



CARISMA

A multi-disciplinary research centre focussed on understanding, modelling, quantification, management and control of RISK

- **School of Information Systems, Computing and Mathematics**
- **Brunel Business School**
- **School of Engineering**
- **School of Social Sciences**



TECHNICAL REPORT

CTR/83-08

**Equity portfolio risk (volatility) estimation
using market information
and sentiment**

Leela Mitra, Gautam Mitra, Dan diBartolomeo

Equity portfolio risk (volatility) estimation
using market information
and sentiment



Equity portfolio risk (volatility) estimation using market information and sentiment

Leela Mitra ^{*†} Gautam Mitra ^{*†} Dan diBartolomeo ^{‡ §}

December 1, 2008

Contents

1	Introduction and background	2
2	Model description	5
3	Updating model volatility using quantified news	7
4	Computational experiments	8
5	Discussion and conclusions	11
6	Acknowledgements	11
A	Sentiment analytics overview	12

^{*}CARISMA (Centre for Analysis of Risk and Optimisation Modelling Applications), Brunel University, Uxbridge, United Kingdom, UB8 3PH

[†]OptiRisk Systems, OptiRisk R&D House, One Oxford Road, Uxbridge, Middlesex, UB9 4DA UNITED KINGDOM

[‡]Northfield Information Services Inc., 184 High Street, Boston, MA 02110

[§]Visiting professor at CARISMA

Abstract

Multifactor models are often used as a tool to describe equity portfolio risk. Naturally, risk is dependent on the market environment and investor sentiment. Traditional factor models fail to update quickly as market conditions change. It is desirable that the risk model updates to incorporate new information as it becomes available and for this reason diBartolomeo & Warrick introduce a factor model that uses option implied volatility to improve estimates of the future covariance matrix. We extend this work to use both quantified news and implied volatility to improve risk estimates as the market sentiment and environment changes.

1 Introduction and background

Equity portfolio management problems require fund managers to make decisions about what portfolio to hold (ex-ante) without knowing what equity returns will be. Though the future returns are uncertain, market participants try to understand the nature of the uncertainty and make decisions based on their beliefs about the market environment.

Traditionally portfolio managers have used variants of Markowitz mean-variance analysis to determine the optimal portfolio to hold and this is still fairly standard practise in industry. Mean-variance portfolio decision models fall in the more general group of mean-risk models, where portfolio risk and expected return are traded-off when making asset choices. Variance and standard deviation both measure the spread of a distribution about its mean. The variance of a portfolio can be easily calculated from the covariances of the pairs of asset returns and the asset weights used in the portfolio. Hence, variance is predominantly used in portfolio formation.

In contrast to computing the asset variances and covariances directly using historical data, multifactor models provide an accurate and efficient way to provide these estimates. They decompose an asset's return into return derived from exposure to common factors and an asset specific component. The common factors can be understood as representing different risk (uncertainty) aspects, which all the assets are exposed to at varying degrees (factor sensitivities). By describing a group of asset returns through a set of common key factors, the size of the estimation problem is significantly reduced. The problem we now face is to estimate the covariance matrix of common sources of risk, the variances of the specific returns and estimates of each security's factor exposures. These models capture the natural intuition that firms with similar characteristics will behave similarly.

Active portfolio managers seek to incorporate their investment insight to "beat the market". An accurate description of asset price uncertainty is key to our ability to outperform the market. Tetlock, Saar-Tsechansky, and Macskassy (2007) develop a fundamental factor model that incorporates news as a factor. Investor's perceptions of the riskiness of an asset are determined by their knowledge about the company and its prospects, that is their "information sets". He notes these are determined from three main sources, analysts forecasts, quantifiable publicly disclosed accounting variables and linguistic descriptions of firm's current and future profit generating activities. If the first two sources of information are incomplete or biased, the third may give us relevant information for equity prices. We seek to extract an improved understanding of equity price uncertainty using a quantified measure of market sentiment to update a traditional factor model. This may give us the tools to make improved portfolio (management) decisions.

There are three main types of multifactor models.

Macroeconomic factor models use economic variables (or functions of economic variables) as factors. They model asset prices as responses to these external influences, capturing the natural idea that there is a relationship between equity prices and the economic environment. Typical factors include, unexpected changes in inflation, changes in oil prices, returns in the bond market, etc... The factors are observable time series. Model calibration involves estimating the unknown factor sensitivities β_{ki} , residual variances σ_i^2 and factor covariance matrix Ω_f . This done using time series regression. Chen, Ross and Roll (1986) is a well known example of such a model. Sharpe's (1970) single factor model can also be regarded as a special case of this

type of model.

Fundamental models use firm specific attributes which are not related to the economic environment. These could include factors based on the firm's structure, such as, size, dividend yield, industry classification. Or they could include factors relating to the market, such as, volatility and momentum. There are two well known approaches ; Fama-French (1992 1993) and BARRA.

