Final Exam

High-Frequency Financial Econometrics

Guilherme Salomé

December 3, 2018

The purpose of this exam is to test your knowledge of the contents discussed in class and in the projects, and to test your ability to work with new concepts.

Read carefully the instructions below.

The exam is Due on December 5th by 11:59 PM. You must push your local repository with all the required files (all MATLAB files and finished report in pdf) to GitHub before the deadline. The repository must contain all functions and scripts used in solving the exam. It should also contain a `main.m` file that generates all required plots for the exam when run. Finally, it should have a `report.pdf` containing your answers to the exam questions.

All results must be interpreted. Half of the work in the exam is doing the computations. The other half of the work is interpreting the results. You must interpret results regardless of whether the exercise explicitly asked for it or not.

All plots in the report must be self-contained. Self-contained means that a reader who only sees your figure (image and caption, but not the surrounding text) can understand what you are plotting. This translates to all plots having axis titles, correct units on the axis, and a caption that summarizes what is plotted.

The exam makes use of stock data. Refer to this page to get access to the data (requires Duke login). You must complete the exercises below for both of your stocks using 5 min. data, unless stated otherwise.

The data files follow the .csv format and contain the prices of different assets. The name of the file represents the ticker symbol for a given stock. For example, the file `AAPL.csv` contains data for Apple's stock. Each file has 3 columns (no headers): date, time and price. The first column of a file contains the date of a given price in the `YYYYMMDD` format. For example, a date of `20070103` means January 3rd of 2007. The second column contains the time of a given price in the `HHMM` or `HHMMSS` format. For example, a time of `935` means that the price in the 3rd column was recorded at 9:35 am. If the value is `93500`, it means the price was recorded at 9:35:00 am (this is only for 5 seconds data). The last column contains the price in dollars of the stock at the given date and time.

Students must uphold the Duke Community Standard. Projects with excessive overlap with other student's answers will receive a zero grade.

The exam is open book. You can use all texts, lecture notes and previous Matlab scripts you developed. Note that if your previous Matlab scripts were not completely correct and you use them when solving the Final Exam, leading to results that are not completely correct, then you may loose points in your solution.

Exam is strictly individual. You must not talk or communicate with each other or anyone else about the exam or course material. This restriction includes e-mail, voice,

phone, text, or any other means of communication. The answers you turn in must be your own. Failure to do so will lead to a zero grade.

By turning in the exam you agree to the conditions above.

Good luck!

Exercise 1 - Value at Risk

Suppose it is 9:00 the morning of September 16, 2008, in the middle of the financial crisis. You work for a mutual fund manager who has a \$100 million position in a particular stock, and the manager is worried about a very large loss over the upcoming trading day starting in the morning and ending at 16:00. Specifically, the manager asks you for the probability that the value of the position (ignoring rare jumps) will decline by 2 percent or worse over the day and for the probability that the value of the position will decline by 4 percent or worse over the day. The position size of \$100 million is very realistic and the potential losses of 2 or 4 percent translate to \$2 and \$4 million, respectively. The manager is asking you for:

$$\mathbb{P}(r_t^c \leq -q) \text{ for } q = 2\% \text{ or } q = 4\%$$

where $r_t^c = \sum_{i=1}^{n} r_{t,i}^c$ is the return over the day and $r_t^c \overset{d}{\sim} \mathcal{N}(0, IV_t)$.

A.

Identify the correct value of t in your data set for the date September 16, 2008. Remember you are in the morning of September 16 2008. Is it possible to estimate IV_t for that date? Why or why not?

B.

We do not know IV_t so we need to forecast its value. Given the previous values of the truncated variance $(TV_{t-1}, TV_{t-2}, \ldots)$, what would be a reasonable forecast for IV_t? Justify how you would find such forecast. To simplify the analysis, we will base our forecast on the white-noise model and use TV_{t-1} as the forecast for IV_t. Report the value of TV_{t-1} based on a separation threshold of $\alpha = 5$. How do you interpret the value of TV_{t-1}?

