EXPLORING THE STRUCTURE OF CORRELATION

Jason Wei, Yoan Hassid, Joachim Edery, Forrest White, and Kevin Hsu

MS&E 444 – SPRING 2010

1. INTRODUCTION

Our paper begins by presenting an overview of various stylized facts that we expected to see in the structure of correlation . For instance, through studying the Epps effect we were able to confirm that our 15 minute intra-day close prices for stocks from the S&P500 were significant and could be used for further studies. Once the stylized facts had been established, we tried to address the question of: *how does one find an accurate model for the correlation but while reducing dimensionality of the model*?

We approached this problem by first looking at the common 1 factor model, which is known to be a relatively good predictor, and which essentially reduces the dimensionality of the problem from $O(n^2)$ to just $O(n)$. However, we noticed that the 1 factor model tends to do estimate the correlation poorly in certain sectors, and so we expanded the 1 factor model into a multi-factor model through analyzing the effects of the less significant components (i.e. the 2nd factor, 3rd factor, etc).

Finally, we acknowledged the fact that the standard correlation equation (as used in the 1 factor model) is meant to measure *linear* correlation between two stochastic processes, and that the true relationship between correlations of two stocks may or may not follow a linear one. Thus, we utilized the more general notion of a *copula* to try and understand the potentially non-linear relationships between two stocks.

2. STYLIZED FACTS

2.1. EPPS EFFECT

The Epps (1979) effect is a well-known phenomenon in financial markets with high frequency data. It shows that empirical correlations virtually disappear at high frequencies, while being significant at daily frequencies. It is associated with non-synchronous trading where prices do not arise simultaneously across markets but are separated by a few seconds. The Epps effect can be explained by the fact that physically, the formation of a price is not a continuous process but a pure jump process, and also by the delay of information transmission in the market.

The consequence of the Epps effect is that correlations on returns which are based on intervals below a certain limit are unreliable. Empirical studies have shown that the threshold is of the order 10/15 minutes. Therefore, before exploring the dynamics of correlation at different time scales, we have to be sure that the data are statistically significant.

To study the impact of the Epps effect, we used 15 minutes close prices for the S&P 500 from 01/03/2007 to 01/30/2009. We computed for a range of time scales (15min to 30*15min) the correlation of each pair of the index using different methods:

$$\begin{cases} \rho_{ij} = \rho_{ij}^{emp} \\ \rho_{ij} = \rho_{iM} \times \rho_{jM} \\ \rho_{ij} = \dfrac{\sigma_M^2}{\frac{1}{n}\sum_{k=1}^{n} \sigma_k^2} \end{cases}$$

The first equation corresponds to the actual empirical correlation computed on the whole data set. The second equation corresponds to the 1st factor approximation (described in PCA studies) where M represents the market (in our case we chose the average of the n stocks, n=100, as the market index). The correlation between stock *i*, and stock *j* is then the product of their correlation with the market. The third equation is the simplest approximation to correlation with all pair wise correlations assumed to be equal – it gives an idea of the average correlation in an index.

In terms of dimension complexity, the first problem is of order $\frac{n(n-1)}{2}$, while the second problem is of order n, and the third of order 1.

With this, we then computed the distribution of correlation in the index and studied the impact of time scale on the mean and standard deviation. The results show that the 1st factor model is a good approximation of the average empirical correlation at all time scales, but it also underestimates the standard deviation of the distribution. The "implied" correlation, however, is a poor approximation and highly underestimates the correlation of an index.

We also noticed an Epps effect for all measures of correlation as the average varies from 0.38 to 0.42 and the standard deviation from 0.095 to 0.118. Nevertheless, it seems that the correlation for 15 minutes returns is still significant and can be used for further studies.

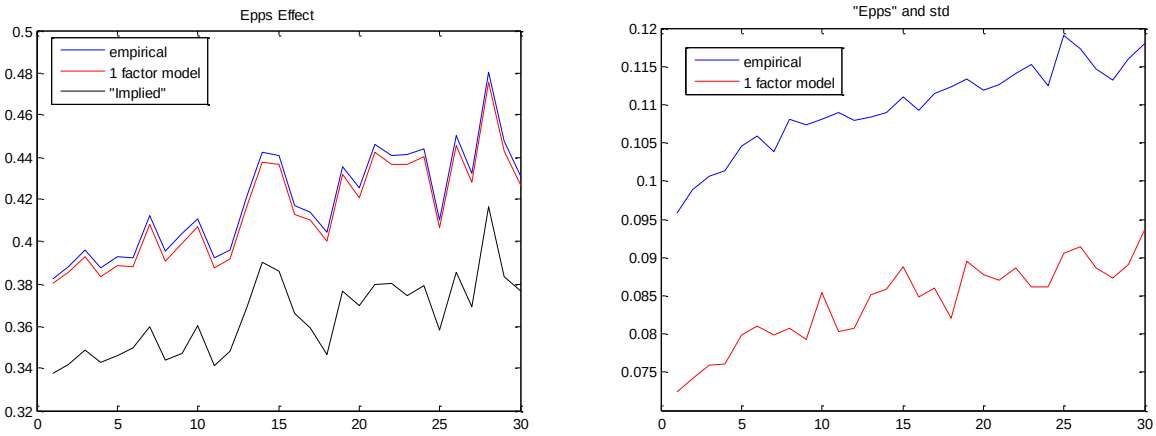


Figure 1 – Epps effects for Empirical, One Factor Model, and Implied Correlations

2.2. DISTRIBUTION OF CORRELATION

Now that we have studied the influence of time scale on the mean and standard deviation of the distribution of correlation for different models, we now wish to gain a better understanding of its shape and deformation.

As expected, the 1 factor approximation underestimates the tails of the distribution. To get an idea of the shape, we fit a normal distribution and a t-distribution to the empirical and 1 factor distributions.

The results show that a t-distribution fits the empirical data better due to the fat tails, whereas a normal distribution fits the 1 factor model better. Therefore, the underlying assumption by using the 1st factor approximation instead of the empirical correlations is that pair wise correlations are *normally distributed*.

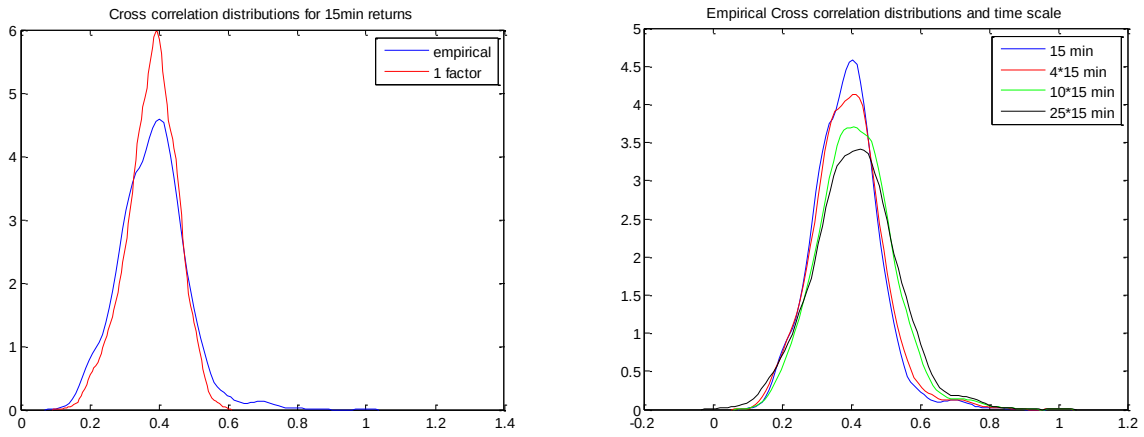


Figure 2 –Empirical vs. 1 factor model correlation (left), Empirical cross correlation with varying time scales (right)

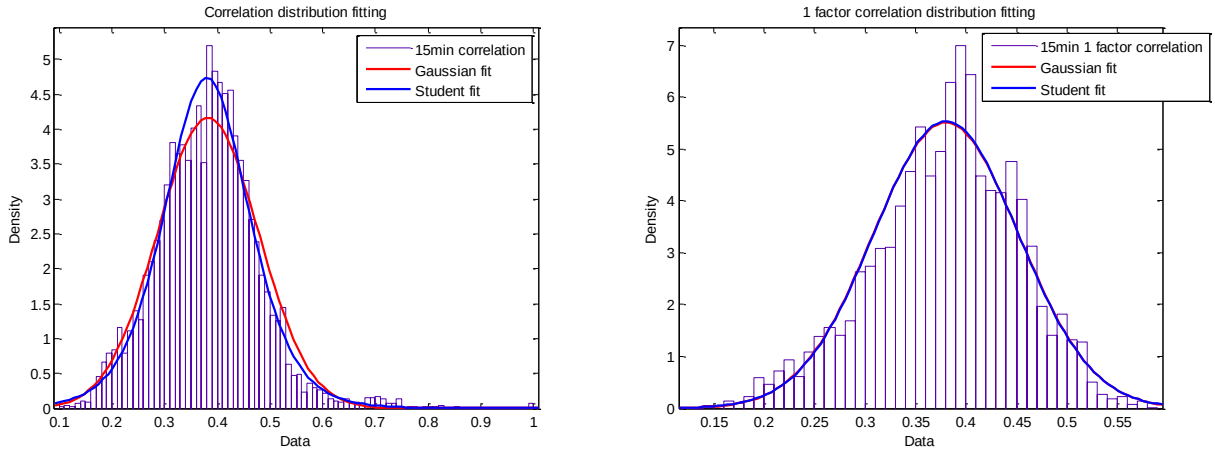
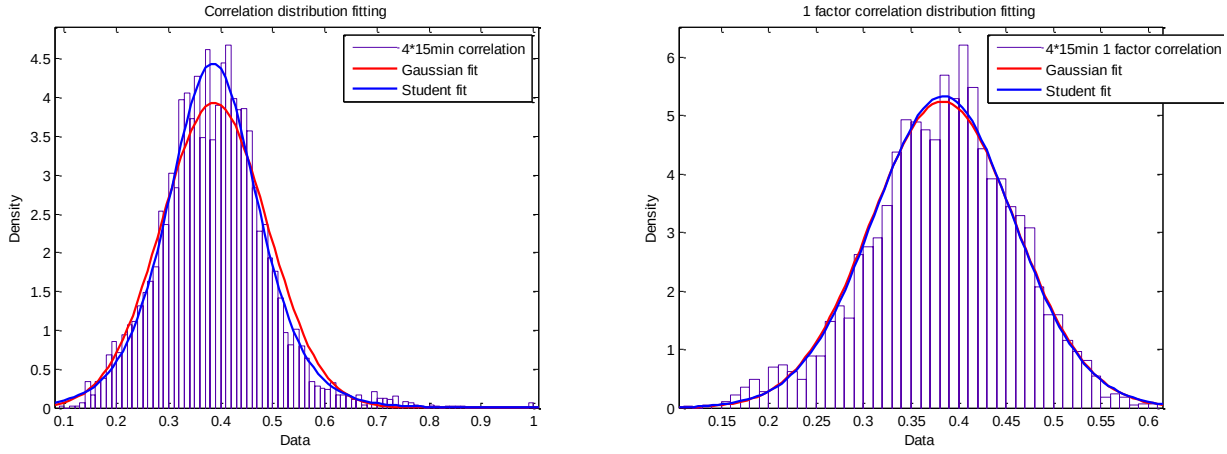


