Avd. Matematisk statistik

KTH Teknikvetenskap

"General"Homework in the course SF2945 Time series analysis.

Choose your own time series

Part 1.

To be finished at latest 19 November 2008.

Choose a time series according to you own interests. If you have to write the data by hand into a file it is enough to use around 100 values. You may get the data from any source except from the book Brockwell and Davis *Introduction to Time Series and Forecasting*. Further it must be real data and *not* randomly generated data. You might for instance find data on the web.

Your "report" on Part 1 must contain a description of the data (including where you found it) and a plot of the raw data set.

Part 2.

To be finished at latest 10 December 2008.

Your problem is to, during the course, do three different analyses of the data. You may very well do these analyses at the same time when you do the "ordinary" homeworks. The idea behind this homework is to realize all problems which may – and will – occur when you are working with real data not chosen mainly to fit the methods. **This is the reason why you must choose your time series at an early stage of the course.**

It is recommended that you do all homeworks in groups of three.

Avd. Matematisk statistik

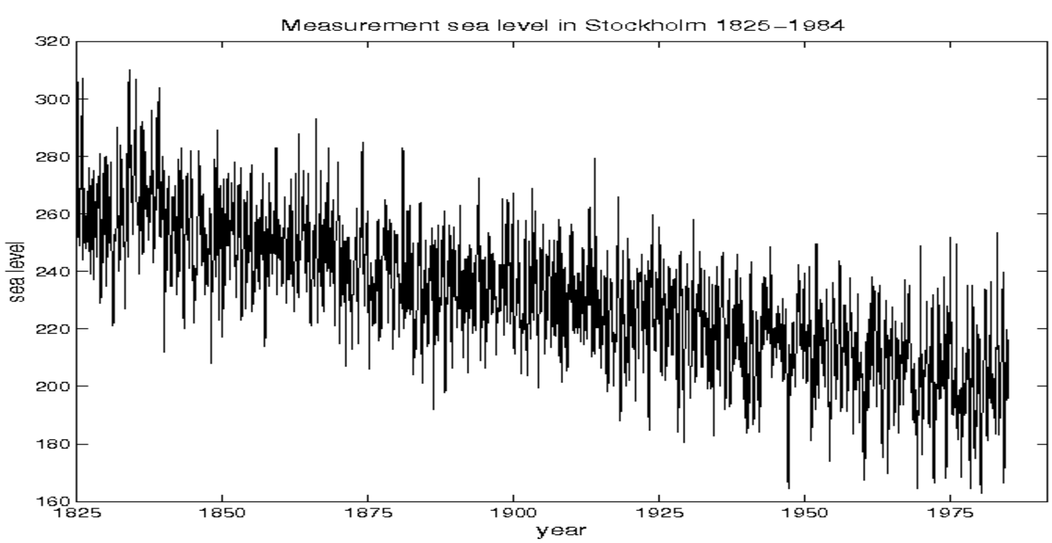
KTH Teknikvetenskap

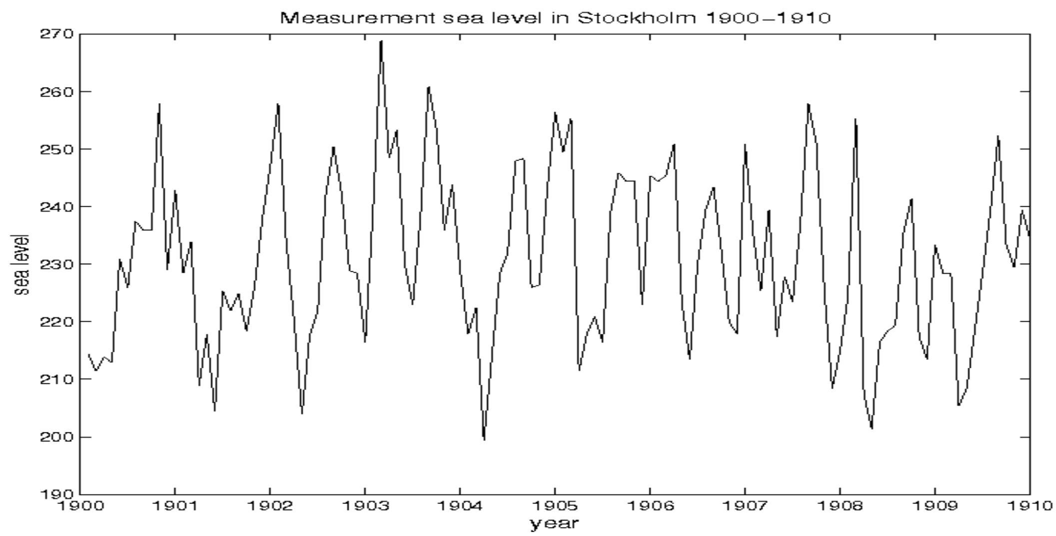
Homework 1.a. in the course SF2945 Time series analysis

