PERIODICA POLYTECHNICA SER. SOC. MAN. SCI. VOL. 11, NO. 2, PP. 201–214 (2003)

ESTIMATION OF FUTURE VOLATILITY

Csaba SOCZO

Department of Finance and Accounting
Budapest University of Technology and Economics
H-1111, Budapest Stoczek u. 2. St. ép., Hungary
Phone: +36 1 463-2379, Fax: +36 1 463-2745

Received: April 23, 2003

Abstract

Estimation of market risk due to price and interest rate movements is inevitable for financial institutions with huge trading portfolios. A basic indicator of the risk exposure could be derived from the standard deviation or volatility of the financial instruments. Accounting for the effect of diversification, other parameters should be determined for the risk analysis of portfolios. Correlations could describe the connection between pairs of basic financial instruments, so they could be used for the risk analysis of portfolios. We are interested in the risk level of assets in the future, therefore we need the future estimates of these risk parameters.

These parameters could be assessed in different ways. First we can assume, that their future expected values could be determined from past data. However, there is another way for the volatility, in case of there is option trading for the specific instrument. In this case implied volatility could be calculated from the option price, which could be an estimate for future volatility. In case if both methods are available, we need some hints, which is better. On the other hand, the estimation of correlations involves filling in missing values from efficiency standpoints, when incomplete time series are available. We argue that data filling raises several problems, which depreciate the supposed benefits of this approach.

Keywords: market risk assessment, implied volatility, future volatility, data filling.

1. Introduction

The purpose of this paper is to investigate estimation methods for volatility and correlations of financial instruments. As we know these parameters are crucial for the risk assessment of portfolios. Future volatility could be determined by the implied volatility defined by the option pricing models, or it could be calculated from time series. Early tests indicated that past-realized volatility has better predictor power than implied volatility. Current analysis doubts this result; moreover they found an adversary relationship, because it has been shown that the implied volatility is a much better estimator. On the other hand, correlations should also be calculated for the risk assessment of portfolios. If the holiday schedules of financial markets are different, we can experience days when some market prices are missing. In this case traditional estimator functions could not be used, therefore a new approach is needed to involve those days when at least one price data is missing. Many methods were introduced to handle this problem, however many of them provide biased and/or inefficient estimates. The Expectation Maximization method is a great theoretical framework for data filling and it is offered for financial time series also. However this approach raises some problems, which will be addressed in the last sections.

The first chapter presents a theory describing the stochastic process of financial assets' price movements. The random walk theory is a widely used model for price movements, which contains two parameters to describe the process. One of these parameters is the standard deviation or volatility. There are different ways to estimate this coefficient such as the estimate of the population standard deviation, the Exponentially Weighted Moving Average Model or the GARCH (1,1) model. These are introduced in the second section. It is common in these models that they try to estimate future volatility with past-realized volatility.

The next section presents some attributes of options and an option pricing theory, the BS model. Here are described the main variables of this approach. The fourth section presents the consequence of the uniqueness of one of these variables. The volatility could not be observed directly although there are numerous models to estimate its value from historical data. Another approach is to extract implied volatility from the current option price and the values of the other variables of the option-pricing model.

The fifth section presents some current results related to the relationship between implied volatility, past realized volatility and future volatility. These results are surprising considering earlier studies, although the analysis of the sampling method explains the difference. The BS model involves some annoying assumptions, but we would like to present some current findings that give quite good estimate for implied volatility compared to other, more sophisticated models.

The last sections are devoted to the missing data problem in case of the estimation of correlation. There are several methods to handle this difficulty, however, they usually raise further problematic issues.

2. Stochastic Process of a Financial Asset's Price

Modern theories of finance use random walk models to describe the price dynamics of financial assets like shares or stock indices [8]. These models could be demonstrated by the following equation:

$$\Delta x = a(x,t)\Delta t + b(x,t)\varepsilon(\Delta t)^{0.5}, \tag{1}$$

where Δx is the asset's absolute return, Δt is the change in time and ε is a random drawing from a standardized normal distribution, and a, b are the parameters of the process. If they are constant the process is called Generalized Wiener process. However, an Ito process permits these parameters to change in x and t. The linear trend of the mean (the grows rate) could be identified as a, while the volatility of the share's return exhibits b $(\Delta t)^{0.5}$. The basic models of the stochastic process of share price movements suppose constant parameters in time and the volatility is considered constant. In this case the price dynamics could be described by the following formula:

$$\Delta x = \mu x \Delta t + \sigma x \varepsilon (\Delta t)^{0.5} \quad \text{or} \quad \frac{\Delta x}{x} = \mu \Delta t + \sigma \varepsilon (\Delta t)^{0.5}, \tag{2}$$

where $\mu \Delta t$ is the expected value of return and $\sigma(\Delta t)^{0.5}$ is the standard deviation (or volatility) of the share's return. It is clear that y possesses normal distribution with the earlier explained parameters. Using Ito's Lemma, it could be proven, that the natural logarithm of the share price follows a Generalized Wiener process as well, which is a consequence of $\frac{\Delta x}{x} = \mu \Delta t + \sigma \varepsilon (\Delta t)^{0.5}$, (2). This means that the share price has lognormal distribution [8]:

$$\Delta \ln x = (\mu - \sigma^2/2)\Delta t + \sigma \varepsilon (\Delta t)^{0.5}. \tag{3}$$

Although the assumption of constant volatility could sound reasonable, many tests show that it is not adequate [8, 10, 16] . This involves primary deviation of estimates calculated by models based on the assumption of stationariness compared to theoretical values and significant inconsistencies are implied by these kinds of models.

3. Estimating Current and Future Volatilities from Historical Data

The simplest estimation of volatility is based on the assumption that volatility is constant in time. In this case an unbiased estimation of the volatility is the estimate of the population's variance:

$$\sigma_n^2 = \frac{1}{T-1} \sum_{t=1}^{T} y_{n-t}^2, \tag{4}$$

where y_{n-t} is the daily return. This approach implies that future volatility equals to the current estimate.

According to many empirical tests, volatility is a function of time. There are many sophisticated methods to estimate future volatility, which consider the fact of changing volatility, such as the Exponentially Weighted Moving Average Model (EWMA). It is an efficient method to estimate future volatility. In this case a simple formula gives the estimation of current volatility:

$$\sigma_n^2 = \lambda \sigma_{n-1}^2 + (1-\lambda) y_{n-1}^2, \tag{5}$$

where σ_n is the estimation of volatility for day n, based on the calculation at the end of day $n-1$, λ is a parameter between 0 and 1, σ_{n-1} is the volatility estimate a day ago and y_{n-1} is the most recently observed daily return. According to MORGAN [11], $\lambda = 0.94$ gives forecasts which are the closest to the real volatility.

The GARCH (1,1) model provides a more general description of volatility movements. In this case the current volatility is estimated by the following formula:

$$\sigma_n^2 = \omega + \alpha y_{n-1}^2 + \beta \sigma_{n-1}^2, \tag{6}$$

where ω, α and β are the parameters of the estimation. It is clear that this model deviates from the earlier method in the ω constant term. This means that using this model we assume that the EWMA model is a biased estimate of the current volatility. The practical advantage of GARCH (1,1) is called mean reversion. Although the volatility moves randomly, it tends to be pulled back to a long run average level over time. The parameters could be calculated by the maximum likelihood methods.

The estimate of future volatility is defined by the GARCH (1,1) model according to the following formula [8]:

$$E(\sigma_{n+k}^2) = V + (\alpha + \beta)^k (\sigma_n^2 - V), \tag{7}$$

where $V = \omega/(1 - \alpha - \beta)$. The EWMA model's result could be derived from this formula. In this case $\alpha + \beta = 1$, and $\omega = 0$, therefore the expected future volatility equals the current estimate.

The models presented above suppose that future volatility could be estimated from past volatility.

4. Options, a Special Type of Derivatives

Derivatives are financial instruments whose prices (and price dynamics) depend on an other asset, the so-called underlying asset. Options are very important representatives of these kinds of contracts although their pricing methodology is quite complicated. Black, Scholes and Merton provided an analytical solution of option pricing with the help of Ito's Lemma. The Black-Scholes (BS) model provides analytical price formulas of European call and put options. According to this model, the option price is a function of the following variables [8]:

- current underlying asset price,
- risk free interest rate,
- volatility of the underlying asset during the life of the option,
- exercise (or strike) price,
- time to the maturity of the option.