Figure 3 – Correlation Distribution fitting with Gaussian and Student T

Another interesting result is the fact that when the time scale increases, the Gaussian fit of the empirical distribution of correlations becomes more and more accurate. This can be shown by the fact that the parameter of degree of freedom of the Student fit becomes larger as the time scale of returns increases, demonstrating a normal behavior.



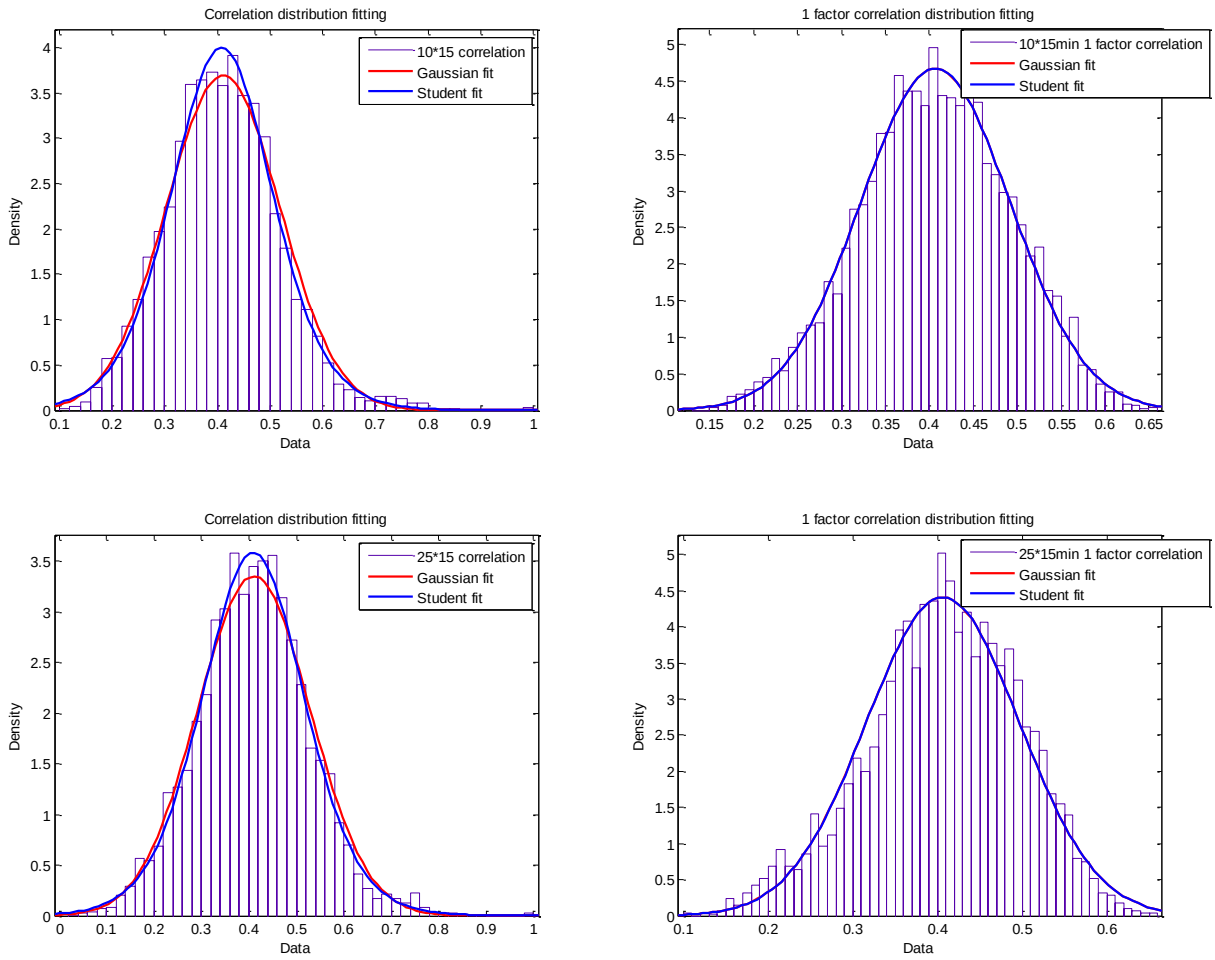


Figure 4 – Convergence of Gaussian and Student fit with respect to increasing time scale

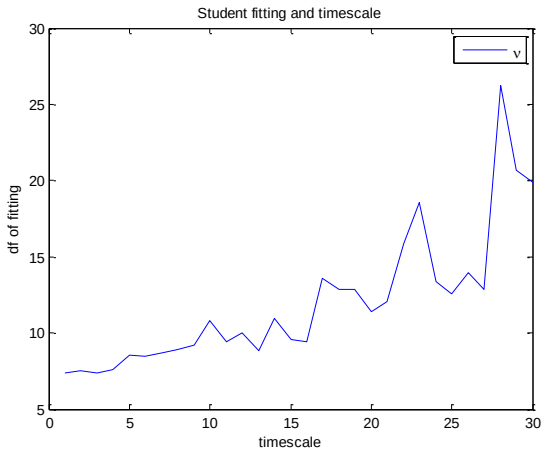


Figure 5 – Increase of Degree of Freedom as timescale increases

From Figure 3, we can clearly see that the student distribution is doing a better job at understanding the tails of the correlation. In the graph on the right of Figure 3, we show the distribution of the model values of the correlation given the one factor model. The Gaussian and the student fit are almost perfectly superimposed on one another, meaning that there is no tail in the values. Furthermore, the one factor model underestimates the tail of the correlation:

on average it is good but it fails to explain the big values in the correlation which can come from the fact that some stocks may be related to each other (i.e. within the same sector). However, as we increase the time decay when computing the log returns (i.e. 15min to one day) the tails diminish and we can expect the one factor model to become a more and more accurate predictor.

2.3. ASYMMETRY IN CORRELATIONS

Here we have explored the stylized fact of asymmetry in correlations during periods of extreme returns. We expect to observe higher correlations for extreme negative return periods than for extreme positive return periods. In order to observe the correlations for extreme return periods, we took daily data from the 15 minute data and applied an exceedance correlation function to see how correlations vary with changes in a threshold level on returns. The exceedance correlation function that was used is given as follows.

$$\rho_{ij}^{+}(\theta) = \frac{\left\langle \tilde{r}_i \tilde{r}_j \right\rangle_{>\theta} - \left\langle \tilde{r}_i \right\rangle_{>\theta} \left\langle \tilde{r}_j \right\rangle_{>\theta}}{\sqrt{\left(\left\langle \tilde{r}_i^2 \right\rangle_{>\theta} - \left\langle \tilde{r}_i^2 \right\rangle_{>\theta} \right) \left(\left\langle \tilde{r}_j^{\,2} \right\rangle_{>\theta} - \left\langle \tilde{r}_j \right\rangle_{>\theta}^2 \right)}}$$

This is the positive exceedance correlation function that is a function of the threshold level, θ. Correlations between two stocks are computed only when the returns of both stocks are greater than θ. As θ is increased, we observe that the average correlations also increase. In order to compute the negative exceedance correlation function, only data for returns less than θ are used. The following figure shows the results when applying the exceedance correlation function to our 15 minute data.

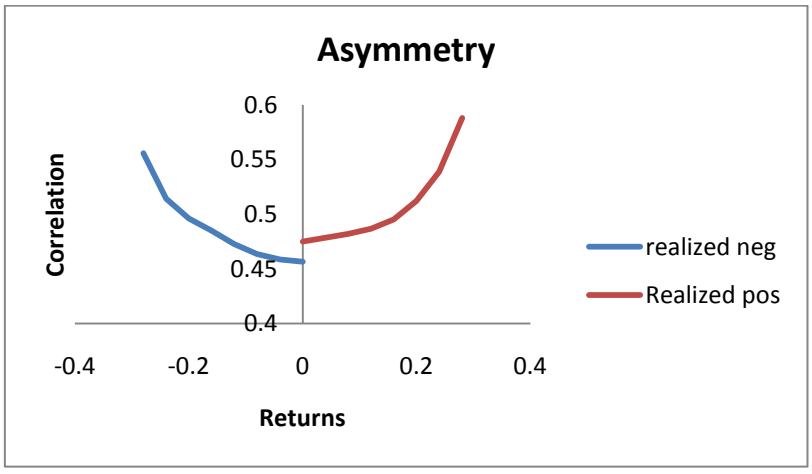


Figure 6 – Asymmetry study results using exceedance correlation function

As we can see from the figure, the correlations for the negative returns are actually lower than those of the positive returns, which is not what we were expecting to see. This may be due to insufficient data in that the period we have analyzed is not long enough. We posit that if we increase the length of the time period we may achieve more consistent results.