To be finished and uploaded at bilda.kth.se at latest 19 November 2008.

This homework consists of making a *classical decomposition* of data of sea level in Stockholm, 1825 to 1984.

The sea level in Stockholm has been measured since 1774, the longest sea level measurement series in the world. From 1825 there exists a complete series of data from every month. This series will be used here. At the beginning the measurement was done at Slussen, now the instrument is placed on Skeppsholmen. The first figure gives the total series from 1825 to 1984, the second one is a ten years series.





The trend in the series is an effect of the land elevation.

The home work can be done in e.g. MATLAB.

MATLAB has an extensive toolbox, Identication Toolbox, which can be used for time series, but we will not do that. Instead, necessary m-files, not existing in MATLAB, can be found in www.math.kth/matstat/gru/SF2945/. In this homework, the m-files *acf.m, acvf.m, diffd., smoothma.m, smoothpf.m, seascomp.m, ljungbox.m and ranktest.m* can be used. The data can be found in the MATLAB data-file *sealevel.mat* or in the text-files *sealevel.dat* and *sldate.dat*. Save them in a appropriate library, where MATLAB can find them.

The text-file *sldate.dat* also contains the months and years of the data, as first row and column. *sealevel.dat* only contains the data and is easier to read into MATLAB. You can use the command *fscanf* to read the textfile data: However it is easier to use the .mat-file directly.

fid=fopen('sealevel.dat','r');
sl=fscanf(fid,'%f');

will read the data into the column vector *sl*. You may need to specify the path to the *sealevel.dat* file in the first command.

For classical decomposition, see Brockwell, section 1.5.

1. The total data vector *sl* consists of 1920 data. Select 600 of them, representing 50 consecutive years. Begin at a random time. Store the selected data in appropriate vector. Use then the m-file $acf.m$ to compute and plot the autocorrelation function of the selected data. The command *y=acf(x)* will compute the autocorrelation function of the time series x. Remember that the index n represents the correlation $\rho(n+1)$ as the index of vectors in MATLAB begin with 1 and not 0.

The command *y=acf(x,plott)* where *plott* is an arbitrary number also draws a plot, with lines $\pm 1.96/\sqrt{n}$, see Brockwell, example 1.4.6 page 20 for explanation. Comment your plot. What conclusions can you draw. Does the plot say something about a period?

2. For estimating trend and season factor, use first Method S1, described on page 31 in Brockwell. What do you think is an appropriate period p?

3. Estimate the seasonal component s_t and the deseasonalized data vector d_t.

The MATLAB-command *[d,s]=seascomp(x,p)* does that, where x is the original data vector and p the period. Plot s_t and d_t.

4. Test the deseasonalized data d for remaining trend, by using the *Rank Test* described in Brockwell page 37. Use the m-file *ranktest*, which compute, in Brockwells terminology,

$$\frac{|P - \mu_P|}{\sigma_P}$$

What is the result of the test? Shall the hypothesis "no trend" be rejected?

