6 Multivariate models

6.1 Basics of multivariate modeling

6.1.1 Random vectors and their distributions

Joint and marginal distributions

- Let $\boldsymbol{X} = (X_1, \dots, X_d) : \Omega \to \mathbb{R}^d$ be a d-dimensional *random vector* (representing risk-factor changes, risks, etc.).
- The *(joint) distribution function (df) H of $\boldsymbol{X}$* is

$$H(\boldsymbol{x}) = F_{\boldsymbol{X}}(\boldsymbol{x}) = \mathbb{P}(\boldsymbol{X} \le \boldsymbol{x}) = \mathbb{P}(X_1 \le x_1, \dots, X_d \le x_d), \quad \boldsymbol{x} \in \mathbb{R}^d.$$

- The *jth margin* or *marginal df F_j of $\boldsymbol{X}$* is

$$\begin{aligned} &F_j(x_j) = \mathbb{P}(X_j \le x_j) \\ &= \mathbb{P}(X_1 \le \infty, \dots, X_{j-1} \le \infty, X_j \le x_j, X_{j+1} \le \infty, \dots, X_d \le \infty) \\ &= H(\infty, \dots, \infty, x_j, \infty, \dots, \infty), \quad x_j \in \mathbb{R},\ j \in \{1, \dots, d\}. \end{aligned}$$

(interpreted as a limit).

- Similarly for *k-dimensional margins*. Suppose we partition $\boldsymbol{X}$ into $(\boldsymbol{X}_1^\top, \boldsymbol{X}_2^\top)^\top$, where $\boldsymbol{X}_1 = (X_1, \dots, X_k)^\top$ and $\boldsymbol{X}_2 = (X_{k+1}, \dots, X_d)^\top$, then the marginal distribution function of $\boldsymbol{X}_1$ is

$$F_{\boldsymbol{X}_1}(\boldsymbol{x}_1) = \mathbb{P}(\boldsymbol{X}_1 \le \boldsymbol{x}_1) = H(x_1, \dots, x_k, \infty, \dots, \infty).$$

- *H is absolutely continuous* if

$$H(\boldsymbol{x}) = \int_{-\infty}^{x_d} \dots \int_{-\infty}^{x_1} h(z_1, \dots, z_d)\, dz_1 \dots dz_d = \int_{(-\boldsymbol{\infty}, \boldsymbol{x}]} h(\boldsymbol{z})\, d\boldsymbol{z}$$

  for some $h \ge 0$ then known as the *(joint) density of $\boldsymbol{X}$ (or H)*. Similarly, the *jth marginal df F_j is absolutely continuous* if $F_j(x) = \int_{-\infty}^{x} f_j(z)\, dz$ for some $f_j \ge 0$ then known as the *density of X_j* (or *F_j*).

- In case h exists, $F_j(x_j) = \int_{-\infty}^{x_j} \int_{(-\boldsymbol{\infty}, \boldsymbol{\infty})} h(\boldsymbol{z})\, d\boldsymbol{z}_{-j}\, dz_j = \int_{-\infty}^{x_j} f_j(z_j)\, dz_j$, so that $f_j(z_j)$ can be recovered from h via

$$\underbrace{\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty}}_{d-1\text{-many}} h(z_1, \dots, z_{j-1}, z_j, z_{j+1}, \dots, z_d)\, dz_1 \dots dz_{j-1} dz_{j+1} \dots dz_d.$$

- Existence of a joint density $\Rightarrow$ Existence of marginal densities for all k-dimensional marginals, $1 \le k \le d-1$. The converse is false in general (counter-examples can be constructed with copulas; see later).
- By replacing integrals by sums, one obtains similar formulas for the discrete case, in which we call densities *probability mass functions*.
- Sometimes it's convenient to work with the *survival function $\bar{H}$ of $\boldsymbol{X}$*, given by

$$\bar{H}(\boldsymbol{x}) = \bar{F}_{\boldsymbol{X}}(\boldsymbol{x}) = \mathbb{P}(\boldsymbol{X} > \boldsymbol{x}) = \mathbb{P}(X_1 > x_1, \dots, X_d > x_d), \quad \boldsymbol{x} \in \mathbb{R}^d,$$

with corresponding *jth marginal survival function $\bar{F}_j$*

$$\begin{aligned}\bar{F}_j(x_j) &= \mathbb{P}(X_j > x_j)\\ &= \bar{H}(-\infty, \dots, -\infty, x_j, -\infty, \dots, -\infty), \quad x_j \in \mathbb{R},\ j \in \{1, \dots, d\}.\end{aligned}$$

- Note that, unlike for $d = 1$, $\bar{H}(\boldsymbol{x}) \ne 1 - H(\boldsymbol{x})$ in general.

Conditional distributions and independence

- A multivariate model for risks in the form of a joint df, survival function or density, implicitly describes their *dependence structure*. We can then make statements about condtional probabilities.
- As before, consider $\boldsymbol{X} = (\boldsymbol{X}_1^\top, \boldsymbol{X}_2^\top)^\top \sim H$. The *conditional df of $\boldsymbol{X}_2$ given $\boldsymbol{X}_1 = \boldsymbol{x}_1$* is $F_{\boldsymbol{X}_2|\boldsymbol{X}_1}(\boldsymbol{x}_2 \,|\, \boldsymbol{x}_1) = \mathbb{P}(\boldsymbol{X}_2 \le \boldsymbol{x}_2 \,|\, \boldsymbol{X}_1 = \boldsymbol{x}_1) = \mathbb{E}[\mathbb{1}_{\{\boldsymbol{X}_2 \le \boldsymbol{x}_2\}} \,|\, \boldsymbol{X}_1 = \boldsymbol{x}_1]$, where $\mathbb{E}[\,\cdot\,|\,\cdot\,]$ denotes conditional expectation (not discussed here).
- A useful identity for conditional dfs is

$$
\begin{aligned}
&\int_{(-\infty, \boldsymbol{x}_1]} F_{\boldsymbol{X}_2|\boldsymbol{X}_1}(\boldsymbol{x}_2 \,|\, \boldsymbol{z})\, dF_{\boldsymbol{X}_1}(\boldsymbol{z}) \\
&= \int_{\mathbb{R}^d} \mathbb{1}_{\{\boldsymbol{z} \le \boldsymbol{x}_1\}} \mathbb{E}[\mathbb{1}_{\{\boldsymbol{X}_2 \le \boldsymbol{x}_2\}} \,|\, \boldsymbol{X}_1 = \boldsymbol{z}]\, dF_{\boldsymbol{X}_1}(\boldsymbol{z}) \\
&= \mathbb{E}[\mathbb{1}_{\{\boldsymbol{X}_1 \le \boldsymbol{x}_1\}} \mathbb{E}[\mathbb{1}_{\{\boldsymbol{X}_2 \le \boldsymbol{x}_2\}} \,|\, \boldsymbol{X}_1]] = \mathbb{E}[\mathbb{E}[\mathbb{1}_{\{\boldsymbol{X}_1 \le \boldsymbol{x}_1, \boldsymbol{X}_2 \le \boldsymbol{x}_2\}} \,|\, \boldsymbol{X}_1]] \\
&\underset{\text{property}}{\overset{\text{tower}}{=}} \mathbb{E}[\mathbb{1}_{\{\boldsymbol{X}_1 \le \boldsymbol{x}_1, \boldsymbol{X}_2 \le \boldsymbol{x}_2\}}] = H(\boldsymbol{x}),
\end{aligned}
$$

where the second-last equality holds by the tower property of conditional expectations.

- If $\boldsymbol{x}_1 \to \infty$, then $F_{\boldsymbol{X}_2}(\boldsymbol{x}_2) = \int_{\mathbb{R}^d} F_{\boldsymbol{X}_2|\boldsymbol{X}_1}(\boldsymbol{x}_2 \,|\, \boldsymbol{z})\, dF_{\boldsymbol{X}_1}(\boldsymbol{z})$. Furthermore, if H has a density h, then $f_{\boldsymbol{X}_2}(\boldsymbol{x}_2) = \int_{\mathbb{R}^d} f_{\boldsymbol{X}_2|\boldsymbol{X}_1}(\boldsymbol{x}_2 \,|\, \boldsymbol{z})\, dF_{\boldsymbol{X}_1}(\boldsymbol{z})$.
- If H has density h and $f_{\boldsymbol{X}_1}$ denotes the density of $\boldsymbol{X}_1$, then

$$\begin{aligned} h(\boldsymbol{x}_1, \boldsymbol{x}_2) &= \frac{\partial^2}{\partial \boldsymbol{x}_2 \partial \boldsymbol{x}_1} H(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{\partial}{\partial \boldsymbol{x}_2} F_{\boldsymbol{X}_2|\boldsymbol{X}_1}(\boldsymbol{x}_2 \,|\, \boldsymbol{x}_1) f_{\boldsymbol{X}_1}(\boldsymbol{x}_1) \\ &= f_{\boldsymbol{X}_2|\boldsymbol{X}_1}(\boldsymbol{x}_2 \,|\, \boldsymbol{x}_1) f_{\boldsymbol{X}_1}(\boldsymbol{x}_1). \end{aligned}$$

  We call

$$f_{\boldsymbol{X}_2|\boldsymbol{X}_1}(\boldsymbol{x}_2 \,|\, \boldsymbol{x}_1) = \frac{h(\boldsymbol{x}_1, \boldsymbol{x}_2)}{f_{\boldsymbol{X}_1}(\boldsymbol{x}_1)} \tag{19}$$

  the *conditional density of $\boldsymbol{X}_2$ given $\boldsymbol{X}_1 = \boldsymbol{x}_1$*. In this case, the conditional df $F_{\boldsymbol{X}_2|\boldsymbol{X}_1}(\boldsymbol{x}_2 \,|\, \boldsymbol{x}_1)$ is given by

$$F_{\boldsymbol{X}_2|\boldsymbol{X}_1}(\boldsymbol{x}_2 \,|\, \boldsymbol{x}_1) = \int_{-\infty}^{x_{k+1}} \dots \int_{-\infty}^{x_d} \frac{h(\boldsymbol{x}_1, \boldsymbol{z})}{f_{\boldsymbol{X}_1}(\boldsymbol{x}_1)}\, dz_{k+1} \dots dz_d.$$

- Inspired by (19), we call *$\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ independent* if

$$H(\boldsymbol{x}_1, \boldsymbol{x}_2) = F_{\boldsymbol{X}_1}(\boldsymbol{x}_1) F_{\boldsymbol{X}_2}(\boldsymbol{x}_2), \quad \forall\, \boldsymbol{x}_1, \boldsymbol{x}_2.$$

- If H has density h, then $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent if

$$h(\boldsymbol{x}_1, \boldsymbol{x}_2) = f_{\boldsymbol{X}_1}(\boldsymbol{x}_1) f_{\boldsymbol{X}_2}(\boldsymbol{x}_2), \quad \forall\, \boldsymbol{x}_1, \boldsymbol{x}_2.$$

  In this case, $f_{\boldsymbol{X}_2 | \boldsymbol{X}_1}(\boldsymbol{x}_2 \,|\, \boldsymbol{x}_1) = h(\boldsymbol{x}_1, \boldsymbol{x}_2)/f_{\boldsymbol{X}_1}(\boldsymbol{x}_1) = f_{\boldsymbol{X}_2}(\boldsymbol{x}_2)$.

- The components $X_1, \dots, X_d$ of $\boldsymbol{X}$ are *(mutually) independent* if

$$H(\boldsymbol{x}) = \prod_{j=1}^{d} F_j(x_j), \quad \forall\, \boldsymbol{x},$$

  or, if H has density h,

$$h(\boldsymbol{x}) = \prod_{j=1}^{d} f_j(x_j), \quad \forall\, \boldsymbol{x}.$$

Moments and characteristic function

- If $\mathbb{E}|X_j| < \infty$, $j \in \{1, \dots, d\}$, the *mean vector of* $\boldsymbol{X}$ is defined by
$$\boldsymbol{\mu} = \mathbb{E}\boldsymbol{X} = (\mathbb{E}X_1, \dots, \mathbb{E}X_d).$$
One can show: $X_1, \dots, X_d$ independent $\Rightarrow \mathbb{E}[X_1 \cdots X_d] = \prod_{j=1}^d \mathbb{E}[X_j]$
- If $\mathbb{E}[X_j^2] < \infty$ for all j, the *covariance matrix of* $\boldsymbol{X}$ is defined as
$$\Sigma = \operatorname{Cov}[\boldsymbol{X}] = \mathbb{E}[(\boldsymbol{X} - \mathbb{E}\boldsymbol{X})(\boldsymbol{X} - \mathbb{E}\boldsymbol{X})^\top].$$
Its (i, j)th element is
$$\begin{aligned}\sigma_{ij} = \Sigma_{ij} = \operatorname{Cov}[X_i, X_j] &= \mathbb{E}[(X_i - \mathbb{E}X_i)(X_j - \mathbb{E}X_j)]\\ &= \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j];\end{aligned}$$
the diagonal elements are $\sigma_{jj} = \operatorname{Var}[X_j]$, $j \in \{1, \dots, d\}$.
- The *cross covariance matrix* between two (admissible) random vectors $\boldsymbol{X}, \boldsymbol{Y}$ is defined as $\operatorname{Cov}[\boldsymbol{X}, \boldsymbol{Y}] = \mathbb{E}[(\boldsymbol{X} - \mathbb{E}\boldsymbol{X})(\boldsymbol{Y} - \mathbb{E}\boldsymbol{Y})^\top]$. Note that $\operatorname{Cov}[\boldsymbol{X}, \boldsymbol{X}] = \operatorname{Cov}[\boldsymbol{X}]$.