2.4. BETAS VS. CORRELATION

Another stylized fact is the Betas vs. Correlation fact. Below we plotted the empirical correlations as well as the model correlations for stocks with the same betas.

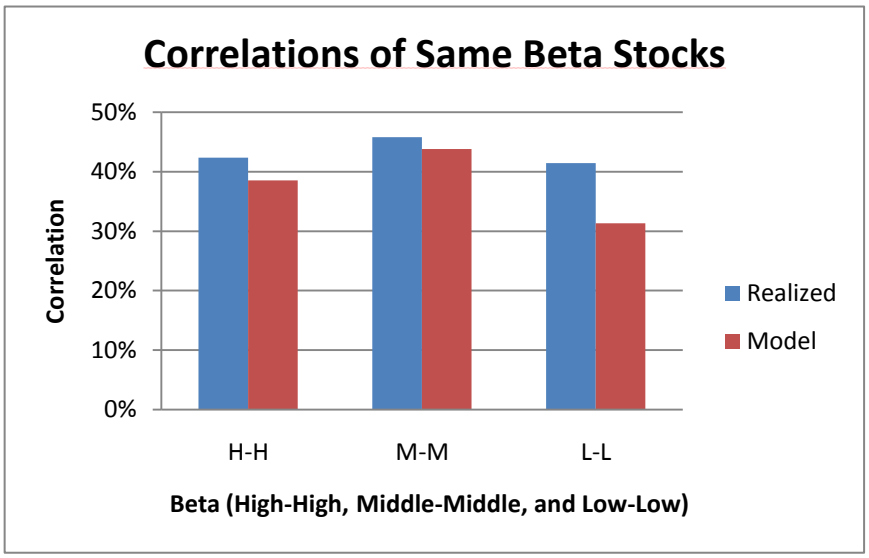


Figure 7 – Beta vs. Correlation analysis

We can see that the model always underestimates the correlation for the stocks with the same betas. These stocks are most likely linked to each other by their sectors which the one factor model fails to account for.

Given these results, we can also gain insight into a general idea behind the structure of correlation: namely, that for any result which concerns the stocks *on average* or which is supposed to work for *any* stock (i.e. the asymmetry or the absolute returns vs. correlation effect), we can expect the one factor model to explain it. In the one factor model, it is enough to look at the correlation of each stock with the market, and therefore reduce the dimensionality to n correlation coefficients instead of $n(n-1)/2$ if we had considered the pair wise correlations for each stock. However, if we want to understand the sub-structure in the correlation distribution, we need to add more information to our model, and so we can potentially look at the other factors (see Section 3).

2.5. MEMORY EFFECT AND MULTIFRACTALITY OF CROSS CORRELATIONS

Financial markets are a complex physical system with a lot of interactions. Above we showed some stylized facts that we can observe such as the fat tails distribution of returns. Now we use the cross-correlation function as an indicator of interaction between stocks. In this section, we introduce the concepts of i*nstantaneous cross-correlation* (IC) and *average instantaneous cross-correlation* (AIC) to consider correlation at a single time step. Therefore, the IC and AIC describe the current interaction between stocks with local information. We aim to study the dynamics of IC and AIC series using fractal analysis in order to detect memory in highly turbulent system.

To do this, we first define the *normalized price return* to compare different stocks by:

$$R_i(t) = \frac{r_i(t) - \langle r_i \rangle}{\sigma_i} \text{ for stock } i \text{ at time } t$$

Then, the IC and AIC are defined by:

$$IC_{ij}(t) = R_i(t) \times R_j(t)$$

$$AIC(t) = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} IC_{ij}(t)$$

To study the memory effect of each time series, we performed a *Detrented Fluctual Analysis* (DFA) because it can be applied to *non-stationary* time series. The method was implemented as follows.

First we consider a fluctuating dynamic time series $A(t')$. Here $A = IC$ or $A = AIC$. We construct the time series $B(t')$ such that:

$$B(t') = \sum_{t''=1}^{t'} A(t'')$$

Then, we divide the total time interval into N_t windows with size t, (so that $t \times N_t = total\ interval$) and fit $B(t')$ to a linear function $B_t(t')$ in each function.

The DFA function of the kth window bow is defined by:

$$f_k(t)^2 = \frac{1}{t} \sum_{t'=(k-1)t+1}^{kt} [B(t') - B_t(t')]^2$$

And the qth order of the overall detrented fluctuation is :

$$F_q(t)^q = \frac{1}{N_t} \sum_{k=1}^{N_t} [f_k(t)]^q$$

$$F_0(t) = exp\{\frac{1}{N_t}\sum_{k=1}^{N_t} ln[f_k(t)]\}$$

In general, $F_q(t)$ will obey a power-law behavior such that:

$$F_q(t) \sim t^{H_q} \qquad (1)$$

In fractal analysis, one usually looks for the stochastic behavior of some geometrical or topological properties of signal fluctuations around the local trend. One expects a power-law relationship between the quantity describing such fluctuations and the length of time window t along which the fluctuation is being measured.

For q = 2, H2 is known as the Hurst exponent, the quantity described is the variance. It is referred to as the "index of dependence" and is the relative tendency of a time series to either strongly regress to the mean or cluster in a direction. The Hurst exponent has the following interpretation:

- $0.5 < H_2 < 1$: $A(t')$ is long-range correlated in time, in our study it means that periods of high correlation are usually followed by period of high correlation and inversely
- $0 < H_2 < 0.5$: $A(t')$ is anti-correlated, so periods of high correlation are usually followed by periods of low correlation
- $H_2 = 0.5$: corresponds to the Gaussian white noise, no memory in the signal
- $H_2 > 1$: the time series is unstable

The relation (1) reveals the independence of scaling properties of a system from the scale if Hq is constant. If Hq is not a constant, the signal presents a multi-fractal feature, and the signal is then very turbulent and harder to study. It shows that not only is there clustering present in the correlation but that there is also inhomogeneity in the magnitude. For such signals, we introduce the scaling exponent function $\tau(q)$ to reveal the multifractality,

$$\tau(q) = qH_q - 1$$

The local singularity exponent α and its spectrum $f(\alpha)$ are :

$$\alpha = \frac{d\tau(q)}{dq}$$

$$f(\alpha) = q\alpha - \tau(q)$$

The difference between the maximum and the minimum of the local singularity: $\Delta\alpha = \alpha_{max} - \alpha_{min}$, is used to quantify the width of the extracted multifractal spectrum; the

larger $\Delta\alpha$, the stronger the multifractality. The presence of long memory dynamics in asset prices provides evidence against the weak form of market efficiency.

2.6. RESULTS OF FRACTAL ANALYSIS

For the S&P 500 during the period considered, we found that the AIC was characterized by a long-range dependence. Therefore, periods of increase in correlation tend to be followed by other periods of increase.

There is also a trend in the time series. This property is consistent across all time scales of returns, except for daily returns and more, and is probably due to a lack of data.

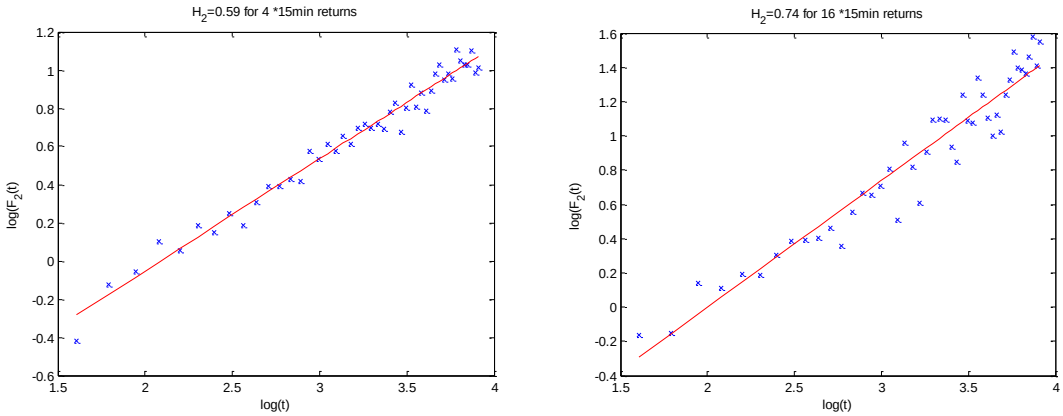


Figure 8 – Long range correlation trend for varying time scales

For pair wise instantaneous correlation, we found that the trend is more difficult as the behavior becomes closer to Gaussian. We considered the distribution of H2 in the index:

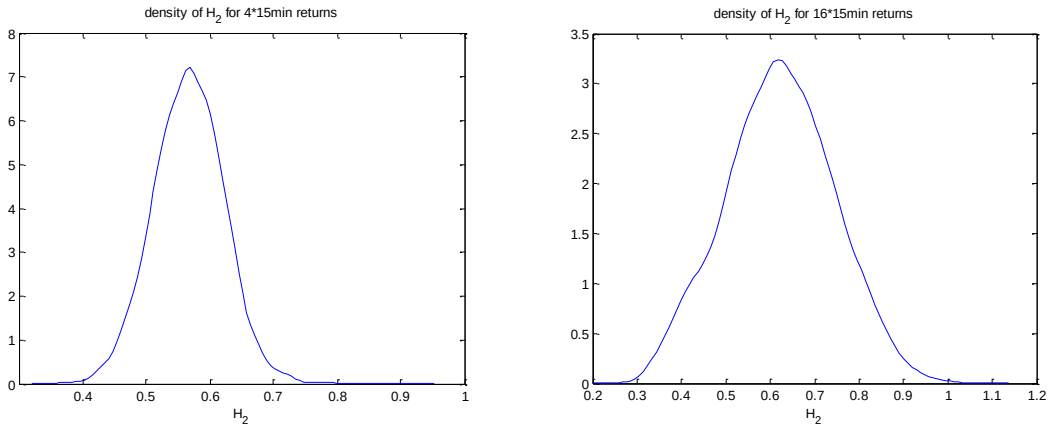


Figure 9 – Density of H2 for varying time scales

Finally, we showed that the time series presents a multifractal behavior and that the signals are therefore rich and complex with heteroscedastic properties. The shape of the spectrum for AIC (more stretched to the right) is an indicator of inhomogeneity in magnitude of correlation. In this case, periods of increase in correlation tend to exhibit a higher magnitude of change:

