702
2012-06-29    0.010853
2012-07-31    0.001986
Freq: BM
```

Of course, in this simple case (no dividends or other adjustments to take into account) these could have been computed from the daily percent changed by resampling with aggregation (here, to periods):

```
In [83]: m_rets = (1 + returns).resample('M', how='prod', kind='period') - 1

In [84]: m_rets['2012']
Out[84]:
Date
```

```
2012-01    0.127111
2012-02    0.188311
2012-03    0.105284
2012-04   -0.025969
2012-05   -0.010702
2012-06    0.010853
2012-07    0.001986
Freq: M
```

If you had dividend dates and percentages, including them in the total return per day would look like:

```
returns[dividend_dates] += dividend_pcts
```

Group Transforms and Analysis

In Chapter 9, you learned the basics of computing group statistics and applying your own transformations to groups in a dataset.

Let's consider a collection of hypothetical stock portfolios. I first randomly generate a broad *universe* of 2000 tickers:

```
import random; random.seed(0)
import string

N = 1000
def rands(n):
    choices = string.ascii_uppercase
    return ''.join([random.choice(choices) for _ in xrange(n)])
tickers = np.array([rands(5) for _ in xrange(N)])
```

I then create a DataFrame containing 3 columns representing hypothetical, but random portfolios for a subset of tickers:

```
M = 500
df = DataFrame({'Momentum' : np.random.randn(M) / 200 + 0.03,
                'Value' : np.random.randn(M) / 200 + 0.08,
                'ShortInterest' : np.random.randn(M) / 200 - 0.02},
                index=tickers[:M])
```

Next, let's create a random industry classification for the tickers. To keep things simple, I'll just keep it to 2 industries, storing the mapping in a Series:

```
ind_names = np.array(['FINANCIAL', 'TECH'])
sampler = np.random.randint(0, len(ind_names), N)
industries = Series(ind_names[sampler], index=tickers,
                    name='industry')
```

Now we can group by `industries` and carry out group aggregation and transformations:

```
In [90]: by_industry = df.groupby(industries)

In [91]: by_industry.mean()
Out[91]:
           Momentum  ShortInterest     Value
```

```
industry
FINANCIAL  0.029485      -0.020739  0.079929
TECH       0.030407      -0.019609  0.080113

In [92]: by_industry.describe()
Out[92]:
                   Momentum  ShortInterest       Value
industry
FINANCIAL count  246.000000     246.000000  246.000000
          mean     0.029485      -0.020739    0.079929
          std      0.004802       0.004986    0.004548
          min      0.017210      -0.036997    0.067025
          25%      0.026263      -0.024138    0.076638
          50%      0.029261      -0.020833    0.079804
          75%      0.032806      -0.017345    0.082718
          max      0.045884      -0.006322    0.093334
TECH      count  254.000000     254.000000  254.000000
          mean     0.030407      -0.019609    0.080113
          std      0.005303       0.005074    0.004886
          min      0.016778      -0.032682    0.065253
          25%      0.026456      -0.022779    0.076737
          50%      0.030650      -0.019829    0.080296
          75%      0.033602      -0.016923    0.083353
          max      0.049638      -0.003698    0.093081
```

By defining transformation functions, it's easy to transform these portfolios by industry. For example, standardizing within industry is widely used in equity portfolio construction:

```
# Within-Industry Standardize
def zscore(group):
    return (group - group.mean()) / group.std()

df_stand = by_industry.apply(zscore)
```

You can verify that each industry has mean 0 and standard deviation 1:

```
In [94]: df_stand.groupby(industries).agg(['mean', 'std'])
Out[94]:
           Momentum       ShortInterest       Value
               mean  std           mean  std   mean  std
industry
FINANCIAL         0    1              0    1      0    1
TECH             -0    1             -0    1     -0    1
```

Other, built-in kinds of transformations, like `rank`, can be used more concisely:

```
# Within-industry rank descending
In [95]: ind_rank = by_industry.rank(ascending=False)

