International Journal of Business, Human and Social Sciences
ISSN: 2517-9411
Vol:3, No:1, 2009

Numerical Optimization within Vector of Parameters Estimation in Volatility Models

J. Arneric, and A. Rozga

***Abstract*—**In this paper usefulness of quasi-Newton iteration procedure in parameters estimation of the conditional variance equation within BHHH algorithm is presented. Analytical solution of maximization of the likelihood function using first and second derivatives is too complex when the variance is time-varying. The advantage of BHHH algorithm in comparison to the other optimization algorithms is that requires no third derivatives with assured convergence. To simplify optimization procedure BHHH algorithm uses the approximation of the matrix of second derivatives according to information identity. However, parameters estimation in a/symmetric GARCH(1,1) model assuming normal distribution of returns is not that simple, i.e. it is difficult to solve it analytically. Maximum of the likelihood function can be founded by iteration procedure until no further increase can be found. Because the solutions of the numerical optimization are very sensitive to the initial values, GARCH(1,1) model starting parameters are defined. The number of iterations can be reduced using starting values close to the global maximum. Optimization procedure will be illustrated in framework of modeling volatility on daily basis of the most liquid stocks on Croatian capital market: Podravka stocks (food industry), Petrokemija stocks (fertilizer industry) and Ericsson Nikola Tesla stocks (information's-communications industry).

***Keywords*—**Heteroscedasticity, Log-likelihood Maximization, Quasi-Newton iteration procedure, Volatility.

I. Introduction

MAXIMUM likelihood estimation (MLE) is usually concerned in parameters evaluation in models with nonstacionary variance (heteroscedasticity). Maximum likelihood estimation chooses coefficient estimates that maximize the likelihood of the sample data being observed.

Likelihood function, for linear regression model, is defined as joint density function for observed output variables y_1, y_2, ..., y_n. According to the assumption that observations are normally distributed, likelihood function is defined as:

$$L(\beta)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}\left(\frac{\varepsilon_i}{\sigma}\right)^2} \quad (1)$$

From above definition the joint density function is given as a product of all normally distributed variables y_i. For practical reasons function (1) is transformed into monotone increasing function, by taking it's natural logarithm:

$$\ln L(\beta)=-\frac{n}{2}\ln\left(2\pi\sigma^2\right)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\varepsilon_i^2 \quad (2)$$

Speaking statistically, it is easy to take expectations and variance of the sums, rather then products. Function defined in (2) is called log-likelihood function [6]. By taking partial derivatives of log-likelihood function with respect to parameters β_0, β_1, ..., β_k and σ^2, and setting them equal to zero, results in the same estimation of vector β as in OLS case (Ordinary Least Squares).

Estimators given by maximization of log-likelihood function are equivalent to OLS estimators if and only if *i.i.d.* assumption is introduced (independently and identically distributed variables). Speaking statistically, assumption that variables y_i have normal distribution with constant variance is equivalent to the assumption that variables ε_i have standard normal distribution with zero mean and variance equal to unity, i.e. $\varepsilon_i \sim N(0,1)$.

However, in modeling financial time series with high frequencies, the assumption of constant variance is unrealistic. Therefore, it is assumed that variance is time-varying (heteroscedasticity). It is well-known that returns from financial instruments such as exchange rates, equity prices and interest rates measured over short time intervals, i.e. daily or weekly, are characterized by volatility clustering and ARCH effects. Models which are used to account daily volatility (standard deviation of returns) are GARCH(p,q) models [1]. Autocorrelation of the squared returns suggests high dependency between them, i.e. ARCH effect exists. This means that volatility is conditioned on its past information's.

Assuming that σ_t is time-varying, log-likelihood function can be expressed as:

$$\ln L(\beta)=-\frac{T}{2}\ln(2\cdot\pi)-\frac{1}{2}\sum_{t=1}^{T}\ln\sigma_t^2-\frac{1}{2}\sum_{t=1}^{T}\left(\frac{\varepsilon_t^2}{\sigma_t^2}\right) \quad (3)$$

By taking first derivatives of function (3), and after some rearrangement:

J. Arneric, and A. Rozga are with Department of Quantitative Methods, Faculty of Economics, University of Split, Croatia (email: jarneric@efst.hr, rozga@efst.hr).

$$\frac{\partial \ln L(\beta)}{\partial \beta} = \frac{1}{2}\sum_{t=1}^{T}\left[\frac{\varepsilon_t^2}{\sigma_t^2} - 1\right]\frac{1}{\sigma_t^2}\cdot\frac{\partial \sigma_t^2}{\partial \beta}, \tag{4}$$

and setting the system of equations (4) equal to zero, becomes to complex to solve, i.e. it is difficult to solve it analytically. Therefore numerical approach is needed.