5. Estimate the trend in the deseasonalized data vector d by using the m-file smoothpf.

[c,m,z]=smoothpf(x,grad) will estimate the trend in data vector x by an appropriate polynom of degree *grad*. The coefficients of the polynom in *coeff* will appear in *c*. on. m will be the trend vector and z estimated residual $x-m$. Try a linear and a quadratic trend. Plot the residuals. Does a quadratic trend seem to be a better model than a linear one?

6. Test the the residual series z in 5 for independence by using Ljung-Box test, page 36 in Brockwell. Use h=40. The test statistics can be computed by the m-file *ljungbox*. What is the result? Use 1% significance level.

7. Make plots of the original data, the deseasonalized data d_t and the estimated seasonal component s_t.

Avd. Matematisk statistik

KTH Teknikvetenskap

Homework 1.b. in the course SF2945 Time series analysis

To be finished and uploaded at bilda.kth.se at latest 19 November 2008.

This homework consists of simulation and analysis of some ARMA-processes. You can use the following m-files.

simarma simulates an ARMA(p, q) process. The m-file first simulates normal distributed observations, and from them an ARMA-process is simulated. The syntax is // *y=simarma(fi, theta,n, s2, seed)*
where *fi* is the vector (in Brockwell notation, see e.g. page 83) *fi*=$[\phi_1\phi_2 \ldots \phi_p]$ and *theta*=$[\theta_1\ \theta_2 \ldots \theta_q]$. n is the number of observations to be generated and *s2* the white noise variance σ^2. *seed* is a seed for the random generation and must be set for every simulation. This makes it possible to regenerate the same series. In all simulations you have to report the seed. y will be a vector with all simulated observations.

To generate an AR-process you just have to put *theta*=0 and if *fi*=0 a MA-process is simulated. We want to simulate causal processes, which implies that a stationary solution exists.

causal tells you if the defined process is causal. The syntax is *causal(fi)*, where *fi* as above.

predarma gives prediction of an ARMA-process. The syntax is
[y, se2]=predarma(x,fi,theta,sigma2,h),
where x is the vector of known observations, *fi* and *theta* as above. *sigma2* is the white noise variance and h is the forward time step, the time ahead that prediction is made. y is the predicted value and *se2* the mean square error of the prediction. See chapter 3.3. In Brockwell notation, $x = [X_1, X_2, \ldots, X_n]$ and $y = P_n X_{n+h}$.

innov is used in *predarma*.

armaacvf gives the autocovariance function of an ARMA process. Do not mix this up with the m-file *acvf*, which gives the sample autocovariance function, based on a time series. The syntax is *y=armaacvf(fi,theta,N,sigma2)*. *fi* and *theta* as above. *sigma2* is the white noise variance σ^2 and N is maximal lag, that is the covariance is computed up to $h = N$. The values of the autocovariance function is found in the vector y. For autocovariance function of an ARMA process see Brockwell section 3.2. The autocorrelation function is, as you know, $\rho(h) = \dfrac{\gamma(h)}{\gamma(0)}$, so dividing y above with its first element, gives the autocorrelation function.

psi is used in *armaacvf*. *psi* computes the ψ-parameters in the linear representation of the process, see Brockwell page 51.

roots2ar computes the AR parameters from the roots in the generating polynomial. *arroots* is the inverse, it gives the roots of the generating function.

AR(2)-process

The defining equation is

$$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} = Z_t$$

Read example 3.2.4 page 91 in Brockwell. A case is missing in that example, namely $\xi_1 = \xi_2,\ |\xi| = |\xi_1| = |\xi_2| > 1$. In that case the process is causal and the covariance function is

$$\sigma^2 \xi^{|h|} \left(\frac{|h|}{\xi^2 - 1} + \frac{\xi^2(3\xi^2 - 1)}{(\xi^2 - 1)^2} \right)$$

1. Choose two complex conjugate roots of the generating function, so that the resulting process is causal. Choose the roots so that their modulus (absolute values) are near 1. Compute ϕ_1 and ϕ_2 according to example 3.2.4 and simulate 100 observations of the process and draw a plot. The simplest way to compute the coefficients from the roots is to use the Matlab function *roots2ar*. You can calculate with complex numbers in Matlab as usual, just write $a + bi$ where a and b are real numbers. You can also use polar coordinates, $r * exp(-\phi i)$ where r and ϕ are real numbers.