For the BARRA approach, it is assumed that the factor sensitivities are observable but the factor realisations are unobservable. The factor realisations are derived through cross-sectional regression or from returns on portfolios based on the observed asset characteristics known as *single factor (also factor mimicking) portfolios*.

The Fama-French(1992) approach estimates the parameters using a two step process. First the factor realisations are determined. For a particular asset specific characteristic, such as size, the assets are sorted based on the value of the characteristic. Then a hedged portfolio is formed which is long in the top quintile of the sorted assets and short in the bottom quintile. The observed return on the hedged portfolio at time t is the observed factor realisation. The process is repeated for each asset specific characteristic. In the second step the factor exposures are determined using N time series regressions (for N assets under consideration).

In **Statistical** factor models both the factors realisations and exposures are unobservable. Model calibration involves using the sample covariance matrix of observed returns, which is decomposed into a factor component and a specific component. The factor exposures are estimated in this process. Methods used for calibration of these models include maximum likelihood factor analysis and principal components analysis.

The three types of factor models differ on what sources of risk they consider and how they are calibrated. They give different ways to describe the return variability and they can be shown to be rotations of each other, see Connor(1995). To assess which model is most appropriate Shiekh (1995) notes "we prefer a procedure that is robust (less liable to spurious correlations), capable of explaining the variability in returns (common sources of risk are captured), dynamic (able to change as the determinants of risk change)."

Statistical models use historical correlations to determine a set of orthogonal factors. The advantage of this is that they can evolve over time to pick up new conditions without the need to identify changes in the factor structure. However these factor are opaque and it is difficult to identify them with interpretable sources of risk. Though methods have been suggested for identifying the statistical factor loadings with fundamental stock attributes (see Wilding 2005), this is often given as a disadvantage to these models. Another common criticism is that statistical models can pick up random, chance correlations between assets. Further a choice needs to be made of how many factors to include in the model.

Fundamental and macroeconomic models pick up correlations between assets due to common interpretable factors. Macroeconomic models are sometimes criticised as they do not capture any aspects that do not relate to the economy. Fundamental models are popular in industry and they use characteristics which portfolio managers understand well. However, a choice of which factors to use needs to be made. Also the factors often have common characteristics and it is difficult to separate their effects on return variability. diBartolomeo & Warrick(2005) note this makes them less effective at predicting future conditions.

None of these three models are dominant. Scowcroft (2006) comments " there is little or no consensus on which factors to use or how the models should be estimated." He finds in his study that the quality of factor information used in a model has a significant influence on the quality of the model. Hence the choice of model should be influenced by the information the model builder has available and the quality of this information. In particular, when there is sparse knowledge or data for factors a statistical model may be the most appropriate choice. He also notes that hybrid models which use both fundamental and statistical factors may be effective.

It is often argued that fundamental models are dynamic because they can capture the changing risk struc-

ture when a company's situation changes, by updating the relevant factor exposures. For statistical and macroeconomic models exposures (hence changes in the risk structure) are only updated when further data becomes available and the models are re-calibrated, hence they are updated more slowly and are not dynamic.

However, all the models have a single period structure and are based on independent, identically distributed distributions and do not allow for changing levels of volatility over time. As the operating environment changes these models calibration parameters are updated, but it takes some time for the models to adapt. Levels of risk can change quickly over time as market participants react to the arrival of new information. This new information can be split into two parts. The first is unexpected news. The second type of information is announcements. In this case the time of the announcement is known but the content is unknown. Conditional heteroskedasticity models (GARCH and ARCH) are one way to describe time varying volatility. However, these models are not directly linked to the market sentiment. It is also difficult to incorporate GARCH processes to describe volatility for a large number of assets. In particular the relationship between the different assets needs to be described. BARRA has used GARCH processes to improve their factor and asset-specific variance estimates.

diBartolomeo and Warrick (2005) note that to account for the lack of historical data to estimate returns over longer periods, daily volatility predictions are often scaled up by the square root of time, which implicitly assumes an independent and identical distribution of security returns over time. However, this approach is not compatible with GARCH processes wherein volatility is presumed to vary over time, and returns are presumed not to be independent from period to period.

Also as diBartolomeo et al. (2005) describes these models can display counter-intuitive behaviour that fails to account for the way announcements effect markets. If the market expects an announcement about a particular company, trading volatility may fall as investors wait to see what the content of the announcement is. When the content of the announcement becomes known traders will react quickly to this information and volatility will jump. However, having reacted to the market announcement traders will then reduce their level of trading for this stock and volatility levels will fall again. A GARCH model describes volatility clustering behaviour. The current volatility is described in terms of the previous period's volatility. So high volatility in one period will influence the model to predict higher volatility for the period. Similiarly a period of low volatility will influence the model's prediction of volatility for the following period. As the market is quiet prior to the announcement date, the GARCH model will predict low volatility on the announcement date, when in fact volatility will be high. Then the model adjusts to predict high volatility on the following day when volatility will then fall.

