Automated Market Making

2014 Morgan Stanley Prize for Excellence

Jing Guo
PhD Candidate in Operation Research, Columbia University
December 31st, 2014

1 Problem and Model Formulation

1.1 Fundamental Value

Fundamental Value (FV) is the intrinsic value of a listed security determined through fundamental analysis without reference to its market value. I assume that at any time point t, bid/ask price of the security fluctuates around its time-t fundamental value. When bid size is significantly larger, i.e., $b_s >> a_s$, its FV is fairly close to the ask price, as most ask orders are executed already, and vice versa. Generally, FV is not necessary between its bid and ask price, further not between its buy and sell price, which potentially gives buy/sell signals.

1.2 FV Estimation

I estimate the security's fundamental value at time t by the following rule-of-thumb formula

$$\widehat{f}_t = \frac{b_{s,t}^{\alpha} a_{p,t} + a_{s,t}^{\alpha} b_{p,t}}{b_{s,t}^{\alpha} + a_{s,t}^{\alpha}}, \tag{1}$$

where $a_{p,t}$, $b_{p,t}$, $a_{s,t}$ and $b_{s,t}$ are the bid/ask price and bid/ask size at time t. α is a tuning parameter. This formula is consistent with the intuition that, when bid size $b_{s,t} >> a_{s,t}$, the FV estimate is fairly close to its ask price. I assume the logarithmic of $\widehat{f}_t$ is a random variable around its log fundamental value f_t at time t:

$$\log(\widehat{f}_t) = \log(f_t) + \sigma_t \, \varepsilon_t, \tag{2}$$

where I assume ε_t are i.i.d. standard-normally distributed. And σ_t is estimated as

$$\widehat{\sigma}_t^2 = \widehat{\sigma}^2(\log(\widehat{f}_t)) \approx \widehat{\sigma}^2(\widehat{(\log(f_t))}) \approx \frac{b_{s,t}^{2\alpha} a_{p,t} + a_{s,t}^{2\alpha} b_{p,t}}{(b_{s,t}^{\alpha} + a_{s,t}^{\alpha})^3} \cdot a_{s,t}^{\alpha} b_{s,t}^{\alpha} \cdot (\log(a_{p,t}) - \log(b_{p,t}))^2. \tag{3}$$

The last approximation is derived from the case when $\log(a_p)$ and $\log(b_p)$ are bernoulli distributed. This approximation is also consistent with the intuition that when log-spread is high, the estimation of FV becomes less accurate, i.e., we have larger estimation variance.

1.3 Log Difference FV State Process Dynamics

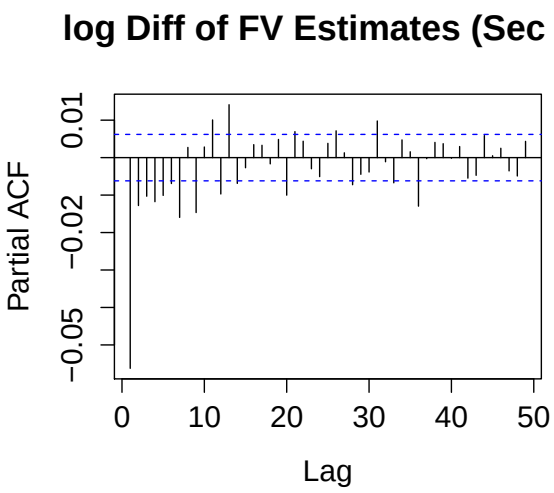
I assume the underlining logarithmic FV process $\log(f_t)$ follows some time series dynamics to be estimated. Define

$$y_t =: \log(\widehat{f}_t) - \log(\widehat{f_{t-1}}) \tag{4}$$

as the observation process, and log difference FV process

$$x_t =: \log(f_t) - \log(f_{t-1}) \tag{5}$$

Figure 1: PACF of log difference of security A FV estimates



as the state process. I assume that x_t satisfies the following ARMA(2,1) dynamics with time-dependent factor:

$$y_t = x_t + \widehat{\sigma}_t^2 \, \varepsilon_t, \tag{6}$$

$$x_t = \beta \cdot \text{factor}_t + \text{ar}_1 \cdot x_{t-1} + \text{ar}_2 \cdot x_{t-2} + z_t + \text{ma}_1 \cdot z_{t-1}, \tag{7}$$

where factor_t is the time-t factor to be determined. Figure 1 and 2 show the PACF and ACF of log difference series of the FV estimates of security A. Thresholds for p and q that the PCAF/ACF is significantly different from zero are 2 and 1, respectively. Therefore, it is reasonable the assume the ARMA(2,1) time series structure for the log difference of the FV process of some security.

