3 Empirical properties of financial data

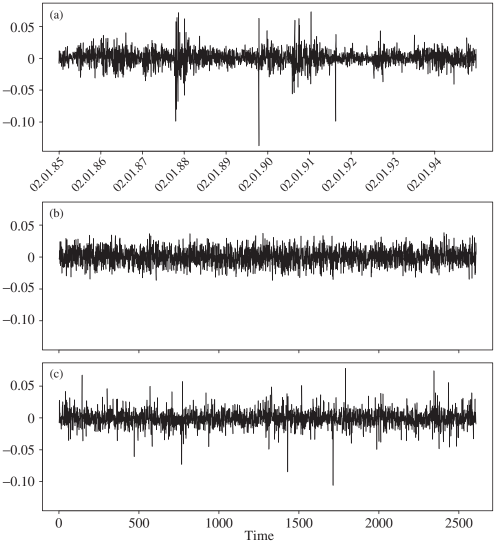
3.1 Stylized facts of financial return series

- The stylized facts are a collection of empirical observations and related inferences, which apply to many time series of risk-factor changes (e.g. log-returns on equities, indices, exchange rates, commodity prices).
- The best-known stylized facts apply to daily log-returns (also to intra-daily, weekly, monthly). Tick-by-tick (high-frequency) data have their own stylized facts (not discussed here) and annual return (low-frequency) data are more difficult to investigate (data sparseness; non-stationarity).
- Consider discrete-time risk-factor changes $X_t = Z_t - Z_{t-1}$ for a log-price or rate $Z_t = \log S_t$. In this case

$$X_t = \log(S_t/S_{t-1}) \approx S_t/S_{t-1} - 1 = (S_t - S_{t-1})/S_{t-1};$$

the former is often called a *(log-)return*, the latter a *relative return*.

3.1.1 Volatility Clustering

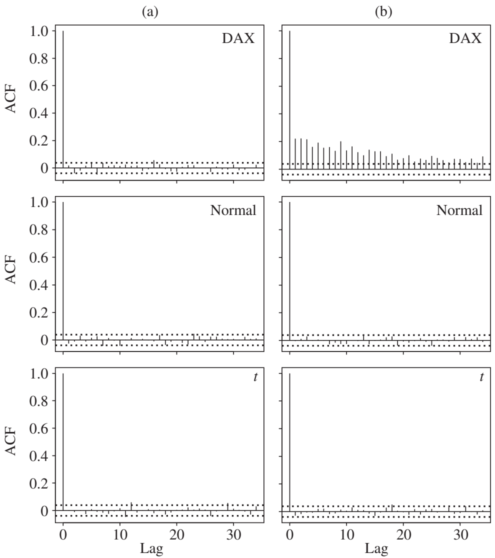


(a) Log-returns for the DAX index from 1985-01-02 to 1994-12-30 ($n = 2608$).

(b) Simulated iid data from a fitted normal with $\hat{\mu} = \bar{X}_n$, $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ show too few extremes.

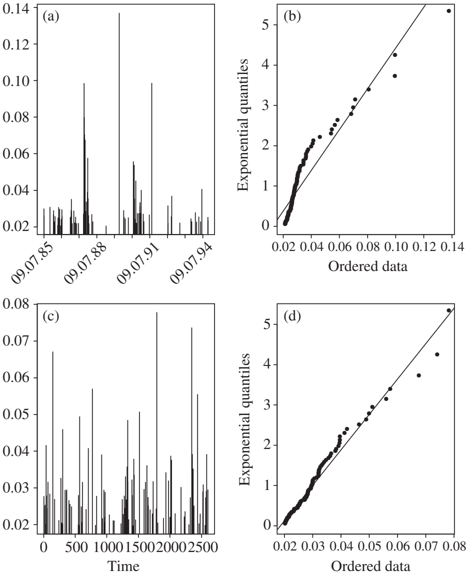
(c) Simulated iid data from a fitted $t_{3.8}$. Better range of values but still no volatility clustering ($=$ tendency for extreme returns to be followed by extreme returns).