II. Numerical Optimization Procedure

A numerical approach is needed when variance σ_t^2 is described through conditional variance equation according to GARCH(1,1) model:

$$\begin{aligned} r_t &= \varepsilon_t; \quad \varepsilon_t = u_t\sqrt{\sigma_t^2} \\ \sigma_t^2 &= \alpha_0 + \alpha_1\cdot\varepsilon_{t-1}^2 + \beta_1\cdot\sigma_{t-1}^2 \end{aligned} \tag{5}$$

In relations (5) Engle sets the multiplicative structure of innovation process $\varepsilon_t = u_t\sqrt{\sigma_t^2}$ assuming $u_t \sim i.i.d.(0,1)$.

Numerically, the maximum can be found by "walking up" the likelihood function until no further increase can be found. The interest of this paper is to estimate parameters of the GARCH(1,1) model by maximization of the log-likelihood function using BHHH algorithm, and to define starting values very close to global maximum. Each iteration moves to a new value of the parameters at which $\ln L(\beta)$ is higher than at the previous step.

To determine the best value of β_{i+1}, a second-order Taylor's approximation of $\ln L(\beta_{i+1})$ around $\ln L(\beta_i)$ is used:

$$\begin{aligned} \ln L(\beta_{i+1}) = \ln L(\beta_i) + (\beta_{i+1} - \beta_i)^T g_i + \\ + \frac{1}{2}(\beta_{i+1} - \beta_i)^T H_i (\beta_{i+1} - \beta_i) \end{aligned} \tag{6}$$

Now we find the value of β_{i+1} that maximizes approximation in (6):

$$\begin{aligned} &\frac{\partial \ln L(\beta_{i+1})}{\partial \beta_{i+1}} = g_i + H_i(\beta_{i+1} - \beta_i) = 0 \\ &H_i(\beta_{i+1} - \beta_i) = -g_i \\ &\beta_{i+1} - \beta_i = H_i^{-1} g_i \\ &\beta_{i+1} = \beta_i + \left(-H_i^{-1}\right) g_i \end{aligned} \tag{7}$$

The Newton procedure uses this formula [13]. The step from the current value of β_i to the new value is $\left(-H_i^{-1}\right)g_i$ the gradient vector multiplied by the negative of the inverse of the Hessian. The scalar λ is introduced in the Newton iterative formula to assure that each step of the procedure provides an increase in $\ln L(\beta)$. The adjustment is performed separately in each iteration:

$$\beta_{i+1} = \beta_i + \lambda_i\left(-H_i^{-1}\right)g_i \tag{8}$$

The vector $\left(-H_i^{-1}\right)g_i$ is called direction, denoted as d_i, and λ is called the step size. Classical modified Newton iterative procedure specified in (8) is often referred as Newton-Raphson algorithm when Hessian is determined analytically [3]. Even so, calculation of the Hessian is usually computation-intensive, i.e. analytical Hessian is rarely available. Therefore, alternative to calculation of inverse Hessian matrix is it's approximation.

Suppose that the log likelihood function has regions that are not concave. In these areas, the classical modified Newton procedure can fail to find an increase. If the function is convex at β_i, then the Newton procedure moves in the opposite direction to the slope of the log-likelihood function and $-H_i^{-1}$ is positive definite.

Therefore Newton-Rapshon algorithm has two main disadvantages:

- calculation of the Hessian is computation-intensive
- the procedure does not guarantee an increase in each step if the log-likelihood function is not globally concave.