The focus of the present study is to investigate the relationship between news and market volatility of asset prices. Jalen (2008) finds there is a relatively strong correlation between asset price volatility and news sentiment. Ederington and Lee (1993) studies the impact of information releases on market level uncertainty on interest rate and foreign exchange futures markets.

Security and market volatility vary over time as conditions change and new information becomes available to investors. Option traders respond quickly to new information that impacts expectations of future volatility because option prices are directly dependent on such volatility expectations. As such, changes in the level of option-implied volatility can be used as a measure of the extent to which market participants believe current conditions that affect volatility are different from their typical state. Hence these models should capture their considered behavior and should help give more sensible estimates of future volatility.

An alternative way to account for changes in market conditions that are manifested as time varying volatility is through the use of quantified news. For example, if on a typical trading day there are ten to fifteen news wire service stories about Firm X, and today there are two hundred news wire service stories about Firm X, we can assert that there is a significantly greater than usual amount of information being imparted to investors about this firm. As such, more substantial share price movements may result than would be typical. We might even be able to analyze whether the content of the news stories would be considered broadly negative or positive with respect to the operations or valuation of Firm X. In essence, the volume

and nature of textual news can be used like option-implied volatility to very rapidly adjust our expectations of future volatility for a particular firm or an entire market.

For a review of both GARCH and implied volatility models that describe the impact of information arrival on volatility levels see diBartolomeo & Warrick (2005).

2 Model description

The model provides updated estimates of portfolio volatility using information about changes to the market environment. We describe in this section a slightly modified form of the model outlined in diBartolomeo & Warrick (2005) which updates traditional factor risk estimates using option implied volatility. This model is extended in the following section with quantified news inputs.

The model is described in two parts. The first is a “basic” statistical factor model. In the second part the factor variance estimates are updated to account for changes in the option implied volatility levels. The asset covariance matrix is re-estimated, using the updated factor variances, to give an improved set of risk estimates.

We construct a statistical factor model applying traditional principal component analysis to extract orthogonal factors.¹ For a general factor model, the variance of each asset is given as a linear combination of the factor variances and an asset specific variance.

$$V_{kt} = \sum_{i=1}^F \sum_{j=1}^F \beta_{kit} \beta_{kjt} \sigma_{it} \sigma_{jt} \rho_{ijt} + \sigma_{s(k)t}^2$$

Sets and indices

$k \in (1, \dots, N_1)$ denotes the asset universe,

$t \in (1, \dots, T)$ denotes time points considered,

$i, j \in (1, \dots, F)$ denotes the factors,

Parameters

V_{kt} denotes the variance for asset k at time point $t \in (1, \dots, T)$,

β_{kit} denotes the factor sensitivity (exposure) to factor i for asset k at time point t ,

σ_{it} denotes the factor variance for factor i at time point t ,

ρ_{ijt} denotes the correlation between factor i and factor j at time point t

$\sigma_{s(k)t}^2$ denotes the asset specific variance for asset k at time point t .

In this case, as the factors are orthogonal, this simplifies to,

$$V_{kt} = \sum_{i=1}^F \beta_{kit}^2 \sigma_{it}^2 + \sigma_{s(k)t}^2 \quad (1)$$

These (asset) variances can be updated by considering the relationship between the implied volatility and factor model volatility. However, as diBartolomeo and Warrick (2005) note implied volatility is often a biased estimator of the actual asset volatility. This is because option pricing methods such as Black-Scholes (1973) are based on the assumption that option positions can be hedged continuously at no cost. In the real-world hedging is costly and positions are hedged periodically not continuously. Positions which are assumed to be risk-less under this pricing framework actually do carry some risk. Traders compensate for this and bias their risk estimates upward. To avoid this problem diBartolomeo and Warrick (2005) consider the re-

¹The computational experiments are carried out using the component assets of the Eurostoxx 50, so the number of time periods $T > N$ number of assets and we are able to carry out a principal component analysis on the sample covariance matrix without the problem of the matrix becoming non-singular.

relationship between the changes in the implied volatility and changes in the basic factor model volatility levels.

We can only study the relationship of those assets for which we have option implied volatility data. We can update our conditional estimate of the asset variance at time t , derived from the principal component model to

$$V_{\ell t}^* = V_{\ell t} M_{\ell t} \quad (2)$$

where $M_{\ell t}$ is an adjustment at time t , defined as,

$$M_{\ell t} = \left[\prod_{r=0}^{w-1} \frac{I_{\ell t-r}}{I_{\ell t-r-1}} \right] \div \left[\prod_{r=0}^{w-1} \frac{V_{\ell t-r}}{V_{\ell t-r-1}} \right] \quad (3)$$

Set and indices

$\ell \in (1, \dots, N_2)$ denotes the assets for which option implied volatility data is available,

Parameters

$I_{\ell t}$ denotes the option implied variance observed for security ℓ at time point t ,

$V_{\ell t}^*$ denotes the updated variance for security ℓ at time point t ,

w denotes the period considered for updating information.