Caution! If you chose two conjugate complex roots and compute ϕ_1 and ϕ_2 according to example 3.2.4 they can be complex with small imaginary part, due to numerical imprecision. In that case take the real part of it.

Plot the autocorrelation function of the process, and also compute the sample autocorrelation function based on the simulations. Plot this sample correlation function and compare the plots.

2. Choose two different real roots of the characteristic function, and make plots of simulations, and autocorrelation functions as in 1.

3. Compare the autocorrelation plots in 1 and 2. What is the main qualitative difference?

4. Consider the case 1, two conjugate complex roots. The autocovariance function is given in (3.2.12) on page 91. The expression for the autocorrelation function is quite simple. Motivate that it is

$$r^{-|h|} \frac{\sin(|h|\theta + \psi)}{\sin(\psi)}$$

where ψ as in (3.2.13) and the notations as in the example.

If one wants to model a process with strong positive dependence, that is with autocorrelation function near 1 for a range of h:s, how should r and θ in (3.2.12) be chosen? Choose a θ and try with two or three different r, so that you have processes with various strong positive dependence. Plot simulations of the processes (100 values) and plot the autocorrelation functions and the sample autocorrelations functions.

5. Simulate 10 observations $x_1, x_2, \ldots, x_{10}$ of a ARMA(2,2) process with parameters $\phi_1 = 0.4,\ \phi_2 = 0.5,\ \theta_1 = 1,\ \theta_2 = 0.8$. Check that the process is causal. Make prediction up to time 20 from these 10 observations. Plot the prediction series and the root mean square errors of the predictions, $\sqrt{E\big((X_{10+h} - P_{10}X_{10+h})^2\big)}, h = 1, 2, \ldots, 10$.

6. The process in 5 is a normal process. What is the expected value of $X_{11} - \widehat{X}_{11}$? Show that the mean square error $E\big((X_{11} - \widehat{X}_{11})^2\big)$ is equal to the variance $V(X_{11} - \widehat{X}_{11})$. See page 65 in Brockwell. Note that the mean square error is computed in 6. Show that

$$P(\widehat{X}_{11} - \lambda_{\alpha/2} r_{11} \leq X_{11} \leq \widehat{X}_{11} + \lambda_{\alpha/2} r_{11}) = 1 - \alpha$$

where r_{11} is the root mean square error $\sqrt{E(X_{11} - \widehat{X}_{11})^2}$ of $\widehat{X}_{11}$.
The interval

$$(\widehat{X}_{11} - \lambda_{\alpha/2} r_{11}, \widehat{X}_{11} + \lambda_{\alpha/2} r_{11})$$

is a *prediction interval* for X_{11} with confidence level $1 - \alpha$. Compute this interval.

Avd. Matematisk statistik

KTH Teknikvetenskap

Homework 2.a. in the course SF2945 Time series analysis.

To be finished and uploaded at bilda.kth.se at latest 10 December 2008.

This homework consists of estimation parameters of AR- and ARMA-processes.

The data you shall use, are in the files *temp.dat*, *timech.dat*, and *el.dat*

The data in *el.dat* is the monthly production of Australia's production of electricity in GWh Jan 1956 - August 1995. The data in *temp.dat* are the monthly mean temperature in New York City, Jan 46 - Dec 59. The data in *timech.dat* are the annual change in earth's rotation 1821-1970 (10^{-5} s). The new m-files are *burg*, *pacf*, *yuwaest*, *pergram*, *specdens*, *specarma*, *boxcox* and *boxcoxf*.