- If $\mathbb{E}[X_j^2] < \infty$, $j \in \{1, \dots, d\}$, the *correlation matrix of $\boldsymbol{X}$* is defined as the matrix $P = \mathrm{Cor}[\boldsymbol{X}]$ with (i, j)th element

$$\rho_{ij} = P_{ij} = \mathrm{Cor}[X_i, X_j] = \frac{\mathrm{Cov}[X_i, X_j]}{\sqrt{\mathrm{Var}[X_i]\,\mathrm{Var}[X_j]}}$$

  which is in $[-1, 1]$ with $\rho_{ij} = \pm 1$ if and only if $X_j \overset{\text{a.s.}}{=} aX_i + b$ for some $a \gtrless 0$ and $b \in \mathbb{R}$. This follows from the Cauchy–Schwarz inequality $|\langle X_i, X_j \rangle| \le \sqrt{\langle X_i, X_i \rangle \langle X_j, X_j \rangle}$ applied to the inner product $\langle X_i, X_j \rangle := \mathbb{E}[X_i X_j]$.
- X_i, X_j $(i \neq j)$ independent $\underset{\not\Leftarrow}{\Rightarrow}$ $\mathrm{Cov}[X_i, X_j] = 0$. The only known distribution for which uncorrelatedness implies independence is the multivariate normal distribution.
- **Some properties:**

  1) For all $A \in \mathbb{R}^{k \times d}$, $\boldsymbol{b} \in \mathbb{R}^k$:

     - $\mathbb{E}[A\boldsymbol{X} + \boldsymbol{b}] = A\mathbb{E}\boldsymbol{X} + \boldsymbol{b} = A\boldsymbol{\mu} + \boldsymbol{b}$;

- $\mathrm{Cov}[A\boldsymbol{X}+\boldsymbol{b}] = A\,\mathrm{Cov}[\boldsymbol{X}]A^\top = A\Sigma A^\top$; if $k=1$ ($A=\boldsymbol{a}^\top$),

$$\boldsymbol{a}^\top\Sigma\boldsymbol{a} = \mathrm{Cov}[\boldsymbol{a}^\top\boldsymbol{X}] = \mathrm{Var}[\boldsymbol{a}^\top\boldsymbol{X}] \ge 0, \quad \boldsymbol{a}\in\mathbb{R}^d, \tag{20}$$

  i.e., covariance matrices are *positive semidefinite* (and, trivially, symmetric).

- $\mathrm{Cov}[\boldsymbol{X}_1+\boldsymbol{X}_2] = \mathrm{Cov}[\boldsymbol{X}_1]+\mathrm{Cov}[\boldsymbol{X}_2]+\mathrm{Cov}[\boldsymbol{X}_1,\boldsymbol{X}_2]+\mathrm{Cov}[\boldsymbol{X}_2,\boldsymbol{X}_1]$

2) If Σ is a *positive definite matrix* (i.e., $\boldsymbol{a}^\top\Sigma\boldsymbol{a}>0$ for all $\boldsymbol{a}\in\mathbb{R}^d\backslash\{\boldsymbol{0}\}$), then Σ is invertible (since pos. def. $\Rightarrow \Sigma\boldsymbol{a}\neq\boldsymbol{0} \Rightarrow \operatorname{rank}\Sigma = d - \dim\ker\Sigma = d-\dim\{\boldsymbol{b}:\Sigma\boldsymbol{b}=\boldsymbol{0}\} = d$ (Rank-nullity Theorem)).

3) A symmetric, positive definite (positive semidefinite) Σ can be written as

$$\Sigma = AA^\top \tag{21}$$

for a lower triangular matrix A with $A_{jj}>0$ ($A_{jj}\ge 0$) for all j. L is known as *Cholesky factor* (also denoted by $\Sigma^{1/2}$) of the *Cholesky decomposition* (21).

- Properties of $\boldsymbol{X}$ can often be shown with the *characteristic function*

$$\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \mathbb{E}[\exp(i\boldsymbol{t}^\top \boldsymbol{X})], \quad \boldsymbol{t} \in \mathbb{R}^d.$$

$X_1, \dots, X_d$ are independent $\Leftrightarrow \phi_{\boldsymbol{X}}(\boldsymbol{t}) = \prod_{j=1}^d \phi_{X_j}(t_j)$ for all $\boldsymbol{t}$.

Proposition 6.1

A symmetric matrix Σ is a covariance matrix if and only if it is positive semidefinite.

Proof.

"$\Rightarrow$" As we have seen in (20), a covariance matrix Σ is positive semidefinite.

"$\Leftarrow$" Let Σ be positive semidefinite with Cholesky factor A. Let $\boldsymbol{X}$ be a random vector with $\operatorname{Cov} \boldsymbol{X} = I_d = \operatorname{diag}(1, \dots, 1)$ (e.g., $X_j \overset{\text{ind.}}{\sim} \mathrm{N}(0,1)$). Then $\operatorname{Cov}[A\boldsymbol{X}] = A \operatorname{Cov}[\boldsymbol{X}]A^\top = AA^\top = \Sigma$, i.e., Σ is a covariance matrix (namely that of $A\boldsymbol{X}$). $\square$

6.1.2 Standard estimators of covariance and correlation

- Assume $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n \sim H$ (daily/weekly/monthly/yearly risk-factor changes) to be serially uncorrelated (i.e., multivariate white noise) with $\boldsymbol{\mu} = \mathbb{E}\boldsymbol{X}_1$, $\Sigma = \operatorname{Cov}\boldsymbol{X}_1$, $P = \operatorname{Cor}\boldsymbol{X}_1$.
- Non-parametric method-of-moments-like estimators of $\boldsymbol{\mu}, \Sigma, P$ are

$$\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i$$ (*sample mean*; unbiased; `colMeans`)

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^\top$$ (*sample cov. mat.*; unbiased; `var`)

$$R = (R_{ij}) \text{ for } R_{ij} = \frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}}$$ (*sample cor. matrix*; unbiased; `cor`)

- $\bar{\boldsymbol{X}}$ and R are also MLEs; $\frac{n-1}{n}S$ is the MLE for Σ.

- Clearly, $\bar{\boldsymbol{X}}$ is unbiased. Since the $\boldsymbol{X}_i$'s are uncorrelated,

$$\mathrm{Cov}[\bar{\boldsymbol{X}}] = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Cov}[\boldsymbol{X}_i] = \frac{1}{n}\mathrm{Cov}[\boldsymbol{X}_1] = \frac{1}{n}\Sigma.$$

- S is unbiased since

$$\begin{aligned}\mathbb{E}S &= \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}[(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^\top] \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}[(\boldsymbol{X}_i - \boldsymbol{\mu})(\boldsymbol{X}_i - \boldsymbol{\mu})^\top - (\bar{\boldsymbol{X}} - \boldsymbol{\mu})(\bar{\boldsymbol{X}} - \boldsymbol{\mu})^\top] \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(\Sigma - \mathrm{Cov}\,\bar{\boldsymbol{X}}) \underset{\mathrm{Cov}[\bar{\boldsymbol{X}}]=\frac{\Sigma}{n}}{=} \frac{n}{n-1}(1-1/n)\Sigma = \Sigma.\end{aligned}$$

- Check that $S = \frac{1}{n-1}ABA^\top$ for $A = (\boldsymbol{X}_1, \dots, \boldsymbol{X}_n) \in \mathbb{R}^{d\times n}$ and $B = I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top \in \mathbb{R}^{n\times n}$ (where $\mathbf{1}_n = (1, \dots, 1) \in \mathbb{R}^n$). Since $\mathrm{rank}\,A \le \min\{n, d\}$, $\mathrm{rank}\,B = n - \dim\ker B = n - 1$, and $\mathrm{rank}(ABA^\top) \le \min\{\mathrm{rank}\,A, \mathrm{rank}\,B\}$, it follows that $\mathrm{rank}\,S \le \min\{d, n-1\}$. If $n \le d$,

S is not invertible. To obtain a positive definite (and thus invertible) estimator of Σ, see, e.g., `Matrix::nearPD()`.

- Further properties of $\bar{\boldsymbol{X}}, S, R$ depend on H.

6.1.3 The multivariate normal distribution

Definition 6.2 (Multivariate normal distribution)

$\boldsymbol{X} = (X_1, \dots, X_d)$ has a *multivariate normal* (or *Gaussian*) *distribution* if

$$\boldsymbol{X} \stackrel{\text{d}}{=} \boldsymbol{\mu} + A\boldsymbol{Z}, \tag{22}$$

where $\boldsymbol{Z} = (Z_1, \dots, Z_k)$, $Z_j \stackrel{\text{ind.}}{\sim} \mathrm{N}(0,1)$, $A \in \mathbb{R}^{d \times k}$, $\boldsymbol{\mu} \in \mathbb{R}^d$.

- $\mathbb{E}\boldsymbol{X} = \boldsymbol{\mu} + A\mathbb{E}\boldsymbol{Z} = \boldsymbol{\mu}$
- $\mathrm{Cov}[\boldsymbol{X}] = \mathrm{Cov}[\boldsymbol{\mu} + A\boldsymbol{Z}] = A\,\mathrm{Cov}[\boldsymbol{Z}]A^\top = AA^\top =: \Sigma$

Proposition 6.3 (Characteristic function)
Let $\boldsymbol{X}$ be as in (22) and $\Sigma = AA^\top$. Then the cf of $\boldsymbol{X}$ is

$$\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \mathbb{E}[\exp(i\boldsymbol{t}^\top \boldsymbol{X})] = \exp\Big(i\boldsymbol{t}^\top\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{t}^\top\Sigma\boldsymbol{t}\Big), \quad \boldsymbol{t} \in \mathbb{R}^d.$$

Proof.

- $Z_1 \sim \mathrm{N}(0,1)$ has density $\varphi(x) = \exp(-x^2/2)/\sqrt{2\pi}$ which satisfies

  i) $\varphi(x) = \varphi(-x)$;

  ii) $\varphi'(x) = -x\varphi(x)$.

  By Euler's Formula, the characteristic function $\phi_{Z_1}(t)$ of Z_1 is given by

  $$\phi_{Z_1}(t) = \int_{-\infty}^{\infty} (\cos(tx) + i\sin(tx))\varphi(x)\,dx \underset{\text{i)}}{=} \int_{-\infty}^{\infty} \cos(tx)\varphi(x)\,dx.$$

  Hence,

  $$\phi'_{Z_1}(t) = \int_{-\infty}^{\infty} \sin(tx)(-x)\varphi(x)\,dx \underset{\text{ii)}}{=} \int_{-\infty}^{\infty} \sin(tx)\varphi'(x)\,dx \underset{\text{by parts}}{=} -t\phi_{Z_1}(t).$$

We also know that $\phi_{Z_1}(0) = 1$. This initial value problem has the unique solution $\phi_{Z_1}(t) = \exp(-t^2/2)$.

- Now let $\tilde{\boldsymbol{t}}^\top = \boldsymbol{t}^\top A$. Then

$$\begin{aligned}
\phi_{\boldsymbol{X}}(\boldsymbol{t}) &= \mathbb{E}[\exp(i\boldsymbol{t}^\top(\boldsymbol{\mu} + A\boldsymbol{Z}))] = \exp(i\boldsymbol{t}^\top\boldsymbol{\mu})\mathbb{E}[\exp(i\tilde{\boldsymbol{t}}^\top\boldsymbol{Z})] \\
&\overset{\text{ind.}}{=} \exp(i\boldsymbol{t}^\top\boldsymbol{\mu})\prod_{j=1}^{d}\mathbb{E}[\exp(i(\tilde{t}_j Z_j))] = \exp\left(i\boldsymbol{t}^\top\boldsymbol{\mu} - \frac{1}{2}\sum_{j=1}^{d}\tilde{t}_j^2\right) \\
&= \exp\left(i\boldsymbol{t}^\top\boldsymbol{\mu} - \frac{1}{2}\tilde{\boldsymbol{t}}^\top\tilde{\boldsymbol{t}}\right) = \exp\left(i\boldsymbol{t}^\top\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{t}^\top AA^\top\boldsymbol{t}\right) \\
&= \exp\left(i\boldsymbol{t}^\top\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{t}^\top\Sigma\boldsymbol{t}\right) \qquad \square
\end{aligned}$$

- We see that the multivariate normal distribution is characterized by $\boldsymbol{\mu}$ and Σ, hence the notation $\boldsymbol{X} \sim \mathrm{N}_d(\boldsymbol{\mu}, \Sigma)$.
- $\mathrm{N}_d(\boldsymbol{\mu}, \Sigma)$ can be characterized by univariate normal distributions. To see this we first need the following theoretical result.

Theorem 6.4 (Cramér–Wold)
Let $\boldsymbol{X}, \boldsymbol{X}_n$, $n \in \mathbb{N}$, be random vectors. Then

$$\boldsymbol{X}_n \underset{n\uparrow\infty}{\overset{\mathsf{d}}{\to}} \boldsymbol{X} \quad \Longleftrightarrow \quad \boldsymbol{a}^\top \boldsymbol{X}_n \underset{n\uparrow\infty}{\overset{\mathsf{d}}{\to}} \boldsymbol{a}^\top \boldsymbol{X} \quad \forall\, \boldsymbol{a} \in \mathbb{R}^d$$

Proof.

"$\Rightarrow$" This follows directly from the Continuous Mapping Theorem with the continuous map being $g(\boldsymbol{x}) = \boldsymbol{a}^\top \boldsymbol{x}$.

"$\Leftarrow$" $\phi_{\boldsymbol{X}_n}(\boldsymbol{t}) = \mathbb{E}[\exp(i \cdot 1 \cdot \boldsymbol{t}^\top \boldsymbol{X}_n)] = \phi_{\boldsymbol{t}^\top \boldsymbol{X}_n}(1) \underset{n\uparrow\infty}{\to} \phi_{\boldsymbol{t}^\top \boldsymbol{X}}(1) = \phi_{\boldsymbol{X}}(\boldsymbol{t})$ for all $\boldsymbol{t}$. The result then follows by the Lévy Continuity Theorem. $\square$

Corollary 6.5
Let $\boldsymbol{X}, \boldsymbol{Y}$ be two random vectors. Then

$$\boldsymbol{X} \overset{\mathsf{d}}{=} \boldsymbol{Y} \quad \Longleftrightarrow \quad \boldsymbol{a}^\top \boldsymbol{X} \overset{\mathsf{d}}{=} \boldsymbol{a}^\top \boldsymbol{Y} \quad \forall\, \boldsymbol{a} \in \mathbb{R}^d.$$

Proposition 6.6 (Characterization of $\mathrm{N}_d(\boldsymbol{\mu}, \Sigma)$)

$\boldsymbol{X} \sim \mathrm{N}_d(\boldsymbol{\mu}, \Sigma) \iff \boldsymbol{a}^\top \boldsymbol{X} \sim \mathrm{N}(\boldsymbol{a}^\top \boldsymbol{\mu}, \boldsymbol{a}^\top \Sigma \boldsymbol{a}) \quad \forall\, \boldsymbol{a} \in \mathbb{R}^d.$

Proof.

"$\Rightarrow$"

$$\begin{aligned}\phi_{\boldsymbol{a}^\top \boldsymbol{X}}(t) &= \mathbb{E}[\exp(it\boldsymbol{a}^\top \boldsymbol{X})] \\ &= \phi_{\boldsymbol{X}}(t\boldsymbol{a}) = \exp\left(i(t\boldsymbol{a})^\top \boldsymbol{\mu} - \frac{1}{2}(t\boldsymbol{a})^\top \Sigma (t\boldsymbol{a})\right) \\ &= \exp\left(it(\boldsymbol{a}^\top \boldsymbol{\mu}) - \frac{1}{2}t^2(\boldsymbol{a}^\top \Sigma \boldsymbol{a})\right).\end{aligned}$$

Uniqueness of characteristic functions $\Rightarrow \boldsymbol{a}^\top \boldsymbol{X} \sim \mathrm{N}(\boldsymbol{a}^\top \boldsymbol{\mu}, \boldsymbol{a}^\top \Sigma \boldsymbol{a})$.