C.

For both of your stocks, estimate the probabilities described in the exercise. Also compute the confidence intervals around the TV_{t-1} to compute the 95% lower and upper bounds for the estimated probabilities. Keep in mind the unit of the returns $r_{t,i}^c$ and the unit of q. Fill in the table below: Comment on the estimation accuracy. If the width of the confidence interval is in the range 0.01–0.06, then that level of accuracy would likely suffice for most applications. However, if the width is something like 0.20–0.30, then the estimates would be too imprecise for any practical application.

		Estimate of $\mathbb{P}(r_t^c \leq -q)$		
Stock	q	Estimate	Lower bound	Upper bound
1	2			
1	4			
2	2			
2	2			

D.

Above, we fixed the loss (2% or 4%) and determined the probability of the loss. In risk management, it is common to turn the problem around by fixing the probability at some $p \in [0, 1]$ and determine the associated loss in dollars that could occur with probability fixed at p. This value is known as the Value at Risk (VaR). Keep in mind that if r_t^c is expressed as a percent, that is, $r_t^c \pm 1$ percent or $r_t^c \pm 2$ percent, the actual dollar gain or loss on an investment of $V = \$100$ million is $\pm \frac{r_t^c}{100} V$. We seek to compute a number Q such that:

$$\mathbb{P}\left(\frac{r_t^c}{100} \times V \leq Q\right) = p$$

Let's fix $p = 0.01$ (1 percent), so that for a random variable Z that is standard Gaussian:

$$\mathbb{P}(Z \leq Q) = 0.01 \text{ and } Q = -2.326$$

To get the Value at Risk, we need to put the probability in terms of a standard normal variable, and then solve for Q with $V = 100$. We would denote this value of Q for the $p = 0.01$ level VaR on $V = 100$ by $Q_{0.01}(100)$. Complete the table below for each of your stocks: Comment briefly o the accuracy of the VaR estimates. Would your boss find

		Estimate of VaR at $p = 0.01$ (1 percent)		
Stock	p	VaR	Lower bound VaR	Upper bound VaR
1				
2				

them helpful or uselessly imprecise?

Exercise 2 - Options and the Black-Scholes Model

The purpose of this exercise is to evaluate the BLS option pricing model by first computing the predicted option price using the realized variance for volatility, and then comparing the BLS model value for the option to the market price. Since the realized variance is the best estimate of volatility over the trading day, perhaps RV will do well in the options setting. Let's see what the data say.

In the data folder on Sakai you will find various files containing data on various options contracts. Each data file contains about 30–60 observations on the closing prices of put options with various strikes and expirations, along with some other data needed to implement the BLS formula. Each row of the file corresponds to a particular put option on the ETF SPY, for the S&P 500. The columns contain the following data:

Column	Description
1	YYYYMMDD date of the trading day
2	HHMM always 1600 (4 PM) since these are closing prices
3	S_0 closing price of SPY
4	interest rate in percent (0.02 means 0.02%)
5	dividend yield on SPY as a percent
6	closing price of the Put option in dollars (1.14 means \$1.14)
7	strike price K
8	days until expiration (36 days)
9	YYYYMMDD of expiration date (unused here)

In class, we discussed the BLS formula for the call option. It turns out there is generally more trading activity and economic interest in puts than calls. The likely reason is that the put option is a bit like insurance, in the sense that it guarantees the buyer of the put the minimum price of K, should the buyer happen to hold the stock and the stock price has collapsed at maturity, $K/S_T > 1$. Of course the trader on the other side of the put option, i.e., the trader who writes the put option, loses $K - S_T$ per contract in this case. The writer of the put is very much like an insurance company.

A.

Select a single file containing options data. Specify which file you chose below. Write a function to import the data to a matrix. Change the units to the appropriate units to be used to calculate option prices from the Black-Scholes model.

B.