The option implied volatility (equivalently variance) levels update faster than the factor model estimates, so the changes in this relationship should give us an improved estimate of future risk. For each asset ℓ we have,

$$V_{\ell t}^* = \sum \beta_{\ell it}^2 (\sigma_{it}^*)^2 + \sigma_{s(\ell)t}^2 \quad (4)$$

where $(\sigma_{it}^*)^2$ are new factor variances implied by the updated asset variances. We solve this set of simultaneous equations to derive the updated factor variances, $(\hat{\sigma}_{it}^*)^2$, which minimise the mean squared error, subject to the condition these values are non-negative. We also introduce the further constraint,

$$(\sigma_{it}^*)^2 \geq (\sigma_{i-1t}^*)^2 \quad (5)$$

to allow for the structure that is expected of the principal component factors. Though factor volatility may rise suddenly as market conditions change, there are few economic circumstances where it would be expected to decline dramatically from one time period to the next. It is also prudent to assume that it would not decline substantially, hence we introduce the constraint,

$$(\sigma_{it}^*)^2 \geq p_1 (\sigma_{i-1}^*)^2 \quad (6)$$

In equation (4) the asset specific variances are taken as the previous period's known values. Once the updated factor variances are derived, the asset specific variances can be re-calculated as,

$$\sigma_{s(\ell)t}^2 = V_{\ell t}^* - \sum \beta_{\ell it}^2 (\hat{\sigma}_{it}^*)^2 \quad (7)$$

As with the factor variances, we do not expect the asset specific variances to decline substantially from one period to the next and we set,

$$\sigma_{s(\ell)t}^2 = \max[\sigma_{s(\ell)t}^2, \sigma_{s(\ell)(t-1)}^2 \times p_2] \quad (8)$$

The updated factor variance estimates are used to re-estimate all the assets variances and covariances. We do not need the relationship (4) to be given for all the assets in the asset universe. As a result we do not have to directly identify which changes in the option implied volatility impact which factors and to what extent. These changes are derived implicitly, by considering the relationship of the changes between the factor model variance estimates and the option implied variance estimates.

3 Updating model volatility using quantified news

There is a strong, yet complex relationship between market sentiment and news. Traders and other market participants digest news rapidly and update their asset positions accordingly. Most traders have access to newswires at their desks. However, for models to incorporate news directly and automatically, we require quantitative inputs whereas raw news is qualitative data.

RavenPack have developed linguistic analytics which process the textual input of news stories to determine quantitative sentiment scores. In particular, they classify individual stories by the market aspects to which they relate; they also assign sentiment indicators which define a story as positive, negative or neutral. These methods are then applied to derive specific scores about different market entities such as a company or an industry sector. Scores which indicate the relative sentiment for a stock over time have been produced; for further details of how these scores are calculated and more specific details of their methodology, see Appendix A.

The score for an individual company varies over time but this time series is defined over time points with uneven intervals, as news stories arrive unexpectedly. We wish to use the information about the changing market sentiment to update our beliefs about factor volatility. The score, a_{nt} , measures the market sentiment about company n , at time t . ($n \in (1, \dots, N_2)$ denotes the assets for which option implied volatility and market sentiment data is available.) If this score varies significantly over time, market beliefs about the company are changing quickly, which indicates rising volatility of the stock.

We calculate the average value of the score over 15 minute intervals and then calculate the variance of these values over one day, b_{nt} . It is assumed that the working day starts from 16:00 the previous day and finishes at 15:59 of the current day. Unlike market data which is only available for the hours that the markets are open, news data is published outside market hours. Finally we calculate S_{nt} as the cumulative sum, of the variances of scores, over the past seven days.

$$S_{nt} = \sum_{r=0}^6 b_{n \ t-r} \quad (9)$$

If a particular company is in the news and its sentiment is changing significantly over time this could indicate its volatility has risen. The following day this company could become “old” news and its score may not vary much, however, there is no reason to believe its volatility has suddenly dropped. Cumulating over seven days allows us to account for this. We use seven days so that weekend news is always included. Excluding the weekend entirely seems inappropriate as markets will account for news published then. However, weekend news may not be processed in the same way as weekday news, so it seems to be inappropriate to include it for some days but not others. S_{nt} is defined so that it incorporates a directional change (up or down) in asset volatility and also the size of the change.