"$\Leftarrow$" Let $\boldsymbol{Y} \sim \mathrm{N}_d(\boldsymbol{\mu}, \Sigma)$. We have just seen that $\boldsymbol{a}^\top \boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{a}^\top \boldsymbol{\mu}, \boldsymbol{a}^\top \Sigma \boldsymbol{a})$ for all $\boldsymbol{a} \in \mathbb{R}^d$, so $\boldsymbol{a}^\top \boldsymbol{X} \stackrel{\mathrm{d}}{=} \boldsymbol{a}^\top \boldsymbol{Y}$ for all $\boldsymbol{a} \in \mathbb{R}^d$. By Corollary 6.5, $\boldsymbol{X} \stackrel{\mathrm{d}}{=} \boldsymbol{Y} \sim \mathrm{N}_d(\boldsymbol{\mu}, \Sigma)$. □

Consequences:

- Margins: $\boldsymbol{X} \sim \mathrm{N}_d(\boldsymbol{\mu}, \Sigma) \underset{\text{see copulas}}{\overset{\boldsymbol{a}=\boldsymbol{e}_j}{\substack{\Rightarrow \\ \not\Leftarrow}}} X_j \sim \mathrm{N}(\mu_j, \sigma_{jj}^2), \quad j \in \{1, \dots, d\}.$
- Sums: $\boldsymbol{X} \sim \mathrm{N}_d(\boldsymbol{\mu}, \Sigma) \overset{\boldsymbol{a}=\boldsymbol{1}}{\Rightarrow} \sum_{j=1}^d X_j \sim \mathrm{N}(\sum_{j=1}^d \mu_j,\ \sum_{i,j} \sigma_{ij}).$

Proposition 6.7 (Density)

Let $\boldsymbol{X} \sim \mathrm{N}_d(\boldsymbol{\mu}, \Sigma)$ with $d \le k$, $\operatorname{rank} A = d$ ($\Rightarrow$ Σ pos. definite, invertible). Then $\boldsymbol{X}$ has density

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}\sqrt{\det \Sigma}} \exp\Bigl(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\Bigr), \quad \boldsymbol{x} \in \mathbb{R}^d.$$

Proof. Let $\boldsymbol{X} \overset{\mathrm{d}}{=} \boldsymbol{\mu} + A\boldsymbol{Z}$ with $\operatorname{rank} A = d$, $\boldsymbol{Z} = (Z_1, \dots, Z_d)$, $Z_i \overset{\text{ind.}}{\sim} \mathrm{N}(0,1)$, $i \in \{1, \dots, d\}$. The density of $\boldsymbol{Z}$ is

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) = \prod_{j=1}^d f_{Z_j}(z_j) = \frac{1}{(2\pi)^{d/2}} \exp\Bigl(-\frac{1}{2}\boldsymbol{z}^\top \boldsymbol{z}\Bigr), \quad \boldsymbol{z} \in \mathbb{R}^d.$$

By the Density Transformation Theorem,

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = f_{T(\boldsymbol{Z})}(\boldsymbol{x}) = f_{\boldsymbol{Z}}(T^{-1}(\boldsymbol{x}))\left|\det \frac{d}{d\boldsymbol{x}}T^{-1}(\boldsymbol{x})\right|.$$

With $T(\boldsymbol{z}) = \boldsymbol{\mu} + A\boldsymbol{z}$, we have $T^{-1}(\boldsymbol{x}) = A^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$ and $\frac{d}{d\boldsymbol{x}}T^{-1}(\boldsymbol{x}) = A^{-1}$, and thus

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}(A^{-1}(\boldsymbol{x} - \boldsymbol{\mu}))^\top (A^{-1}(\boldsymbol{x} - \boldsymbol{\mu}))\right)|\det(A^{-1})|.$$

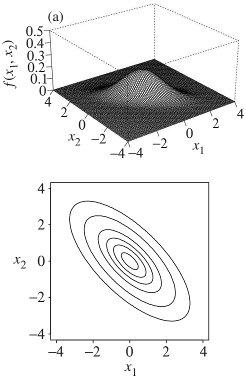
Now $(A^{-1})^\top A^{-1} = (A^\top)^{-1}A^{-1} = (AA^\top)^{-1} = \Sigma^{-1}$ and $\det(A^{-1}) = 1/\det(A) = 1/\sqrt{\det(A)\det(A^\top)} = 1/\sqrt{\det \Sigma}$ and thus the result follows.

$\square$

Consequences:

- Sets of the form $S_c = \{\boldsymbol{x} \in \mathbb{R}^d : (\boldsymbol{x} - \boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) = c\}$, $c > 0$, describe points of equal density. Contours of equal density are thus ellipsoids. Whenever a multivariate density $f_{\boldsymbol{X}}(\boldsymbol{x})$ depends on $\boldsymbol{x}$ only through the quadratic form $(\boldsymbol{x}-\boldsymbol{\mu})^\top\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$, it is the density of an elliptical distribution (see later).

- The components of $\boldsymbol{X} \sim \mathrm{N}_d(\boldsymbol{\mu}, \Sigma)$ are mutually independent if and only if Σ is diagonal, i.e., if and only if the components of $\boldsymbol{X}$ are uncorrelated.



Left: $\mathrm{N}_d\big(\boldsymbol{\mu} = \binom{0}{0}, \Sigma = \begin{pmatrix} 1 & -0.7 \\ -0.7 & 1 \end{pmatrix}\big)$; Right: $t_{\nu=4}(\boldsymbol{\mu}, \frac{\nu-2}{\nu}\Sigma)$ (same mean and covariance matrix as on the left-hand side)

The definition of $\mathrm{N}_d(\boldsymbol{\mu}, \Sigma)$ in terms of a stochastic representation ($\boldsymbol{X} \stackrel{\mathrm{d}}{=} \boldsymbol{\mu} + A\boldsymbol{Z}$) directly justifies the following sampling algorithm; see also `mvtnorm::rmvnorm(, method="chol")`.

Algorithm 6.8 (Sampling $\mathrm{N}_d(\boldsymbol{\mu}, \Sigma)$)

Let $\boldsymbol{X} \sim \mathrm{N}_d(\boldsymbol{\mu}, \Sigma)$ with Σ positive definite.

1) Compute the Cholesky factor A of Σ; see, e.g., Press et al. (1992).
2) Generate $Z_j \stackrel{\text{ind.}}{\sim} \mathrm{N}(0,1)$, $j \in \{1, \dots, d\}$ (R: done with inversion!).
3) Return $\boldsymbol{X} = \boldsymbol{\mu} + A\boldsymbol{Z}$, where $\boldsymbol{Z} = (Z_1, \dots, Z_d)$.

Further useful properties of multivariate normal distributions

- **Linear combinations**

  If $\boldsymbol{X} \sim \mathrm{N}_d(\boldsymbol{\mu}, \Sigma)$ and $B \in \mathbb{R}^{k \times d}, \boldsymbol{b} \in \mathbb{R}^k$, then

$$\begin{aligned} B\boldsymbol{X} + \boldsymbol{b} &= B(\boldsymbol{\mu} + A\boldsymbol{Z}) + \boldsymbol{b} = (B\boldsymbol{\mu} + \boldsymbol{b}) + BA\boldsymbol{Z} \\ &\sim \mathrm{N}_k(B\boldsymbol{\mu} + \boldsymbol{b}, BA(BA)^\top) = \mathrm{N}_k(B\boldsymbol{\mu} + \boldsymbol{b}, B\Sigma B^\top). \end{aligned}$$

Special case (see variance-covariance method; or Proposition 6.6): $\boldsymbol{b}^\top \boldsymbol{X} \sim \mathrm{N}(\boldsymbol{b}^\top \boldsymbol{\mu}, \boldsymbol{b}^\top \Sigma \boldsymbol{b})$

- **Marginal dfs**

  Let $\boldsymbol{X} \sim \mathrm{N}_d(\boldsymbol{\mu}, \Sigma)$ and write $\boldsymbol{X} = (\boldsymbol{X}_1^\top, \boldsymbol{X}_2^\top)^\top$, where $\boldsymbol{X}_1 \in \mathbb{R}^k$, $\boldsymbol{X}_2 \in \mathbb{R}^{d-k}$, and $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^\top, \boldsymbol{\mu}_2^\top)^\top$, $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$. Then

  $$\boldsymbol{X}_1 \sim \mathrm{N}_k(\boldsymbol{\mu}_1, \Sigma_{11}) \quad \text{and} \quad \boldsymbol{X}_2 \sim \mathrm{N}_{d-k}(\boldsymbol{\mu}_2, \Sigma_{22}).$$

  Proof. Choose $B = \begin{pmatrix} I_k & 0 \\ 0 & 0 \end{pmatrix}$ and $B = \begin{pmatrix} 0 & 0 \\ 0 & I_{d-k} \end{pmatrix}$, respectively.

- **Conditional distributions**

  Let $\boldsymbol{X}$ be as before and Σ be positive definite. Then

  $$\boldsymbol{X}_2 \,|\, \boldsymbol{X}_1 = \boldsymbol{x}_1 \sim \mathrm{N}_{d-k}(\boldsymbol{\mu}_{2.1}, \Sigma_{22.1}),$$

  where $\boldsymbol{\mu}_{2.1} = \boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1)$ and $\Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$.

  Proof. Via conditional densities or as follows: Consider $\boldsymbol{Z} = A\boldsymbol{X}_1 + \boldsymbol{X}_2$ with $A = -\Sigma_{21}\Sigma_{11}^{-1}$. Note that $(\boldsymbol{Z}, \boldsymbol{X}_1) = \begin{pmatrix} A & I_{d-k} \\ I_k & 0 \end{pmatrix} \boldsymbol{X}$ is jointly normal. Since $\boldsymbol{Z} = \boldsymbol{X}_2 + A\boldsymbol{X}_1$, we know that $(\boldsymbol{X}_2 \,|\, \boldsymbol{X}_1 = \boldsymbol{x}_1) \stackrel{\text{d}}{=} \boldsymbol{Z} - A\boldsymbol{x}_1$ is multivariate normal (since $\boldsymbol{Z}$ is). We have left to show that the formulas for $\boldsymbol{\mu}_{2.1}$ and $\Sigma_{22.1}$ hold. Since $A = -\Sigma_{21}\Sigma_{11}^{-1}$,

$$\operatorname{Cov}[\boldsymbol{Z}, \boldsymbol{X}_1] = \operatorname{Cov}[\boldsymbol{X}_2, \boldsymbol{X}_1] + A\operatorname{Cov}[\boldsymbol{X}_1] = \Sigma_{21} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{11} = 0,$$

hence $\boldsymbol{Z}$ and $\boldsymbol{X}_1$ are independent. Therefore

$$\begin{aligned}\boldsymbol{\mu}_{2.1} = \mathbb{E}[\boldsymbol{X}_2 \,|\, \boldsymbol{X}_1 = \boldsymbol{x}_1] &\underset{\text{ind.}}{=} \mathbb{E}[\boldsymbol{Z}] - \mathbb{E}[A\boldsymbol{X}_1 \,|\, \boldsymbol{X}_1 = \boldsymbol{x}_1] \\ &= \mathbb{E}\boldsymbol{X}_2 + A\mathbb{E}\boldsymbol{X}_1 - A\boldsymbol{x}_1 = \boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1),\end{aligned}$$

$$\begin{aligned}\Sigma_{22.1} &= \operatorname{Cov}[\boldsymbol{X}_2 \,|\, \boldsymbol{X}_1 = \boldsymbol{x}_1] = \operatorname{Cov}[\boldsymbol{Z} - A\boldsymbol{X}_1 \,|\, \boldsymbol{X}_1 = \boldsymbol{x}_1] \\ &\underset{\text{ind.}}{=} \operatorname{Cov}[\boldsymbol{Z} - A\boldsymbol{x}_1] = \operatorname{Cov}\boldsymbol{Z} = \operatorname{Cov}[\boldsymbol{X}_2 + A\boldsymbol{X}_1] \\ &= \operatorname{Cov}[\boldsymbol{X}_2] + A\operatorname{Cov}[\boldsymbol{X}_1]A^\top + \operatorname{Cov}[\boldsymbol{X}_2, \boldsymbol{X}_1]A^\top + A\operatorname{Cov}[\boldsymbol{X}_1, \boldsymbol{X}_2] \\ &= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{11}(-\Sigma_{21}\Sigma_{11}^{-1})^\top + \Sigma_{21}(-\Sigma_{21}\Sigma_{11}^{-1})^\top - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}.\end{aligned}$$

Noting that $(\Sigma_{21}\Sigma_{11}^{-1})^\top = (\Sigma_{11}^{-1})^\top\Sigma_{21}^\top = (\Sigma_{11}^\top)^{-1}\Sigma_{12} = \Sigma_{11}^{-1}\Sigma_{12}$, the form of $\Sigma_{22.1}$ easily follows. $\square$

- **Quadratic forms**

  Let $\boldsymbol{X} \sim \mathrm{N}_d(\boldsymbol{\mu}, \Sigma)$, Σ positive definite with Cholesky factor A. Furthermore, let $\boldsymbol{Z} = A^{-1}(\boldsymbol{X} - \boldsymbol{\mu})$. Then $\boldsymbol{Z} \sim N_d(\boldsymbol{0}, I_d)$. Moreover,

  $$(\boldsymbol{X} - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{X} - \boldsymbol{\mu}) = \boldsymbol{Z}^\top \boldsymbol{Z} \sim \chi_d^2, \tag{23}$$

  which is useful for (goodness-of-fit) testing of $\mathrm{N}_d(\boldsymbol{\mu}, \Sigma)$; see later.

  Proof. Clear via linearity and definition, and the definition of χ_d^2.