Each line of the options data contains data at the market close time. Compute the realized variance for each of those days. Convert the values to the appropriate units.

C.

The data file has the closing market price for each put, $put_{mkt,i}$, along with the information needed to compute the BLS model-implied value, $put_{BLS,i}$. The value of $i = 1, 2, \ldots, N_t$ where N_t is the total number of available option prices for day t (the number of lines in a single file). Make a plot of the market price of the puts against their moneyness (strike-to-underlying ratio). Remember that moneyness can be computed as:

$$\text{moneyness} = \frac{K}{S_0}$$

Limit the values of the moneyness to be between 0.85 and 1.10.

D.

Use the function `BLScallput.m` to compute the prices $put_{BLS,i}$ using the realized variance (in the proper units) as the standard deviation input of the model. Add these prices to the plot from the previous item. Make sure to use a different color for the plot and add a legend.

E.

As a general rule, analysts and financial economists have found that BLS tends to underprice (under value) far out of the money put options, and usually, but by no means always, correctly prices put options close to the money: 0.98 < strike-to-underlying ratio < 1.02. Is your plot of prices against moneyness consistent with this finding?

F.

Choose 10 other options data files and repeat exercise E. above for each of them.

G.

Suppose you work for a firm in the financial services industry, and your job is "to make money," a topic that seems very important to students in this class. In the previous items you have collected BLS model prices along with the associated market prices. If you take the model seriously, then whenever $put_{BLS,i} < put_{mkt,i}$ the market is overvaluing the option relative to the model. The correct trading strategy would be to sell (write) the put options and thereby collect the option premium $put_{mkt,i}$, which could be held as cash while you wait for the option to expire worthlessly according to the BLS model. Would you feel comfortable going to your boss and indicating that the firm could make a lot of money by selling (writing) a $10 million of out-of-the-money put options on SPY? (Short answer)

Exercise 3 - Options and the Hull-White Model

In the previous exercise we saw that the Black-Scholes model implies that the value for out-of-the-money put options using realized volatility could be well below the market value. A problem with this model is that it treats volatility as constant until the expiration date, while we know that is not the case. The Hull-White (HW) options model accounts for random volatility, and it is very easy to implement. Let V denote the random annualized volatility between now and expiration of the option. The Hull-White price of an option is the expected value of the BLS price using V averaged over the distribution of V. A reasonable model is that V is log-normal. Thus, suppose that $\log V \overset{d}{\sim} \mathcal{N}\left(\log 40 - \frac{1.20^2}{2}, 1.20^2\right)$. The term being subtracted in the distribution is to ensure that $\mathbb{E}[V] = 40$. The Hull-White model can be implemented via simulation as follows:

- Simulate the volatility:

$$\tilde{V}_i = \mathrm{e}^{\log 40 - \frac{1.20^2}{2} + 1.20\tilde{Z}_i} \text{ where } \tilde{Z}_i \overset{d}{\sim} \mathcal{N}(0,1)$$

- Convert the volatility to appropriate units:

$$\tilde{\sigma}_i \equiv \frac{\tilde{V}_i}{100}$$

- Repeat the two steps above for $i = 1, 2, \ldots, 1000$.

- Compute the Hull-White price via:

$$\text{Hull-White Put price} = \frac{1}{1000}\sum_{i=1}^{1000} \text{BLS put price}(S_0, K, r, T, \tilde{\sigma}_i, q)$$

A.

Consider a 90-day put option with the following characteristics:

$S_0 = 50$ (current stock price in dollars)
$K =$ strike price in dollars, various values below
$r = 0.013$ (interest rate in decimals)
$T = 90/365$ (time to expiration in years)
$\sigma = 0.40$ (annualized volatility in decimals)
$q = 0.025$ (dividend yield in decimals)

Compute the BLS implied price for this put option (strike prices below) and its price implied by the Hull-White model. Fill the table below:

	Put Price implied by	
K	Black-Scholes	Hull-White
42.5		
44		
48		
50		
52		
55		

B.