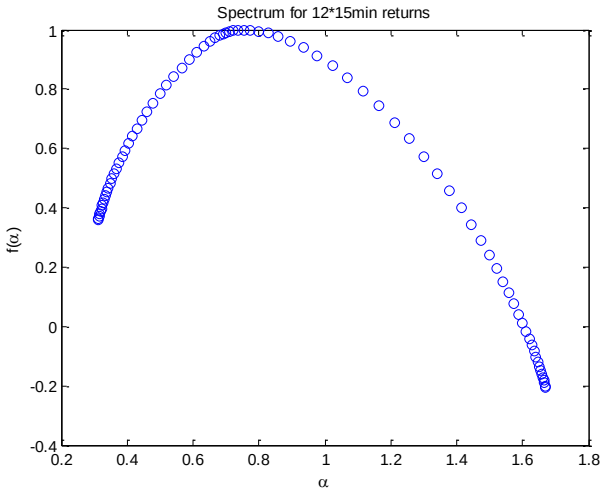


Figure 10 – AIC spectrum indicating inhomogeneity in magnitude of correlation

Fractal analysis seems to be a promising direction for further studies on the behavior of correlation. It confirms the memory effect for the average correlation and can be a tool to help detect and predict different patterns in correlation. However the complexity of multifractality has to be taken into account and a study of spectra and its interpretation is necessary.

3. FACTOR MODELS

3.1 MOTIVATION

The model we decided to focus on is the *factor model*. This model became very popular with the development of statistical arbitrage and the paper of Avellaneda and Lee: *Statistical arbitrage in the US equity market*. Given a set of data – in this case we considered the time-series of the hourly log-returns of the stocks: $\log(\frac{S_{t+1}}{S_t})$ – the *principal component analysis* (PCA) is used to find the most important factors of explanations of these data. For a more detailed explanation of the factor model we direct the reader to Avellaneda and Lee.

3.2 THE ONE FACTOR MODEL

Let's begin by looking at the case where k=1, meaning we only consider the first component of the PCA. The model is:

$$X_i(t) \cong P_{i1}V_1(t).$$

Where V_1 is a common time-series that can approximate every stock, and P_{i1} is a proportion for each stock. Based on the Sharpe theory, we can identify the first score to the general market component and the first loadings to the betas.

In order to verify this, we plot the first score against the market average of all the stocks. After accounting for a multiplying factor, here is our comparison:

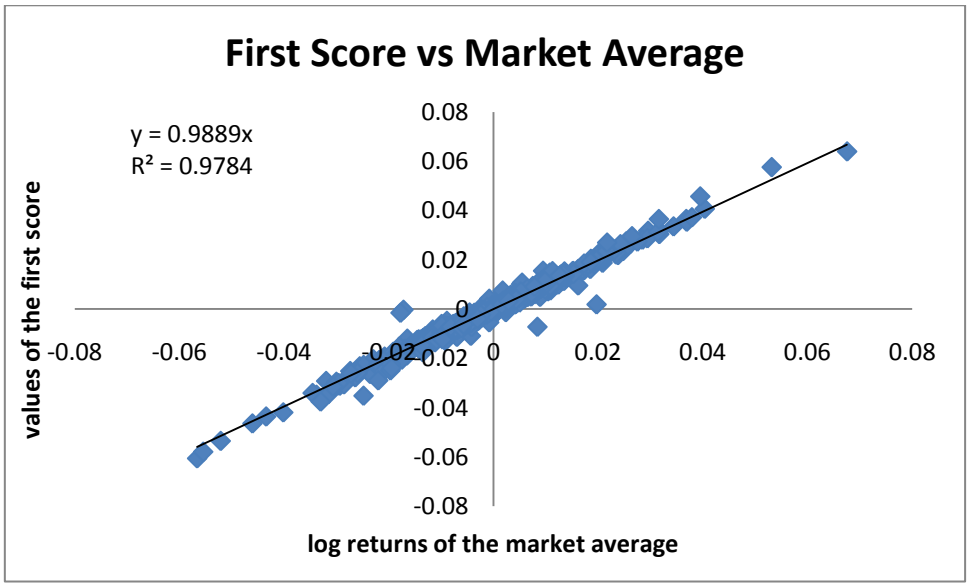


Figure 11 – First score vs. market average of stocks

We can see that the first score is indeed following the market. Then, we can identify the coefficient of the first vector of P as the usual betas. We note that this model is almost identical to linearly projecting each stock onto the market – in this case the S&P500.

3.3 Multi-Factor Model Selection

To find a basis for my factor model selection, I used the PCA of the log returns of the data over a time period of 2 weeks, and looked at both the screeplot of the PCA components as well as the average loading size for each component. From the screeplot:

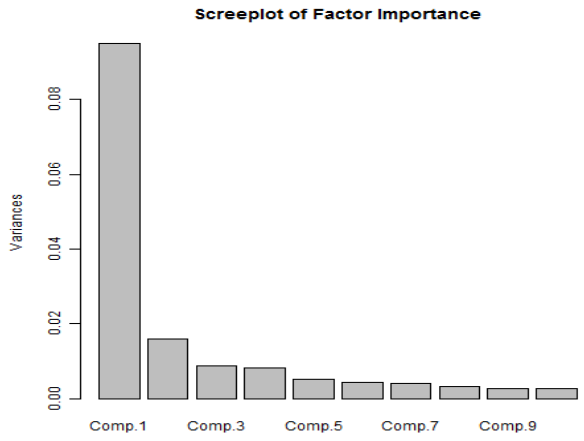


Figure 12 – Screeplot of PCA of half-month log returns of data

I saw that the first component (more commonly known as the *market component*) captured a significant portion of the variance of the data. Referencing Figure 12, we see how the screeplot makes clear the motivation behind the one factor model. However, in hopes of achieving even better models, we analyzed the other components to see if they had any significant effects on modeling the correlation between two stocks. From this we were faced with the question of how many components to look at and consider in our multi-factor models.

To determine this we first saw that from the screeplot, any component after the first nine components contributed almost nothing to the variance, and so, generously, we decided to use the first nine components in our model selection. Then, to determine which of the nine components we should use to model each sector, we looked at the average loadings of each component for each sector. In theory, the average loading is supposed to tell us the relative "importance" of that component for that particular sector, and so graphing these average loadings we got the following nine graphs:

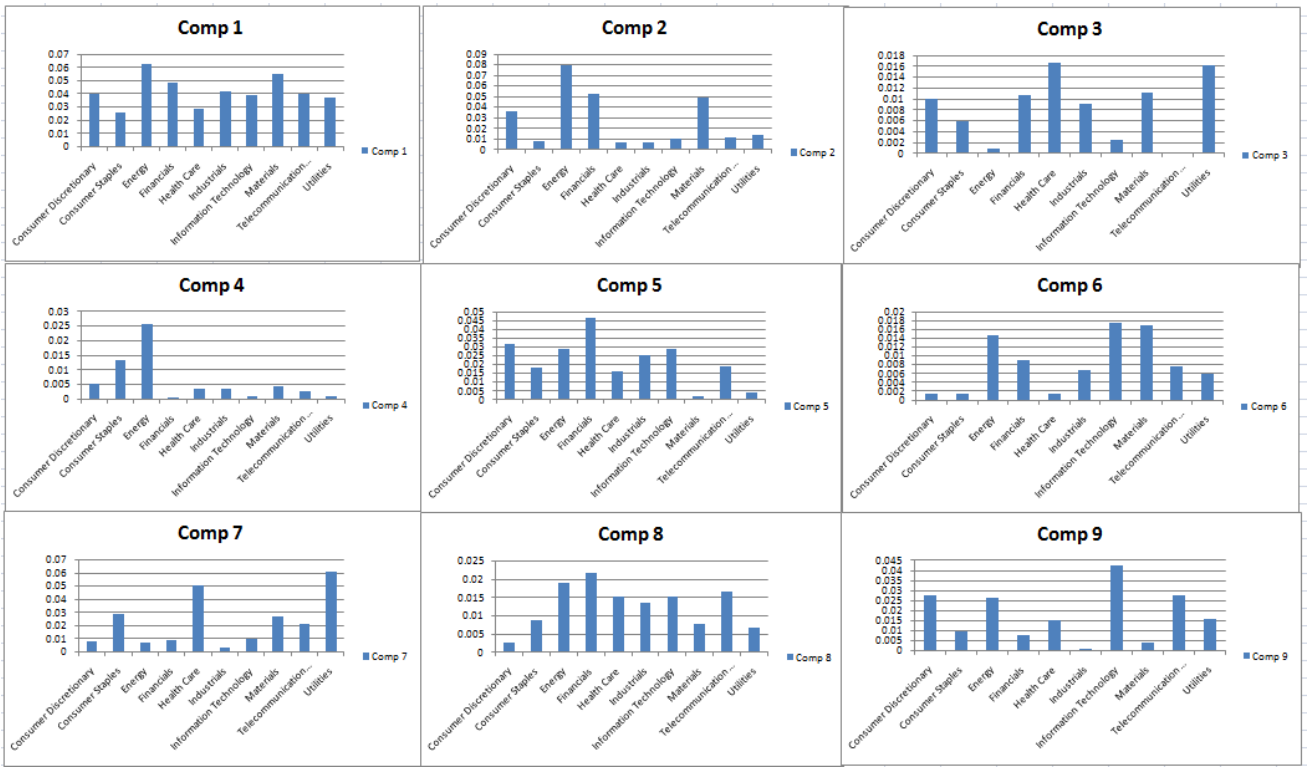


Figure 13 – Absolute value of the average loadings of nine components from PCA

With these graphs, for each sector we then looked at which components seemed to be of significance. However, there were two problems with this approach:

1. We didn't have a strong understanding of how strongly each component contributed to the model relative to the other components. The best we can do is look at the screeplot of the PCA (Figure 12) and guess at the relative importance of each component – empirically, though, this fact was unclear and was something we needed to experiment with.
2. There is no known metric for what exactly determines a "good" component from a "bad" component for each sector. Our metric was to just take the three or four of the top sectors for each component, and compile our model selection that way. However, potentially it could even be the case that this metric changes with each component (i.e. it might make sense that for components which capture a higher amount of the variance, that we would be more liberal in including those components in our model selection), so again this was another point that our group needed to experiment with.