Estimated *autocorrelation function (ACF)* $\rho(h) = \mathrm{corr}(X_0, X_h),\ h \in \mathbb{Z}$



- Estimated
  (a) ACF of $(X_t)_{t\in\mathbb{Z}}$
  (b) ACF of $(|X_t|)_{t\in\mathbb{Z}}$
- Non-zero ACF at lag 1 implies a tendency for a return to be followed by a return of equal sign; not the case here.
  $\Rightarrow$ Predicted return ≈ 0
- For iid process $(X_t)_{t\in\mathbb{Z}}$ $\rho_X(h) = \rho_{|X|}(h) = I_{\{h=0\}}$; not the case here. (Can confirm with Ljung–Box tests.)
- $(Z_t)_{t\in\mathbb{Z}}$ not a random walk $(S_t)_{t\in\mathbb{Z}}$ not GBM.

Concerning clustering of extremes, consider the 100 largest losses of the...



- (a) ... DAX index
  (c) ... simulated fitted $t_{3.8}$
- (b), (d) Q-Q plots of waiting times between these large losses (should be $\mathrm{Exp}(\lambda)$ for iid data; see EVT chapter).
- The DAX data shows shorter and longer waiting times than the iid data
  $\Rightarrow$ clustering of extremes.

3.1.2 Non-normality and heavy tails

Formal statistical tests of normality

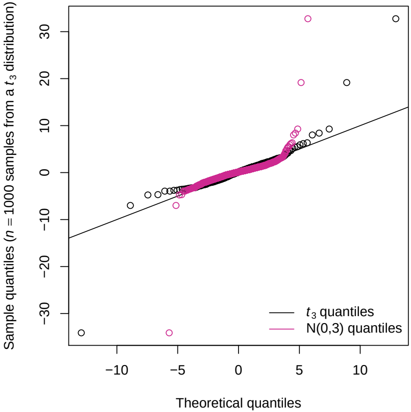
- For general univariate df F:
  - Kolmogorov–Smirnov (test statistic $T_n = \sup_x |\hat{F}_n(x) - F(x)|$)
  - Cramér–von Mises ($T_n = n\int_{-\infty}^{\infty} (\hat{F}_n(x) - F(x))^2 \, dF(x)$)
  - Anderson–Darling ($T_n = n\int_{-\infty}^{\infty} \frac{(\hat{F}_n(x)-F(x))^2}{F(x)(1-F(x))} \, dF(x)$; recommended by D'Agostino and Stephens (1986))
- For $F = \mathrm{N}(\mu, \sigma^2)$:
  - Shapiro–Wilk (idea: quantify Q-Q plot in one number)
  - D'Agostino (based on skewness and kurtosis as Jarque–Bera)
  - Jarque–Bera test: Compares skewness $\beta = \frac{\mathbb{E}((X-\mu)^3)}{\sigma^3}$ and kurtosis $\kappa = \frac{\mathbb{E}((X-\mu)^4)}{\sigma^4}$ with sample versions. The test statistic is

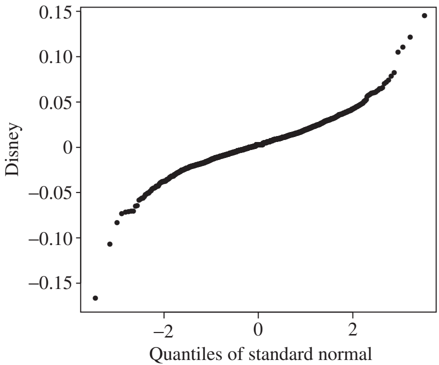
$$T_n = \frac{n}{6}\Big(\hat{\beta}^2 + \frac{1}{4}(\hat{\kappa} - 3)^2\Big) \underset{n \text{ large}}{\overset{H_0}{\sim}} \chi_2^2.$$

Graphical tests

- Suppose we want to graphically test whether $X_1, \dots, X_n \sim F$ for some df F based on realizations of iid $X_1, \dots, X_n$.
- Let $X_{(1)} \le \dots \le X_{(n)}$ denote the corresponding order statistics and note that $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n I_{\{X_i \le x\}} = \frac{1}{n}\sum_{i=1}^n I_{\{X_{(i)} \le x\}}$, $x \in \mathbb{R}$, i.e. the order statistics contain all relevant information about $X_1, \dots, X_n$.
- Possible graphical tests (see also the appendix):
  - P-P plot: Plot $\{(p_i, F(X_{(i)})) : i = 1, \dots, n\}$, where $p_i \underset{n>10}{=} \frac{i-1/2}{n} \approx \frac{i}{n} \approx \hat{F}_n(X_{(i)})$. If $F \approx \hat{F}_n$, the points lie roughly on a line with slope 1; this also applies to Q-Q plots.
  - Q-Q plot: Plot $\{(F^{\leftarrow}(p_i), X_{(i)}) : i = 1, \dots, n\}$ (tail differences better visible).