The only parameter in the pricing formula that cannot be directly observed is the volatility of the stock price. This variable could be estimated by the methods described in the previous section.

The Black-Scholes model is a convenient analytical approach to define an option's price, however, it involves some inconsistent assumptions. A main limitation of the Black-Scholes model is that it is able to define prices only for European option. Most options on the U.S. financial market are American in nature, and the BS model will miss any values associated with premature exercise.

There are other option valuation models, which could handle early exercises and/or stochastic volatility. Unfortunately, these are numerical approaches and their usage could be very costly and time consuming. Although, the biases of the BS model are relevant, there are some tests demonstrating its surprising feasibility compared to more sophisticated methods.

5. Defining Implied Volatility

According to the earlier section, the only parameter of the BS model, which could not be observed directly, is the expected volatility during the life of the option.

Two different approaches are available to estimate this volatility value. The first was described in Section 3, when historical data of the underlying asset was used for the estimation. The other approach is to use the option pricing formula backward, calculating the implied volatility of the pricing formula belonging to a certain option price level observed on the market. This is an iterative method so we should change the volatility in the pricing formula until the calculated price equals to the current market value. The implied volatility level could be calculated for different exercise prices and maturities. An earlier research related to implied volatility shows that its value changes for different exercise prices [8]. This could be concluded from the so-called volatility smile figures, which present the implied volatility as a function of the strike price. This pattern was typical before the 1987 crash. After this event the pattern is more similar to a ‘sneer’ [4]. Ignoring the type of the figure, an important conclusion could be drawn that the BS pricing model’s assumptions are not right. The first consequence is that the volatility is not stable in time. The second is that the probability distribution of the underlying asset’s relative change is not normal distribution. From the view point of our discussion, the first statement is important. The analysis of the implied volatility as a function of maturity also suggests that the expected volatility changes in time.

There is a couple of models which were developed to handle this phenomenon but they are numerical and very computer sensitive. This discussion is restricted to the BS model, because using this theory surprising results could be developed.

6. Relation between Implied and Realized Volatility

In the earlier sections two approaches were demonstrated to estimate future volatility. The first method considers historical volatility to be a predictor of future volatility while the other uses implied volatility of option prices. It is interesting to investigate which value serves as a better estimator of the future volatility. An early research showed that implied volatility is a biased and inefficient forecast and it does not involve more information beyond the past realized volatility. However, some current papers suggest that it would be a mistake to eliminate implied volatility as an efficient predictor moreover it seems to be a better estimator than the past realized volatility. This surprising result could be explained by the different sampling methods, which will be presented later.

Investigating the relationship between future, past realized and implied volatility, the following formula could be established:

$$h_t = \alpha_0 + \alpha_i i_t + \alpha_h h_{t-1} + \varepsilon_t, \tag{8}$$

where h_t denotes the realized volatility for period t (which is aimed to be estimated by past realized and implied volatility), i_t is the implied volatility at the beginning of t, and h_{t-1} is the past realized volatility for period $t-1$, ε_t is the error term and α_0, α_i, α_h are constant parameters. This form of relationships permits expected future volatility to be dependent on both implied and past-realized volatility. Using OLS[1] estimation method, the parameters of the regression equation could be derived from historical data.

The analysis by CHRISTEEN et al. (1998) is based on S&P 500 index options. They calculated the regression parameters for the regression equation above. Additionally, they determined the regression equations when the implied volatility or the past realized volatility was used alone as a predictor variable. The surprising result is that implied volatility is a much better predictor than past realized volatility in both cases. Moreover, the slope coefficient of implied volatility is more than twice of the coefficient of past volatility for the regression equation containing both explanatory variables. This result means that the implied volatility defined by the BS model has more predictive power than the past-realized volatility and it contains more information about future volatility.