In [96]: ind_rank.groupby(industries).agg(['min', 'max'])
Out[96]:
           Momentum       ShortInterest       Value
                min  max            min  max    min  max
industry
FINANCIAL         1  246              1  246      1  246
TECH              1  254              1  254      1  254
```

In quantitative equity, "rank and standardize" is a common sequence of transforms. You could do this by chaining together `rank` and `zscore` like so:

```
# Industry rank and standardize
In [97]: by_industry.apply(lambda x: zscore(x.rank()))
Out[97]:
<class 'pandas.core.frame.DataFrame'>
Index: 500 entries, VTKGN to PTDQE
Data columns:
Momentum         500  non-null values
ShortInterest    500  non-null values
Value            500  non-null values
dtypes: float64(3)
```

Group Factor Exposures

Factor analysis is a technique in quantitative portfolio management. Portfolio holdings and performance (profit and less) are decomposed using one or more *factors* (risk factors are one example) represented as a portfolio of weights. For example, a stock price's co-movement with a benchmark (like S&P 500 index) is known as its *beta*, a common risk factor. Let's consider a contrived example of a portfolio constructed from 3 randomly-generated factors (usually called the *factor loadings*) and some weights:

```
from numpy.random import rand
fac1, fac2, fac3 = np.random.rand(3, 1000)

ticker_subset = tickers.take(np.random.permutation(N)[:1000])

# Weighted sum of factors plus noise
port = Series(0.7 * fac1 - 1.2 * fac2 + 0.3 * fac3 + rand(1000),
              index=ticker_subset)
factors = DataFrame({'f1': fac1, 'f2': fac2, 'f3': fac3},
                    index=ticker_subset)
```

Vector correlations between each factor and the portfolio may not indicate too much:

```
In [99]: factors.corrwith(port)
Out[99]:
f1    0.402377
f2   -0.680980
f3    0.168083
```

The standard way to compute the factor exposures is by least squares regression; using `pandas.ols` with `factors` as the explanatory variables we can compute exposures over the entire set of tickers:

```
In [100]: pd.ols(y=port, x=factors).beta
Out[100]:
f1           0.761789
f2          -1.208760
f3           0.289865
intercept    0.484477
```

As you can see, the original factor weights can nearly be recovered since there was not too much additional random noise added to the portfolio. Using groupby you can compute exposures industry by industry. To do so, write a function like so:

```
def beta_exposure(chunk, factors=None):
    return pd.ols(y=chunk, x=factors).beta
```

Then, group by industries and apply that function, passing the DataFrame of factor loadings:

```
In [102]: by_ind = port.groupby(industries)

In [103]: exposures = by_ind.apply(beta_exposure, factors=factors)

In [104]: exposures.unstack()
Out[104]:
                 f1        f2        f3  intercept
industry
FINANCIAL  0.790329 -1.182970  0.275624   0.455569
TECH       0.740857 -1.232882  0.303811   0.508188
```

Decile and Quartile Analysis

Analyzing data based on sample quantiles is another important tool for financial analysts. For example, the performance of a stock portfolio could be broken down into quartiles (four equal-sized chunks) based on each stock's price-to-earnings. Using pandas.qcut combined with groupby makes quantile analysis reasonably straightforward.

As an example, let's consider a simple trend following or *momentum* strategy trading the S&P 500 index via the SPY exchange-traded fund. You can download the price history from Yahoo! Finance:

```
In [105]: import pandas.io.data as web

In [106]: data = web.get_data_yahoo('SPY', '2006-01-01')

In [107]: data
Out[107]:
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 1655 entries, 2006-01-03 00:00:00 to 2012-07-27 00:00:00
Data columns:
Open         1655  non-null values
High         1655  non-null values
Low          1655  non-null values
Close        1655  non-null values
Volume       1655  non-null values
Adj Close    1655  non-null values
dtypes: float64(5), int64(1)
```

Now, we'll compute daily returns and a function for transforming the returns into a trend signal formed from a lagged moving sum:

```
px = data['Adj Close']
returns = px.pct_change()
```

```
def to_index(rets):
    index = (1 + rets).cumprod()
    first_loc = max(index.notnull().argmax() - 1, 0)
    index.values[first_loc] = 1
    return index

def trend_signal(rets, lookback, lag):
    signal = pd.rolling_sum(rets, lookback, min_periods=lookback - 5)
    return signal.shift(lag)
```

Using this function, we can (naively) create and test a trading strategy that trades this momentum signal every Friday:

```
In [109]: signal = trend_signal(returns, 100, 3)

In [110]: trade_friday = signal.resample('W-FRI').resample('B', fill_method='ffill')

In [111]: trade_rets = trade_friday.shift(1) * returns
```

We can then convert the strategy returns to a return index and plot them (see Figure 11-1):

```
In [112]: to_index(trade_rets).plot()
```

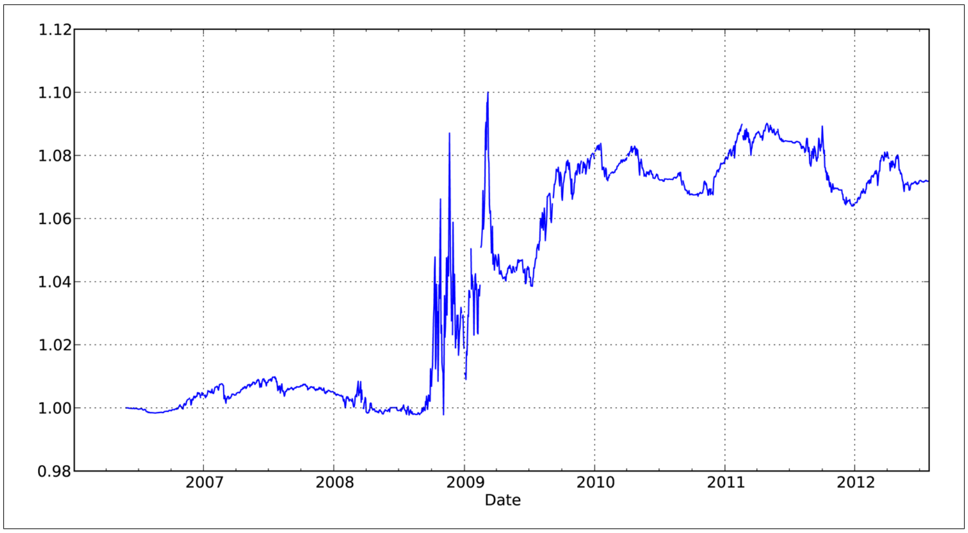

Figure 11-1. SPY momentum strategy return index

Suppose you wanted to decompose the strategy performance into more and less volatile periods of trading. Trailing one-year annualized standard deviation is a simple measure of volatility, and we can compute Sharpe ratios to assess the reward-to-risk ratio in various volatility regimes:

```
vol = pd.rolling_std(returns, 250, min_periods=200) * np.sqrt(250)
```

```
def sharpe(rets, ann=250):
    return rets.mean() / rets.std()  * np.sqrt(ann)
```

Now, dividing `vol` into quartiles with `qcut` and aggregating with `sharpe` we obtain:

```
In [114]: trade_rets.groupby(pd.qcut(vol, 4)).agg(sharpe)
Out[114]:
[0.0955, 0.16]    0.490051
(0.16, 0.188]     0.482788
(0.188, 0.231]   -0.731199
(0.231, 0.457]    0.570500
```

These results show that the strategy performed the best during the period when the volatility was the highest.

More Example Applications

Here is a small set of additional examples.

Signal Frontier Analysis

In this section, I'll describe a simplified cross-sectional momentum portfolio and show how you might explore a grid of model parameterizations. First, I'll load historical prices for a portfolio of financial and technology stocks:

```
names = ['AAPL', 'GOOG', 'MSFT', 'DELL', 'GS', 'MS', 'BAC', 'C']
def get_px(stock, start, end):
    return web.get_data_yahoo(stock, start, end)['Adj Close']
px = DataFrame({n: get_px(n, '1/1/2009', '6/1/2012') for n in names})
```

We can easily plot the cumulative returns of each stock (see Figure 11-2):

```
In [117]: px = px.asfreq('B').fillna(method='pad')

In [118]: rets = px.pct_change()