- **Convolutions**

  Let $\boldsymbol{X} \sim \mathrm{N}_d(\boldsymbol{\mu}, \Sigma)$ and $\boldsymbol{Y} \sim \mathrm{N}_d(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma})$ be independent. Then

  $$\boldsymbol{X} + \boldsymbol{Y} \sim \mathrm{N}_d(\boldsymbol{\mu} + \tilde{\boldsymbol{\mu}}, \Sigma + \tilde{\Sigma}).$$

  Proof. By independence, $\phi_{\boldsymbol{X}+\boldsymbol{Y}}(\boldsymbol{t})$ factors into

  $$= \phi_{\boldsymbol{X}}(\boldsymbol{t})\phi_{\boldsymbol{Y}}(\boldsymbol{t}) = \exp\Big(i\boldsymbol{t}^\top \boldsymbol{\mu} - \frac{1}{2}\boldsymbol{t}^\top \Sigma \boldsymbol{t}\Big) \exp\Big(i\boldsymbol{t}^\top \tilde{\boldsymbol{\mu}} - \frac{1}{2}\boldsymbol{t}^\top \tilde{\Sigma} \boldsymbol{t}\Big)$$
  $$= \exp\Big(i\boldsymbol{t}^\top (\boldsymbol{\mu} + \tilde{\boldsymbol{\mu}}) - \frac{1}{2}\boldsymbol{t}^\top (\Sigma + \tilde{\Sigma})\boldsymbol{t}\Big),$$

  which is the cf of $\mathrm{N}_d(\boldsymbol{\mu} + \tilde{\boldsymbol{\mu}}, \Sigma + \tilde{\Sigma})$. $\square$

Further properties:

1) Univariate t_ν distribution: $Z \sim \mathrm{N}(0,1)$, $W \sim \chi^2_\nu$ independent $\Rightarrow$ $X = Z/\sqrt{W/\nu} \sim t_\nu$. With $Y = \mu + \sigma X$, one has $\mathbb{E}Y = \mu$ if $\nu > 1$ and $\mathrm{Var}[Y] = \frac{\nu}{\nu-2}\sigma$ if $\nu > 2$.

   Generalization of χ^2_ν to $\nu > 0$: $\chi^2_\nu = \Gamma(\nu/2, 1/2)$ where $\Gamma(\alpha,\beta)$ has density $f(x) = \beta^\alpha x^{\alpha-1}e^{-\beta x}/\Gamma(\alpha)$ (β is the rate; see also R).

2) If $\boldsymbol{X}_1,\dots,\boldsymbol{X}_n \overset{\text{ind.}}{\sim} \mathrm{N}_d(\boldsymbol{\mu},\Sigma)$ with $\operatorname{rank}\Sigma = d$, then

$$\sum_{i=1}^n (\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^\top \sim \mathrm{W}_d(\Sigma, n-1) \quad \text{(\textit{Wishart distr.})} \tag{24}$$

   and $\bar{\boldsymbol{X}}$ and (24) are independent. For $X = (\boldsymbol{X}_1^\top,\dots,\boldsymbol{X}_n^\top)^\top$, one has $X^\top X = \sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^\top \sim \mathrm{W}_d(\Sigma, n)$; special case: $\mathrm{W}_1(1,n) = \chi^2_n$.

6.1.4 Testing multivariate normality

By Proposition 6.6,

$$\boldsymbol{X}_1,\dots,\boldsymbol{X}_n \overset{\text{ind.}}{\sim} \mathrm{N}_d(\boldsymbol{\mu},\Sigma) \;\Rightarrow\; \boldsymbol{a}^\top\boldsymbol{X}_1,\dots,\boldsymbol{a}^\top\boldsymbol{X}_n \overset{\text{ind.}}{\sim} \mathrm{N}(\boldsymbol{a}^\top\boldsymbol{\mu},\boldsymbol{a}^\top\Sigma\boldsymbol{a}).$$

This can be tested statistically (for some $\boldsymbol{a}$) with various goodness-of-fit tests (e.g., Q-Q plots) known for univariate normality (however, for $\boldsymbol{a}=\boldsymbol{e}_j$, $j\in\{1,\dots,d\}$, we would only test normality of the margins, not joint normality). Alternatively, (23) could be used to test joint normality.

Univariate tests

Formal statistical tests (see `fBasics::NormalityTests`)

- For general univariate df F:
  - Kolmogorov–Smirnov (`stats::ks.test()`)
  - Cramér–von Mises (for normal df: `nortest::cvm.test()`)

- Anderson–Darling (recommended by D'Agostino and Stephens (1986); `ADGofTest::ad.test()`)

- For $\mathrm{N}(\mu, \sigma^2)$:
  - D'Agostino (`fBasics::dagoTest()`, `moments::agostino.test()`)
  - Shapiro–Wilk (`stats::shapiro.test()`)
  - Jarque–Bera (`tseries::jarque.bera.test()`, `moments::jarque.test()`)

Graphical tests

Let $X_1, \dots, X_n$ be iid, $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{X_i \le x\}}$ the corresponding empirical distribution function (edf). Suppose we want to graphically test whether $X_1, \dots, X_n \sim F$ for some df F based on given realizations $x_1, \dots, x_n$. Let $x_{(1)} \le \dots \le x_{(n)}$ denote the corresponding ordered statistics. Possible options are:

- P-P plot: Plot $\{(p_i, F(x_{(i)})) : i = 1, \dots, n\}$, where $p_i :=$ `ppoints(n)[i]` $\approx \frac{i-1/2}{n}$.
- Q-Q plot: Plot $\{(F^-(p_i), x_{(i)}) : i = 1, \dots, n\}$ (differences in tails better visible).

Justification:

1) Glivenko–Cantelli: $\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)| \underset{n\uparrow\infty}{\overset{\text{a.s.}}{\to}} 0$

2) $\hat{F}_n(x) \underset{n\to\infty}{\to} F(x)\ \forall\, x \in C(F) \Leftrightarrow \hat{F}_n^-(u) \underset{n\to\infty}{\to} F^-(u)\ \forall\, u \in C(F^-)$; see van der Vaart (2000, Lemma 21.2)

By 1), the first (and thus the 2nd) part of 2) holds. Hence, for the true underlying F, $x_{(i)} = \hat{F}_n^-(i/n) \approx \hat{F}_n^-(p_i) \underset{2)}{\approx} F^-(p_i)$.

Interpretation: If F is (reasonably close to) the underlying unknown df, P-P and Q-Q plots resemble lines close to $y = x$ (possibly after standardization to mean 0 and variance 1).

Multivariate tests

Formal statistical tests

- Multivariate Shapiro–Wilk (`mvnormtest::mshapiro.test()`)
- Mardia's test (`dprep::mardia()`):
  - According to (23), if $\boldsymbol{X} \sim \mathrm{N}_d(\boldsymbol{\mu}, \Sigma)$ with Σ positive definite, then $(\boldsymbol{X} - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{X} - \boldsymbol{\mu}) \sim \chi^2_d$.
  - Let $D_i^2 = (\boldsymbol{X}_i - \bar{\boldsymbol{X}})^\top S^{-1} (\boldsymbol{X}_i - \bar{\boldsymbol{X}})$ denote the *squared Mahalanobis distances* and $D_{ij} = (\boldsymbol{X}_i - \bar{\boldsymbol{X}})^\top S^{-1} (\boldsymbol{X}_j - \bar{\boldsymbol{X}})$ the *Mahalanobis angles*.
  - Let $b_d = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n D_{ij}^3$ and $k_d = \frac{1}{n} \sum_{i=1}^n D_i^4$. Under the null hypothesis one can show that asymptotically for $n \to \infty$,

$$\tfrac{n}{6} b_d \sim \chi^2_{d(d+1)(d+2)/6}, \quad \frac{k_d - d(d+2)}{\sqrt{8d(d+2)/n}} \sim \mathrm{N}(0,1),$$

which can be used for testing; see Joenssen and Vogel (2014).

Graphical test

- Due to $\bar{\boldsymbol{X}}$ and S, the D_i^2's are not exactly following a χ_d^2 anymore. It turns out that $\frac{n}{(n-1)^2} D_i^2 \overset{H_0}{\sim} \mathrm{Beta}(d/2, (n-d-1)/2)$; see Gnanadesikan and Kettenring (1972). Check this with a Q-Q plot. For large n, the approximate χ_d^2 distribution is fine.

Example 6.9 (Multivariate (non-)normality of 10 Dow Jones stocks)

- We apply Mardia's test (of multivariate skewness and kurtosis) to daily/weekly/monthly/quarterly log-returns of 10 (of the 30) Dow Jones stocks from 1993–2000.
- We also compare D_i^2 data to a χ_{10}^2 using a Q-Q plot.

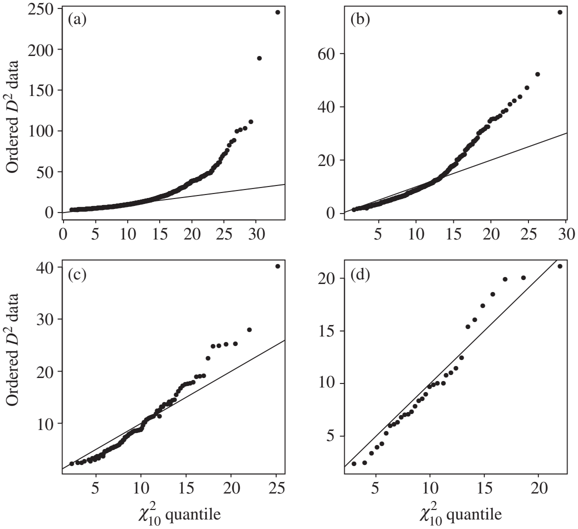
Mardia's (asymptotic) test based on the multivariate measures of skewness and kurtosis:

	Daily	Weekly	Monthly	Quarterly
n	2020	416	96	32
b_{10}	9.31	9.91	21.10	50.10
p-value	0.00	0.00	0.00	0.02
k_{10}	242.45	177.04	142.65	120.83
p-value	0.00	0.00	0.00	0.44

Conclusion: Daily/weekly/montly data: Evidence against joint normality

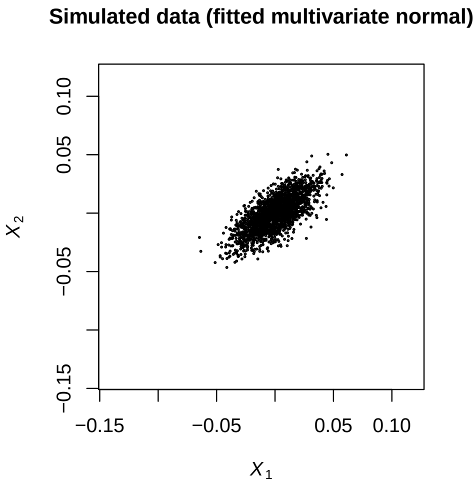
Quarterly data: CLT effect seems to take place (but too little data to say more); still evidence against joint normality.

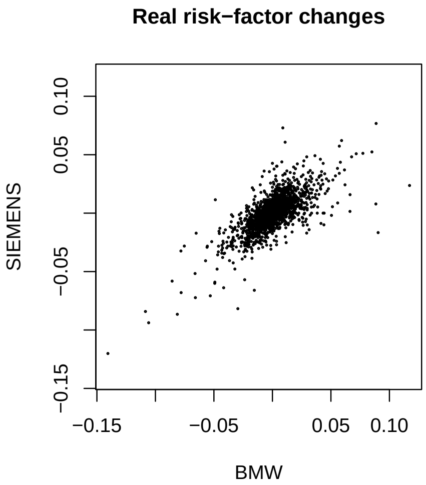
Q-Q plot of D_i^2 data against a χ^2_{10} distribution:
(a) daily data; (b) weekly data; (c) monthly data; and (d) quarterly data



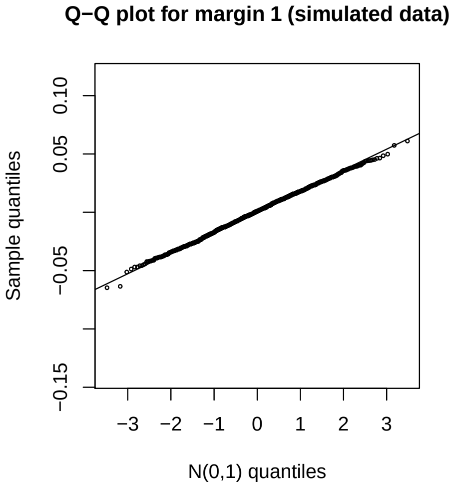
Example 6.10 (Simulated data vs BMW–Siemens)

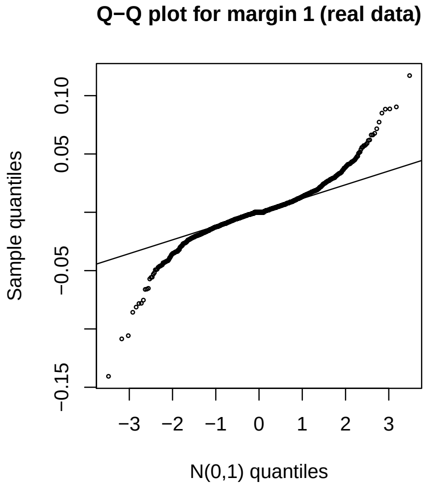
Is the BMW–Siemens data (see Section 3.2.2) jointly normal?



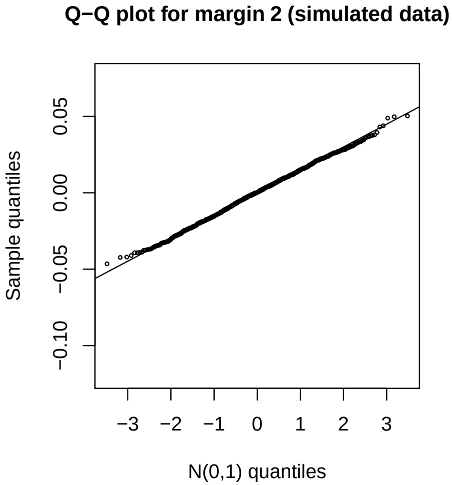


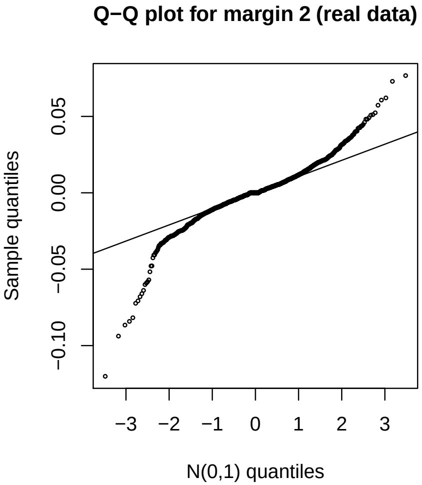
Considering the first margin only:



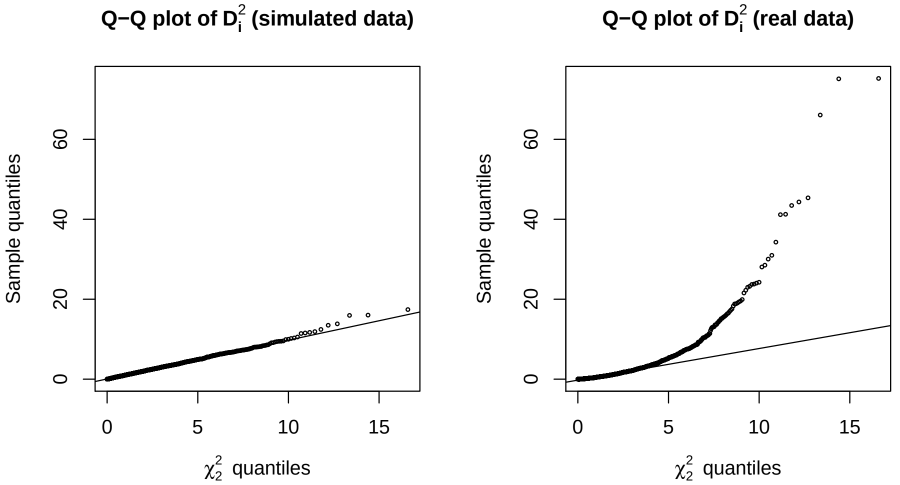


Considering the second margin only:





Q-Q plot of the simulated (left) or real (right) D_i^2's against a χ_2^2:



Advantages of $N_d(\boldsymbol{\mu}, \Sigma)$

- Inference "easy".
- Distribution is determined by $\boldsymbol{\mu}$ and Σ.
- Linear combinations are normal ($\Rightarrow$ VaR_α and ES_α calculations (for portfolios, for example) are easy).
- Marginal distributions are normal.
- Conditional distributions are normal.
- Quadratic forms are known.
- Convolutions are normal.
- Sampling is straightforward.
- Independence and uncorrelatedness are equivalent.