Interpreting Q-Q plots (S-shape hints at heavier tails than $\mathrm{N}(\mu, \sigma^2)$):





Daily returns typically have kurtosis $\kappa > 3$ (*leptokurtic*; narrower center, heavier tails than $\mathrm{N}(\mu, \sigma^2)$ for which $\kappa = 3$). They have power-like tails rather than exponential.

3.1.3 Longer-interval return series

- By going from daily to weekly, monthly, quarterly and yearly data, these effects become less pronounced (returns look more iid, less heavy-tailed).
- The (non-overlapping) h-period log-return at $t \in \{h, 2h, \dots, \lfloor \frac{n}{h} \rfloor h\}$ is
$$X_t^{(h)} = \log\Big(\frac{S_t}{S_{t-h}}\Big) = \log\Big(\frac{S_t}{S_{t-1}}\frac{S_{t-1}}{S_{t-2}}\dots\frac{S_{t-h+1}}{S_{t-h}}\Big) = \sum_{k=0}^{h-1} X_{t-k}$$
A Central Limit Theorem (CLT) effect takes place (less heavy-tailed, less evidence of serial correlation).
- Problem: the larger h, the less data are available.
- Possible remedy: Consider overlapping returns
$$\Big\{X_t^{(h)} : t \in \Big\{h, h+k, \dots, h + \Big\lfloor \frac{n-h}{k} \Big\rfloor k\Big\}\Big\}, \quad 1 \le k < h.$$
$\Rightarrow$ More data but serially dependent now.

To summarize, we can infer the following stylized facts about univariate financial return series:

(U1) Return series are not iid although they show little serial correlation;

(U2) Series of absolute or squared returns show profound serial correlation;

(U3) Conditional expected returns are close to zero;

(U4) Volatility (conditional standard deviation) appears to vary over time;

(U5) Extreme returns appear in clusters;

(U6) Return series are leptokurtic or heavy-tailed (power-like tail).

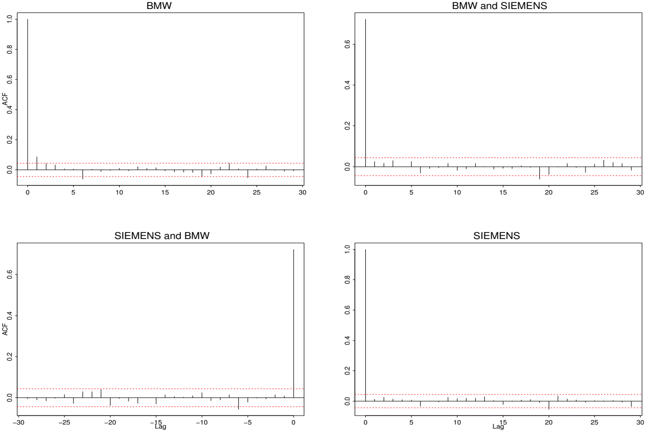
3.2 Multivariate stylized facts

Consider multivariate log-return data $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$.