In [119]: ((1 + rets).cumprod() - 1).plot()
```

For the portfolio construction, we'll compute momentum over a certain lookback, then rank in descending order and standardize:

```
def calc_mom(price, lookback, lag):
    mom_ret = price.shift(lag).pct_change(lookback)
    ranks = mom_ret.rank(axis=1, ascending=False)
    demeaned = ranks - ranks.mean(axis=1)
    return demeaned / demeaned.std(axis=1)
```

With this transform function in hand, we can set up a strategy backtesting function that computes a portfolio for a particular lookback and holding period (days between trading), returning the overall Sharpe ratio:

```
compound = lambda x : (1 + x).prod() - 1
daily_sr = lambda x: x.mean() / x.std()
```

```
def strat_sr(prices, lb, hold):
    # Compute portfolio weights
    freq = '%dB' % hold
    port = calc_mom(prices, lb, lag=1)

    daily_rets = prices.pct_change()

    # Compute portfolio returns
    port = port.shift(1).resample(freq, how='first')
    returns = daily_rets.resample(freq, how=compound)
    port_rets = (port * returns).sum(axis=1)

    return daily_sr(port_rets) * np.sqrt(252 / hold)
```

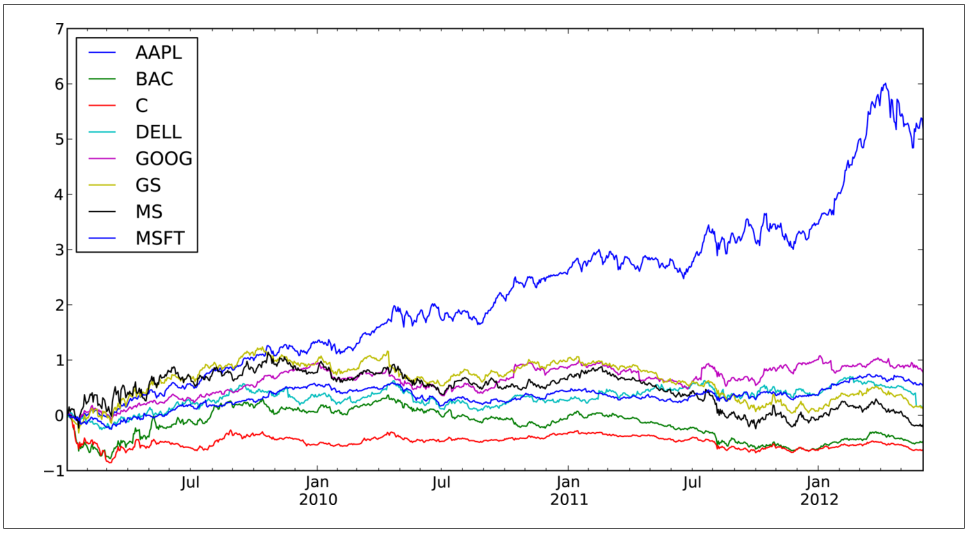

Figure 11-2. Cumulative returns for each of the stocks

When called with the prices and a parameter combination, this function returns a scalar value:

```
In [122]: strat_sr(px, 70, 30)
Out[122]: 0.27421582756800583
```

From there, you can evaluate the `strat_sr` function over a grid of parameters, storing them as you go in a `defaultdict` and finally putting the results in a DataFrame:

```
from collections import defaultdict

lookbacks = range(20, 90, 5)
holdings = range(20, 90, 5)
dd = defaultdict(dict)
for lb in lookbacks:
    for hold in holdings:
        dd[lb][hold] = strat_sr(px, lb, hold)
```

```
ddf = DataFrame(dd)
ddf.index.name = 'Holding Period'
ddf.columns.name = 'Lookback Period'
```

To visualize the results and get an idea of what's going on, here is a function that uses matplotlib to produce a heatmap with some adornments:

```
import matplotlib.pyplot as plt

def heatmap(df, cmap=plt.cm.gray_r):
    fig = plt.figure()
    ax = fig.add_subplot(111)
    axim = ax.imshow(df.values, cmap=cmap, interpolation='nearest')
    ax.set_xlabel(df.columns.name)
    ax.set_xticks(np.arange(len(df.columns)))
    ax.set_xticklabels(list(df.columns))
    ax.set_ylabel(df.index.name)
    ax.set_yticks(np.arange(len(df.index)))
    ax.set_yticklabels(list(df.index))
    plt.colorbar(axim)
```

Calling this function on the backtest results, we get Figure 11-3:

```
In [125]: heatmap(ddf)
```

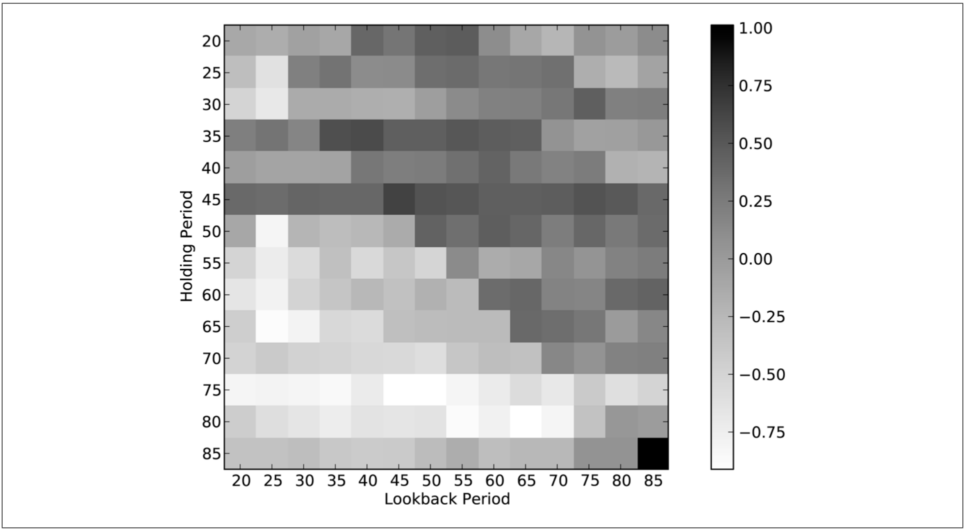

Figure 11-3. Heatmap of momentum strategy Sharpe ratio (higher is better) over various lookbacks and holding periods

Future Contract Rolling

A *future* is an ubiquitous form of derivative contract; it is an agreement to take delivery of a certain asset (such as oil, gold, or shares of the FTSE 100 index) on a particular date. In practice, modeling and trading futures contracts on equities, currencies, commodities, bonds, and other asset classes is complicated by the time-limited nature of each contract. For example, at any given time for a type of future (say silver or copper futures) multiple contracts with different *expiration dates* may be traded. In many cases, the future contract expiring next (the *near* contract) will be the most liquid (highest volume and lowest bid-ask spread).

For the purposes of modeling and forecasting, it can be much easier to work with a *continuous* return index indicating the profit and loss associated with always holding the near contract. Transitioning from an expiring contract to the next (or *far*) contract is referred to as *rolling*. Computing a continuous future series from the individual contract data is not necessarily a straightforward exercise and typically requires a deeper understanding of the market and how the instruments are traded. For example, in practice when and how quickly would you trade out of an expiring contract and into the next contract? Here I describe one such process.

First, I'll use scaled prices for the SPY exchange-traded fund as a proxy for the S&P 500 index:

```
In [127]: import pandas.io.data as web

# Approximate price of S&P 500 index
In [128]: px = web.get_data_yahoo('SPY')['Adj Close'] * 10

In [129]: px
Out[129]:
Date
2011-08-01    1261.0
2011-08-02    1228.8
2011-08-03    1235.5
...
2012-07-25    1339.6
2012-07-26    1361.7
2012-07-27    1386.8
Name: Adj Close, Length: 251
```

Now, a little bit of setup. I put a couple of S&P 500 future contracts and expiry dates in a Series:

```
from datetime import datetime
expiry = {'ESU2': datetime(2012, 9, 21),
          'ESZ2': datetime(2012, 12, 21)}
expiry = Series(expiry).order()
```

`expiry` then looks like:

```
In [131]: expiry
Out[131]:
ESU2    2012-09-21 00:00:00
ESZ2    2012-12-21 00:00:00
```

Then, I use the Yahoo! Finance prices along with a random walk and some noise to simulate the two contracts into the future:

```
np.random.seed(12347)
N = 200
walk = (np.random.randint(0, 200, size=N) - 100) * 0.25
perturb = (np.random.randint(0, 20, size=N) - 10) * 0.25
walk = walk.cumsum()

rng = pd.date_range(px.index[0], periods=len(px) + N, freq='B')
near = np.concatenate([px.values, px.values[-1] + walk])
far = np.concatenate([px.values, px.values[-1] + walk + perturb])
prices = DataFrame({'ESU2': near, 'ESZ2': far}, index=rng)
```

`prices` then has two time series for the contracts that differ from each other by a random amount:

```
In [133]: prices.tail()
Out[133]:
               ESU2     ESZ2
2013-04-16  1416.05  1417.80
2013-04-17  1402.30  1404.55
2013-04-18  1410.30  1412.05
2013-04-19  1426.80  1426.05
2013-04-22  1406.80  1404.55
```

One way to splice time series together into a single continuous series is to construct a weighting matrix. Active contracts would have a weight of 1 until the expiry date approaches. At that point you have to decide on a roll convention. Here is a function that computes a weighting matrix with linear decay over a number of periods leading up to expiry:

```
def get_roll_weights(start, expiry, items, roll_periods=5):
    # start : first date to compute weighting DataFrame
    # expiry : Series of ticker -> expiration dates
    # items : sequence of contract names

    dates = pd.date_range(start, expiry[-1], freq='B')
    weights = DataFrame(np.zeros((len(dates), len(items))),
                        index=dates, columns=items)

    prev_date = weights.index[0]
    for i, (item, ex_date) in enumerate(expiry.iteritems()):
        if i < len(expiry) - 1:
            weights.ix[prev_date:ex_date - pd.offsets.BDay(), item] = 1
            roll_rng = pd.date_range(end=ex_date - pd.offsets.BDay(),
                                     periods=roll_periods + 1, freq='B')

            decay_weights = np.linspace(0, 1, roll_periods + 1)
            weights.ix[roll_rng, item] = 1 - decay_weights
            weights.ix[roll_rng, expiry.index[i + 1]] = decay_weights
        else:
            weights.ix[prev_date:, item] = 1

        prev_date = ex_date
    return weights
```

The weights look like this around the `ESU2` expiry:

```
In [135]: weights = get_roll_weights('6/1/2012', expiry, prices.columns)

In [136]: weights.ix['2012-09-12':'2012-09-21']
Out[136]:
            ESU2  ESZ2
2012-09-12   1.0   0.0
2012-09-13   1.0   0.0
2012-09-14   0.8   0.2
2012-09-17   0.6   0.4
2012-09-18   0.4   0.6
2012-09-19   0.2   0.8
2012-09-20   0.0   1.0
2012-09-21   0.0   1.0
```

Finally, the rolled future returns are just a weighted sum of the contract returns:

```
In [137]: rolled_returns = (prices.pct_change() * weights).sum(1)
```

Rolling Correlation and Linear Regression

Dynamic models play an important role in financial modeling as they can be used to simulate trading decisions over a historical period. Moving window and exponentially-weighted time series functions are an example of tools that are used for dynamic models.

Correlation is one way to look at the co-movement between the changes in two asset time series. pandas's `rolling_corr` function can be called with two return series to compute the moving window correlation. First, I load some price series from Yahoo! Finance and compute daily returns:

```
aapl = web.get_data_yahoo('AAPL', '2000-01-01')['Adj Close']
msft = web.get_data_yahoo('MSFT', '2000-01-01')['Adj Close']

aapl_rets = aapl.pct_change()
msft_rets = msft.pct_change()
```

Then, I compute and plot the one-year moving correlation (see Figure 11-4):

```
In [140]: pd.rolling_corr(aapl_rets, msft_rets, 250).plot()
```

One issue with correlation between two assets is that it does not capture differences in volatility. Least-squares regression provides another means for modeling the dynamic relationship between a variable and one or more other predictor variables.