The trouble with the implied volatility defined by the BS model is associated with the slope coefficient for the implied volatility. In both cases the author found that it has a smaller value than 1. This fact could be a consequence of an errors-in-variable problem. Therefore we can conclude that the implied volatility is biased. The other fact that past-realized volatility remained a significant predictor when both types of volatility were included in the regression equation indicates that implied volatility is not efficient. To handle this problem CHRISTEEN et al. (1998) used an alternative specification. In this case the future implied volatility was estimated by implied volatility and past volatility from the earlier period, similarly to a GARCH(1,1) model according to the next equation:

$$i_t = \beta_0 + \beta_i i_{t-1} + \beta_h h_{t-1} + \varepsilon_{it}. \tag{9}$$

This model could be applied to test the predictor variables of implied volatility. The analysis shows that both past implied and realized volatility have the same significant predictor power on implied volatility. Handling the problem of biased

[1] Ordinary Least Squares method, for more details see any introductory statistical book on regression analysis.

and inefficient properties of implied volatility, implied volatility values were substituted with regressed values. (This is called instrumental variables procedure.) The earlier model was used to define the replacing values. Using this data series and the model of future volatility (8), new slope coefficients could be derived. It was found that the coefficient of replaced implied volatility series is nearly one while the coefficient for past-realized volatility is almost zero. This means that the new approach eliminated the bias and inefficiency of implied volatility.

It was already mentioned that earlier tests of implied volatility as an estimator of future realized volatility did not provide favourable results like that. This major difference could be explained by the different sampling method. A simple way to assess volatility is the estimate of the population standard deviation as it was described in section 1. For example, consider an option with time to expiration of T. The past (p) volatility for the prior period of P and the future realized (f) volatility could be calculated by the following equations:

$$\sigma^2_{p,T,t} = \frac{1}{P-1}\sum_{k=t-P}^{t} y_k^2 \tag{10}$$

$$\sigma^2_{f,T,t} = \frac{1}{T-t-1}\sum_{k=t}^{T} y_k^2, \tag{11}$$

where y_k is the daily rate of return. If we use daily consecutive series of historical data to define the coefficients for *Eq.* (8), the next term in the daily time series of past volatility is as follows

$$\sigma^2_{p,T,t+1} = \frac{1}{P-1}\sum_{k=t-P+1}^{t+1} y_k^2 \tag{12}$$

This involves that the two volatility estimates are almost identical. Therefore even if there is no correlation between the daily true volatilities, this type of analysis could prove autocorrelation. This argument is true for the future volatility values too. It is easy to ascertain that this type of estimation error could cause a wrong regression model.

Testing the method of generating volatility series from historical data, the estimation shows that the slope coefficient is much smaller for the implied volatility compared to the past volatility when only one of them is used as a predictor variable. If both were present in the regression *Eq.* (8) the coefficient for implied volatility was almost zero. That is the reason why non-overlapping data were used for the analysis referenced in this paper. The historical values were selected, so the time periods covered by successive options did not overlap.

7. Defending the BS Model

In the earlier sections some weaknesses of the BS model were mentioned. The model is unable to handle early exercise and it supposes constant volatility during the life of the option. However, American options permit early exercise and volatility is quite unstable in time. Therefore, the BS model could be biased. There are some numerical procedures to handle these conditions, for example the finite difference method [7] or the implied tree approach [4] with deterministic volatility function. These models introduce some arbitrary additional assumptions, for example the local volatility rate is supposed to be a deterministic function of asset price and time.

Some current tests show that the BS model performs very well compared to the numerical procedures thought to be more accurate and it appears more reliable. This fact confirms the old phrase that 'the simpler is better' [4]. The analysis proved that the reason of the weak performance could be assigned to the high instability of the deterministic volatility functions.

8. Portfolios

In the earlier chapters we have considered only individual assets, however, financial institutions usually handle large portfolios with many different types of financial instruments. In this situation the separate risk assessment of each individual instrument does not provide sufficient information about the risk level of the whole portfolio because of the well-known phenomenon of diversification. According to common sense, it is less risky to have different types of shares for example, because there is still chance to gain on the others in case of suffering losses on some instruments. Therefore, the price movements should be modelled based on their collective behaviour. According to simpler approaches, this problem could be supported by correlations between price movements of different instruments:

$$\text{Correlation}\left(y_i, y_j\right) = \overline{\left(y_i - \overline{y_i}\right)\left(y_j - \overline{y_j}\right)}, \tag{13}$$

where the line above the variables denote their mean values. The concept of correlation should be familiar from basic statistics courses, and it has large absolute value (near to one) in case if the relationship between two different assets are strong, and its sign is positive if the deviations from the mean values have probably the same direction for both variables. In case of no connection the correlation has zero value. The distribution of the whole portfolio's price changes could be calculated quite easily with the correlation matrix supposing normal distribution for the individual instruments and their aggregate value [8, 10, 11].