3.2.1 Correlation between series

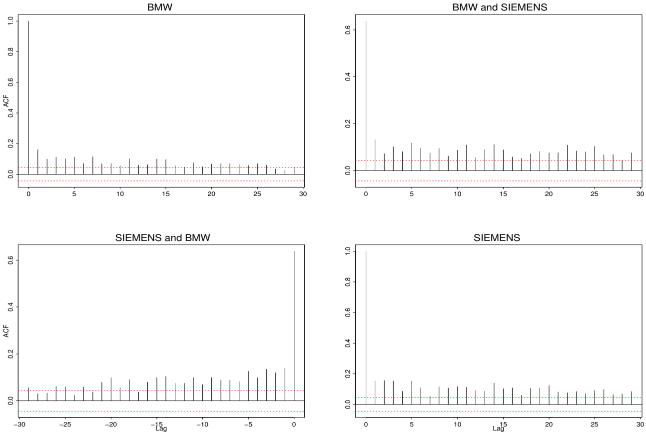
- By (U1), the returns of stock A at t and $t+h$ show little correlation. The same applies to the returns of stock A at t and stock B at $t+h$, $h > 0$. Stock A and stock B on day t may be correlated due to factors that affect the whole market (*contemporaneous dependence*).
- Correlations of returns at t vary over time (difficult to detect whether changes are continual or constant within regimes; fit different models for changing correlation, then make a formal comparison).
- Periods of high/low volatility are typically common to more than one stock $\Rightarrow$ Returns of large magnitude in A at t may be followed by returns of large magnitude in A and B at $t+h$.

Estimated correlations between/within series:



2000 values from period 1985-01-23 to 1994-09-22

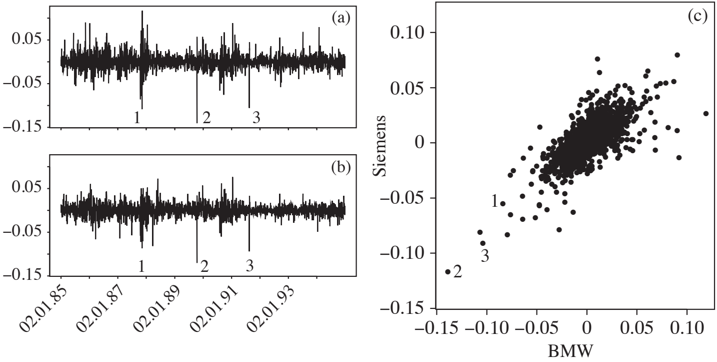
Estimated correlations between/within series of absolute values:



2000 values from period 1985-01-23 to 1994-09-22

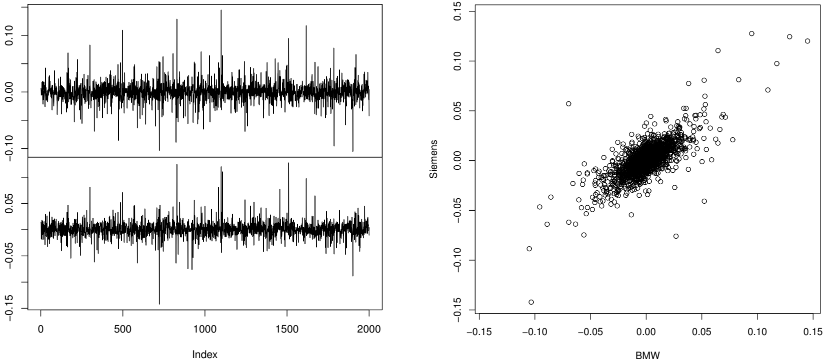
3.2.2 Tail dependence

(BMW, Siemens) log-returns from 1985-01-23 to 1994-09-22 ($n = 2000$)



In volatile/extreme periods, dependence is stronger (1: 1987-10-19 Black Monday (DJ drop by 22%); 2: 1989-10-16 Monday demonstrations in Leipzig (Wende); 3: 1991-08-19 coup against soviet president M. Gorbachev).

Simulated log-returns from a fitted bivariate t distribution ($n = 2000$; $\rho = 0.72$, $\nu = 2.8$ both fitted to (BMW, Siemens))



- The multivariate t distribution can replicate joint large gains/losses but in a symmetric way.
- The multivariate normal distribution cannot replicate such behaviour, known as tail dependence; see Chapter 7.

To summarize, we can infer the following stylized facts about multivariate financial return series:

(M1) Multivariate return series show little evidence of cross-correlation, except for contemporaneous returns (i.e. at the same t);

(M2) Multivariate series of absolute returns show profound cross-correlation;

(M3) Correlations between contemporaneous returns vary over time;

(M4) Extreme returns in one series often coincide with extreme returns in several other series (e.g. tail dependence).