3 Empirical properties of financial data

3.1 Stylized facts of financial return series

3.2 Multivariate stylized facts

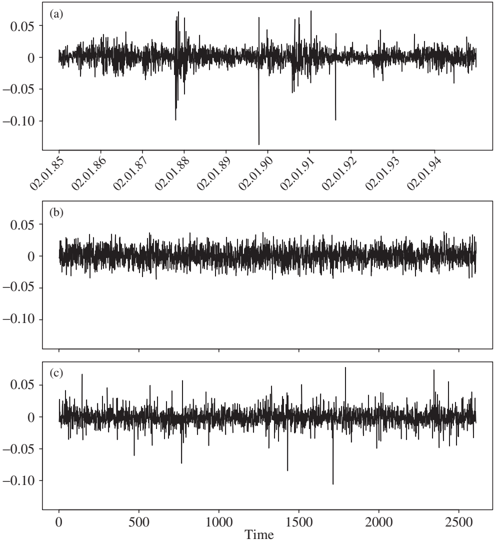
3.1 Stylized facts of financial return series

- 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 *simple return*.

3.1.1 Volatility Clustering

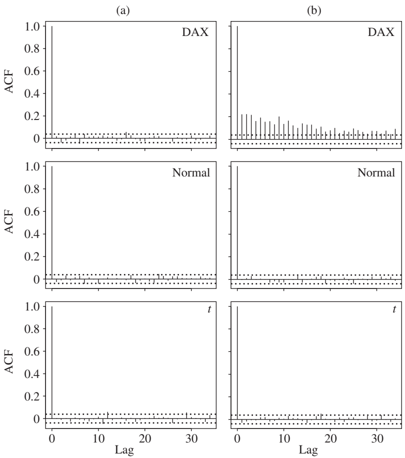


(a) Log-returns for the DAX index from 1985-01-02 to 1994-12-30 ($n = 2608$).

(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.

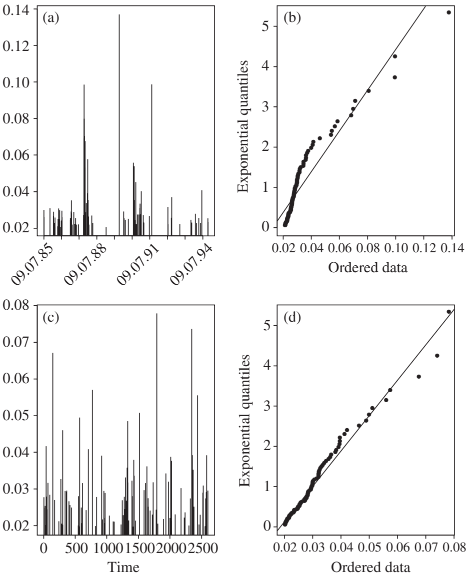
(c) Simulated iid data from a fitted $t_{3.8}$. Better range of values but still no volatility clustering ($=$ tendency for extreme returns to be followed by extreme returns).

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



- (a) Sample ACF of $(X_t)_{t\in\mathbb{Z}}$
  (b) and $(|X_t|)_{t\in\mathbb{Z}}$ (similarly: $(X_t^2)_{t\in\mathbb{Z}}$)
- (a): Positive ACF at lag 1 would imply a tendency for returns to be followed by returns of equal sign $\Rightarrow$ not the case $\Rightarrow$ predicted returns ≈ 0
- (b): Positive ACF at lag 1 would imply a tendency for large (small) returns to be followed by large (small) returns $\Rightarrow$ the case for the DAX data $\Rightarrow$ returns cluster (not iid).
- $(X_t)_{t\in\mathbb{Z}}$ not iid $\Rightarrow$ $(Z_t)_{t\in\mathbb{Z}}$ not a Brownian motion $\Rightarrow$ $(S_t)_{t\in\mathbb{Z}}$ not a 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);
- The DAX data shows longer and shorter waiting times than the iid data, so 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 ($T_n = n\int_{-\infty}^{\infty}(\hat{F}_n(x) - F(x))^2 \,\mathrm{d}F(x)$)
  - Anderson–Darling ($T_n = n\int_{-\infty}^{\infty}\frac{(\hat{F}_n(x)-F(x))^2}{F(x)(1-F(x))}\,\mathrm{d}F(x)$; recommended by D'Agostino and Stephens (1986))
- For $F = \mathrm{N}(\mu, \sigma^2)$:
  - Shapiro–Wilk (idea: quantify Q-Q plot in one number, biased by n)
  - 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 $T_n = \frac{n}{6}\big(\hat{\beta}^2 + \frac{1}{4}(\hat{\kappa} - 3)^2\big) \underset{n\,\text{large}}{\overset{H_0}{\sim}} \chi^2_2$.