Stationary distributions are usually not appropriate for modelling financial instruments, because of the alternation of calm and noisy periods. Therefore, market prices are heteroscedastic with changing volatility. This phenomenon could be grabbed with different autoregressive and heteroscedastic models for volatility as it was presented earlier, but similar approaches could be introduced for correlations as well [5, 8]. It is clear that historical data is the only way to get information about the correlations of financial instruments, because implied correlations could not be calculated from derivative prices.

9. Correlation Estimation in Case of Missing Data

Calculating the correlation between two instruments, it is an important condition to analyse their price movements for the same time periods, for which daily changes are usually used. This requirement raises an important question, because different financial markets could have different holiday schedules. Therefore, there could be certain days with no information about some of the assets. This means that traditional estimation function for correlation$\left(\text{Correlation}(y_i, y_j) = \sum_t y_{i,t} y_{j,t} - \overline{y_i y_j}\right)$ could not be used, because this formula would contain missing values. Several methods are suggested to handle this problem. They range from the simplest of throwing out cases with one missing value at least to the more advanced methods of modelling missing data. It is clear that eliminating days when there is at least one missing data is not very efficient, because many data are thrown away which is wasting of scarce information. In finance, returns are usually modelled, which is a ratio of two consecutive prices. This deepens the problem of data missing, because one missing price data could lead to two holes in the time series of returns. Sophisticated modelling methods such as the Expectation Maximization (EM) method are suggested for filling in missing financial data. It is offered also by the RiskMetrics [11] methodology, which is an industry bench-mark. In the next chapter, this concept will be presented briefly for a better understanding.

10. Filling in Missing Data

For the risk analysis of a portfolio of different instruments, the historical time series of the underlying instruments are needed, regarding their price movements and financial returns in the past. Working with N different instruments through T trading days, the historical data could be represented in a matrix of N times T. Each row (or case) of this matrix contains the returns of all the instruments for the same day. However, the return matrix could have empty cells for realistic cases, because markets of financial instruments could have different holidays. Therefore, can be introduced a response R_{ij} matrix which contains information about the availability of return data. A cell in the response matrix is 1 if the corresponding cell in the return matrix contains data, and it is zero otherwise. Generally, the missing mechanism could be considered a random process when the contents of the response matrix are random variables. From the point of data filling we are interested in how the return data determines whether values are missing or not. Therefore, we can introduce a conditional probability for the contents of the response matrix depending on the return values:

$$P\left(R_{ij}=1 \mid Y\right)=P\left(R_{ij}=1 \mid Y_{\text{obs}}, Y_{\text{mis}}\right), \tag{14}$$

where Y_{obs} represents the values which were observed, while Y_{mis} denotes the ones which are missing. Different missing data processes could be defined depending on the structure of the probability function. For our purpose the missing process should be at least random, which means that the missing process depends mostly on the observed values. This condition is quite favorable, because the missing process should be investigated otherwise. Furthermore, the data missing in connection with holidays do not depend on any market value, but on the holiday schedule.

The EM method is a quite sophisticated model-based approach to analyse samples with missing data. For the risk analysis of portfolios, distributions and distribution parameters of returns as random variables should be assessed, so we are usually not interested in the expected values of missing data. This approach is quite advanced, because the desired parameters are estimated using the conditional distribution of the missing data depending on the observed values for the same day. This means that we do not have to generate point estimates for the missing data, which would decrease volatility. According to the EM methodology, the conditional distribution functions of missing values could be used to calculate the expected value of the traditional statistical estimator functions for the desired parameters. For this process an initial estimation is needed for the distribution parameters, which should be used to calculate the conditional distribution functions for the missing values. This initial estimation could be refined in the next step, therefore this method provides solution through an iterative process.