According to the CAPM model, factor_t is the market return. I construct my own *market* based on the PCA loading matrix on the correlation matrix of security log returns. The first principle component loading coefficients are used as the weights of securities in the market. By specifying *market*=1 in my codes, users can see the model fitting and P&L results using *market* return as the factor.

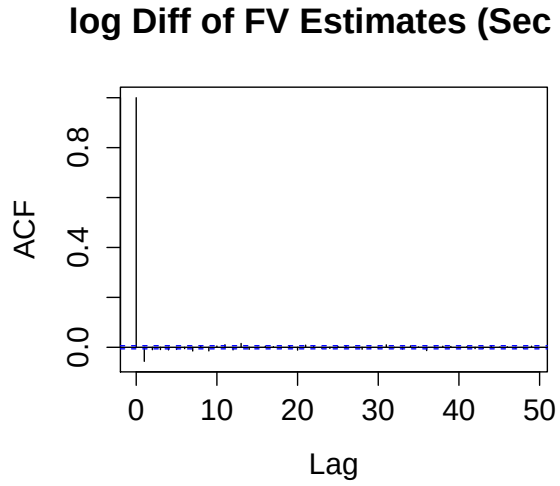
The weakness of using self-constructed market return is that there is too much noise in this constructed market, for which our model fitting is less accurate. It is empirically superior to use some other security price as the factor. I take the first principle component of logarithmic FV estimates and look for the security that influences the my target security most. The following table 1 is the PCA analysis of log FV estimates return of security A, B, C, D and E.

Table 1: Importance of components of log FV estimates of five securities (t=1:10,000):

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Cumulative Proportion	0.2141	0.4145	0.6124	0.8065	1.000

```
Loadings:
   Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
 A  0.484  0.270  0.475  0.566 -0.383
 B         -0.948 0.258  0.139 -0.128
```

Figure 2: ACF of log difference of security A FV estimates




```
 C  0.523        -0.388 -0.475 -0.585
 D  0.509 0.482  -0.481  0.526
 E  0.483 -0.146 -0.570  0.451  0.467
```

```
  [1] "corr matrix on log difference:"
             A            B           C           D          E
A  1.000000000 -0.005346592 0.020607527 0.029740669 0.01753387
B -0.005346592  1.000000000 0.001203938 0.003299628 0.00229523
C  0.020607527  0.001203938 1.000000000 0.024054703 0.03088039
D  0.029740669  0.003299628 0.024054703 1.000000000 0.01884054
E  0.017533869  0.002295230 0.030880389 0.018840541 1.00000000
```

Empirically, I take the log difference of security A price as the factor for security C and D, security D for security B, and set security E as the factor for security A. This choice of factors gives me the best trading P&L outcomes. See the detailed choice of factors in table 5.

1.4 Heteroskedasticity Analysis

There are heterogenous-variance model (ARCH, GARCH) available for time series analysis. Figure 3 plots the log difference of the FV estimates of security A. We can observe that the fluctuation levels of the process are roughly constant. Therefore, it is valid to just apply homogenous-variance model (ARMA) .

1.5 Price impact analysis

Some recent literature (for example, [1]) develops a linear model showing that prices changes are mainly driven by the imbalance between supply and demand at the best bid and ask prices. [1] suggests that this relationship is linear and varies every half an hour. I analyze the correlation between bid/ask prices changes and bid/ask size changes of security A in table 2.

