3 Empirical properties of financial data

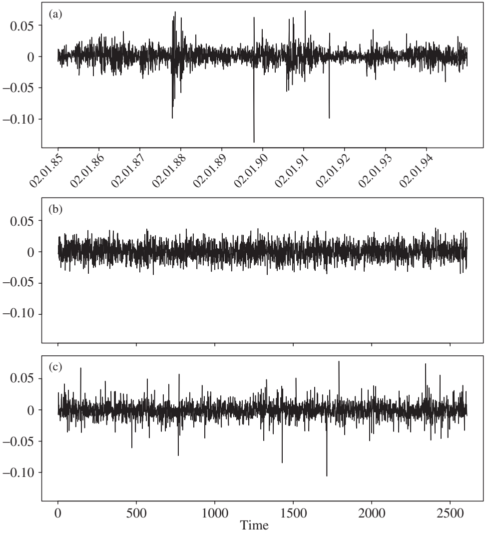
3.1 Stylized facts of financial return series

- Stylized facts are a collection of empirical observations and inferences drawn of such, which apply to many time series of risk-factor changes (e.g. log-returns on equities, indices, exchange rates, commodity prices).
- Consider discrete-time risk-factor changes $X_t = Z_t - Z_{t-1}$, e.g. $Z_t = \log S_t$, in which case

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

  is often called a *(log-)return*.
- Stylized facts often 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 sparcity; non-stationarity).

3.1.1 Volatility Clustering

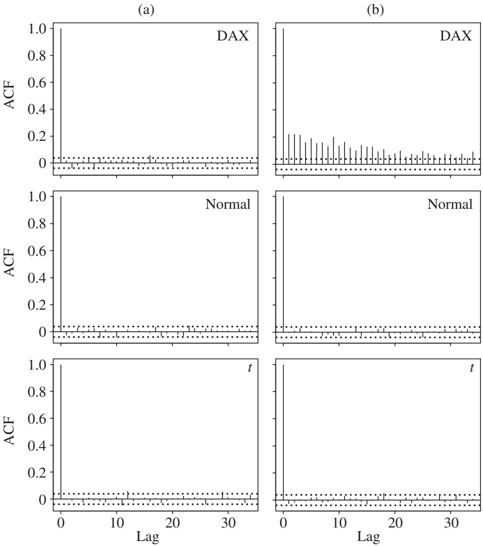


(a) Log-returns for the DAX index from 1985-01-02 to 1994-12-30 ($n = 2608$)

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

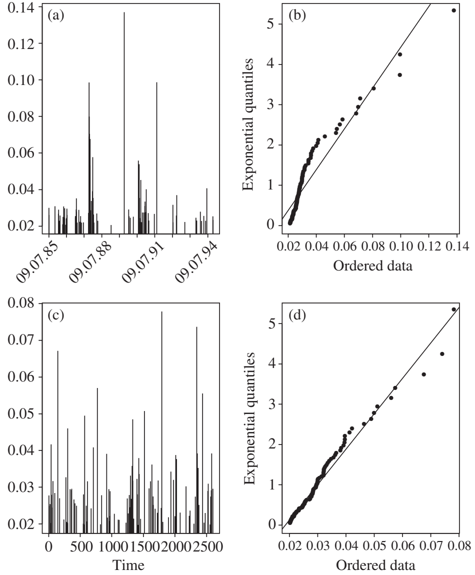
(c) Simulated iid data from a fitted $t_{3.8}$ (num. max. of log-likelihood; still no volatility clustering $=$ tendency for extreme returns to be followed by extreme returns)

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



- (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 $= 0$
- iid data $(X_t)_{t\in\mathbb{Z}}$ implies $\rho_X(h) = \rho_{|X|}(h) = I_{\{h=0\}}$; not the case here; confirm with a Ljung–Box test (H_0: ACF at first h lags $= 0$)
- Not a random walk

100 largest negative log-returns (losses) of...



- (a) ... DAX index
  (c) ... simulated fitted $t_{3.8}$
- (b), (d) Q-Q plots of (theoretical) waiting times between extreme losses (should be $\mathrm{Exp}(\lambda)$ for iid data; see EVT) against empirical ones.
- 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
  - Cramér–von Mises
  - Anderson–Darling (recommended by D'Agostino and Stephens (1986))
- For $\mathrm{N}(\mu, \sigma^2)$:
  - D'Agostino
  - Shapiro–Wilk
  - Jarque–Bera test: Compares $\beta = \frac{\mathbb{E}((X-\mu)^3)}{\sigma^3}$ (skewness) and $\kappa = \frac{\mathbb{E}((X-\mu)^4)}{\sigma^4}$ (kurtosis) with sample versions. The test statistic is

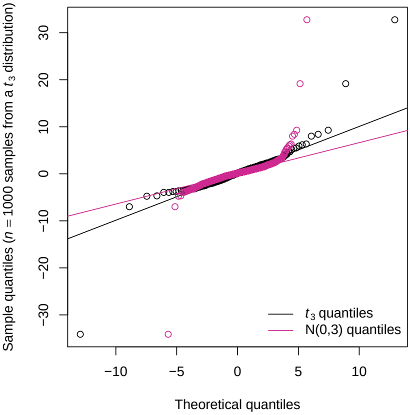
$$T = \frac{n}{6}(\hat{\beta}^2 + \frac{1}{4}(\hat{\kappa} - 3)^2),$$

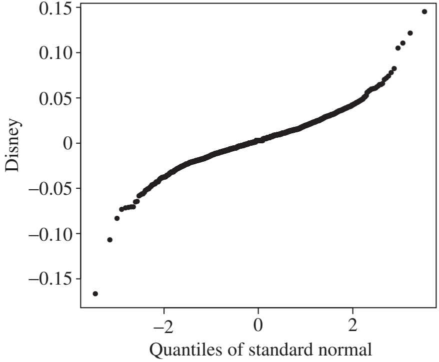
which is asymptotically χ^2_2 distributed under H_0: data is normal.