And so, acknowledging these potential issues, we made the following selections for our multi-factored models:

	Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6	Comp 7	Comp 8	Comp 9
Health	**x**		**x**				**x**		
Utilities	**x**		**x**				**x**		
Finance	**x**	**x**			**x**			**x**	
consumer d	**x**				**x**				**X**
consumer s	**x**			**x**			**x**		
industrials	**x**		**x**					**x**	
info tech	**x**				**x**	**x**			**X**
materials	**x**	**x**	**x**			**x**			
telecom	**x**							**x**	**x**
energy	**x**	**x**		**x**	**x**	**x**		**x**	

Figure 14 – Model Selection using PCA components

Once our models were selected, we calculated the pair wise correlations for each stock using the log returns of the half-month data and over a sliding window of 30 half-months. Similarly, we then computed the model correlations according to our factors. From the sliding window, we then had several vectors of equal length with the empirical correlations and the model correlations over the two year period. To determine the goodness of the fit we computed the *residual sum of squares* (RSS) between the empirical correlation vector and each model correlation vector, and then plotted the RSS of each pair of stocks in a heatmap.

What we hoped for was that for a given sector and corresponding model selection, we would see a concentration of low RSS values in that given sector, meaning that the modeled correlations stayed relatively close to the empirical correlations, followed by a concentration of high RSS values for those sectors that required components which were not selected. For instance, for the Consumer Discretionary sector, since we used factors 1, 5, and 9, we would expect the Energy and Finance sectors to do extremely poorly as they both require the 2nd factor (amongst others). Below are the sector model selections that performed relatively well:

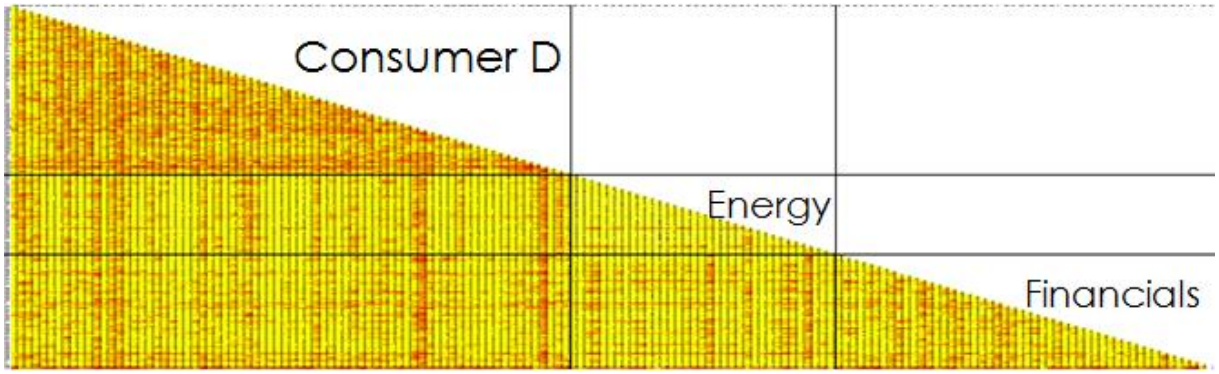


Figure 15 – Consumer Discretionary model selection heat map

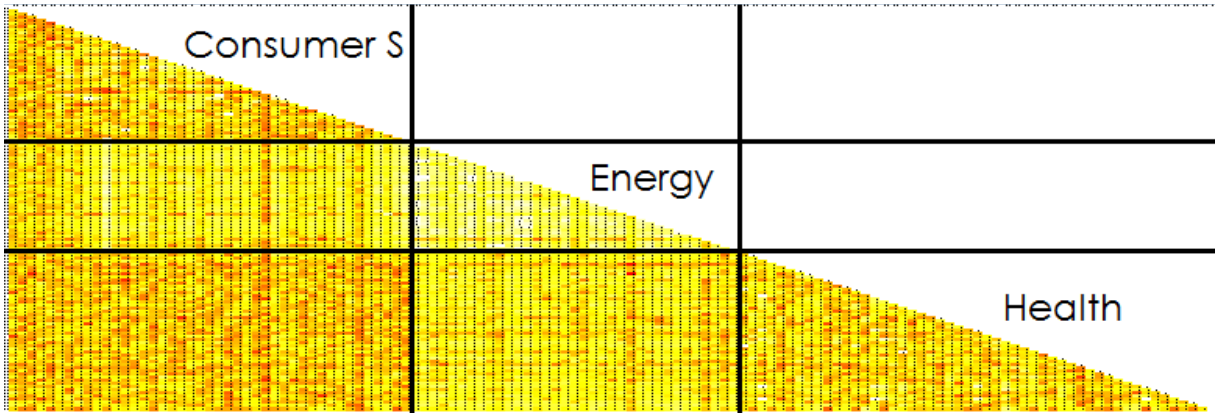


Figure 16 – Health model selection heat map

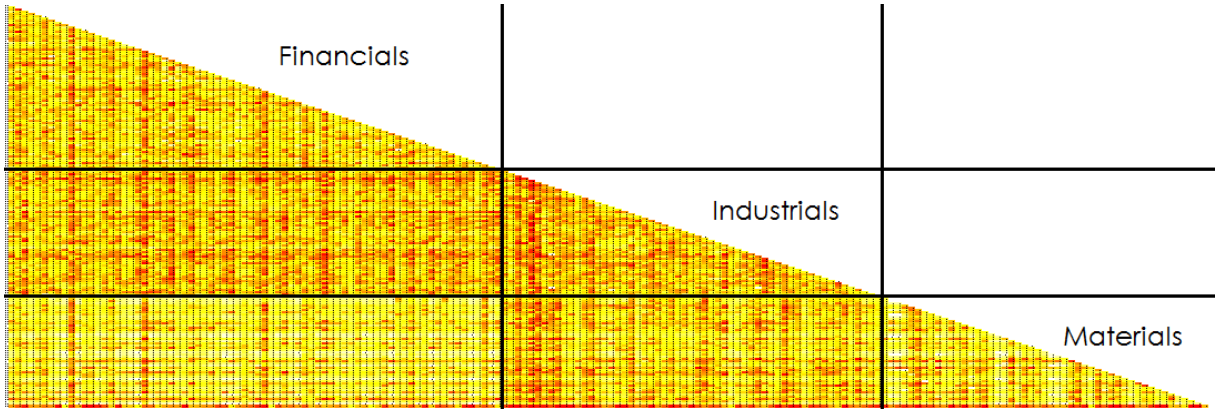


Figure 17 – Industrials model selection heat map

However, many of our models did not perform quite as well. Nonetheless, through their failures we were still able to gain some valuable insights into the two questions we posed earlier. For instance, when running our model for the Information Technology sector and the Materials sector, we got the following heat maps:

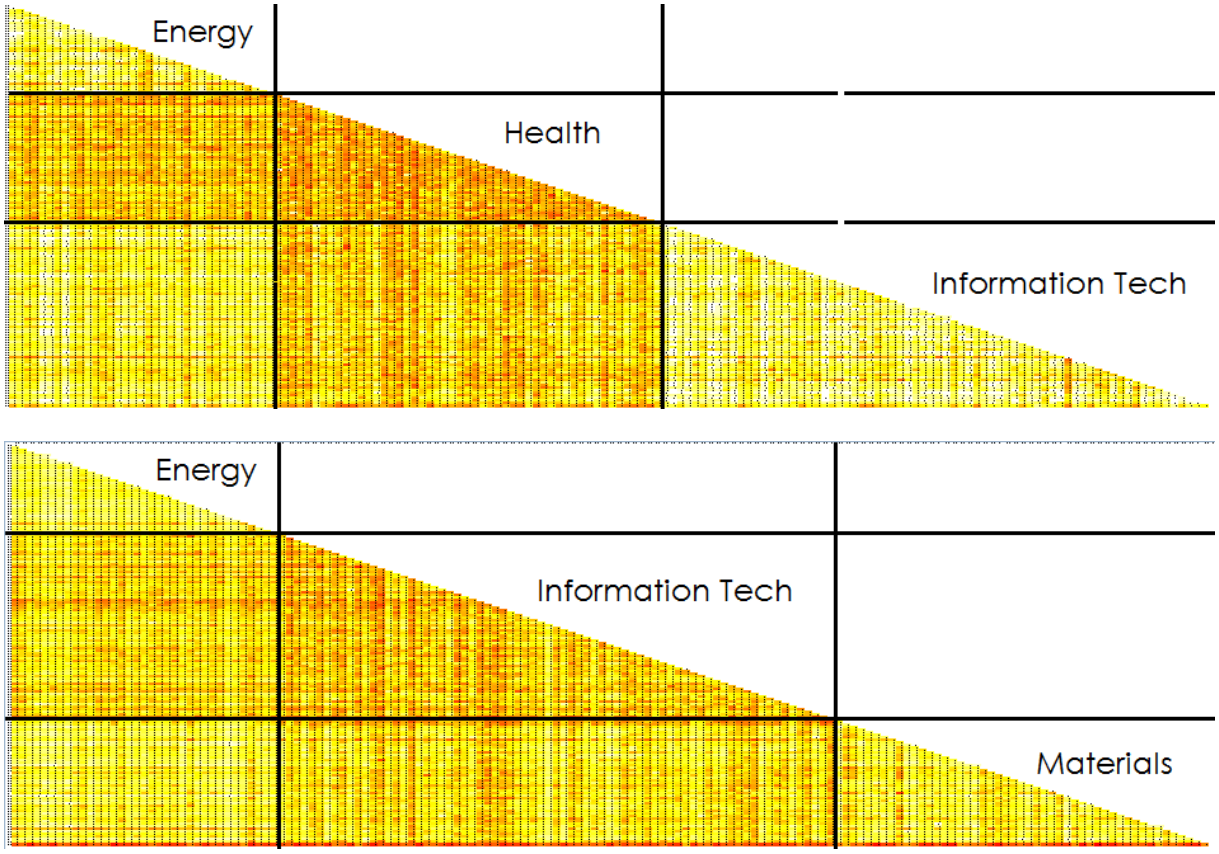


Figure 18 – Information Technology heat map (top) and Materials heat map (bottom)