We can see that there is significant price impact between bid/ask prices changes and bid/ask size changes. However, this relationship becomes insignificant when there is some lag. Moreover, the correlation becomes

Figure 3: Log difference of security A FV estimates

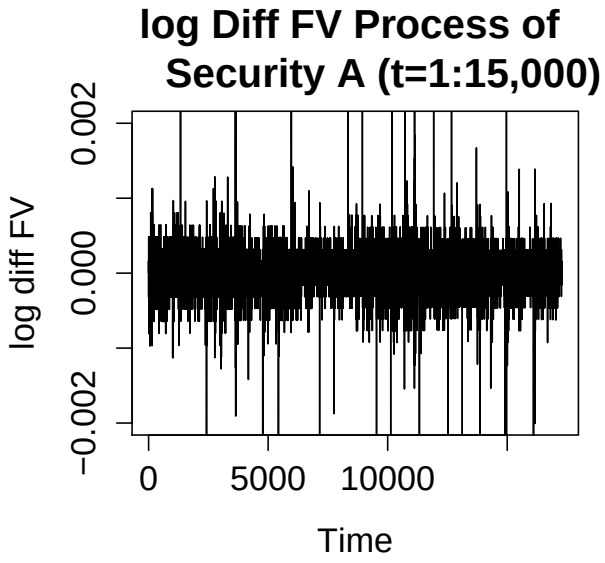


Table 2: Price impact analysis of security A (t=1:10,000)

Correlation Table	Δb_{p0} Vs Δb_{s0}	Δa_{p0} Vs Δa_{s0}
lag=0	-0.29707	0.38838
lag=-1	0.07767	-0.1018
lag=1	0.06186	-0.06982

much smaller as sample size increases. Because we only have information up to the current time, I do not consider price impact in my model.

2 Buy/Sell Signal Criteria and Model Estimation

2.1 Buy/Sell Signal Criteria

Denote buy/sell execution price at time t as $P_{buy,t}$ and $P_{sell,t}$

$$P_{buy,t} = b_{p,t} + (a_{p,t} - b_{p,t}) \cdot \sin(\frac{\pi}{2} \cdot \frac{b_{s,t}}{b_{s,t} + a_{s,t}}) \tag{8}$$

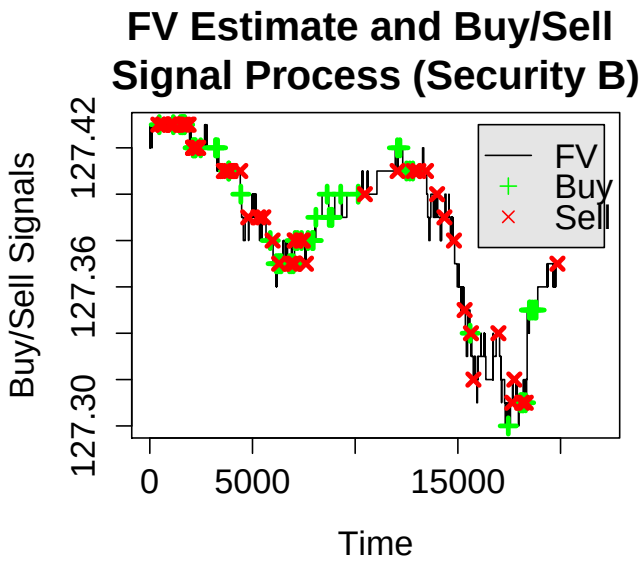
$$P_{sell,t} = a_{p,t} - (a_{p,t} - b_{p,t}) \cdot \sin(\frac{\pi}{2} \cdot \frac{a_{s,t}}{b_{s,t} + a_{s,t}}) \tag{9}$$

And $\mathrm{E}(x_t|y_t)$ and $\mathrm{Var}(x_t|y_t)$ are as the posterior expectation and posterior variance of hidden state x_t given by Kalman Filter technique.