1. Consider first monthly temperatures in the variable *temp*. Plot the temperatures. Plot the periodogram (see Brockwell p 121). Make one plot with no smoothing (no weights), and one with the weights [3 3 2 1] ([W(0) W(1) W(2) W(3)]=[3 3 2 1]/15), see Brockwell p 125). Use the m-file *pergram*. *pergram(x,w,plott)* gives the periodogram. If two arguments, the the plot is drawn with no weight function if w is a scalar (number). If w is a vector, this is the weights $w(0), w(1), \ldots$. The weight is symmetric and automatically normalized (you need not to enter weights with sum 1).

You shall produce the two plots and explain why the period seems to be 12 in the plots.

2. Compute the autocovariance function of the temp series. Use the m-file acvf. Then compute the partial autocovariance function. *pacf(g,n,plott,cl)* computes the PACF (see Brockwell p 94). *pacf(g)* computes the PACF up to lag equal to the number of observations, g is the ACVF. *pacf(g,n)* computes PACF up to lag n. *pacf(g,n,plott)* where plott is an arbitrary number also draws a plot. If you put the second argument n = [] then the PACF is

computed and plotted up to lag equal to number of observations. At last, *pacf(g,n,plott,cl)*, where *cl* is an arbitrary number, also draws 95% probability bounds (see Brockwell e.g. p 97). The PACF plot suggests an AR-model. Fit an appropriate order p (see Brockwell example 3.2.6 and 3.2.9 on p 95 and 99).

You shall produce the PACF plot with probability lines, give the order of the AR process you think is appropriate and motivate your choice of the order.

3. Use *yuwaest* to estimate the parameters in the model in 2. The syntax is

[fi,s2,C] = yuwaest(y,p) where

y is the time series,

fi is the autoregressive model parameter ,

$s2$ is the estimated WN variance and

C is the estimated covariance matrix of the estimated parameter vector fi. The diagonal in C thus contains the estimated variances of the estimated parameters.

Give the estimates and approximate 95% confidence limits of the parameter ϕ_1.

You shall give the estimations of the ϕ parameters and the estimated WN variance σ^2. Give also the confidence interval above.

4. Compute and plot the spectral density for the AR(p) process fitted in 3. Use *specarma*. Type *help specarma for the syntax.* Find the dominating period from the maximum of the spectral density.

You shall produce the plots and again explain the period.

5. Plot the data in timech. There seems to be a linear trend, which suggests that an ARIMA model could be appropriate (Brockwell chapter 6). We shall study the difference series $Y_t = (1 - B)X_t$. Use the m-file *diffd* to construct the series y_t. Compute the autocovariance function of y_t and after that the PACF. Model this series as coming from a suitable AR-process. Estimate the parameters, and construct 95% confidence limits for one of the parameters.

You shall produce a plot of the PACF, explain your choice of model, and give the estimated parameters together with one confidence interval. You shall also express the ARIMA model of the original time series $\{X_t\}$ that is fitted (see Brockwell section 6.1), that is in the form

$$X_t = a_1 X_{t-1+} a_2 X_{t-2} + \cdots + b_1 Z_t + \cdots + b_q Z_{t-q+1}$$

The parameters $a_1, a_2, \ldots$ shall be written as numbers and observe that X_t only appears on the left side.

6. The data file el.mat (or el.dat) contains Australia monthly production of electricity in GWh Jan 1956 - August 1995. Plot the time series. You see that the production fluctuates more and more and thus seems to be non-stationary. One can use a transformation of the data to make the series

more compatible with a stationary series, se Brockwell page 188. A Box-Cox transformation is given by the logarithm of the data or by a power of the data. In most common cases the variance or the standard deviation of the data seems to be proportional to the mean. To stabilize the fluctuations one in these cases use the square root or the logarithm of the data as the transformation, in Brockwells notation, λ is 0 or 1/2.

The matlab file *boxcox* is a script that gives a plot in which you can enter the data and the λ value. You can, by using the slide, see how the fluctuations stabilizes when entering a new λ, which can be done manually or by a slide. The script *boxcox* uses the m-file *boxcoxf*. Make an appropriate transformation and then by some method eliminate the trend and the seasonality. Plot the estimated dual (see homework 1 clause 5.)