Consider the adjustment using the implied volatility described in equation (3); in a comparable way we define a second adjustment based on the news sentiment information.

$$M_{nt}^S = \left[\prod_{r=0}^{w-1} \frac{S_{n \ t-r}}{S_{n \ t-r-1}} \right] \div \left[\prod_{r=0}^{w-1} \frac{V_{n \ t-r}}{V_{n \ t-r-1}} \right] \quad (10)$$

We derive updated factor and asset specific variances, using first the option implied data and then the news sentiment data. We denote the updated factor and asset specific variances, which are determined using the option implied data as $(^O\sigma_i^*)^2$ and $(^O\sigma_{s(t)}^*)^2$ respectively. Likewise, $(^N\sigma_i^*)^2$ and $(^N\sigma_{s(t)}^*)^2$ are given from news sentiment data. The time subscripts have been dropped to aid readability.

We combine these variances to give risk estimates based on both sources of information. The combined factor variances are defined as

$$({}^C\sigma_i^*)^2 = q({}^O\sigma_i^*)^2 + (1 - q)({}^N\sigma_i^*)^2 \quad (11)$$

where $0 \leq q \leq 1$.

The asset specific variances are updated as

$$({}^C\sigma_{s(t)}^*)^2 = q({}^O\sigma_{s(t)}^*)^2 + (1 - q)({}^N\sigma_{s(t)}^*)^2 \quad (12)$$

In the case that $({}^O\sigma_{s(t)}^*)^2$ is defined but $({}^N\sigma_{s(t)}^*)^2$ is not (this is the case when option implied data is available for a stock but news sentiment data is not available), we use $({}^O\sigma_{s(t)}^*)^2$ and vica versa.

4 Computational experiments

Two separate computational studies were undertaken. The first covers the period in January 2008 when equity markets were starting to decline; the component stocks of the EURO STOXX 50 index are considered. The second study covers a period in September 2008 when the global economy was beginning to move into recession; the component stocks of the Dow Jones 30 are considered.

STUDY I

The first study covers the period 17 January 2008 to 23 January 2008 when sentiment worsened and option implied volatility measures surged. Over this period worldwide stocks markets fell significantly. From 2003 equity markets had been growing steadily, but at the end of 2007 they started to decline and sentiment started to fall. Over January 2008 market sentiment worsened further. This was driven by a few key events. In the US, George Bush announced a stimulus plan for the economy. The Fed cut interest rates by 75 basis points, the largest cut since October 1984. In Europe Societe Generale was hit by the fraud scandal of alleged rogue trader Jerome Kerviel. Asian stock markets also fell in this period.

The table below shows the volatility values of a portfolio of finance stocks weighted by their market capitalisations. The first column shows the values predicted by the “basic” factor model, the second the values from the model updated using only option implied volatility data($q=1$), the third the values from the model updated using a combination of the option implied volatility data and sentiment data($q=0.5$) and the final column the values for the model updated using only sentiment data($q=0$).

In this study 25 factors were used as this explained 90% of the historic volatility; also $p_1 = 90\%$ and $p_2 = 75\%$.

Table 1: Volatility of portfolio of EURO STOXX 50 finance stocks

Dates	Volatility under “basic” statistical model	Volatility under model updated by option implied volatility ($q=1$)	Volatility under model updated by option implied volatility and market sentiment ($q=0.5$)	Volatility under model updated by market sentiment ($q=0$)
17 1 2008	19.065	19.130	20.853	22.430
18 1 2008	19.032	21.564	21.619	21.625
21 1 2008	19.319	26.575	28.845	30.845
22 1 2008	21.187	26.759	28.911	30.829
23 1 2008	21.453	26.212	27.869	29.370

On 21 January 2008 there was a sharp decline on non-US stock markets. (The US market was closed) It is reasonable to assume that stock volatility rose on this date. The portfolio volatility estimate from the model updated using option implied data is higher than that from the “basic” model and it rises significantly on 21 January. The estimate from the market sentiment(news) model is higher and this value rises earlier than the option implied model, though there is still a significant increase on 21 January. This could indicate

that the model is picking up increased volatility at an earlier date than the option implied volatility. This seems sensible as news and market sentiment changes precedes changes in actual price volatility. (Traders first process news and then trade on their knowledge and beliefs.) Hence, this type of model can provide us with an “early” indication or warning that volatility is rising. The improved volatility model accounts for rapid changes in market sentiment which results in relatively large movements in equity portfolio risk.

It should be noted that sentiment indicators have the potential to be used, not only to adjust the expected return variance of an investment, but also higher moments such as skew and kurtosis. To the extent that such higher moment expectations arise from our process, their influence on the variance forecast can be incorporated by standard mathematical means such as the Cornish-Fisher expansion.