And so we saw that when running the Information Technology model selection by using factors 1, 5, 6, and 9, we actually did extremely poorly in explaining the Information Technology companies (and did a good job explaining the Health sector). However, when running the Materials model selection by using factors 1, 2, 3, and 6, we saw that we actually did a good job explaining the Information Technology sector, but a poor one when trying to explain the Materials sector.

What this suggests is that the 2nd and 3rd factors are of significantly greater importance in the model selection process than the 5th and 9th factors (as the 1st and 6th were used in both models). Furthermore, just like how we were extremely liberal in selecting the 1st factor for each of our models (in fact we included the market factor in all of our models), this suggests that the greater the importance of a component, the more liberal we should be when including that component in our model.

Finally, I end with our observation that the Energy sector cannot be modeled well by using these PCA components. By looking at Figures 3, 4, and 6, we see that in those models the Energy sector performed relatively poorly. Then, by looking at the model we chose for Energy, which included almost all of the factors, we got the following heat map:

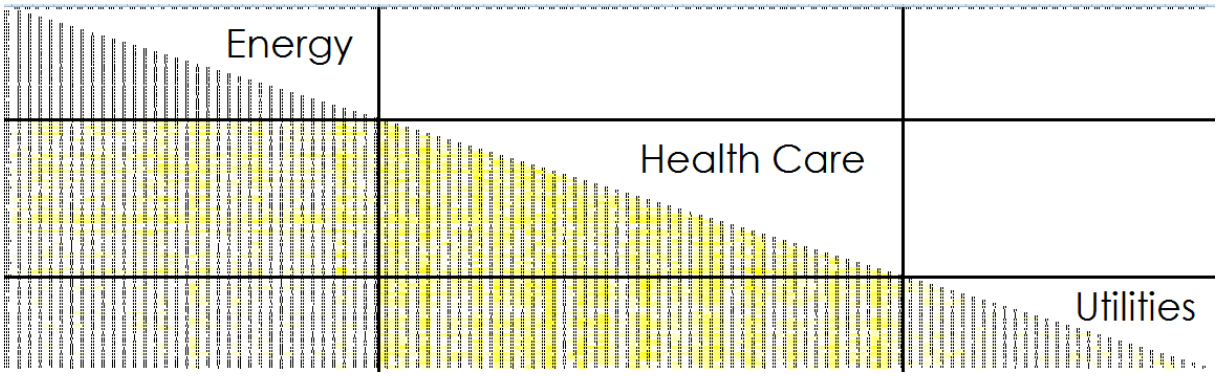


Figure 19 – Energy model selection heat map

And so we see that even with all of those factors the Energy sector still performed extremely poorly. Furthermore, we see that the Utilities and Health Care sectors even performed poorly under this set of factors. This suggests that including unnecessary factors may actually *detract* from your model. Thus, looking at all of our multi-factor models, we conclude with the following points:

- The first few components can play a significantly greater role in our model selections than those components which capture a smaller amount of the variance.
- When selecting components to include in a model, the more important the component, the more liberal one should be when deciding to include that component.
- Adding too many components in your model selection can potentially hurt your model and cause it to be less accurate.
- These factors do seem to make a difference. Though the specific factors we chose for each model were not always the "right" ones, there does seem to be some kind of relationship between each set of factors and the various sectors. Simply looking at the figures above we see very well defined cutoffs between the sectors in our heat maps which suggests that the values in the heatmap are not random.

Though our model selections were only able to explain 3 of the 10 sectors, by using these refined heuristics, we think that there is much potential in exploring these multi-factor models and trying to derive some kind of meaning for factors other than the first. Namely, our results suggest that we should try running the same pair wise tests but with *fewer* components.

Why fewer? Using the fact that the first couple of components are of significantly greater importance than the later factors, along with the fact that adding too many components can hurt your model, this tells us that perhaps we should only use the first few components (i.e. Factors 1 through 4) to achieve the best results. This is a positive result as it actually *encourages* us to continue to reduce the dimensionality of our model.

4. COPULAS

4.1 DEFINITION AND MOTIVATION

The copula is a function describing the correlation between data. Given the margins of two random variables X and Y for example, the copula function C will transform the cdf of each random variable into the cdf of the pair:

$$F_{XY}(x, y) = C(F_X(x), F_Y(y))$$

We can extend this definition to N random variables. As $F_X(x)$ can itself be considered as a random variable following a standard uniform law, the copula function can also be seen as the multivariate distribution of standard uniform random variables. Such a copula always exists, as *Sklar's theorem* asserts that for any multivariate distribution function, there is a copula that binds the margins to give the joint distribution.

In practice, the copula is useful when we don't know the multivariate distribution of the random variables and we want to create a model for it. The multivariate distribution is indeed the fundamental function to know in the pricing of options and that's usually the goal that we can reach by the mean of a copula function. On the other hand when we know this multivariate distribution, thus the margins, we could easily compute the copula. For examples and calculations regarding copulas, see Appendix A.

4.2 THE GAUSSIAN COPULA IN PRACTICE

The correlation matrix could take any form, but in practice we assume that it has the following form:

$$P = \begin{bmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{bmatrix}$$

This is a drastic simplification: we assume that the correlation coefficient is the same for all pairs of stocks. This is equivalent to saying that:

For any stock X_i, given the market X_M and a set of independent variables (Y_i) we have :

$$X_i = \rho X_M + \sqrt{1 - \rho^2}\, Y_i$$

This looks like the one-factor model but with only one free parameter here in the correlation matrix. More specifically, this means that each stock has the same beta, whereas we had N betas in the previous one-factor model (namely each value of corr(Xi , Xm) could be defined). This simplification is makes calibration and computations easy, and that's why options tend to

be quoted this way. Unfortunately, this model has two dramatic weaknesses: it underestimates the big correlated movements as the Gaussian copula is somehow modeling the distribution as a multivariate Gaussian with no tail, and it also ignores the repartition of the correlation coefficients.

To show these problems, let's consider two stocks with a strong correlation: Apple and Microsoft. The empirical correlation is indeed 45%. We compute the normalized log returns, as well as the empirical joint distribution function between them.

Assume $\tilde{A}$ and $\tilde{M}$ are the normalized log returns of Apple and Microsoft, in the following table, at the intersection between the row -1 and the column -0.4 you can see the value of the density $f(\tilde{A} = -1, \tilde{M} - 0.4)$. The color shows the intensity of the value.