III. Log-likelihood Maximization within BHHH Algorithm

Berdnt, Hall, Hall and Hausman (1974) proposed using information identity in the numerical search for the maximum of the log-likelihood function [4]. In particular, iterative procedure is defined as:

$$\begin{aligned} \beta_{i+1} &= \beta_i + \lambda_i\cdot d_i \\ d_i &= -H_i^{-1} g_i \\ -H_i &= \sum_{t=1}^{T} g_t g_t^T \\ g_i &= \sum_{t=1}^{T} g_t \end{aligned} \tag{9}$$

According to the relations in (9) information identity means that the asymptotic variance-covariance matrix of a maximum likelihood estimator is equal to the variance-covariance matrix of the gradient of the likelihood function [4]. By the central limit theorem, the asymptotic distribution of $\hat{\beta}$ is multivariate normal with mean vector β_p and variance matrix equal to inverse of negative expected Hessian:

$$Var(\hat{\beta}) = \left[-E\left(H(\beta_p)\right)\right]^{-1} \tag{10}$$

In other words, the unbiased estimation of variance-covariance matrix can be approximated as the inverse of the outer product of gradients (OPG) as follows:

$$Var\left(\hat{\beta}\right)=\frac{T}{T-1}\left(\sum_{t=1}^{T}g_t\left(\hat{\beta}\right)g_t\left(\hat{\beta}\right)^T\right)^{-1} \quad (11)$$

After presentation of some properties of OPG estimators, numerical optimization procedure of BHHH algorithm could be summarized in following steps:

1. determine initial vector of parameters β_{init}, and convergence criteria $tol=0.0001$,
2. at current iteration calculate a direction vector $d_i=\left[-H\left(\beta_i\right)\right]^{-1}$, while $-H\left(\beta_i\right)$ is calculated by the outer of the gradients,
3. calculate a new vector $\beta_{i+1}=\beta_i+\lambda d_i$, where λ is scalar. Start with $\lambda=1$. If $f\left(\beta_i+d_i\right)>f\left(\beta_i\right)$ try with $\lambda=2$. If $f\left(\beta_i+2d_i\right)>f\left(\beta_i+d_i\right)$ try with $\lambda=4$, etc. until lambda is found for which $f\left(\beta_i+\lambda d_i\right)$ is in maximum,
4. if convergence criteria is satisfied algorithm stops, if not repeat steps from 2 to 4.

IV. Vector Of Parameters Estimation In Volatility Models

Since the introduction by Engle [7] of the ARCH(p) model (Autoregressive Conditional Heteroscedasticity) and it's generalization, i.e. GARCH(1,1) model by Bollerslev [5] a wide range of extensions and modifications have been developed.

It has been shown that ARCH(p) process with infinite number of parameters is equivalent to generalized ARCH process, i.e. GARCH(p,q) process which is very well approximated by simple GARCH(1,1). As the time lag increases in an ARCH(p) model it becomes more difficult to estimate parameters. Besides it is recommended to use parsimonious model as GARCH(1,1) that is much easier to identify and estimate (model has just one lagged square error and one autoregressive term).

When there is asymmetric volatility clustering Glosten, Jagannathan and Runkle proposed asymmetric GARCH(1,1) model, in which leverage effect is measured with parameter associated to dummy variable:

$$\sigma_t^2=\alpha_0+\left(\alpha_1+\gamma_1 d_{t-1}\right)\varepsilon_{t-1}^2+\beta_1\sigma_{t-1}^2$$
$$d_{t-1}=\begin{cases}1 & if \quad \varepsilon_{t-1}<0\\ 0 & if \quad \varepsilon_{t-1}\geq 0\end{cases} \quad (12)$$

From expression (12) it can be seen that good news from previous day, when $d_{t-1}=0$, influences conditional variance by parameter α_1, and a bad news, when $d_{t-1}=1$, effects conditional variance by sum of parameters $\alpha_1+\gamma_1$.If is founded that parameter $\gamma_1>0$ and statistically significant, then negative shocks have larger effects on volatility than positive shocks, assuming that other estimated parameters are nonnegative (simple test to investigate the leverage effect is to test the significance of first-order autocorrelation coefficient between lagged returns and current squared returns. If there is asymmetric information influence it is expected for this coefficient to be negative).

In Table I parameters estimation of a/symmetric GARCH(1,1) model, using BHHH algorithm, are presented.