STUDY II

Over 2008 global equity markets continued to fall. This was heavily influenced by the severe loss of liquidity in credit markets and the banking system. Many large and well established investment and commercial banks suffered bankruptcy or were propped up by governments. Volatility for financial stocks over September and October 2008 was particularly high. Specific events that contributed to the volatility of the financial sector include Lehman’s filing for bankruptcy, Bank of America’s announcement of its intention to purchase Merrill Lynch, the Fed’s announcement of AIG rescue, Lloyds takeover of HBOS and on 19 September restrictions imposed on short selling of financial stocks. This second study covers the period 18 September 2008 to 24 September 2008.

Table 2 shows the volatility for a portfolio of three finance stocks with equal weights on each stock: Bank of America, CitiGroup and J.P. Morgan Chase. Similarly Table 3 shows the figures for a portfolio of three non-finance stocks: Johnson and Johnson, Kraft Foods and Coca Cola. The first columns show the values predicted by the “basic” factor model, the second the values from the model updated using only option implied volatility data ($q=1$), the third the values from the model updated using a combination of the option implied volatility data and sentiment data ($q=0.5$) and the final columns the values for the model updated using only sentiment data ($q=0$).

In this study 14 factors were used as this explained 90% of the historic volatility; also $p_1 = 90\%$ and $p_2 = 75\%$.

Table 2: Volatility of portfolio of Dow Jones 30 finance stocks

Dates	Volatility under “basic” statistical model	Volatility under model updated by option implied volatility ($q=1$)	Volatility under model updated by option implied volatility and market sentiment ($q=0.5$)	Volatility under model updated by market sentiment ($q=0$)
18 9 2008	56.031	71.023	70.326	69.622
19 9 2008	57.949	67.770	72.765	77.439
22 9 2008	61.719	66.302	71.014	75.433
23 9 2008	62.270	62.766	67.557	72.030
24 9 2008	62.279	59.531	63.968	68.118

In most cases there is higher volatility for the finance portfolio when the volatility estimate is updated using option implied data and likewise are found to increase when the news sentiment data is processed. On comparing the estimates for the finance and non-finance companies we see that the finance stocks volatility has risen significantly more than the non-finance stocks. This seems a sensible result for this period, given the market conditions and the news. Figure 1 shows the changes in prices from August to the end of October. It can be seen that the prices for the financial stocks show higher variation and this increases during September.

The differences between the volatility estimates using news sentiment and option implied volatility serves to highlight the complex nature of news and the way it impacts markets. This study and the scores used are based on the relative volume of negative and positive news items over a period of time. However, they do

Table 3: Volatility of portfolio of Dow Jones 30 non-finance stocks

Dates	Volatility under "basic" statistical model	Volatility under model updated by option implied volatility ($q=1$)	Volatility under model updated by option implied volatility and market sentiment ($q=0.5$)	Volatility under model updated by market sentiment ($q=0$)
18 9 2008	13.751	15.474	15.274	15.196
19 9 2008	13.912	14.907	15.392	16.214
22 9 2008	13.935	15.109	15.819	15.933
23 9 2008	14.316	15.159	16.593	16.709
24 9 2008	14.360	14.169	16.021	16.443

not account for how different news items may impact market prices and volatility differently. We note the importance of using a variety of sources of information when updating risk estimates.

These computational experiments are illustrative; in order to further exploit the value of quantified news, substantial additional work is needed to refine the process by which news indicators are used to form conditional volatility forecasts and to tune the adjustments to subsequent realisations. A formal Bayesian framework for such inclusion is described in Shah (2008), <http://www.northinfo.com/documents/286.pdf>.