Empirical dis	-2	-1.8	-1.6	-1.4	-1.2	-1	-0.8	-0.6	-0.4	-0.2	-2.776E-16	0.2	0.4	0.6	0.8	1	1.2	1.4	1.6	1.8	2
-2	0	0	0.000279252	0.00084	0.0002793	0.0002793	0.0002793	0.0008378	0	0	0.0005585	0	0.0003	0	0.0002793	0.0005585	0	0	0	0	0
-1.8	0	0.000558503	0.000279252	0.00028	0.0002793	0.001117	0.0005585	0.0008378	0.0003	0	0.001117	0.0006	0.0006	0	0	0	0	0	0	0	0
-1.6	0	0.000837755	0.000558503	0.00028	0.0005585	0.0002793	0.0005378	0.0013963	0.0006	0.0017	0.0008378	0.0006	0	0	0	0	0	0	0	0	0
-1.4	0.000837755	0.000279252	0.000279252	0.00028	0.0002793	0.0016755	0.0008378	0.0005585	0.0003	0	0.0005585	0.0006	0.0003	0.0002793	0.0002793	0	0	0	0	0	0
-1.2	0.000558503	0.000558503	0.000837755	0.00084	0.0016755	0.0005585	0.0027925	0.0016755	0.0017	0.0017	0.0025133	0.0011	0.0008	0.0002793	0.0002793	0.0002793	0.0002793	0	0	0	0.0002793
-1	0.000279252	0.000558503	0	0.00112	0.0016755	0.001117	0.0030718	0.0033351	0.0036	0.0022	0.001117	0.0008	0.002	0.0016755	0.0005585	0	0	0	0	0.0002793	0
-0.8	0.000279252	0.000558503	0.000279252	0.00251	0.0008378	0.003351	0.002234	0.0072605	0.005	0.0056	0.0047473	0.0036	0.0011	0.0013963	0.0005585	0	0	0	0.0002793	0	0.0002793
-0.6	0.000279252	0	0.000279252	0.0014	0.003351	0.003351	0.0047473	0.0083775	0.0089	0.0067	0.0072605	0.0056	0.0034	0.0016755	0.0005585	0.001117	0.0002793	0.0002793	0.0002793	0.0002793	0
-0.4	0	0	0.000837755	0.00028	0.0025133	0.002234	0.0069813	0.007819	0.0067	0.0156	0.0128456	0.0092	0.005	0.003351	0.0025133	0	0	0.0005585	0.0002793	0	0.0005585
-0.2	0.000279252	0.000279252	0.000558503	0.00084	0.0005585	0.004468	0.0053058	0.0083775	0.0134	0.0288	0.0240156	0.0179	0.0087	0.0047473	0.0027925	0.0013963	0.002234	0.0002793	0	0.0002793	0
-2.776E-16	0	0.000558503	0.000279252	0.00084	0.0025133	0.0019548	0.0041888	0.0083775	0.0154	0.0293	0.0393745	0.0293	0.0173	0.0094946	0.0036303	0.0025133	0.0019548	0.0002793	0.0008378	0	0
0.2	0	0.000279252	0	0.00028	0.0005585	0.0016755	0.0036303	0.0036303	0.0134	0.0165	0.0273667	0.0232	0.017	0.007819	0.0047473	0.0041888	0.002234	0.0016755	0.0005585	0.0008378	0.0002793
0.4	0.000279252	0	0	0.00084	0.0008378	0.0008378	0.0016755	0.002234	0.0056	0.0106	0.0173136	0.0134	0.0126	0.0106116	0.0061435	0.0039095	0.001117	0.0008378	0.0008378	0.0005585	0.0002793
0.6	0	0.000279252	0.000558503	0	0.0002793	0.001117	0.0027925	0.0019548	0.0017	0.0039	0.0094946	0.0103	0.0078	0.0058643	0.0041888	0.003351	0.0019548	0.0016755	0.0008378	0.0002793	0.0002793
0.8	0	0	0.000279252	0	0.0008378	0.0002793	0.0008378	0.0016755	0.0017	0.0034	0.0030718	0.0056	0.0073	0.0058643	0.0036303	0.0019548	0.0019548	0.0005585	0.0005585	0	0.0005585
1	0	0	0	0	0.0005585	0.0005585	0.0005585	0.0002793	0.0008	0.0008	0.0016755	0.0036	0.0017	0.004468	0.002234	0.0016755	0.001117	0.0008378	0.0005585	0.0002793	0.0002793
1.2	0	0	0	0	0	0.0002793	0	0.001117	0.0017	0.0017	0.002234	0.0014	0.0031	0.0027925	0.002234	0.0008378	0.0008378	0.0002793	0.0005585	0.0002793	0
1.4	0	0	0	0	0	0.0002793	0.0002793	0	0.0003	0	0.001117	0	0.0011	0.0019548	0.0005585	0.0008378	0.0002793	0.001117	0.0013963	0.0005585	0.0002793
1.6	0	0.000279252	0	0	0	0	0.0002793	0	0.0003	0.0003	0.0008378	0	0.0006	0.0002793	0	0.0008378	0	0.0002793	0.0002793	0.0002793	0.0002793
1.8	0.000279252	0	0.000558503	0	0	0.0002793	0	0.0002793	0.0003	0.0006	0	0.0011	0.0003	0	0.0005585	0.0005585	0	0.0002793	0	0	0.0002793
2	0	0	0	0	0	0	0	0	0	0	0.0005585	0.0006	0.0006	0.0002793	0.0005585	0	0.0008378	0.0002793	0.0005585	0.0008378	0

Figure 20 – Empirical joint distribution of normalized log returns for APPL and MSFT

If we then fit a Gaussian copula function to the empirical cdf, we get an optimal correlation value of 60% and the following density:

Gaussian cc	-2	-1.8	-1.6	-1.4	-1.2	-1	-0.8	-0.6	-0.4	-0.2	-2.776E-16	0.2	0.4	0.6	0.8	1	1.2	1.4	1.6	1.8	2
-2	0.000659142	0.000819681	0.000957155	0.00105	0.0010806	0.0010448	0.0009485	0.0008086	0.0006	0.0005	0.0003434	0.0002	0.0001	8.28E-05	4.544E-05	2.341E-05	1.133E-05	5.148E-06	2.196E-06	8.8E-07	3.31E-07
-1.8	0.000819681	0.001058781	0.001284218	0.00146	0.0015643	0.001571	0.0014814	0.0013118	0.0011	0.0009	0.0006244	0.0004	0.0003	0.0001687	9.616E-05	5.147E-05	2.587E-05	1.221E-05	5.411E-06	2.252E-06	8.8E-07
-1.6	0.000957155	0.001284218	0.001617957	0.00191	0.0021264	0.0022181	0.0021727	0.0019984	0.0017	0.0014	0.001066	0.0008	0.0005	0.0003228	0.0001911	0.0001063	5.547E-05	2.719E-05	1.252E-05	5.411E-06	2.196E-06
-1.4	0.001049521	0.001462659	0.001914109	0.00235	0.0027141	0.0029408	0.0029921	0.0028586	0.0026	0.0022	0.001709	0.0013	0.0009	0.0005799	0.0003566	0.000206	0.0001117	5.687E-05	2.719E-05	1.221E-05	5.148E-06
-1.2	0.001080616	0.001564296	0.002126365	0.00271	0.003253	0.0036612	0.0038693	0.0038398	0.0036	0.0031	0.0025726	0.002	0.0014	0.0009783	0.000625	0.0003749	0.0002112	0.0001117	5.547E-05	2.587E-05	1.133E-05
-1	0.001044776	0.001570964	0.002218098	0.00294	0.0036612	0.0042801	0.0046984	0.0048431	0.0047	0.0043	0.0036364	0.0029	0.0022	0.0015498	0.0010283	0.0006407	0.0003749	0.000206	0.0001063	5.147E-05	2.341E-05
-0.8	0.000948521	0.001481444	0.002172677	0.00299	0.0038693	0.0046984	0.0053574	0.0057361	0.0058	0.0054	0.0048267	0.004	0.0031	0.0023053	0.0015889	0.0010283	0.000625	0.0003566	0.0001911	9.616E-05	4.544E-05
-0.6	0.000808615	0.001311825	0.001998394	0.00286	0.0038398	0.0046431	0.0057361	0.0063794	0.0067	0.0065	0.0060159	0.0052	0.0042	0.0032201	0.0023053	0.0015498	0.0009783	0.0005799	0.0003228	0.0001687	8.28E-05
-0.4	0.000647305	0.001090782	0.001725991	0.00256	0.0035781	0.0046878	0.0057671	0.0066622	0.0072	0.0074	0.0070408	0.0063	0.0053	0.0042235	0.0031407	0.0021931	0.001438	0.0008854	0.0005119	0.0002779	0.0001417
-0.2	0.000486572	0.00085167	0.001399806	0.00216	0.0031309	0.0042607	0.0054446	0.0065331	0.0074	0.0078	0.0077377	0.0072	0.0063	0.0052018	0.004018	0.0029143	0.0019849	0.0012694	0.0007623	0.0004299	0.0002276
-2.776E-16	0.000343445	0.00062442	0.001066028	0.00171	0.0025726	0.0036364	0.0048267	0.0060159	0.007	0.0077	0.0079851	0.0077	0.007	0.0060159	0.0048267	0.0036364	0.0025726	0.001709	0.001066	0.0006244	0.0003434
0.2	0.000227635	0.000429886	0.000762326	0.00127	0.0019849	0.0029143	0.004018	0.0052018	0.0063	0.0072	0.0077377	0.0078	0.0074	0.0065331	0.0054446	0.0042607	0.0031309	0.0021604	0.0013998	0.0008517	0.0004866
0.4	0.000141674	0.000277909	0.000518899	0.00089	0.001438	0.0021931	0.0031407	0.0042235	0.0053	0.0063	0.0070408	0.0074	0.0072	0.0066622	0.0057671	0.0046878	0.0035781	0.0025645	0.001726	0.0010908	0.0006473
0.6	8.27972E-05	0.000169703	0.000322775	0.00056	0.0009783	0.0015498	0.0023053	0.0032201	0.0042	0.0052	0.0060153	0.0065	0.0067	0.0063794	0.0057361	0.0048431	0.0038398	0.0028586	0.0019984	0.0013118	0.0008086
0.8	4.54372E-05	9.61643E-05	0.00019112	0.00036	0.000625	0.0010283	0.0015889	0.0023053	0.0031	0.004	0.0048267	0.0054	0.0058	0.0057361	0.0053574	0.0046984	0.0038693	0.0029921	0.0021727	0.0014814	0.0009485
1	2.34142E-05	5.14727E-05	0.00010625	0.00021	0.0003749	0.0006407	0.0010283	0.0015498	0.0022	0.0029	0.0036364	0.0043	0.0047	0.0048431	0.0046984	0.0042801	0.0038612	0.0029408	0.0022181	0.001571	0.0010448
1.2	1.13297E-05	2.58709E-05	5.54723E-05	0.00011	0.0002112	0.0003749	0.000625	0.0009783	0.0014	0.002	0.0025726	0.0031	0.0036	0.0038398	0.0038693	0.0038612	0.003253	0.0027141	0.0021264	0.0015643	0.0010806
1.4	5.1479E-06	1.22101E-05	2.71943E-05	5.7E-05	0.0001117	0.000206	0.0003566	0.0005799	0.0009	0.0013	0.001709	0.0022	0.0026	0.0028586	0.0029921	0.0029408	0.0027141	0.0023521	0.0019141	0.0014627	0.0010495
1.6	2.1964E-06	5.41123E-06	1.25185E-05	2.7E-05	5.547E-05	0.0001063	0.0001911	0.0003228	0.0005	0.0008	0.001066	0.0014	0.0017	0.0019984	0.0021727	0.0022181	0.0021264	0.0019141	0.001618	0.0012842	0.0009572
1.8	8.79967E-07	2.25188E-06	5.41123E-06	1.2E-05	2.587E-05	5.147E-05	9.616E-05	0.0001687	0.0003	0.0004	0.0006244	0.0009	0.0011	0.0013118	0.0014814	0.001571	0.0015643	0.0014627	0.0012842	0.0010588	0.0008197
2	3.31049E-07	8.79967E-07	2.1964E-06	5.1E-06	1.133E-05	2.341E-05	4.544E-05	8.28E-05	0.0001	0.0002	0.0003434	0.0005	0.0006	0.0008086	0.0009485	0.0010448	0.0010806	0.0010495	0.0009572	0.0008197	0.0006591

Figure 21 – Gaussian copula function of normalized log returns for APPL and MSFT

We can see that the fit is not very accurate because the correlation coefficient has been overestimated so as to explain the tails of the empirical distribution (60% in the model, 45% with the empirical data). The repartition of the color in the Gaussian copula is much smoother than in the empirical data, meaning that the copula cannot really explain the tails independently of the mean: the shape of the entire distribution has to be modified to fit them. However, the center seems well explained and it shows that the Gaussian copula may be good for small returns.