TABLE I
Parameters Estimation Using BHHH Algorithm

Parameter estimation of the GARCH(1,1) model [a]	Stocks		
	Podravka	Petrokemija	Ericsson NT
α_0	0.0000376	0.0002552	0.0000476
α_1	0.3009699	0.4147483	0.2801536
γ_1	0.1532592	-	-
β_1	0.419941	0.2214572	0.221457
Number of iterations	6	10	12

[a] Parameter values are estimated using initial vector of parameters as: $\beta_{init}=\left[0.00001 \quad 0.2 \quad 0.1 \quad 0.8\right]^{\tau}$

From Table I it can be seen that asymmetric information influence is present in Podravka stocks, while parameter γ_1 is omitted in other two stocks (there was no leverage effect in Petrokemija and Ericsson NT stocks). Number of iterations is much reduced assuming initial vector of parameters very close to global maximum: $\beta_{init}=\left[0.00001 \quad 0.2 \quad 0.1 \quad 0.8\right]^{\tau}$. The reason we introduced these starting values lies in expectation that volatility reacts at low intensity on past market movements (α_1), and that conditional volatility decays slowly, i.e. long time is needed for shocks to die out ($\alpha_1+\beta_1$). In initial vector it is also incorporated the assumption that the bad news effects the volatility by 50% more than the good news.

On example of Podravka stocks, if we have not assumed these starting values, by default software would use $\beta_{init}=\left[0.0002473 \quad 0 \quad 0 \quad 0\right]^{\tau}$, which would result in portion of not concave likelihood function at null iteration. STATA 9.1 package was used in parameter estimation with initial values $\alpha_1=\gamma_1=\beta_1=0$, while α_0 is calculated as variance of returns from observed sample [8]. Hence, eleven iterations would be needed in comparison to 6 iterations.

In Fig. 1 convergences of estimated parameters of GARCH(1,1) model on example of Petrokemija stock from Zagreb Stock Exchange, as well as convergence of log-likelihood are presented.

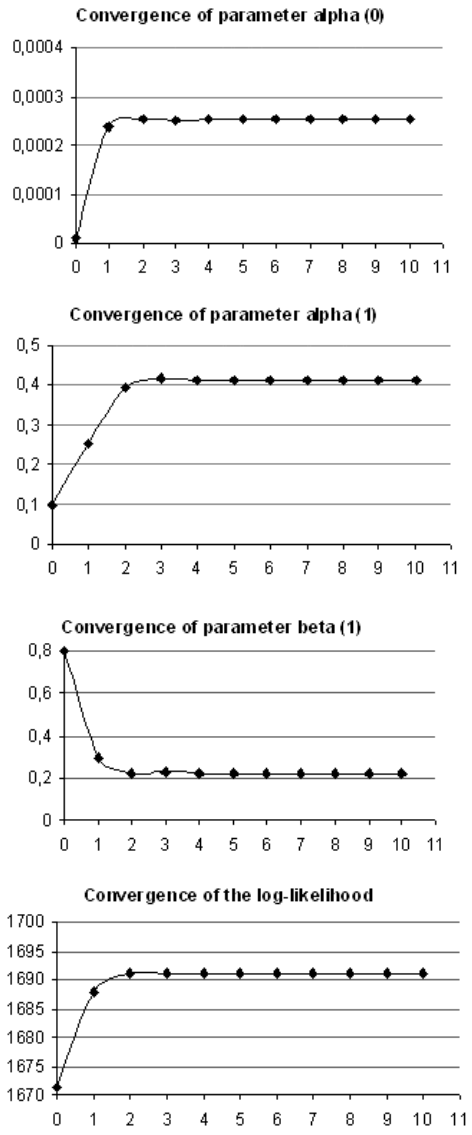

Fig. 1 Convergences of estimated parameters and log-likelihood of GARCH(1,1) model of Petrokemija stock (10 iterations)

Small changes in parameter values, with small increases in log-likelihood function, from one iteration to the next iteration could be evidence that convergence has been achieved. Even so, small changes in β_i and $\ln L(\beta_i)$ accompanied by a gradient vector that is not close to zero indicate that we are not effective in finding the maximum. At each iteration the step size is reduced (stepping backward) or increased (stepping forward) in purpose to calculate new vector for which the maximum of log-likelihood function is increased the most [8]. Step size is reduced when the initial step is bad, and it is increased when the initial step is good.

Table II shows how step size is increased (doubled) in each iteration as long as $\ln L(\beta_i)$ continues to rise.