Drawbacks of $\mathrm{N}_d(\boldsymbol{\mu}, \Sigma)$ for modeling risk-factor changes

1) Tails of univariate (normal) margins are too thin (generate too few extreme events).

2) Joint tails are too thin (generate too few joint extreme events). $\mathrm{N}_d(\boldsymbol{\mu}, \Sigma)$ cannot capture the notion of tail dependence (see later).

3) Very strong symmetry known as radial symmetry: $\boldsymbol{X}$ is called *radially symmetric about* $\boldsymbol{\mu}$ if $\boldsymbol{X} - \boldsymbol{\mu} \overset{\mathrm{d}}{=} \boldsymbol{\mu} - \boldsymbol{X}$. For $\mathrm{N}_d(\boldsymbol{\mu}, \Sigma)$: $\boldsymbol{X} - \boldsymbol{\mu} \overset{\mathrm{d}}{=} A\boldsymbol{Z} \overset{\mathrm{d}}{=} A(-\boldsymbol{Z}) = -A\boldsymbol{Z} \overset{\mathrm{d}}{=} -(\boldsymbol{X} - \boldsymbol{\mu}) = \boldsymbol{\mu} - \boldsymbol{X}$.

In short:

- Elliptical distributions (a generalization of normal mixture distributions) can address 1) and 2) while sharing many of the desirable properties of $\mathrm{N}_d(\boldsymbol{\mu}, \Sigma)$.
- Normal mean-variance mixture distribution can also address 3) (but at the expense of tractability in comparison to $\mathrm{N}_d(\boldsymbol{\mu}, \Sigma)$).

6.2 Normal mixture distributions

Idea: Randomize Σ (and $\boldsymbol{\mu}$) with a non-negative rv W.

6.2.1 Normal variance mixtures

Definition 6.11 (Multivariate normal variance mixtures)

The random vector $\boldsymbol{X}$ has a (multivariate) *normal variance mixture distribution* if

$$\boldsymbol{X} \stackrel{\mathrm{d}}{=} \boldsymbol{\mu} + \sqrt{W} A \boldsymbol{Z}, \tag{25}$$

where $\boldsymbol{Z} \sim \mathrm{N}_k(\boldsymbol{0}, I_k)$, $W \geq 0$ is a rv independent of $\boldsymbol{Z}$, $A \in \mathbb{R}^{d \times k}$, and $\boldsymbol{\mu} \in \mathbb{R}^d$. $\boldsymbol{\mu}$ is called *location vector* and $\Sigma = AA^\top$ *scale* (or *dispersion*) *matrix*.

Observe that $(\boldsymbol{X} \mid W = w) \stackrel{\mathrm{d}}{=} \boldsymbol{\mu} + \sqrt{w} A \boldsymbol{Z} = \mathrm{N}_d(\boldsymbol{\mu}, wAA^\top) = \mathrm{N}_d(\boldsymbol{\mu}, w\Sigma)$; or $(\boldsymbol{X} \mid W) \stackrel{\mathrm{d}}{=} \mathrm{N}_d(\boldsymbol{\mu}, W\Sigma)$. W can be interpreted as a shock affecting the volatilities of all risk factors.

Properties of multivariate normal variance mixtures

Assume rank(A)$= d \le k$ and that Σ is positive definite. Let $\boldsymbol{Y} = \boldsymbol{\mu} + A\boldsymbol{Z}$.

- If $\mathbb{E}\sqrt{W} < \infty$, then $\mathbb{E}[\boldsymbol{X}] \overset{\text{ind.}}{=} \boldsymbol{\mu} + \mathbb{E}[\sqrt{W}]A\mathbb{E}[\boldsymbol{Z}] = \boldsymbol{\mu} + \boldsymbol{0} = \boldsymbol{\mu}$ $(= \mathbb{E}\boldsymbol{Y})$
- If $\mathbb{E}W < \infty$, then

$$\begin{aligned}\mathrm{Cov}[\boldsymbol{X}] &= \mathrm{Cov}[\sqrt{W}A\boldsymbol{Z}] = \mathbb{E}[(\sqrt{W}A\boldsymbol{Z})(\sqrt{W}A\boldsymbol{Z})^\top] \\ &\overset{\text{ind.}}{=} \mathbb{E}[W]\cdot\mathbb{E}[A\boldsymbol{Z}\boldsymbol{Z}^\top A^\top] = \mathbb{E}[W]\cdot A\mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^\top]A^\top \\ &= \mathbb{E}[W]AI_kA^\top = \mathbb{E}[W]\Sigma \underset{\text{in general}}{\neq} \Sigma \quad (= \mathrm{Cov}[\boldsymbol{Y}])\end{aligned}$$

- However, if they exist (i.e., if $\mathbb{E}W < \infty$), $\mathrm{Cor}[\boldsymbol{X}]$ and $\mathrm{Cor}[\boldsymbol{Y}]$ are equal:

  Proof. $\mathrm{Cov}[\boldsymbol{X}] = \mathbb{E}[W]\Sigma \Rightarrow \mathrm{Cov}[X_i, X_j] = \mathbb{E}[W]\Sigma_{ij}$ and $\mathrm{Var}[X_i] = \mathbb{E}[W]\Sigma_{ii}$. This implies that

$$\mathrm{Cor}[X_i, X_j] = \frac{\mathrm{Cov}[X_i, X_j]}{\sqrt{\mathrm{Var}[X_i]\,\mathrm{Var}[X_j]}} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}} = \mathrm{Cor}[Y_i, Y_j]. \qquad \square$$

Lemma 6.12 (Independence in normal variance mixtures)

Let $\boldsymbol{X} = \boldsymbol{\mu} + \sqrt{W} A \boldsymbol{Z}$ with $\mathbb{E}W < \infty$ and $A = I_d$ ($\Rightarrow \operatorname{Cov}[\boldsymbol{X}] = \mathbb{E}[W] \operatorname{Cov}[\boldsymbol{Z}] = \mathbb{E}[W] I_d$ (uncorrelated)). Then

$$X_i \text{ and } X_j \text{ are independent} \iff W \text{ is a.s. constant (i.e., } \boldsymbol{X} \sim \mathrm{N}_d).$$

Proof. W.l.o.g. assume $\boldsymbol{\mu} = \mathbf{0}$.

"$\Rightarrow$" $\mathbb{E}|X_i|\,\mathbb{E}|X_j| \overset{\text{ind.}}{=} \mathbb{E}[|X_i||X_j|] = \mathbb{E}[W|Z_i||Z_j|] \overset{\text{ind.}}{=} \mathbb{E}[W]\,\mathbb{E}|Z_i|\,\mathbb{E}|Z_j|$

$$\underset{\text{Jensen}}{\geq} \mathbb{E}[\sqrt{W}]^2\,\mathbb{E}|Z_i|\,\mathbb{E}|Z_j| \overset{\text{ind.}}{=} \mathbb{E}|\sqrt{W} Z_i|\,\mathbb{E}|\sqrt{W} Z_j| = \mathbb{E}|X_i|\mathbb{E}|X_j|$$

$\Rightarrow$ We must have "$=$" in Jensen's inequality. This hols if and only if W is constant a.s.; so $\boldsymbol{X} \sim \mathrm{N}_d(\mathbf{0}, W I_d)$ in this case.

"$\Leftarrow$" W a.s. constant $\Rightarrow \boldsymbol{X} \sim \mathrm{N}_d(\mathbf{0}, W I_d) \Rightarrow X_i, X_j$ independent. $\square$

Recall: If $\boldsymbol{X} \sim \mathrm{N}_d(\boldsymbol{\mu}, \Sigma)$, then $\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \exp(i\boldsymbol{t}^\top\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{t}^\top\Sigma\boldsymbol{t})$.

Furthermore, $\boldsymbol{X} \,|\, W = w \sim \mathrm{N}_d(\boldsymbol{\mu}, w\Sigma)$ (or: $\boldsymbol{X} \,|\, W \sim \mathrm{N}_d(\boldsymbol{\mu}, W\Sigma)$)

- **Characteristic function:** The cf of a multivariate normal variance mixtures is

$$\begin{aligned}\phi_{\boldsymbol{X}}(\boldsymbol{t}) &= \mathbb{E}[\exp(i\boldsymbol{t}^\top\boldsymbol{X})] = \mathbb{E}[\,\mathbb{E}[\exp(i\boldsymbol{t}^\top\boldsymbol{X}) \,|\, W]\,] \\ &= \mathbb{E}[\exp(i\boldsymbol{t}^\top\boldsymbol{\mu} - \tfrac{1}{2}W\boldsymbol{t}^\top\Sigma\boldsymbol{t})] = \exp(i\boldsymbol{t}^\top\boldsymbol{\mu})\mathbb{E}[\exp(-W\tfrac{1}{2}\boldsymbol{t}^\top\Sigma\boldsymbol{t})].\end{aligned}$$

- **LS transform:** The *Laplace-Stieltjes transform* of F_W is

$$\hat{F}_W(\theta) := \mathcal{LS}[F_W](\theta) := \mathbb{E}[\exp(-\theta W)] = \int_0^\infty e^{-\theta w}\,dF_W(w).$$

Therefore, $\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \exp(i\boldsymbol{t}^\top\boldsymbol{\mu})\hat{F}_W(\frac{1}{2}\boldsymbol{t}^\top\Sigma\boldsymbol{t})$. We thus introduce the notation $\boldsymbol{X} \sim M_d(\boldsymbol{\mu}, \Sigma, \hat{F}_W)$ for a d-dimensional multivariate normal variance mixture.

- **Density:** If Σ is positive definite, $\mathbb{P}(W=0)=0$, the density of $\boldsymbol{X}$ is

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \int_0^\infty f_{\boldsymbol{X}|W}(\boldsymbol{x}\,|\,w)\,dF_W(w)$$
$$= \int_0^\infty \frac{1}{(2\pi)^{d/2}w^{d/2}|\Sigma|^{1/2}}\exp\Bigl(-\frac{(\boldsymbol{x}-\boldsymbol{\mu})^\top\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2w}\Bigr)\,dF_W(w).$$

$\Rightarrow$ Only depends on $\boldsymbol{x}$ through $(\boldsymbol{x}-\boldsymbol{\mu})^\top\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$.

$\Rightarrow$ Multivariate normal variance mixtures are elliptical distributions.

If Σ is diagonal and $\mathbb{E}W<\infty$, $\boldsymbol{X}$ is uncorrelated (as $\mathrm{Cov}[\boldsymbol{X}]=\mathbb{E}[W]\Sigma$) but not independent unless W is constant ($\sqrt{W}$ creates dependence).

- **Linear combinations:** For $\boldsymbol{X}\sim M_d(\boldsymbol{\mu},\Sigma,\hat{F}_W)$ and $\boldsymbol{Y}=B\boldsymbol{X}+\boldsymbol{b}$, where $B\in\mathbb{R}^{k\times d}$ and $\boldsymbol{b}\in\mathbb{R}^k$, we have $\boldsymbol{Y}\sim M_k(B\boldsymbol{\mu}+\boldsymbol{b},B\Sigma B^\top,\hat{F}_W)$.

Proof. Recall that $\phi_{\boldsymbol{X}}(\boldsymbol{t})=\exp(i\boldsymbol{t}^\top\boldsymbol{\mu})\hat{F}_W(\frac{1}{2}\boldsymbol{t}^\top\Sigma\boldsymbol{t})$. Thus,

$\phi_{\boldsymbol{Y}}(\boldsymbol{t})=\mathbb{E}[\exp(i\boldsymbol{t}^\top(B\boldsymbol{X}+\boldsymbol{b}))]=\exp(i\boldsymbol{t}^\top\boldsymbol{b})\cdot\mathbb{E}[\exp(i(B^\top\boldsymbol{t})^\top\boldsymbol{X})]$

$=\exp(i\boldsymbol{t}^\top\boldsymbol{b})\,\phi_{\boldsymbol{X}}(B^\top\boldsymbol{t})=\exp(i\boldsymbol{t}^\top(\boldsymbol{b}+B\boldsymbol{\mu}))\,\hat{F}_W(\frac{1}{2}\boldsymbol{t}^\top B\Sigma B^\top\boldsymbol{t}).$ $\square$

If $\boldsymbol{a}\in\mathbb{R}^d$ ($\boldsymbol{b}=\boldsymbol{0}$, $B=\boldsymbol{a}^\top\in\mathbb{R}^{1\times d}$), $\boldsymbol{a}^\top\boldsymbol{X}\sim M_1(\boldsymbol{a}^\top\boldsymbol{\mu},\boldsymbol{a}^\top\Sigma\boldsymbol{a},\hat{F}_W)$.

- **Sampling:**

Algorithm 6.13 (Simulation of $\boldsymbol{X} = \boldsymbol{\mu} + \sqrt{W} A \boldsymbol{Z} \sim M_d(\boldsymbol{\mu}, \Sigma, \hat{F}_W)$)

1) Generate $\boldsymbol{Z} \sim N_d(\mathbf{0}, I_d)$.
2) Generate $W \sim F_W$ (with LS transform $\hat{F}_W$), independent of $\boldsymbol{Z}$.
3) Compute the Cholesky factor A (such that $AA^\top = \Sigma$).
4) Return $\boldsymbol{X} = \boldsymbol{\mu} + \sqrt{W} A \boldsymbol{Z}$.