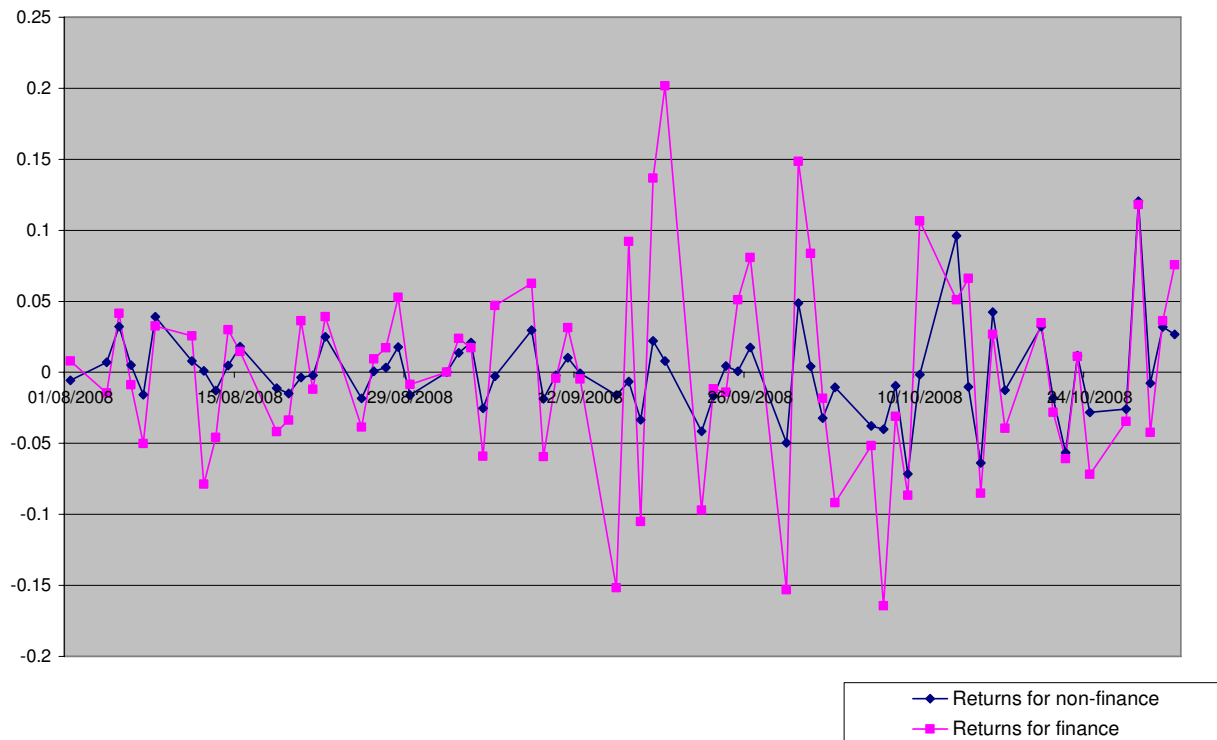


Figure 1: Returns for finance and non-finance portfolio over August to October 2008

5 Discussion and conclusions

In this paper we address the problem of making equity portfolio risk estimates sensitive to changes in the market environment and investor sentiment. Traditional multi factor risk models fail to update quickly as new information becomes available. diBartolomeo & Warrick (2005) use option implied volatility to determine improved estimates of the future covariance matrix. There is a strong, yet complex relationship between market sentiment and news. Traders and other market participants digest news rapidly and update their asset positions accordingly. However, for models to incorporate news directly and automatically, we require quantitative inputs, whereas raw news is qualitative data. RavenPack have developed linguistic analytics which process the textual input of news stories to determine quantitative sentiment scores.

To the extent that we are interested in risk estimation over a relatively short future horizon, conventional factor model methods of estimating security and portfolio risk can be made more responsive to changing conditions by conditioning the forecasts on changes in implied volatility and quantified news. We have presented a tractable method of including both option-implied volatility and quantified news into portfolio risk estimation.

While much research remains to be done to refine our methods, frequent crises in financial markets remind us of the urgency with which all investors, even those with a long term orientation, should be attentive to short term fluctuations in financial market risk. Implicit in the wealth accumulation goals of every investor is the assumption of survival: “To finish first, you first must finish.”

6 Acknowledgements

CARISMA and Ms Leela Mitra gratefully acknowledge the financial sponsorship provided by RavenPack International S.L.. RavenPack also supplied the news sentiment data used in this study see Appendix for further details.

A Sentiment analytics overview

RavenPack has developed linguistic analytics which process the textual input of news stories to determine quantitative sentiment scores. These scores allow us to incorporate information about the volume and nature of news into quantitative models. We give a brief description of how these have been created.

TAGGING PROCESS

As a news story is received from a newswire it is tagged to record the various linguistic aspects. One particular aspect is a story’s “aboutness”. This incorporates the entities to which the story applies, the subjects it covers, and the market to which it is relevant.

This analysis is applied to tens of thousands of stories per day aggregated from RavenPack’s compilation of diverse and respected sources of news.

SENTIMENT CLASSIFIERS

RavenPack’s sentiment classifiers detect story type as a preliminary step to distinguishing the story as being “positive” (POS), “negative” (NEG) or “neutral” (NEU) relative to a specific market or asset class. There are two main methods for detecting sentiment. The Expert Consensus Method uses financial experts’ tagging of several thousand stories as POS, NEG or NEU to train a Bayes Classifier which discerns rules from the training set to imitate the experts’ tagging. The Traditional Method maps specific words or phrases to pre-defined sentiment values.

SCORE CALCULATION

The tagging of individual stories can be used to aggregate sentiment scores of specific companies, such as the components of the EURO STOXX 50. Such scores indicate the relative news sentiment about each stock over time. The scores account for stories about the company and the sector in which it operates, thus creating continuous counts of the relative volume of positive and negative stories. For each company six time series of scores are derived, one based on each of five sentiment classifiers (WLE_SCORE, PCM_SCORE, ECM_SCORE, RCM_SCORE, VCM_SCORE) and one aggregate score (AGG_SCORE). Further descriptions of these classifiers are given below.