Usually, the maximum likelihood method is applied to generate unbiased estimates for distribution parameters. A priori we could have assumptions about the distributions of random variables[2], which could be described by specifying the class of their probability density functions (PDFs) (for example the normal distribution function). These a priori PDFs are called likelihood estimator functions and they are general enough to fit to a wide range of statistical processes. These PDFs contain parameters (Θ), which specify a given distribution in the class, and they are not fixed at this stage yet, furthermore the goal is to find the best parameter combination, which fits the given behaviour most accurately. This means that a PDF class specified by given distribution parameters is expected to describe the statistical behaviour of random variables. Samples are usually used to estimate these parameters. Formally, the PDF could be specified a posteriori for a given sample and distribution class as a likelihood function. Maximizing this function for Θ, we can estimate the parameters for which the observed sample is the most plausible. In case of missing data, the likelihood function contains the response matrix (R) and the distribution parameters of the missing process (Ψ) besides the observed values (supposing Θ and Ψ are independent, and the missing process is at least random) (FIRTH, LITTLE et al. (2002), SOCZO (2002)):

$$f\left(Y_{\text{obs}}, R; \Theta, \Psi\right)=f\left(Y_{\text{obs}}; \Theta\right) f\left(R \mid Y_{\text{obs}}; \Psi\right). \tag{15}$$

[2] In our case, these random variables are financial returns.

The main consequence of this result is that the distribution parameters of the random variables and the missing process could be analysed separately. The reason is that the maximization of the expression above could be accomplished through the individual analysis of the two parts of the right side, because the missing mechanism and the market process are independent. For our analysis we have to investigate only the first part, because it depends only on the desired parameters. Therefore, the optimal Θ could be defined through the maximization of the first part, which relates to the observed values and the distribution of the investigated random variables. This is a quite favorable result showing that the distribution parameters of returns could be estimated without involving the missing process in case of the missing process is at least random, and the distribution parameters of the returns and the missing process are different.

As we explained earlier, the likelihood estimator function could be used to define the likelihood function for a given sample, and the parameters of the distributions could be estimated by maximizing the likelihood function. In case of missing data, the a posteriori likelihood function is the marginal probability density function for the observed values by definition:

$$f\,(Y_{\text{obs}};\Theta)=\int f\,(Y_{\text{obs}},Y_{\text{mis}};\Theta)\ \mathrm{d}Y_{\text{mis}}. \tag{16}$$

This definition is taught in basic statistics meaning that the marginal PDF for a smaller set of variables could be generated through integrating the initial PDF over the eliminated variables. Therefore, the desired parameters could be estimated through the maximization of this marginal PDF for a given sample in case of missing values. However, this function could not be factorized generally, therefore standard estimator expressions for the distribution parameters could not be used. Fortunately, the EM method could be used to estimate the distribution parameters for samples with missing data through an iterative process.

The intuitive base for this model is to calculate the expected value of the initial log likelihood function (LLF) with missing values $(\log(f(Y_{\text{obs}},Y_{\text{mis}};\Theta)))$ over the possible values of missing elements, using the conditional distribution of the missing data based on the observed values. This expected function should be maximized to generate estimates for the distribution parameters. However, this process could be accomplished only if we know the distribution of the random variables a priori. But the latter could be estimated through the maximization of the expected value of the LLF. Therefore, this model contains the desired parameters implicitly which could be handled through an iterative process. At the beginning an initial estimate could be used for the distribution parameters to define the expected LLF for the observed values (expectation – E step). After that, a new estimate could be calculated through maximizing this likelihood function (maximization – M step)[3]

[3] Assuming stationary normal distribution, we are interested in the estimation of $\Theta=(\boldsymbol{\mu},\boldsymbol{\Sigma})$, the mean values ($\boldsymbol{\mu}$) and the covariance matrix ($\boldsymbol{\Sigma}$). In the 'E' part of step $k+1$, the conditional

We do not intend to present a detailed justification of the EM method, because it is not necessary for our conclusions. Therefore we should be satisfied with this heuristic introduction at this point, but mathematically correct and detailed arguments could be found in many great sources in conjunction with this theory [3, 12, 13, 15].

11. Problems Raised by Data Filling in Terms of Financial Time Series

We presented that missing data could make it difficult to calculate efficient estimates for correlation without data elimination, because traditional formulas could not be used. Fortunately many methods were developed such as the EM method, to handle this difficulty. The EM method is a quite sophisticated approach and its main advantage is to consider missing values random variables, whose conditional distribution should be estimated. Therefore, this method does not decrease volatility by substituting missing data with their expected values, using them as if they would have been observed. This means that more efficient and less biased estimates could be generated form the available data. However, the EM method raises several questions, which should be considered in conjunction with its implementation for financial time series.