Example 6.14 ($t_d(\nu, \boldsymbol{\mu}, \Sigma)$ distribution)

1) Generate $\boldsymbol{Z} \sim N_d(\mathbf{0}, I_d)$.
2) Generate $V \sim \chi^2_\nu$ and set $W = \frac{\nu}{V} \sim \mathrm{Ig}(\nu/2, \nu/2)$.
   Alternatively, $W = 1/V$ with $V \sim \Gamma(\nu/2, \text{ rate} = \nu/2)$.
3) Compute the Cholesky factor A (such that $AA^\top = \Sigma$).
4) Return $\boldsymbol{X} = \boldsymbol{\mu} + \sqrt{W} A \boldsymbol{Z}$.

Examples of multivariate normal variance mixtures

- **Multivariate normal distribution**
  $W = 1$ a.s. (degenerate case)

- **Two point mixture**

$$W = \begin{cases} w_1 \text{ with probability } p, \\ w_2 \text{ with probability } 1-p \end{cases} \qquad w_1,\ w_2 > 0,\ w_1 \neq w_2.$$

  Can be used to model ordinary and stress regimes; extends to k regimes.

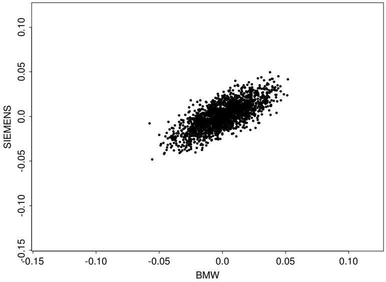
- **Symmetric generalised hyperbolic distribution**
  W has a generalised inverse Gaussian distribution (GIG); see McNeil et al. (2015, p. 187)

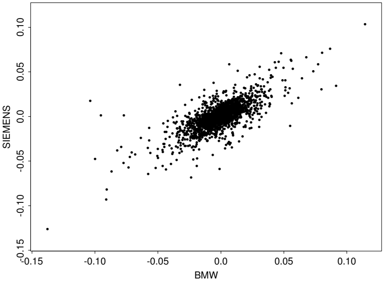
- **Multivariate t distribution**
  W has an inverse gamma distribution $W = 1/V$ for $V \sim \Gamma(\nu/2, \nu/2)$.
  - $\mathbb{E}[W] = \frac{\nu}{\nu-2} \Rightarrow \mathrm{Cov}\,[\boldsymbol{X}] = \frac{\nu}{\nu-2}\Sigma$. For finite variances/correlations, $\nu > 2$ is required. For finite mean, $\nu > 1$ is required.

- The (elliptical) density of the multivariate t distribution is given by

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{\Gamma((\nu+d)/2)}{\Gamma(\nu/2)(\nu\pi)^{d/2}|\Sigma|^{1/2}}\left(1+\frac{(\boldsymbol{x}-\boldsymbol{\mu})^\top\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{\nu}\right)^{-\frac{\nu+d}{2}},$$

  where $\boldsymbol{\mu} \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d\times d}$ is a positive definite matrix, and ν is the degrees of freedom. Notation: $\boldsymbol{X} \sim t_d(\nu, \boldsymbol{\mu}, \Sigma)$.
- $t_d(\nu, \boldsymbol{\mu}, \Sigma)$ has heavier marginal and joint tails than $\mathrm{N}_d(\boldsymbol{\mu}, \Sigma)$.
- BMW–Siemens data: Simulations from fitted $\mathrm{N}_d(\boldsymbol{\mu}, \Sigma)$ and $t_d(3, \boldsymbol{\mu}, \Sigma)$:





6.2.2 Normal mean-variance mixtures

- Radial symmetry implies that all one-dimensional margins of normal variance mixtures are symmetric.
- Often visible in data: Joint losses have heavier tails than joint gains.

Idea: Introduce asymmetry by mixing normal distributions with different means and variances.

$\boldsymbol{X}$ has a (multivariate) *normal mean-variance mixture distribution* if

$$\boldsymbol{X} \stackrel{\mathrm{d}}{=} \boldsymbol{m}(W) + \sqrt{W} A \boldsymbol{Z}, \tag{26}$$

where

- $\boldsymbol{Z} \sim \mathrm{N}_k(\mathbf{0}, I_k)$;
- $W \geq 0$ is a scalar random variable which is independent of $\boldsymbol{Z}$;
- $A \in \mathbb{R}^{d \times k}$ is a matrix of constants;
- $\boldsymbol{m} : [0, \infty) \to \mathbb{R}^d$ is a measurable function.

- Normal mean-variance mixtures add radial asymmetry: Let $\Sigma = AA^\top$ and observe that $\boldsymbol{X} \mid W = w \sim N_d(\boldsymbol{m}(w), w\Sigma)$. In general, they are no longer elliptical and $\mathrm{Cor}(\boldsymbol{X}) \neq \mathrm{Cor}(\boldsymbol{Y})$ (where $\boldsymbol{Y} = \boldsymbol{\mu} + A\boldsymbol{Z}$)

Example 6.15 (Generalized hyperbolic distribution)

- Here, $\boldsymbol{m}(W) = \boldsymbol{\mu} + W\boldsymbol{\gamma}$. Since

$$\mathbb{E}[\boldsymbol{X} \mid W] = \boldsymbol{\mu} + W\boldsymbol{\gamma},$$
$$\mathrm{Cov}[\boldsymbol{X} \mid W] = W\Sigma$$

  one has

$$\begin{aligned} \mathbb{E}\boldsymbol{X} &= \mathbb{E}[\mathbb{E}[\boldsymbol{X} \mid W]] = \boldsymbol{\mu} + \mathbb{E}[W]\boldsymbol{\gamma} \quad \text{if } \mathbb{E}W < \infty, \\ \mathrm{Cov}[\boldsymbol{X}] &= \mathbb{E}[\mathrm{Cov}[\boldsymbol{X} \mid W]] + \mathrm{Cov}[\mathbb{E}[\boldsymbol{X} \mid W]] \\ &= \mathbb{E}[W]\Sigma + \mathrm{Var}[W]\boldsymbol{\gamma}\boldsymbol{\gamma}^\top \quad \text{if } \mathbb{E}[W^2] < \infty. \end{aligned}$$

- If W has a GIG distribution, then $\boldsymbol{X}$ follows a *generalised hyperbolic distribution*. $\boldsymbol{\gamma} = \boldsymbol{0}$ leads to (elliptical) normal variance mixtures; see McNeil et al. (2015, Sections 6.2.3) for details.

6.3 Spherical and elliptical distributions

Empirical examples (see McNeil et al. (2015, Sections 6.2.4)) show that

1) $M_d(\boldsymbol{\mu}, \Sigma, \hat{F}_W)$ (e.g., multivariate t, NIG) provide superior models to $\mathrm{N}_d(\boldsymbol{\mu}, \Sigma)$ for daily/weekly US stock-return data;
2) the more general radially asymmetric normal mean-variance mixture distributions did not seem to offer much of an improvement.

We soon study elliptical distributions, a generalization of $M_d(\boldsymbol{\mu}, \Sigma, \hat{F}_W)$.

6.3.1 Spherical distributions

Definition 6.16 (Spherical distribution)

A random vector $\boldsymbol{Y} = (Y_1, \dots, Y_d)$ has a *spherical distribution* if for every orthogonal $U \in \mathbb{R}^{d\times d}$ (i.e., $U \in \mathbb{R}^{d\times d}$ with $UU^\top = U^\top U = I_d$)

$$\boldsymbol{Y} \stackrel{\mathrm{d}}{=} U\boldsymbol{Y} \quad \text{(distributionally invariant under rotations and reflections)}$$

Theorem 6.17 (Characterization of spherical distributions)

Let $\|\boldsymbol{t}\| = (t_1^2 + \cdots + t_d^2)^{1/2}$, $\boldsymbol{t} \in \mathbb{R}^d$. The following are equivalent:

1) $\boldsymbol{Y}$ is spherical.
2) $\exists$ a *characteristic generator* $\psi : [0, \infty) \to \mathbb{R}$, such that $\phi_{\boldsymbol{Y}}(\boldsymbol{t}) = \mathbb{E}[e^{i\boldsymbol{t}^\top \boldsymbol{Y}}] = \psi(\|\boldsymbol{t}\|^2)$, $\forall\, \boldsymbol{t} \in \mathbb{R}^d$.
3) For every $\boldsymbol{a} \in \mathbb{R}^d$, $\boldsymbol{a}^\top \boldsymbol{Y} \overset{\mathrm{d}}{=} \|\boldsymbol{a}\| Y_1$ (linear combinations are of the same type $\underset{\text{see later}}{\Rightarrow}$ subadditivity of VaR_α for elliptically distr. losses).

Proof. 1) $\Rightarrow$ 2): $\phi_{\boldsymbol{Y}}(\boldsymbol{t}) = \phi_{U\boldsymbol{Y}}(\boldsymbol{t}) = \phi_{\boldsymbol{Y}}(U^\top \boldsymbol{t})$ for all $U \in \mathbb{R}^{d\times d}$ orthogonal. Since U can only change the direction of $\boldsymbol{t}$ but not its length, $\phi_{\boldsymbol{Y}}(\boldsymbol{t})$ only depends on $\|\boldsymbol{t}\|$, i.e., the length of $\boldsymbol{t}$ $\Rightarrow$ we can define $\psi(\|\boldsymbol{t}\|^2) = \phi_{\boldsymbol{Y}}(\boldsymbol{t})$.

2) $\Rightarrow$ 3): $\phi_{Y_1}(t) = \phi_{\boldsymbol{Y}}(t\boldsymbol{e}_1) \underset{2)}{=} \psi(t^2)$ $(*)$. Now $\phi_{\boldsymbol{a}^\top \boldsymbol{Y}}(t) = \phi_{\boldsymbol{Y}}(t\boldsymbol{a}) \underset{2)}{=} \psi(t^2\|\boldsymbol{a}\|^2) = \psi((t\|\boldsymbol{a}\|)^2) \underset{(*)}{=} \phi_{Y_1}(t\|\boldsymbol{a}\|) = \phi_{\|\boldsymbol{a}\| Y_1}(t)$

3) $\Rightarrow$ 1): $\phi_{U\boldsymbol{Y}}(\boldsymbol{t}) = \mathbb{E}[\exp(i(U^\top \boldsymbol{t})^\top \boldsymbol{Y})] \underset{U^\top \boldsymbol{t}=:\boldsymbol{a}}{=} \mathbb{E}[\exp(i\boldsymbol{a}^\top \boldsymbol{Y})] \underset{3)}{=} \mathbb{E}[\exp(i\|\boldsymbol{a}\|Y_1)]$

$= \mathbb{E}[\exp(i\|\boldsymbol{t}\|Y_1)] \underset{3)}{=} \mathbb{E}[\exp(i\boldsymbol{t}^\top \boldsymbol{Y})] = \phi_{\boldsymbol{Y}}(\boldsymbol{t})$ □

Due to the above characterizations, we introduce the notation $\boldsymbol{Y} \sim S_d(\psi)$.

Theorem 6.18 (Stochastic representation)

$\boldsymbol{Y} \sim S_d(\psi)$ if and only if

$$\boldsymbol{Y} \overset{\text{d}}{=} R\boldsymbol{S}, \tag{27}$$

for independent *radial part* $R \geq 0$ and $\boldsymbol{S} \sim \mathrm{U}(\{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| = 1\})$.

Proof. Let Ω_d be the characteristic generator of $\boldsymbol{S}$.

"$\Rightarrow$" $\boldsymbol{Y} \sim S_d(\psi) \Rightarrow \phi_{\boldsymbol{Y}}(\|\boldsymbol{t}\|\boldsymbol{u}) \underset{2)}{=} \psi(\|\boldsymbol{t}\|^2\boldsymbol{u}^\top\boldsymbol{u}) = \psi(\|\boldsymbol{t}\|^2)$ for all $\boldsymbol{u} \in \mathbb{R}^d$: $\|\boldsymbol{u}\| = 1$. Replacing $\boldsymbol{u}$ by $\boldsymbol{S}$ and integrating leads to $\psi(\|\boldsymbol{t}\|^2) = \mathbb{E}_{\boldsymbol{S}}[\phi_{\boldsymbol{Y}}(\|\boldsymbol{t}\|\boldsymbol{S})] = \mathbb{E}_{\boldsymbol{S}}[\mathbb{E}_{\boldsymbol{Y}}[e^{i\|\boldsymbol{t}\|\boldsymbol{S}^\top\boldsymbol{Y}}]] \underset{\text{Fubini}}{=} \mathbb{E}_{\boldsymbol{Y}}[\mathbb{E}_{\boldsymbol{S}}[e^{i\|\boldsymbol{t}\|\boldsymbol{S}^\top\boldsymbol{Y}}]] = \mathbb{E}_{\boldsymbol{Y}}[\phi_{\boldsymbol{S}}(\|\boldsymbol{t}\|\boldsymbol{Y})] \underset{2)}{=} \mathbb{E}_{\boldsymbol{Y}}[\Omega_d(\|\boldsymbol{t}\|^2\boldsymbol{Y}^\top\boldsymbol{Y})]$. We thus obtain that

$$\phi_{\boldsymbol{Y}}(\boldsymbol{t}) \underset{2)}{=} \psi(\|\boldsymbol{t}\|^2) \underset{R:=\|\boldsymbol{Y}\|}{=} \mathbb{E}_R[\Omega_d(\|\boldsymbol{t}\|^2 R^2)] = \int_0^\infty \Omega_d(\|\boldsymbol{t}\|^2 r^2)\, dF_R(r)$$
$$\underset{2)}{=} \int_0^\infty \phi_{\boldsymbol{S}}(r\boldsymbol{t})\, dF_R(r) = \phi_{R\boldsymbol{S}}(\boldsymbol{t}) \text{ for all } \boldsymbol{t} \in \mathbb{R}^d.$$

"$\Leftarrow$" Let $\boldsymbol{Z} \sim \mathrm{N}_d(\boldsymbol{0}, I_d)$. Since $\boldsymbol{Z}$ is spherical and $\|\boldsymbol{Z}/\|\boldsymbol{Z}\|\| = \|\boldsymbol{Z}\|/\|\boldsymbol{Z}\| = 1$, $\boldsymbol{S} \overset{\mathrm{d}}{=} \boldsymbol{Z}/\|\boldsymbol{Z}\|$. As such, $\boldsymbol{S}$ itself is spherical, since $U\boldsymbol{S} \overset{\mathrm{d}}{=} U\boldsymbol{Z}/\|\boldsymbol{Z}\| \overset{\mathrm{d}}{=} \boldsymbol{Z}/\|\boldsymbol{Z}\| \overset{\mathrm{d}}{=} \boldsymbol{S}$ for any orthogonal $U \in \mathbb{R}^{d\times d}$. Theorem 6.17 Part 2) implies that $\phi_{\boldsymbol{S}}(\boldsymbol{t}) = \Omega_d(\|\boldsymbol{t}\|^2)$, so $\phi_{R\boldsymbol{S}}(\boldsymbol{t}) = \mathbb{E}[\exp(i\boldsymbol{t}^\top R\boldsymbol{S})] = \mathbb{E}_R[\,\mathbb{E}[\exp(i\boldsymbol{t}^\top R\boldsymbol{S}) \mid R]\,] = \mathbb{E}_R[\phi_{\boldsymbol{S}}(R\boldsymbol{t})] = \mathbb{E}_R[\Omega_d(R^2\|\boldsymbol{t}\|^2)]$, which is a function in $\|\boldsymbol{t}\|^2$ and thus, by 2), $R\boldsymbol{S}$ is spherical. □

Corollary 6.19

If $\boldsymbol{Y} \sim S_d(\psi)$ and $\mathbb{P}(\boldsymbol{Y} = \boldsymbol{0}) = 0$, then $(\|\boldsymbol{Y}\|, \frac{\boldsymbol{Y}}{\|\boldsymbol{Y}\|}) \overset{\mathrm{d}}{=} (R, \boldsymbol{S})$ since

$$(\|\boldsymbol{Y}\|, \tfrac{\boldsymbol{Y}}{\|\boldsymbol{Y}\|}) \overset{\mathrm{d}}{=} (\|R\boldsymbol{S}\|, \tfrac{R\boldsymbol{S}}{\|R\boldsymbol{S}\|}) = (|R|\|\boldsymbol{S}\|, \tfrac{\boldsymbol{S}}{\|\boldsymbol{S}\|}) = (R, \boldsymbol{S}).$$

In particular, $\|\boldsymbol{Y}\|$ and $\boldsymbol{Y}/\|\boldsymbol{Y}\|$ are independent ($\Rightarrow$ goodness-of-fit).