Avd. Matematisk statistik

KTH Teknikvetenskap

Homework 2.b. in the course SF2945 Time series analysis.

To be finished and uploaded at bilda.kth.se at latest 10 December 2008

The purpose of this homework is to understand the ARCH and GARCH models for modeling financial time series.

The data you shall use, is in the file *logret_DEM_USD.mat*.

The data in *logret_DEM_USD.mat* are daily logreturns of foreign exchange rates (FX rates) for the German mark (DEM) quoted against the US dollar (USD). If S_t is the FX rate between the German mark and the US dollar at day t then the logreturn X_t over day t is $X_t = \log(S_{t+1}/S_t)$ ($\log(x)$ denotes the natural logarithm with base e).

To load the data into MATLAB use the command:
load logret_DEM_USD
Then you will get a vector named *dem* containing the logreturns for the DEM/USD exchange rate. Note that you have to run MATLAB form the same directory where you have the file *logret_DEM_USD.mat*.

The construction of ARCH and GARCH time series makes it difficult to explicitly compute interesting quantities that a bank or financial institution might be interested in, for instance the *Value-at-Risk*. Therefore simulation of a fitted model is needed to compute such quantities.

1. Simulate the ARCH(1)-process with different choices of the parameters $\alpha_0 > 0$ and $\alpha_1 > 0$. This can be done using the m-file *archsim*. Simulate samples of different length and different starting positions y_0. The command is $[y, \sigma] = \text{archsim}(a0, a1, y0, n)$, where $a0$, $a1$ are the parameters in the ARCH-model, $y0$ is the starting position and n is the sample length. The vector y is the resulting ARCH-process and σ is the volatility process, $\sigma_t = (\alpha_0 + \alpha_1 X_{t-1}^2)^{1/2}$. Your homework should include the following:

(a) One plot of an ARCH-process $\{Y_t\}$ of length $n = 500$ with parameters α_0 and $\alpha_1 < 1$, starting at $y0 = 0$. You should also include a plot with the corresponding volatility process σ_t.

(b) One plot of an ARCH-process $\{Y_t\}$ of length $n = 500$ with parameters α_0 and $1 < \alpha_1 < 2e^\gamma$, starting at $y0 = 0$.

(c) One plot of an ARCH-process $\{Y_t\}$ of length $n = 500$ with parameters α_0 and $2e^\gamma < \alpha_1$, starting at $y0 = 0$.

Can you see any qualitative differences between these plots?

In (a) and (b) the distribution of the ARCH process will eventually converge to a stationary distribution with (a) finite variance and (b) infinite variance. In (c) it will not converge to a stationary distribution.

2. Simulate the GARCH$(1,1)$-process with different choices of the parameters $\alpha_0 > 0$, $\alpha_1 > 0$ and $\beta_1 > 0$. This can be done using the m-file *garchsim*. Simulate samples of different length and different starting positions y_0. The command is $[y, \sigma] = \text{garchsim}(a_0, a_1, b_1, y_0, \sigma_0, n, \mathit{seed})$, where $a0$, $a1$ and $b1$ are the parameters in the GARCH-model, y_0 is the starting position, σ_0 the volatility at the starting position, n is the sample length and *ssed* is the random number seed. For every simulation you do, give the seed. The vector y is the resulting GARCH-process and σ is the volatility process, $\sigma_t = (\alpha_0 + \alpha_1 X_{t-1}^2 + \beta_1 \sigma_{t-1}^2)^{1/2}$.

Your homework should include the following:

(a) One plot of a GARCH-process $\{Y_t\}$ of length $n = 500$ with parameters α_0, α_1 and β_1, starting at $y_0 = 0$, $\sigma_0 = \alpha_0$, such that $\alpha_1 + \beta_1 < 1$. You should also include a plot with the corresponding volatility process σ_t.

(b) One plot of a GARCH-process $\{Y_t\}$ of length $n = 500$ with parameters α_0, α_1 and β_1, starting at $y_0 = 0$, $\sigma_0 = \alpha_0$, such that $\alpha_1 + \beta_1 > 1$.