```
In [142]: model = pd.ols(y=aapl_rets, x={'MSFT': msft_rets}, window=250)

In [143]: model.beta
Out[143]:
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 2913 entries, 2000-12-28 00:00:00 to 2012-07-27 00:00:00
Data columns:
```

```
MSFT         2913  non-null values
intercept    2913  non-null values
dtypes: float64(2)

In [144]: model.beta['MSFT'].plot()
```

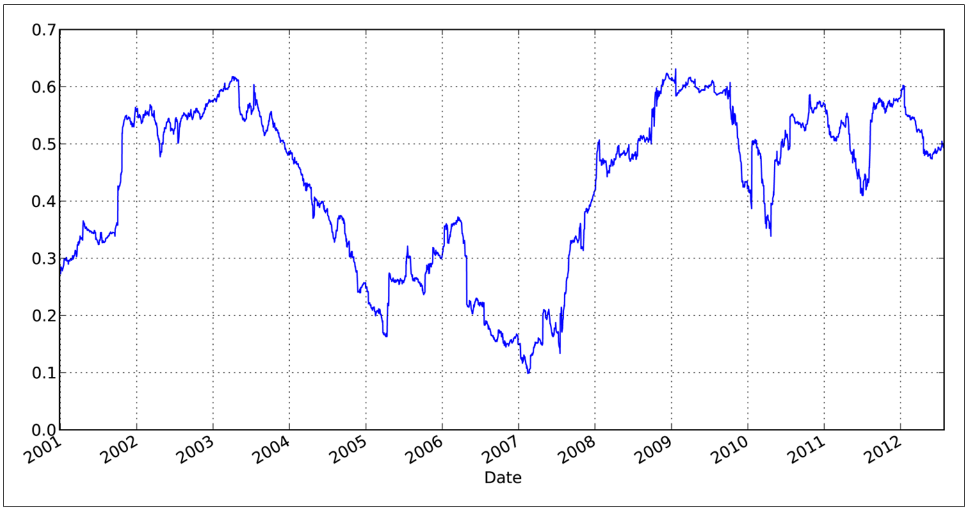

Figure 11-4. One-year correlation of Apple with Microsoft

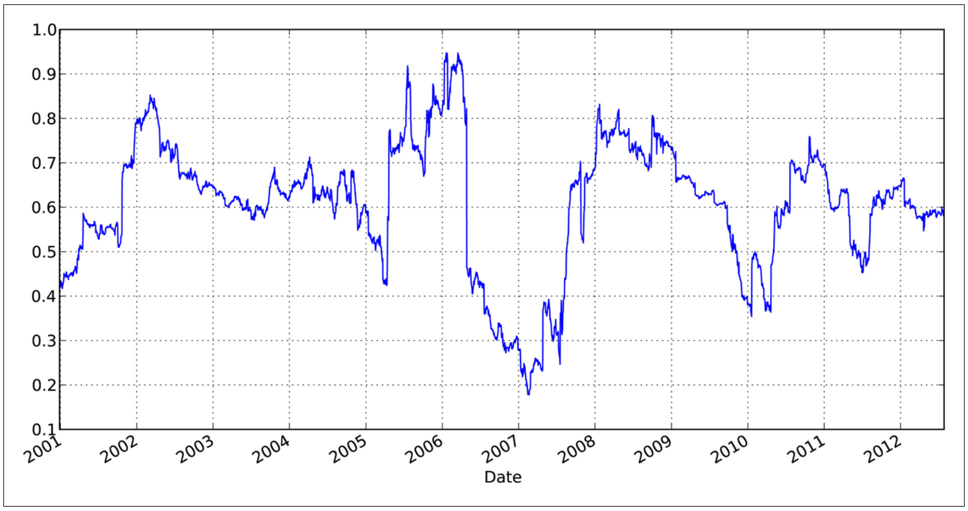

Figure 11-5. One-year beta (OLS regression coefficient) of Apple to Microsoft

pandas's `ols` function implements static and dynamic (expanding or rolling window) least squares regressions. For more sophisticated statistical and econometrics models, see the statsmodels project (*http://statsmodels.sourceforge.net*).