- If $\boldsymbol{Y} \sim S_d(\psi)$ and admits a density $f_{\boldsymbol{Y}}$, then the *inversion formula* $f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} e^{-i\boldsymbol{t}^\top \boldsymbol{y}}\phi_{\boldsymbol{Y}}(\boldsymbol{t})\,d\boldsymbol{t}$ and Theorem 6.17 Part 2) show that for any orthogonal U,

$$\begin{aligned}
f_{\boldsymbol{Y}}(U\boldsymbol{y}) &\underset{\text{inv.}}{=} \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} e^{-i(U^\top \boldsymbol{t})^\top \boldsymbol{y}}\phi_{\boldsymbol{Y}}(\boldsymbol{t})\,d\boldsymbol{t}\\
&\underset{\text{subs.}}{=} \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} e^{-i\boldsymbol{s}^\top \boldsymbol{y}}\phi_{\boldsymbol{Y}}(U\boldsymbol{s})\,d\boldsymbol{s}\\
&\underset{2)}{=} \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} e^{-i\boldsymbol{s}^\top \boldsymbol{y}}\psi((U\boldsymbol{s})^\top U\boldsymbol{s})\,d\boldsymbol{s}\\
&= \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} e^{-i\boldsymbol{s}^\top \boldsymbol{y}}\psi(\boldsymbol{s}^\top \boldsymbol{s})\,d\boldsymbol{s} \underset{\text{backwards}}{=} \cdots = f_{\boldsymbol{Y}}(\boldsymbol{y}).
\end{aligned}$$

  This implies that $f_{\boldsymbol{Y}}(\boldsymbol{y}) = g(\|\boldsymbol{y}\|^2)$ for a function $g : [0,\infty) \to [0,\infty)$ referred to as *density generator*. So $f_{\boldsymbol{Y}}(\boldsymbol{y})$ is constant on hyperspheres in $\mathbb{R}^d$.

- For $\boldsymbol{Y} \sim t_d(\nu, \boldsymbol{0}, I_d)$, $g(x) = \frac{\Gamma((\nu+d)/2)}{\Gamma(\nu/2)(\pi\nu)^{d/2}}(1 + \frac{x}{\nu})^{-(\nu+d)/2}$.

Example 6.20 (Standardized multivariate normal variance mixtures)

- $\boldsymbol{Y} \sim M_d(\mathbf{0}, I_d, \hat{F}_W)$ is spherical (recall: $\boldsymbol{Y} \stackrel{\text{d}}{=} \mathbf{0} + \sqrt{W} I_d \boldsymbol{Z}$) since

$$\phi_{\boldsymbol{Y}}(\boldsymbol{t}) = \mathbb{E}[\exp(i\boldsymbol{t}^\top \sqrt{W}\boldsymbol{Z})] = \mathbb{E}_W[\,\mathbb{E}[\exp(i(\boldsymbol{t}\sqrt{W})^\top \boldsymbol{Z}) \,|\, W]\,]$$
$$= \mathbb{E}[\exp(-\tfrac{1}{2}W\boldsymbol{t}^\top\boldsymbol{t})] = \hat{F}_W(\tfrac{1}{2}\boldsymbol{t}^\top\boldsymbol{t}) = \hat{F}_W(\tfrac{1}{2}\|\boldsymbol{t}\|^2),$$

  so $\boldsymbol{Y} \sim S_d(\psi)$ by Theorem 6.17 Part 2). We see that the characteristic generator of $\boldsymbol{Y}$ is $\psi(t) = \hat{F}_W(t/2)$.

- For $\boldsymbol{Y} \sim N_d(\mathbf{0}, I_d)$, $\psi(t) = \exp(-t/2)$. By Corollary 6.19, simulating $\boldsymbol{S} \sim \mathrm{U}(\{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| = 1\})$ can thus be done via $\boldsymbol{S} \stackrel{\text{d}}{=} \boldsymbol{Y}/\|\boldsymbol{Y}\|$. Fang et al. (1990, pp. 48) show that ψ generates $S_d(\psi)$ for all $d \in \mathbb{N}$ if and only if it is the characteristic generator of a normal mixture.

- Standardized normal variance mixtures $\subseteq$ spherical distributions. They do not coincide, however, since $\boldsymbol{S} \sim \mathrm{U}(\{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| = 1\})$ is spherical but not a normal variance mixture (if it was, $\boldsymbol{S} = \sqrt{W}\boldsymbol{Z}$, so $\sqrt{W}$ would have to scale $Z_1, \dots, Z_d$ differently in order for $\|\boldsymbol{S}\| = 1$).

Example 6.21 (R, $\boldsymbol{S}$, Cov, Cor)

- It follows from $\boldsymbol{Y} \sim N_d(\mathbf{0}, I_d)$ and $R^2 = \boldsymbol{Y}^\top \boldsymbol{Y} \sim \chi^2_d$ that

$$\mathbf{0} = \mathbb{E}\boldsymbol{Y} \underset{\text{Th. 6.18}}{=} \mathbb{E}R\,\mathbb{E}\boldsymbol{S} \;\Rightarrow\; \mathbb{E}\boldsymbol{S} = \mathbf{0},$$

$$I_d = \operatorname{Cov}\boldsymbol{Y} \underset{\text{Th. 6.18}}{=} \mathbb{E}[R^2]\operatorname{Cov}\boldsymbol{S} = d\operatorname{Cov}\boldsymbol{S} \;\Rightarrow\; \operatorname{Cov}\boldsymbol{S} = I_d/d. \qquad (28)$$

- For $\boldsymbol{Y} \sim S_d(\psi)$ with $\mathbb{E}[R^2] < \infty$, it follows that $\operatorname{Cov}\boldsymbol{Y} \underset{\text{Th. 6.18}}{=} \mathbb{E}[R^2]\operatorname{Cov}\boldsymbol{S} = \frac{\mathbb{E}[R^2]}{d} I_d$ and thus $\operatorname{Cor}\boldsymbol{Y} = I_d$.
- For $\boldsymbol{X} = \boldsymbol{\mu} + A\boldsymbol{Y}$ with $\mathbb{E}[R^2] < \infty$ and Cholesky factor A of a covariance matrix Σ, we have $\operatorname{Cov}\boldsymbol{X} = \frac{\mathbb{E}[R^2]}{d}\Sigma$ and $\operatorname{Cor}\boldsymbol{X} = P$ (the correlation matrix corresponding to Σ).
- **Example:** For $\boldsymbol{Y} \sim t_d(\nu, \mathbf{0}, I_d)$, $R^2 = \boldsymbol{Y}^\top\boldsymbol{Y} = W\boldsymbol{Z}^\top\boldsymbol{Z}$ for $\boldsymbol{Z} \sim \mathrm{N}_d(\mathbf{0}, I_d)$. Therefore, $\frac{R^2}{d} = \frac{\boldsymbol{Z}^\top\boldsymbol{Z}/d}{(\nu/W)/\nu} = \frac{\chi^2_d/d}{\chi^2_\nu/\nu} \sim F(d,\nu)$ and thus $\mathbb{E}[R^2/d] = \frac{\nu}{\nu-2}$. It follows that $\boldsymbol{X} \sim t_d(\nu, \boldsymbol{\mu}, \Sigma)$ has $\operatorname{Cov}\boldsymbol{X} = \frac{\nu}{\nu-2}\Sigma$ and $\operatorname{Cor}\boldsymbol{X} = P$ which we already know from Section 6.2.1.

6.3.2 Elliptical distributions

Definition 6.22 (Elliptical distribution)
A random vector $\boldsymbol{X} = (X_1, \dots, X_d)$ has an *elliptical distribution* if

$$\boldsymbol{X} \stackrel{\text{d}}{=} \boldsymbol{\mu} + A\boldsymbol{Y}, \quad \text{(multivariate affine transformation)}$$

where $\boldsymbol{Y} \sim S_k(\psi)$, $A \in \mathbb{R}^{d\times k}$ (*scale matrix* $\Sigma = AA^\top$), and (*location vector*) $\boldsymbol{\mu} \in \mathbb{R}^d$.

- By Theorem 6.18, an elliptical random vector admits the stochastic representation $\boldsymbol{X} \stackrel{\text{d}}{=} \boldsymbol{\mu} + RA\boldsymbol{S}$, with R and $\boldsymbol{S}$ as given in (27).
- The characteristic function of an elliptical random vector $\boldsymbol{X}$ is $\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \mathbb{E}[e^{i\boldsymbol{t}^\top \boldsymbol{X}}] = \mathbb{E}[e^{i\boldsymbol{t}^\top(\boldsymbol{\mu}+A\boldsymbol{Y})}] = e^{i\boldsymbol{t}^\top\boldsymbol{\mu}}\,\mathbb{E}[e^{i(A^\top \boldsymbol{t})^\top \boldsymbol{Y}}] = e^{i\boldsymbol{t}^\top\boldsymbol{\mu}}\,\psi(\boldsymbol{t}^\top\Sigma\boldsymbol{t})$.
  Notation: $\boldsymbol{X} \sim \mathrm{E}_d(\boldsymbol{\mu}, \Sigma, \psi)$ $(= \mathrm{E}_d(\boldsymbol{\mu}, c\Sigma, \psi(\cdot/c))$, $c > 0)$.
- If Σ is positive definite with Cholesky factor A, then $\boldsymbol{X} \sim \mathrm{E}_d(\boldsymbol{\mu}, \Sigma, \psi)$ if and only if $\boldsymbol{Y} = A^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) \sim S_d(\psi)$.

- Normal variance mixture distributions are (all) elliptical (most useful examples) since $\boldsymbol{X} \overset{\text{d}}{=} \boldsymbol{\mu} + \sqrt{W} A\boldsymbol{Z} = \boldsymbol{\mu} + \sqrt{W}\|\boldsymbol{Z}\| A\boldsymbol{Z}/\|\boldsymbol{Z}\| = \boldsymbol{\mu} + RA\boldsymbol{S}$ with $R = \sqrt{W}\|\boldsymbol{Z}\|$ and $\boldsymbol{S} = \boldsymbol{Z}/\|\boldsymbol{Z}\|$. By Corollary 6.19, R and $\boldsymbol{S}$ are indeed independent.
- If $\boldsymbol{X} \sim \mathrm{E}_d(\boldsymbol{\mu}, \Sigma, \psi)$ with $\mathbb{P}(\boldsymbol{X} = \boldsymbol{\mu}) = 0$, then $\boldsymbol{Y} = A^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) \sim S_d(\psi)$. Corollary 6.19 implies that

$$\left(\sqrt{(\boldsymbol{X} - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{X} - \boldsymbol{\mu})}, \frac{A^{-1}(\boldsymbol{X} - \boldsymbol{\mu})}{\sqrt{(\boldsymbol{X} - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{X} - \boldsymbol{\mu})}}\right) \overset{\text{d}}{=} (R, \boldsymbol{S}), \qquad (29)$$

which can be used for testing elliptical symmetry. One can also use the following result for testing.

Proposition 6.23

Let $\boldsymbol{X} \sim E_d(\boldsymbol{\mu}, \Sigma, \psi)$ for positive definite Σ and $\mathbb{E}[R^2] < \infty$ (i.e., $\mathrm{Cov}[\boldsymbol{X}]$ finite). For any $c \geq 0$ such that $\mathbb{P}((\boldsymbol{X} - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{X} - \boldsymbol{\mu}) \geq c) > 0$,

$$\mathrm{Cor}[\boldsymbol{X} \mid (\boldsymbol{X} - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{X} - \boldsymbol{\mu}) \geq c] = \mathrm{Cor}[\boldsymbol{X}].$$

Proof. $\boldsymbol{X}\,|\,((\boldsymbol{X}-\boldsymbol{\mu})^\top\Sigma^{-1}(\boldsymbol{X}-\boldsymbol{\mu})\ge c)\underset{(29)}{\overset{\mathrm{d}}{=}}\boldsymbol{\mu}+RA\boldsymbol{S}\,|\,(R^2\ge c)\overset{\text{ind.}}{=}\boldsymbol{\mu}+\tilde{R}A\boldsymbol{S}$ where $\tilde{R}\overset{\mathrm{d}}{=}(R\,|\,R^2\ge c)$. Therefore, the conditional distribution remains elliptical with scale matrix Σ and thus the claim holds. □

6.3.3 Properties of elliptical distributions

- **Density:** Let Σ be positive definite and $\boldsymbol{Y}\sim S_d(\psi)$ have density generator g. The Density Transformation Theorem implies that $\boldsymbol{X}=\boldsymbol{\mu}+A\boldsymbol{Y}$ has density

$$f_{\boldsymbol{X}}(\boldsymbol{x})=\frac{1}{\sqrt{\det\Sigma}}\,g\big((\boldsymbol{x}-\boldsymbol{\mu})^\top\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\big),$$

  which depends on $\boldsymbol{x}$ only through $(\boldsymbol{x}-\boldsymbol{\mu})^\top\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$, i.e., is constant on ellipsoids (hence the name "elliptical").