As a news item (story) s_t is received on the newswire at time t , it is classified by the WLE classifier as “positive” (POS), “negative” (NEG) or “neutral” (NEU). We define $I_{Pos\ s_t}$ an indicator function which takes value 1 when s_t is POS and 0 otherwise. We define a similar function for NEG. Further the story s_t has a relevance r_{s_t} attached to it. The unscaled score for the company under this classifier is defined as

$$R = \frac{\sum_{q=t-1}^{t-Q} I_{Pos\ s_q} r_{s_q} - \sum_{q=t-1}^{t-Q} I_{Neg\ s_q} r_{s_q}}{\sum_{q=t-1}^{t-Q} r_{s_q}} \quad (13)$$

The times considered are $t - Q, \dots, t - 1$ which are the time points in the 24 hours prior to t when news stories relevant to the company were received. This is scaled to give the score

$$T = \mathbf{sign}(R) \sqrt{|R|} \quad (14)$$

This gives values over the range $[-1, 1]$. By applying the relation $WLE_SCORE = (T + 1) \times 50$ the values are shifted and scaled to lie in the range $[0, 100]$. This computational process is repeated for each classifier to produce four further time series of scores. (PCM_SCORE, ECM_SCORE, RCM_SCORE and VCM_SCORE) A weighted average of these scores is finally used to give AGG_SCORE the value of a in our computation.

SUMMARY OF CLASSIFIERS AND SCORES

WLE_SCORE - A raw score that represents the aggregate news sentiment for the given company over the given time period according to the WLE classifier, which specializes in identifying positive and negative words and phrases in articles about global equities. This sentiment score is based on RavenPack's Traditional Methodology.

PCM_SCORE - A raw score that represents the aggregate news sentiment for the given company over the given time period according to the PCM classifier, which specializes in identifying the sentiment of stories that are about global equity future earnings developments and projections. This sentiment score is based on RavenPack's Expert Consensus Methodology.

ECM_SCORE - A raw score that represents the aggregate news sentiment for the given company over the given time period according to the ECM classifier, which specializes in short commentary and editorials on global equity markets. This sentiment score is based on RavenPack's Expert Consensus Methodology.

RCM_SCORE - A raw score that represents the aggregate news sentiment for the given company over the given time period according to the RCM classifier, which specializes in corporate action announcements. This sentiment score is based on RavenPack's Expert Consensus Methodology.

VCM_SCORE - A raw score that represents the aggregate news sentiment for the given company over the given time period according to the VCM classifier, which specializes in news stories about mergers, acquisitions and takeovers. This sentiment score is based on RavenPack's Expert Consensus Methodology.

AGG_SCORE - An overall interpreted sentiment score based on weightings of WLE_SCORE, PCM_SCORE, ECM_SCORE, RCM_SCORE and VCM_SCORE. This identifies the overall news sentiment for the given company over the given time period.

References

- [1] Black, F., & Scholes, M. 1973 “The Pricing Of Options And Corporate Liabilities” *The Journal of Business*,v81(3), 637-654.
- [2] Chen, N., Roll, R., Ross, S., 1986 “Economic forces and the stockmarket” *The Journal of Business*, 59(3) 383-404
- [3] Connor, G. 1995 “The three types of factor models: A comparison of their explanatory power” *Financial Analysts Journal*, 42-46
- [4] diBartolomeo, D. and Warrick, S. 2005 “Making covariance based portfolio risk models sensitive to the rate at which markets reflect new information ” Ch12 in *Linear Factor models* Edited. Knight, J. and Satchell, S. Elsevier Finance
- [5] Ederington, L., and Lee, J.. 1993 “How Markets Process Information: News Releases and Volatility” *Journal of Finance*, 48 p.11611161
- [6] Fama, F. and French, K. 1992 “The cross-section of expected stock returns” *Journal of Finance*, 47 p. 427-465
- [7] Fama, F. and French, K. 1993 “Common risk factors in the returns of stocks and bonds” *Journal of Financial Economics*, 33 p.3-56
- [8] Goldman Sachs 2008 “Headline Numbers The effects of news on market microstructures” *Source: Goldman Sachs*
- [9] Jalen 2008 “News scores for Euro Stoxx 50” *Source: RavenPack International S.L.*
- [10] RavenPack 2008 “RavenPack’s Analytics Knowledge Base” *Source: RavenPack International S.L.*
- [11] Scowcroft, A. and Sefton, J. 2006 “Understanding factor models” *UBS Investment Research*
- [12] Sharpe, W. 1970 “Portfolio theory and capital markets” *Mc-Graw Hill New York*
- [13] Sheikh A. 1995 “BARRAs risk models” *Internal working paper*
- [14] Sun. X., and Mitra, G. 2008 “The impact of news sentiment on volatilities” *Source: RavenPack International S.L.*
- [15] Tetlock, P. C., Saar-Tsechansky, M., and Macskassy, S., 2007 “More Than Words: Quantifying Language to Measure Firms Fundamentals” *Journal of Finance* Forthcoming