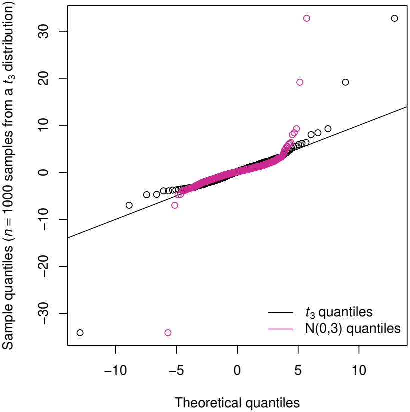
Graphical tests

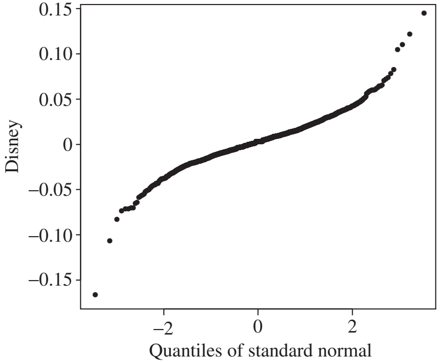
- We can also graphically test whether $X_1, \dots, X_n \sim F$ for some df F based on realizations of iid $X_1, \dots, X_n$.
- Let $X_{(1)} \le \dots \le X_{(n)}$ denote the corresponding order statistics and note that

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n I_{\{X_i \le x\}} = \frac{1}{n}\sum_{i=1}^n I_{\{X_{(i)} \le x\}}, \quad x \in \mathbb{R},$$

  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: For $p_i \underset{n>10}{=} \frac{i-1/2}{n} \approx \frac{i}{n}$, plot $\{(p_i, F(X_{(i)})) : i = 1, \dots, n\}$. If $F \approx \hat{F}_n$, $F(X_{(i)}) \approx p_i$, so the points lie on a line with slope 1.
  - 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 a leptokurtic distribution, i.e., narrower center, heavier tails than $\mathrm{N}(\mu, \sigma^2)$ (kurtosis $\kappa = 3$)):





Daily returns typically have kurtosis $\kappa > 3$.

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).
- Since the h-period log-return at $t \in \{h, 2h, \dots, \lfloor \frac{n}{h} \rfloor h\}$ is

$$X_t^{(h)} = \log\biggl(\frac{S_t}{S_{t-h}}\biggr) = \log\biggl(\frac{S_t}{S_{t-1}} \frac{S_{t-1}}{S_{t-2}} \dots \frac{S_{t-h+1}}{S_{t-h}}\biggr) = \sum_{k=0}^{h-1} X_{t-k},$$

  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.
- One could consider overlapping returns

$$\Bigl\{X_t^{(h)} : t \in \Bigl\{h, h+k, \dots, h + \Bigl\lfloor \frac{n-h}{k} \Bigr\rfloor k \Bigr\}\Bigr\}, \quad 1 \le k < h.$$

  $\Rightarrow$ more data but serially dependent now (typically not a good idea).

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 tails).