Do you see any qualitative difference of these plots?

In (a) the distribution of the GARCH process will eventually converge to a stationary distribution whereas in (b) it will not converge to a stationary distribution.

3. Plot the FX logreturns in *logret_DEM_USD.mat*. This plot should be included in your homework.

4. We shall fit a GARCH$(1,1)$-model to this data set. This can be done using the GARCH parameter estimation functions in MATLAB's FINANCIAL TOOLBOX. Unfortunately it is unavailable for the students. Therefore we give the estimates:

$$\alpha_0 = 8.440 \cdot 10^{-7}, \ \alpha_1 = 0.06434, \ \beta_1 = 0.9195$$

5. By assuming that the estimated parameters are correct and that the observed FX logreturns $\{Y_t\}$ comes from the estimated GARCH$(1,1)$ process we can, if we assume some starting position, compute the volatility process for the observed FX logreturns. Let $y_0 = 0$ and $\sigma_0 = 0.005$. Compute the volatility process $\{\sigma_t\}$ for the observed FX logreturns by:

$$\sigma_t = (\alpha_0 + \alpha_1 Y_{t-1}^2 + \beta_1 \sigma_{t-1}^2)^{1/2}, \qquad Y_0 = 0, \ \sigma_0 = 0.005.$$

Plot $\{\sigma_t\}$ and compare with the FX logreturns. Note how the volatility process increases as the variation of the FX data becomes large. Include the plot in your homework.

6. An important issue for many financial institutions is to compute the *Value-at-Risk* for a cash flow over some period of time. The *Value-at-Risk* is defined as follows. If X is a random variable interpreted as some random amount we have earned (or lost if $X < 0$) at some time T, then $\text{VaR}_p(X)$ is the number x_p such that $\mathbb{P}(X \leq -x_p) = p$. For instance, if $p = 0.01$ then $\text{VaR}_{0.01}(X)$ is the amount $x_{0.01}$ such that the probability of loosing more than $x_{0.01}$ is 0.01. If X has normal distribution with mean 0 and variance 1 then $-\text{VaR}_p(X)$ is the pth quantile of the standard normal distribution. Of particular interest to financial institutions is to compute the *Value-at-Risk* over a 10 day period. This is the objective of this exercise.

Assume that we have observed the FX logreturns in the given data set and that the last date in that series is today, day 500. We want to compute $\text{VaR}_{0.01}(W_{510} - W_{500})$ where W_{510} is the amount held in the German mark day 510 given that $W_{500} = 1000000$ DEM. This means that we want to compute the amount $z_{0.01}$ such that the probability of loosing more than $z_{0.01}$ over the next 10 days is 0.01.

Let X be the 10 day logreturn from today: $X = Y_{501} + Y_{502} + \cdots + Y_{510}$. The profit between day 500 and day 510 is

$$W_{510} - W_{500} = \frac{W_{510} - W_{500}}{W_{500}} \cdot W_{500} = \Big(\frac{W_{510}}{W_{500}} - 1\Big) \cdot W_{500} = (e^X - 1) \cdot W_{500} = (e^X - 1) \cdot 10^6.$$

The value $z_{0.01}$ can be estimated by simulating a large number, $N = 10000$, trajectories of the GARCH process $y_{501}, y_{502}, \ldots, y_{510}$ and determine $z_{0.01}$ as the amount such that $0.01 \cdot N = 100$ of the trajectories have $w_{510} - w_{500} = 10^6 \cdot (\exp\{y_{501} + y_{502} + \cdots + y_{510}\} - 1) \leq z_{0.01}$. A useful function to determine $z_{0.01}$ is *prctile* in the m-file *prctile.m*. In your homework you should include:

A histogram over the outcomes of the $N = 10000$ samples of $W_{510} - W_{500}$. This can be done with the MATLAB function *hist*.
An estimate for $z_{0.01}$ based on your simulations of the $N = 10000$ paths of the GARCH process.
The Matlab code for the simulations and calculation of Value-at-Risk.