4.3 FITTING A T-COPULA

As each stock seems to follow a student distribution, the natural copula seems to be a *student-copula*. The advantage of a student copula compared to a multivariate student distribution is that the multivariate student assumes that all the stocks have the same degree of freedom, meaning the same amount of tail, whereas the copula does not require this assumption. The copula only captures the correlation between the random variables and thus does not depend on the degrees of freedom of each marginal. The results of the fitting are the following; data are daily returns on 10 representative stocks of the S&P.

	ML : Max likelihood	df : degree of freedom		
Market absolute log-returns	ML Gaussian copula	ML T copula	df	Relative difference
< 0.30% (0-20% quantile)	40	45.5	10.4	13.8%
< 0.58% (0-40% quantile	64	74.71	12.9	16.7%
< 1.00% (0-60% quantile)	110	144	9.45	30.9%
< 1.65% (0-80% quantile)	218	280	8.83	28.4%
all	792	974	4.34	23.0%

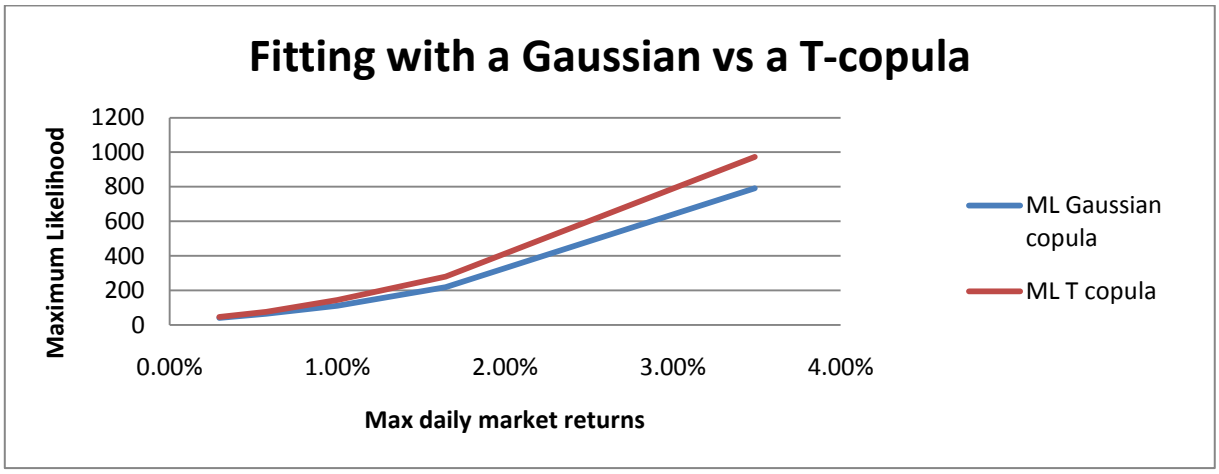


Figure 22 – Quality of fit analysis between Gaussian copula and T-copula

Note we use the maximum likelihood as an estimation of the quality of the fit. We can see that the T-copula is always out performs the Gaussian copula, and that they are only close to each other if we only select the days when the daily market return is below 0.6% (about 15% of relative difference in the fits). This means that in this subset the Gaussian copula is enough as the tails are not important. Furthermore, we can indeed see in the table that the optimal degrees of freedom are above 10, meaning that the T-copula has very low tails. However, if we increase the subset and take days with bigger and bigger daily market returns, we can see in the graph that the gap between the T-copula and the Gaussian copula increases,

meaning that the T-copula is improving (about 25% better). We can also see the decrease in the optimal degree of freedom as we select bigger and bigger returns, meaning the tail becomes more and more important and we are getting further away from a Gaussian copula.

Finally, from these analyses we conclude that the Gaussian copula has many flaws: it cannot account for the increase of correlations with absolute returns, and also cannot account for any asymmetries or clusters in the correlation. The T-copula is able to manage the increase of correlation with absolute returns thanks to the degree of freedom, as well as tails in the correlation. Further improvements to the fitting process could involve using methods like fitting the Kendall correlation matrix, but we leave that for future studies.

APPENDIX A – CALCULATIONS AND EXAMPLES OF COPULAS

Independent copula

The first example is the independent copula. Given X and Y two independent variables, we have:

$$\begin{aligned} F_{XY}(x,y) &= F_X(x) \times F_Y(y) \\ &= C(F_X(x), F_Y(y)) \\ so\ C(a,b) &= a \times b \end{aligned}$$

We can generalize this result to see that the independent-copula of N random variables is simply the product function.

Perfect positive dependence copula

If on the opposite, we always consider the same random variable, we get:

$$\begin{aligned} F_{XXX}(x,y,z) &= P(X \le x, X \le y, X \le z) \\ F_{XXX}(x,y,z) &= P(X \le \min(x,y,z)) \\ &= \min(P(X \le x), P(X \le y), P(X \le z)) \\ &= \min(F_X(x), F_X(y), F_X(z)) \\ &= C(F_X(x), F_X(y), F_X(z)) \\ so\ C(a,b,c) &= \min(a,b,c) \end{aligned}$$

We can generalize this result and get that the perfect positive dependence copula is the minimum function.

Gaussian copula

The last example is the Gaussian copula. Again, as we know the multivariate Gaussian distribution it's easy to compute the Gaussian copula for N random variables:

Define P the correlation matrix, and Σ the covariance matrix

$$\begin{aligned} F_{XYZ}(x,y,z) &= \Phi_\Sigma(x,y,z) \\ F_{XYZ}(x,y,z) &= \Phi_P(\Phi^{-1} \circ \Phi(x), \Phi^{-1} \circ \Phi(y), \Phi^{-1} \circ \Phi(z)) \\ F_{XYZ}(x,y,z) &= C(\Phi(x), \Phi(y), \Phi(z)) \\ so\ C(a,b,c) &= \Phi_P(\Phi^{-1}(\mathrm{a}), \Phi^{-1}(\mathrm{b}), \Phi^{-1}(\mathrm{c})) \end{aligned}$$

BIBLIOGRAPHY

Ang, Andrew, and Geert Bekaert, 1999, "International Asset-Allocation with Time Varying Correlations," NBER working paper 7056, available at www.nber.org/papers/w7056.

Ang, Andrew, Joseph Chen, 2002, "Asymmetric Correlations of Equity Portfolios," The Journal of Financial Economics, 63, 443-494.

Avellaneda, Marco, and Jeong-Hyun Lee, 2008, "Statistical Arbitrage in the U.S. Equity Market," Working paper available at www.ssrn.com/abstract=1153505.

Avellaneda, Marco, and Mike Lipkin, 2009, "A Dynamic Model for Hard-to-Borrow Stocks," Working paper available online at www.ssrn.com/abstract=1357069.

Bekaert, Geert, Guojun Wu, 2000, "Asymmetric Volatility and Risk in Equity Markets," The Review of Financial Studies, 13, 1-42.

Bouchaud, Jean-Philippe, and Marc Potters, 2001, "More stylized facts of financial markets: leverage effect and downside correlations," Physica A, 299, 60-70.

Boudoukh, Jacob, Matthew P. Richardson, Robert F. Whitelaw, 1994, "A Tale of Three Schools: Insights on Autocorrelations of Short-Horizon Stock Returns," The Review of Financial Studies, 7, 539-573.

Boyer, Brian H., Michael S. Gibson, and Mico Loretan, 1999, "Pitfalls in tests for changes in correlations," International Finance Discussion Paper number 597.

Chakraborti, Anirban, Ioane Muni Toke, Marco Patriarca, and Frederic Abergel, 2009, "Econophysics: Empirical facts and agent-based models," Working paper available online at www.arvix.org/pdf/0909.1974.

Christodoulakis, George A., 2007, "Common volatility and correlation clustering in asset returns," The European Journal of Operational Research, 182, 1263-1284.

Cizeau, Pierre, Marc Potters, and Jean-Philippe Bouchaud, 2001, "Correlation structure of extreme stock returns," Quantitative Finance, 1, 217-222.

Karolyi, G. Andrew, and Rene M. Stulz, 1996, "Why do Markets Move Together? An Investigation of U.S.-Japan Stock Return Comovements using ADRS," The Journal of Finance, 51, 951-956.

Lin, Wen-Ling, Robert F. Engle, Takatoshi Ito, 1994, "Do Bulls and Bears Move Across Borders? International Transmission of Stock Returns and Volatility," The Review of Financial Studies, 7, 503-538.

Liu Yanhui, Pierre Cizeau, Martin Meyer, C. K. Peng and H. Eugene Stanley, 1997, "Correlations in economic time series," Physica A, 245, 437-440.

Longin, Francois, and Bruno Solnik, 1995, "Is the correlation in international equity returns constant: 1960-1990?" Journal of International Money and Finance, 14, 3-26.

Longin, Francois, and Bruno Solnik, 2001, "Extreme Correlation of International Equity Markets," The Journal of Finance, 56, 649-676.

Qiu, Tian, Guang Chen, Li-Xin Zhong, Xiao-Wei Lei, 2010, "Memory effect and multifractality of cross-correlations in financial markets," Working paper available at www.arvix.org/pdf/1004.5547.