Plot the prices implied by both models against the moneyness. Are the HW put prices higher than the BLS put prices? Does the random volatility help explain the under-pricing by the BLS model observed in the previous exercise?

C.

Repeat the exercise above using the actual options data from 10 files chosen at random. In this case also add the actual market price to the plots.

Exercise 4 - Microstructure Noise

In the data folder you can find data available at the 5 seconds frequency. Choose two of the various files available to work with in this exercise.

A.

Load the data using the appropriate values for n and T. Implement a function to compute the realized variance using coarse sampling. For now, assume the summation starts at the very first price. The function should take a parameter k_n that specifies the coarse sampling. For example, if we want to compute RV using a 5-minute sampling frequency, we would input $k_n = 60$.

B.

Use the function created in the previous item to make a volatility signature plot for your stock. For each day, compute RV using $k_n = 1, 2, 3, \ldots, 120$. Average the value of RV over the entire year for each k_n. Plot RV (in the appropriate units) against the sampling frequency (scale the axis to units of 1 minute). Using the theory discussed in class about noise interpret the results of the plot.

C.

Estimate the value of σ_χ^2 using the data at the highest frequency. Compute the estimate for each day of the sample. Explain the idea behind the $\hat{\sigma}_\chi^2$ estimator?

D.

When we sample at rate k_n there are $\lfloor n/k_n \rfloor$ terms in the RV sum. Therefore, the relative contribution of the noise to the total return variation RV is:

$$\mathit{contribution}(k_n) \equiv \frac{2\frac{n}{k_n}\sigma_\chi^2}{RV}$$

Using the material from the last two lectures, explain why $\mathit{contribution}(k_n)$ above is a measure of:

$$\frac{\text{Noise}}{IV + \text{Noise}}$$

at frequency k_n.

E.

Compute the average (across days) $\widehat{\mathit{contribution}}(k_n)$ for each frequency $k_n = 1, 2, \ldots, 120$ by substituting σ_χ^2 for $\hat{\sigma}_\chi^2$:

$$\text{average contribution}(k_n) = 100\frac{1}{T}\sum_{t=1}^{T}\text{contribution}(k_n, t)$$

Plot the average contribution against the frequency k_n.

Does the noise dominate RV at the very high frequencies in the sense of the contribution being close to 100%? The noise is considered unimportant at low frequency if it contributes less than 10% of the total variation at 5-min and 8-min. Is the noise unimportant at low frequencies in your data set?

F.

Modify the function that computes RV with coarse sampling so that it starts the summation at a price different from the first one. For example, in the case of coarse sampling from 5-seconds to 30-seconds, we could compute RV starting at 6 different values:

$$
\begin{aligned}
RV_0 &= (Y_6^n - Y_0^n)^2 + (Y_{12}^n - Y_6^n)^2 + \cdots + (Y_{L(0)}^n - Y_{L(0)-6})^2 \\
RV_1 &= (Y_7^n - Y_1^n)^2 + (Y_{13}^n - Y_7^n)^2 + \cdots + (Y_{L(1)}^n - Y_{L(1)-6})^2 \\
&\vdots \\
RV_5 &= (Y_{11}^n - Y_5^n)^2 + (Y_{17}^n - Y_{11}^n)^2 + \cdots + (Y_{L(5)}^n - Y_{L(5)-6})^2
\end{aligned}
$$

G.

Compute the two-scale realized variance estimator for the highest frequency $k_n = 1$ (5-seconds) for each day. Plot TSRV and RV based on 5-min coarse sampling against time.

H.

Compute TSRV for the frequencies $k_n = 1, 2, 3, 4, \ldots, 120$. Compute the average value of TSRV over all days, for each of the sampling frequencies. Add the values of TSRV to the volatility signature plot. Does TSRV deal with the microstructure noise? Is the curve relatively flat even when the sampling frequency increases? How does it compare to the usual RV estimator?