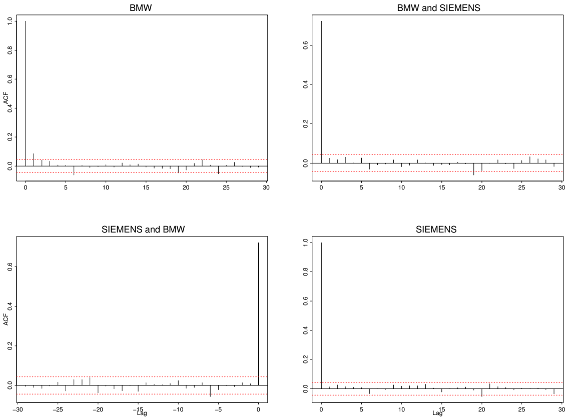
3.2 Multivariate stylized facts

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

3.2.1 Correlation between series

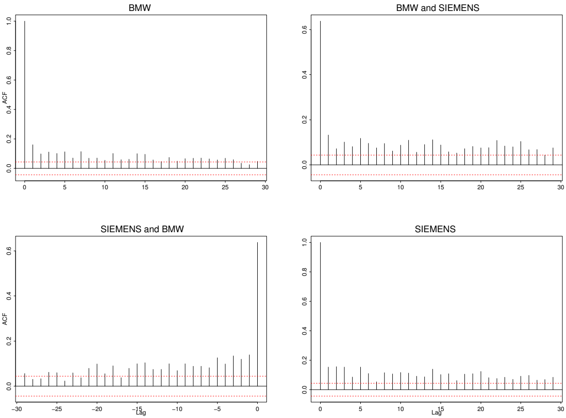
- By (U1), the returns of stock A at t and $t+h$ show little (auto)correlation. The returns of stock A at t and stock B at $t+h$, $h > 0$, also show little cross-correlation. However, Stock A and stock B on day t may be correlated due to factors that affect the whole market (*contemporaneous dependence*).
- These 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, so 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:



Based on 2000 values from period 1985-01-23 to 1994-09-22. Little autocorrelation, little crosscorrelation (at different lags), contemporaneous correlation.

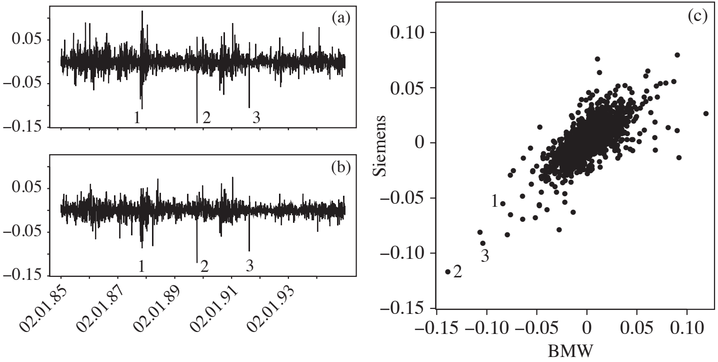
Estimated correlations between/within series of absolute values:



Autocorrelation of absolute returns (indication of volatility clustering). Common to more than one stock.

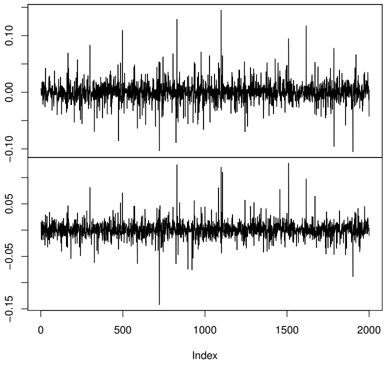
3.2.2 **Tail dependence** (for quantifying joint extremes)

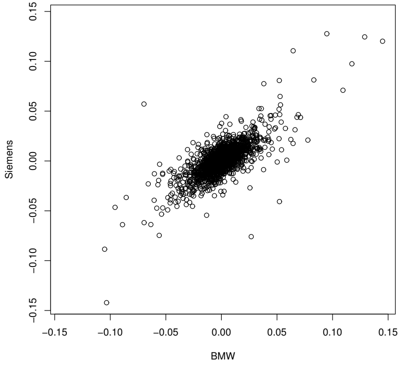
(BMW, Siemens) log-returns from 1985-01-23 to 1994-09-22 ($n = 2000$)



In volatile/extreme(ly bad) periods, dependence seems stronger (1: 1987-10-19 Black Monday (DJ drop by 22%); 2: 1989-10-16 Monday demonstrations (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 (difficult to infer from empirical correlations due to estimation error in small samples);

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

References

D'Agostino, R. B. and Stephens, M. A. (1986), Goodness-of-fit techniques, Dekker.