Modelling financial returns for financial risk analysis means that one missing price data generates two missing returns. The reason is that the missing price data is the numerator of the formula generating the return for the same day, and it is the denominator for the next day's return. The EM method is used for financial returns, and estimates could be provided for the mean values of both missing returns attributed to the same missing price data. Two estimates could be generated for the missing price from the filled in returns and it is not guaranteed that both indirect price

distribution of the missing values should be calculated using the estimates of Θ^k determined in the earlier step:

$$\left\langle Y_{\text{mis},t} \mid Y_{\text{obs},t} \right\rangle_{\Theta^k} = \boldsymbol{\mu}^k_{Y_{\text{mis}},t} + \boldsymbol{\Sigma}^k_{Y_{\text{mis}_t},Y_{\text{obs}_t}} \boldsymbol{\Sigma}^k_{Y_{\text{obs}_t} Y_{\text{obs}_t}}{}^{-1} \left(Y_{\text{obs},t} - \boldsymbol{\mu}^k_{Y_{\text{obs},t}} \right)$$

$$\text{covariance } \left(Y_{\text{mis},t} \mid Y_{\text{obs},t} \right)_{\Theta^k} = \boldsymbol{\Sigma}^k_{Y_{\text{mis}_t},Y_{\text{mis}_t}} - \boldsymbol{\Sigma}^k_{Y_{\text{mis}_t},Y_{\text{obs}_t}} \boldsymbol{\Sigma}^k_{Y_{\text{obs}_t},Y_{\text{obs}_t}}{}^{-1} \boldsymbol{\Sigma}^k_{Y_{\text{obs}_t},Y_{\text{mis}_t}}$$

In the M step, new estimates are generated for the distribution parameters Θ^{k+1} using the traditional estimator functions for each day:

$$\hat{\mu}_i = \frac{1}{T} \sum_{t=1}^{T} Y_{i,t}, \qquad \hat{\Sigma}_{i,j} = \frac{1}{T} \sum_{t=1}^{T} \left(Y_{i,t} - \hat{\mu}_i \right) \left(Y_{j,t} - \hat{\mu}_j \right) = \frac{1}{T} \sum_{t=1}^{T} Y_{i,t} Y_{j,t} - \hat{\mu}_i \hat{\mu}_j .$$

Those parts, which contain missing values, should be substituted with their expected values determined in step E. These steps should be continued until the desired parameters converge. Through this process better estimates could be generated using all the cases, which are not necessarily complete. For a more detailed discussion of this application see [12, 17, 1].

guesses have the same values. Therefore, the model could involve inconsistency in case when both returns are estimated. This problem could be handled via two ways. First, we should choose one of the missing returns to be estimated then the missing price and the other return could be calculated. However, it is arbitrary which day's return to choose, therefore different estimates could be produced. Secondly, we can leave the original prices focusing on the daily returns only. This approach could be appropriate in case we are interested in the distributions of returns to generate estimates for future expected maximum losses for a given confidence level according to the VaR (Value at Risk) concept [8, 10]. In this case the arbitrariness is eliminated which is quite favorable from statistical standpoints.

Another problem relating to the EM method is more significant than the earlier one in respect of financial analysis. This issue is in connection with the assumption of the underlying process of returns. It is a quite basic assumption about financial returns to be stationary normally distributed according to a brown motion. However, it was shown that financial returns have alternating calm and noisy periods, when the volatility for different periods could vary significantly. Furthermore, the probability of extreme events is much higher than models with stationary normal distribution predict and there are more complex models with stochastic volatility preserving local normality. Therefore, the assumption of stationary normal distribution, for which the EM method is quite well implemented, is not appropriate. However, the EM method with stationary normality is suggested for example by the RiskMetrics methodology [11], which is considered an industry benchmark. The basic assumption of RiskMetrics is that log-returns follow normal distribution locally. The volatility could be generated according to the Exponentially Weighted Moving Average Method as it was presented earlier. Therefore, it is strange to use the stochastic volatility model for risk assessment and the EM method with stationary normality simultaneously for data filling. We claim that this assumption provides biased estimates because it has been proved that the volatility of financial returns is not stationary.