Then a *buy* signal is given when

$$\Big(\log(\widehat{f}_t) + \mathrm{E}(x_t|y_t) + \widehat{\sigma}^2(\widehat{\log(f_t)}) + \widehat{\sigma}^2(\widehat{\log(f_{t-1})}) + \mathrm{Var}(x_t|y_t)\Big) / \log(P_{buy,t}) > c, \tag{10}$$

Figure 4: Out-of-sample buy/sell signal and price process of security B (t=10,001:100,000, factor=A)



and a *sell* signal appears when

$$\Big(\log(\widehat{f}_t) + \mathrm{E}(x_t|y_t) - \widehat{\sigma}^2(\widehat{\log(f_t)}) - \widehat{\sigma}^2(\widehat{\log(f_{t-1})}) - \mathrm{Var}(x_t|y_t)\Big) / \log(P_{sell,t}) < c, \tag{11}$$

where c is some threshold parameter. This idea comes from the concept of statistical significance that a significant observation lies out of some range of c standard deviations of my estimate. Intuitively, when my estimated FV is significantly *higher* than the buy price, I have a *buy* signal, and when my estimated FV is significantly *lower* than the buy price, I have a *sell* signal. In practice, I choose c to be 1 to have a balance of the number of trading signals and the accuracy of having a profitable strategy.

Figure 4 shows the buy/sell signal together with the FV estimate price process of security B when A is applied as its factor for time $t = 10,001 : 100,000$. We can see that under my criteria, *buy* signals are mostly given price is moving up, while *sell* signals appear most when price is going down.

2.2 Model Fitting

In this project I apply fast Kalman Filter package "FKF" to estimate model parameters β, ar_1, ar_2, ma_1 and σ. The followings in table 3 are typical estimation results of model parameters

Table 3: Parameter estimation of security B (t=1:10,000, factor=security A):

	β	ar_1	ar_2	ma_1	σ
Estimates	1.0390	.2357	.2068	.2535	1.808e(-8)

3 In-Sample and Out-of-Sample Trading Test

3.1 Code Instruction

```
MSSM.printPCA(tickerList, market)
```

MSSM.printPCA() prints the PCA and correlation matrix of the log-differences of different securities listed in *tickerList*. The first principle loading is used to construct market return with the loading coefficients being market weights. When *market*=0, users are to input the corresponding factor security from keyboard.

```
MSSM.fit(tickerList, m, alpha, type, market)
```

MSSM.fit() fits model parameters illustrated in the previous section using package "FKF" and function *fitPar()* written by me. Here m is the length of the training data. α is a tuning parameter. When *type*=2, I fit the model with ARAM(2,1) on log-difference FV estimates, and *type*=1 is for the model fitting with ARAM(2,1) on the log of FV estimates only. *market*=1 is for the case when the user prefers using *market* return constructed by PCA loadings as the factor, and *market*=0 is used when the factor is specified by user.

```
MSSM.test(tickerList, c, alpha, type, market)
```

MSSM.test() estimates the hidden FV states and gives buy/sell signals according to the criteria given in the previous section using function *test()* written by me. Here c is a tuning parameter. The usage of α, *type* and *market* are similar as before.

3.2 Choice of Parameters and P&L Result

Table 4: Empirical Choice of Parameters

	α	c	log-difference? (*type*)
Choice of Parameter	.5	1	Yes. (*type* = 2)

Table 5: Empirical Choice of Parameters (Cont')

	Security A	Security B	Security C	Security D	Security E
Factor Security	E	D	A	A	A

Table 6: Annualized Sharpe Ratio (FrequencyUnit="mins", in-sample size n=10,000, out-of-sample size n=100,000)

	In-sample	Out-sample
Sharpe Ratio	1.004	0.1134

4 Data and Coding Issues

There are six security price files A, B, C, D, E and F, where A and F are identical. I only work on security A-E.

Moreover, in row 136-140 and row 151 in file backtester.R, code

```
lastOB[lastOB$sym == ticker]
```

could potentially cause problems, and which I change into

```
lastOB[lastOB$sym == ticker, ].
```

References

[1] R Cont, A Kukanov, S Stoikov (2014), The price impact of order book events. *Journal of Financial Econometrics* **12** (1) 47-88.

[2] P Brockwell, R Davis (2014), Introduction to Time Series and Forecasting. *Springer Texts in Statistics*.