TABLE II
DOUBLING STEP SIZE IN FIRST 4 ITERATIONS OF LIKELIHOOD MAXIMIZATION

```
Iteration  0  Step Size = 1.00000  Likelihood =  2.89883
Iteration  0  Step Size = 2.00000  Likelihood = -1.00000e+010
Iteration  1  Step Size = 1.00000  Likelihood =  2.90109
Iteration  1  Step Size = 2.00000  Likelihood =  2.90247
Iteration  1  Step Size = 4.00000  Likelihood =  2.90387
Iteration  1  Step Size = 8.00000  Likelihood =  2.90477
Iteration  1  Step Size = 16.0000  Likelihood =  2.90513
Iteration  1  Step Size = 32.0000  Likelihood =  2.90065
Iteration  2  Step Size = 1.00000  Likelihood =  2.90676
Iteration  2  Step Size = 2.00000  Likelihood =  2.90778
Iteration  2  Step Size = 4.00000  Likelihood =  2.90822
Iteration  2  Step Size = 8.00000  Likelihood =  2.90320
Iteration  3  Step Size = 1.00000  Likelihood =  2.90855
Iteration  3  Step Size = 2.00000  Likelihood =  2.90871
Iteration  3  Step Size = 4.00000  Likelihood =  2.90842
Iteration  4  Step Size = 1.00000  Likelihood =  2.90884
Iteration  4  Step Size = 2.00000  Likelihood =  2.90894
Iteration  4  Step Size = 4.00000  Likelihood =  2.90906
Iteration  4  Step Size = 8.00000  Likelihood =  2.90901
⋮
Convergence is less than tolerance. Convergence reached.
```

The advantage of this approach of doubling step size is that it usually reduces the number of iterations (procedure presented in Table II is obtained using S+FinMetrics module of S-PLUS package, where likelihood is normalized in each step). Procedure stops at last iteration when a convergence criterion is satisfied. In theory the maximum of log-likelihood occurs when the gradient vector is zero. Namely, in practice the calculated gradient is never exactly zero, but can be very close. Therefore, $g_i^T(-H_i)^{-1}g_i$ is often used to evaluate convergence:

$$g_i^T(-H_i)^{-1}g_i < 0.0001 \quad (13)$$

If (13) is satisfied, the iterative process stops and the parameters at current iteration are considered as estimates.

Estimated conditional volatilities of Podravka, Petrokemija and Ericsson stocks are presented in Fig. 2 and 3, from 13 December 2004 to 8 November 2007 (728 trading days).