We can eliminate the inefficiency resulting from data elimination with the EM method, but inefficiency would be replaced with bias, which would be brought in by the assumption of stationary normality. Unfortunately, as far as we know, the EM method is invented fully for stationary normal distribution regarding financial returns but not for more complex models, which are able to describe financial processes better. Therefore, we think that more theoretical development is needed to generate less biased estimates for financial time series when incomplete cases are also used.

12. Conclusion

Measuring market risk, we need to estimate the future behaviour of financial instruments. For this purpose standard deviation and correlations could be quite useful to characterize individual behaviour of and the relation between instruments. We have seen there are two major concepts to estimate future volatility; one way is the assessment from historical data, while the other concepts are utilizing option pricing theory to get expected future volatility from option prices. The earlier test showed the former to be better estimates, however this result could be the reason of inappropriate testing methods. The latest research with better testing approaches revealed that the implied volatility derived from option prices is less biased.

For the risk analysis of portfolios, correlation should also be estimated for which past data could only be used. In case of financial markets having different holiday schedules, correlation calculations could be quite problematic with traditional methods unless the incomplete cases are eliminated. This solution would provide quite inefficient estimates, however, filling in missing data causes additional bias. Therefore, more improvement is needed to implement data filling methods for more realistic financial processes.

References

[1] ALDRICH, J., Lecture Notes for Statistical Theory, `http://www.economics.soton.ac.uk/courses/ec369`.

[2] CHRISTEEN, B. J. – PRABHALA,N. R., The Relation Between Implied and Realized Volatility, *Journal of Financial Economics*, **50** (1998), pp. 125–150.

[3] DELLAERT, F., The Expectation Maximization Method, *working paper*, `http://www.cc.gatech.edu/$ˆ\sim$dellaert.`

[4] DUMAS, B. – FLEMING, J. – WHALEY, R. E., Implied Volatility Function: Empirical Tests, *The Journal of Finance*, **53** (1998), No. 6, pp. 2059–2106.

[5] EMBRECHTS, P. – KLÜPPELBERG, C. – MIKOSCH T., *Modelling Extremal Events*, Springer, 1997.

[6] FIRTH, D., Lecture Notes for Advanced Social Statistics, `http://www.stats.ox.ac.uk/$\sim$firth/advss/.`

[7] HAMID, A Comparison of Implied Volatilities from European and American Futures Option Pricing Models, *Derivatives Quarterly*, 1998/Winter, pp. 42–51.

[8] HULL, J. C., Options, Futures, and Other Derivatives, *Prentice-Hall, Inc*, fourth edition, 1999.

[9] JONES, M. B. – NEUBERGER, A., Option Prices, Implied Price Processes and Stochastic Volatility, *The Journal of Finance*, **55** (2000), No. 2, pp. 839–866.

[10] JORION, P., *Value at Risk*, The McGraw-Hill Companies, Inc., USA, 1997.

[11] MORGAN, J. P./REUTERS,*Riskmetrics™- Technical Document*, Morgan Guaranty Trust Company of New York, USA, 1996.

[12] LITTLE, R. J. A. – RUBIN, D. B., *Statistical Analysis with Missing Data*, John Wiley & Sons, Inc., second edition, New Jersey, USA, 2002.

[13] MINKA, T., Expectation-Maximization as Lower Bound Maximization, *tutorial*, `http://citeseer.nj.nec.com/minka98expectationmaximization.html.`

[14] MORGAN, M., Cleaning Financial Data, *Financial Engineering News*, June/July, 2002.

[15] NEAL, R. – HINTON, G., A View of the EM Algorithm that Justifies Incremental, Sparse, and Other Variants In M. I. Jordan (editor) Learning in Graphical Models, pp. 355–368, Dordrecht: Kluwer Academic Publishers, 1998.

[16] SMITH, C. W. – SMITHSON, C. W. – WILFORD, D. S., *Managing Financial Risk*, Ballinger Publishing Company, 1990.

[17] SOCZO, CS., Problems With Current Methods of Filling in Missing Financial Data, *Information, Knowledge, Competitivity*, Budapest College of Management, 2002.