Graphical tests

- For iid $X_1, \dots, X_n$, the *empirical distribution function (edf)* is defined by $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n I_{\{X_i \le x\}}$, $x \in \mathbb{R}$.
- Suppose we want to graphically test whether $X_1, \dots, X_n \sim F$ for some df F based on realizations $x_1, \dots, x_n$. Let $x_{(1)} \le \dots \le x_{(n)}$ denote the corresponding order statistics. Possible testing options are:
  - P-P plot: Plot $\{(p_i, F(x_{(i)})) : i = 1, \dots, n\}$, where $p_i \approx \frac{i-1/2}{n} \approx \hat{F}_n(x_i)$ (equality for $p_i = i/n \Rightarrow$ If $F \approx \hat{F}_n$, the points lie roughly on a line).
  - Q-Q plot: Plot $\{(F^{\leftarrow}(p_i), x_{(i)}) : i = 1, \dots, n\}$ (tail differences better visible).
- If F is (reasonably close to) the underlying unknown df, P-P and Q-Q plots resemble lines with slope 1.

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). They are typically 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

$$X_t^{(h)} = \log\left(\frac{S_t}{S_{t-h}}\right) = \log\left(\frac{S_t}{S_{t-1}} \cdots \frac{S_{t-h+1}}{S_{t-h}}\right) = \sum_{k=0}^{h-1} X_{t-k}$$

  A CLT effect takes place (less heavy-tailed, less evidence of serial correlation)
- Problem: the larger h, the less data is available
- Possible remedy: overlapping returns $\{X_t^{(h)} : t \in \{h, h+k, \dots, h + \lfloor \frac{n-h}{k} \rfloor k\}\}$ for $1 \le k < h \Rightarrow$ more data but now serially dependent.

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 leptokurticor heavy-tailed (power-like tail).

3.2 Multivariate stylized facts

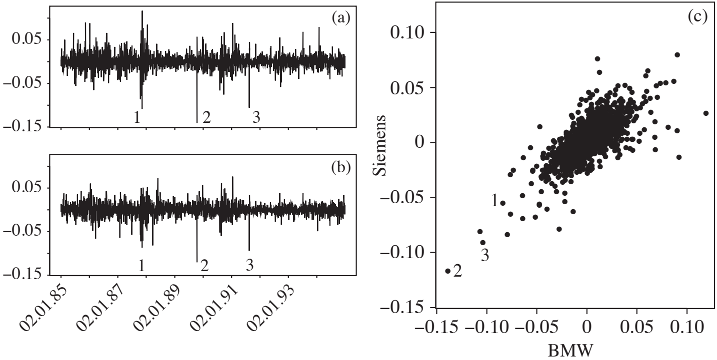
Consider multivariate log-return data $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$.

3.2.1 Correlation between series

- (U1) $\Rightarrow$ Returns of stock A at t and $t + h$ show little correlation, so do 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).
- 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 $t + h$.
- 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).

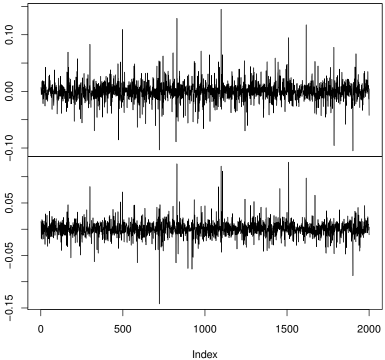
3.2.2 Tail dependence

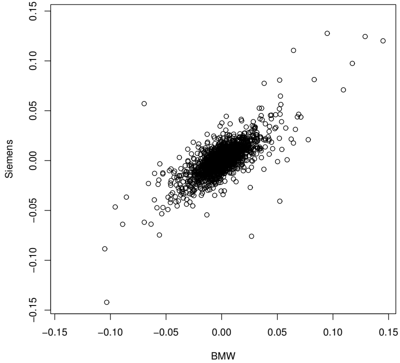
(BMW, Siemens) log-returns from 1985-01-23 to 1994-09-22 ($n = 2000$)



In volatile/extreme periods, dependence seems stronger (1987-10-19 Black Monday (DJ drop by 22%); 1989-10-16 Monday demonstrations in Leipzig (Wende); 1991-08-19 coup against soviet president Mikhail 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 later).

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;

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

(M3) Correlations between contemporaneous returns vary over time (not so easy to infer with empirical correlations due to considerable estimation error in small samples; most reliable way: fit various models and make a formal statistical comparison between the models);

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