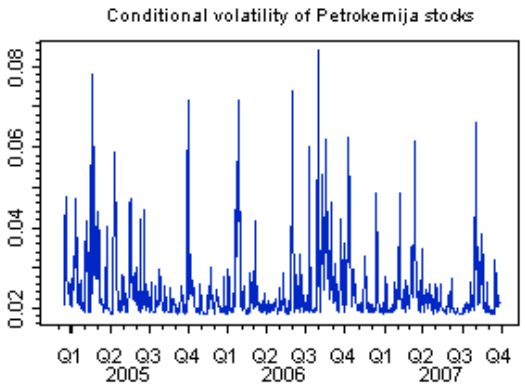

Fig. 2 Conditional volatilitiy of Petrokemija stocks

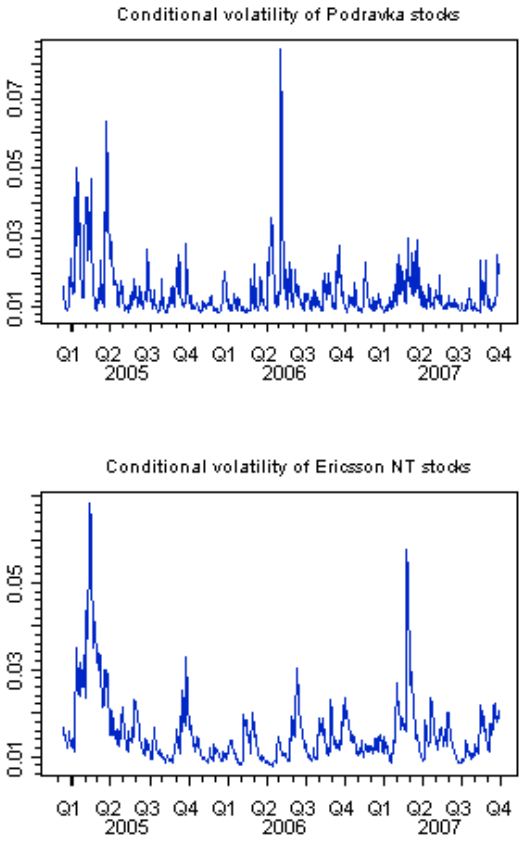

Fig. 3 Conditional volatilities of Podravka and Ericsson NT stocks

All estimated parameters in Table I are significant at empirical p-value less then 5%. Also sum of parameters $\alpha_1 + \beta_1$ in modeling Ericsson volatility indicate that there is high volatility persistence, i.e. conditional variance decays slowly [12]. It means that GARCH(1,1) model is nonstacionary, i.e. it follows integrated GARCH model (IGARCH(1,1) model belongs to familiy of long memory models when long time is needed for shocks in volatility to die out). Even so, parameter α_1 detects low intensity reaction of volatility on past information's.

Because the stationary conditions of estimated model of Podravka and Petrokemija stocks are satisfied the unconditional long-term variance can be calculated. Unconditional long-term standard deviation of Podravka returns is 1.36%, and long-term standard deviation of Petrokemija returns is 2,65%. Also, on example of Podravka stocks, it is evident that bad in formations effects volatility 50.92% more than the good information's.

V. Conclusion

To investigate if local maximum is the global optimum we should use different starting values and observe whether convergence occurs at the same parameter values. Empirical research has showed that initial vector of parameters as null-vector is not appropriate.

Namely, BHHH algorithm has approved to be faster when good initial parameters are used (close to global maximum). To simplify optimization procedure BHHH algorithm uses the approximation of the matrix of second derivatives according to information identity: "at the true value of parameter vector β the expected value of the outer product of the first derivatives is equal to minus the expected value of the second derivatives".

Convergence is assured because the approximation of the inverse of the Hessian matrix is guaranteed to be a positive definite. Even so, convergence problem may arise, because the more parameters in the model are entered the "flatter" the log-likelihood function becomes, and therefore the more difficult it is to maximize.

References

[1] C. Alexander, *Market Models: A Guide to Financial Data Analysis*, John Wiley and & Sons Ltd., New York, 2001.
[2] J. Arnerić, B. Škrabić, and Z. Babić, "Maximization of the likelihood function in financial time series models", in *Proceedings of the International Scientific Conference on Contemporary Challenges of Economic Theory and Practice,* Belgrade, 2007, pp. 1-12.
[3] M. S. Bazarra, H. D. Sherali, and C. M. Shetty, *Nonlinear Programming - Theory and Algorithms (second edition)*, John Wiley and & Sons Ltd., New York, 1993.
[4] E. Berndt, B. Hall, R. Hall, and J. Hausman, "Estimation and Inference in Nonlinear Structural Models", *Annals of Social Measurement*, Vol. 3, 1974, pp. 653-665.
[5] T. Bollerslev, "Generalized Autoregressive Conditional Heteroscedasticity", *Journal of Econometrics*, Vol. 31, 1986, pp. 307-327.
[6] W. Enders, *Applied Econometric Time Series (second edition)*, John Wiley and & Sons Ltd., New York, 2004.
[7] R. Engle, "The Use of ARCH/GARCH Models in Applied Econometrics", *Journal of Economic Perspectives*, Vol. 15, No. 4, 2001, pp. 157-168.
[8] W. Gould, J. Pitblado, and W. Sribney, *Maximum Likelihood Estimation with Stata (third edition)*, College Station, StatCorp, 2006.
[9] C. Gourieroux, and J. Jasiak, *Financial Econometrics: Problems, Models and Methods*, Princeton University Press, 2001.
[10] L. Neralić, *Uvod u Matematičko programiranje 1*, Element, Zagreb, 2003.
[11] J. Petrić, and S. Zlobec, *Nelinearno programiranje*, Naučna knjiga, Beograd, 1983.
[12] P. Posedel, "Properties and Estimation of GARCH(1,1) Model", *Metodološki zvezki*, Vol. 2, No. 2, 2005, pp. 243-257.
[13] R. Schoenberg, "Optimization with the Quasi-Newton Method", *Aptech Systems working paper*, Walley WA, 2001, pp. 1-9.
[14] D. F. Shanno, "Conditioning of quasi Newton methods for function minimization", *Mathematics of Computation*, No. 24, 1970, pp. 145-160.