- **Linear combinations:** For $\boldsymbol{X}\sim \mathrm{E}_d(\boldsymbol{\mu},\Sigma,\psi)$, $B\in\mathbb{R}^{k\times d}$ and $\boldsymbol{b}\in\mathbb{R}^k$,

$$B\boldsymbol{X}+\boldsymbol{b}\sim \mathrm{E}_k(B\boldsymbol{\mu}+\boldsymbol{b},B\Sigma B^\top,\psi).$$

If $\boldsymbol{a} \in \mathbb{R}^d$ (take $\boldsymbol{b} = \boldsymbol{0}$ and $B = \boldsymbol{a}^\top \in \mathbb{R}^{1\times d}$),

$$\boldsymbol{a}^\top \boldsymbol{X} \sim \mathrm{E}_1(\boldsymbol{a}^\top \boldsymbol{\mu}, \boldsymbol{a}^\top \Sigma \boldsymbol{a}, \psi) \quad \text{(as for } \mathrm{N}(\boldsymbol{\mu}, \Sigma)\text{).} \tag{30}$$

From $\boldsymbol{a} = \boldsymbol{e}_j = (0, \dots, 0, 1, 0, \dots, 0)$ we see that all marginal distributions are of the same type.

Proof. Similarly as for multivariate normal variance mixtures,

$$\begin{aligned}\phi_{B\boldsymbol{X}+\boldsymbol{b}}(\boldsymbol{t}) &= \mathbb{E}[\exp(i\boldsymbol{t}^\top(B\boldsymbol{X}+\boldsymbol{b}))] = e^{i\boldsymbol{t}^\top \boldsymbol{b}} \phi_{\boldsymbol{X}}(B^\top \boldsymbol{t}) \\ &= e^{i\boldsymbol{t}^\top(\boldsymbol{b}+B\boldsymbol{\mu})} \psi(\boldsymbol{t}^\top B \Sigma B^\top \boldsymbol{t}).\end{aligned}$$

□

- **Marginal dfs:** As for $\mathrm{N}_d(\boldsymbol{\mu}, \Sigma)$, it immediately follows that $\boldsymbol{X} = (\boldsymbol{X}_1^\top, \boldsymbol{X}_2^\top)^\top \sim \mathrm{E}_d(\boldsymbol{\mu}, \Sigma, \psi)$ satisfies $\boldsymbol{X}_1 \sim \mathrm{E}_k(\boldsymbol{\mu}_1, \Sigma_{11}, \psi)$ and $\boldsymbol{X}_2 \sim \mathrm{E}_{d-k}(\boldsymbol{\mu}_2, \Sigma_{22}, \psi)$.
- **Conditional distributions:** One can show that

  $$\boldsymbol{X}_2 \mid \boldsymbol{X}_1 = \boldsymbol{x}_1 \sim \mathrm{E}_{d-k}(\boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}, \tilde{\psi}),$$

  where the characteristic generator $\tilde{\psi}$ is given in Embrechts et al. (2002). For $\mathrm{N}_d(\boldsymbol{\mu}, \Sigma)$ the characteristic generator remains the same.

- **Quadratic forms:** It follows from (29) that $(\boldsymbol{X}-\boldsymbol{\mu})^\top\Sigma^{-1}(\boldsymbol{X}-\boldsymbol{\mu}) \overset{\mathrm{d}}{=} R^2$. If $\boldsymbol{X} \sim N_d(\boldsymbol{\mu},\Sigma)$, then $R^2 \sim \chi^2_d$; and if $\boldsymbol{X} \sim t_d(\nu,\boldsymbol{\mu},\Sigma)$, then $R^2/d \sim F(d,\nu)$.
- **Convolutions:** Let $\boldsymbol{X} \sim \mathrm{E}_d(\boldsymbol{\mu},\Sigma,\psi)$ and $\boldsymbol{Y} \sim \mathrm{E}_d(\tilde{\boldsymbol{\mu}},c\Sigma,\tilde{\psi})$ be independent. Then

  $$a\boldsymbol{X}+b\boldsymbol{Y} \sim \mathrm{E}_d(a\boldsymbol{\mu}+b\tilde{\boldsymbol{\mu}},\Sigma,\psi^*)$$

  for $a,b\in\mathbb{R}$, $c>0$, and $\psi^*(t)=\psi(a^2t)\tilde{\psi}(b^2ct)$. Therefore, if $a=b=c=1$, then $\boldsymbol{X}+\boldsymbol{Y} \sim \mathrm{E}_d(\boldsymbol{\mu}+\tilde{\boldsymbol{\mu}},\Sigma,\psi(\cdot)\tilde{\psi}(\cdot))$.

  Proof. $\phi_{a\boldsymbol{X}}(\boldsymbol{t})=e^{i\boldsymbol{t}^\top a\boldsymbol{\mu}}\psi(a^2\boldsymbol{t}^\top\Sigma\boldsymbol{t})$ and $\phi_{b\boldsymbol{Y}}(\boldsymbol{t})=e^{i\boldsymbol{t}^\top b\tilde{\boldsymbol{\mu}}}\tilde{\psi}(b^2c\boldsymbol{t}^\top\Sigma\boldsymbol{t})$. By independence of $\boldsymbol{X}$ and $\boldsymbol{Y}$, $\phi_{a\boldsymbol{X}+b\boldsymbol{Y}}(\boldsymbol{t})=\phi_{a\boldsymbol{X}}(\boldsymbol{t})\phi_{b\boldsymbol{Y}}(\boldsymbol{t})$
  $=e^{i\boldsymbol{t}^\top(a\boldsymbol{\mu}+b\tilde{\boldsymbol{\mu}})}\psi^*(\boldsymbol{t}^\top\Sigma\boldsymbol{t})$, so $a\boldsymbol{X}+b\boldsymbol{Y} \sim \mathrm{E}_d(a\boldsymbol{\mu}+b\tilde{\boldsymbol{\mu}},\Sigma,\psi^*)$. $\square$
- We see that many nice properties of $\mathrm{N}_d(\boldsymbol{\mu},\Sigma)$ are preserved.

Proposition 6.24 (Subadditivity of VaR in elliptical models)
Let $L_i = \boldsymbol{\lambda}_i^\top \boldsymbol{X}$, $\boldsymbol{\lambda}_i \in \mathbb{R}^d$, $i \in \{1, \dots, n\}$, with $\boldsymbol{X} \sim E_d(\boldsymbol{\mu}, \Sigma, \psi)$. Then $\mathrm{VaR}_\alpha(\sum_{i=1}^n L_i) \le \sum_{i=1}^n \mathrm{VaR}_\alpha(L_i)$ for all $\alpha \in [1/2, 1]$.

Proof. Consider a generic $L = \boldsymbol{\lambda}^\top \boldsymbol{X} \stackrel{\text{d}}{=} \boldsymbol{\lambda}^\top \boldsymbol{\mu} + \boldsymbol{\lambda}^\top A \boldsymbol{Y}$ for $\boldsymbol{Y} \sim S_k(\psi)$. By Theorem 6.17 Part 3), $\boldsymbol{\lambda}^\top A \boldsymbol{Y} \stackrel{\text{d}}{=} \|\boldsymbol{\lambda}^\top A\| Y_1$, so $L \stackrel{\text{d}}{=} \boldsymbol{\lambda}^\top \boldsymbol{\mu} + \|\boldsymbol{\lambda}^\top A\| Y_1$ (all of the same type). By Translation Invariance and Positive Homogeneity,

$$\mathrm{VaR}_\alpha(L) = \boldsymbol{\lambda}^\top \boldsymbol{\mu} + \|\boldsymbol{\lambda}^\top A\| \,\mathrm{VaR}_\alpha(Y_1). \tag{31}$$

Applying (31) to $L = \sum_{i=1}^n L_i$ and $L = L_i$, $i \in \{1, \dots, n\}$, and using that $\mathrm{VaR}_\alpha(Y_1) \ge 0$ for $\alpha \in [1/2, 1]$, we obtain $\mathrm{VaR}_\alpha(\sum_{i=1}^n L_i)$
$= \mathrm{VaR}_\alpha((\sum_{i=1}^n \boldsymbol{\lambda}_i)^\top \boldsymbol{X}) \underset{(31)}{=} \sum_{i=1}^n \boldsymbol{\lambda}_i^\top \boldsymbol{\mu} + \|\sum_{i=1}^n \boldsymbol{\lambda}_i^\top A\| \,\mathrm{VaR}_\alpha(Y_1)$
$\le \sum_{i=1}^n \boldsymbol{\lambda}_i^\top \boldsymbol{\mu} + (\sum_{i=1}^n \|\boldsymbol{\lambda}_i^\top A\|) \,\mathrm{VaR}_\alpha(Y_1) = \sum_{i=1}^n (\boldsymbol{\lambda}_i^\top \boldsymbol{\mu} + \|\boldsymbol{\lambda}_i^\top A\| \,\mathrm{VaR}_\alpha(Y_1))$
$\underset{(31)}{=} \sum_{i=1}^n \mathrm{VaR}_\alpha(L_i)$. **Note:** For $\boldsymbol{\lambda}_i = \boldsymbol{e}_i$, $\mathrm{VaR}_\alpha(\sum_{i=1}^d X_i) \le \sum_{i=1}^d \mathrm{VaR}_\alpha(X_i)$.

□

6.3.4 Estimating scale and correlation

- Suppose $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n \sim \mathrm{E}_d(\boldsymbol{\mu}, \Sigma, \psi)$. How can we estimate $\boldsymbol{\mu}, \Sigma$ and P? (P is the correlation matrix corresponding to Σ; this always exists)
- $\bar{\boldsymbol{X}}$, S, R may not be the best options for heavy-tailed data (e.g., concerning robustness against contamination).

M-estimators for $\boldsymbol{\mu}, \Sigma$ (see Maronna (1976))

- **Goal:** Improve given estimators $\hat{\boldsymbol{\mu}}, \hat{\Sigma}$.
- **Idea:** Compute improved estimates by downweighting observations with large $D_i = \sqrt{(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})^\top \hat{\Sigma}^{-1} (\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})}$ (these are the ones which tend to distort $\hat{\boldsymbol{\mu}}$, $\hat{\Sigma}$ most).
- This can be turned into an iterative procedure that converges to so-called *M-estimates* of location and scale ($\hat{\Sigma}$ is in general biased).

Algorithm 6.25 (M-estimators of location and scale)

1) Set $k = 1$, $\hat{\boldsymbol{\mu}}^{[1]} = \bar{\boldsymbol{X}}$ and $\hat{\Sigma}^{[1]} = S$.

2) Repeat until convergence:

   2.1) For $i \in \{1, \dots, n\}$ set $D_i = \sqrt{(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}^{[k]})^\top \hat{\Sigma}^{[k]-1} (\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}^{[k]})}$.

   2.2) Update:
$$\hat{\boldsymbol{\mu}}^{[k+1]} = \frac{\sum_{i=1}^n w_1(D_i)\boldsymbol{X}_i}{\sum_{i=1}^n w_1(D_i)},$$
   where w_1 is a weight function, e.g., $w_1(x) = (d+\nu)/(x^2+\nu)$ (or $\mathbb{1}_{x \le a} + (a/x)\mathbb{1}_{x>a}$ for some value a).

   2.3) Update:
$$\hat{\Sigma}^{[k+1]} = \frac{1}{n}\sum_{i=1}^n w_2(D_i^2)(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}^{[k]})(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}^{[k]})^\top,$$
   where w_2 is a weight function, e.g., $w_2(x) = w_1(\sqrt{x})$ (or $(w_1(\sqrt{x}))^2$).

   2.4) Set k to $k+1$.

Estimating P via Kendall's tau

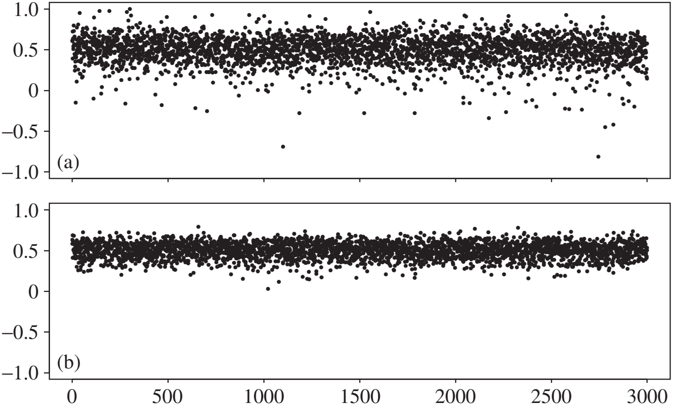
- One can show (see later) that if $\boldsymbol{X} \sim E_d(\boldsymbol{\mu}, \Sigma, \psi)$, then

$$\tau(X_i, X_j) = \frac{2}{\pi} \arcsin(P_{ij}), \quad i \neq j, \tag{32}$$

  where P is the matrix of pairwise correlations corresponding to Σ (always existing, no matter whether the second moments of $X_1, \dots, X_d$ do).
- Estimate $\tau(X_i, X_j)$ by $\hat{\tau}_{ij}$ (see later) and solve $\hat{\tau}_{ij}$ w.r.t. P_{ij} to obtain $\hat{P}_{ij}$ (this does not require estimating variances/covariances).
- $(\hat{P}_{ij})_{ij}$ is not necessarily positive definite. There are various methods for finding a "near" matrix which is positive definite, see, e.g., Higham (2002) (or `Matrix::nearPD()`).

Example 6.26 (Correlation estimation for heavy-tailed data)
Consider $n = 3000$ realizations of independent samples of size 90 from $t_2\left(3, \mathbf{0}, \left(\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right)\right)$ ($\Rightarrow$ linear correlation 0.5).

(a) Pearson's correlation; (b) Inversion of pairwise Kendall's tau estimator



The Kendall's tau transform method produces estimates that show less variation (and thus provides a more efficient way of estimating ρ).

6.4 Dimension reduction techniques

6.4.1 Factor models

Explain the variability of $\boldsymbol{X}$ in terms of common factors.

Definition 6.27 (p-factor model)
$\boldsymbol{X}$ follows a *p-factor model* if

$$\boldsymbol{X} = \boldsymbol{a} + B\boldsymbol{F} + \boldsymbol{\varepsilon}, \tag{33}$$

where

1) $B \in \mathbb{R}^{d\times p}$ is a matrix of *factor loadings* and $\boldsymbol{a} \in \mathbb{R}^d$;
2) $\boldsymbol{F} = (F_1, \dots, F_p)$ is the random vector of *(common) factors* with $p < d$ and existing $\Omega := \operatorname{Cov}[\boldsymbol{F}]$, (*systematic risk*);
3) $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_d)$ is the random vector of *idiosyncratic error terms* with $\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$, $\Upsilon := \operatorname{Cov}[\boldsymbol{\varepsilon}]$ diag., $\operatorname{Cov}[\boldsymbol{F}, \boldsymbol{\varepsilon}] = (0)$ (*idiosync. risk*).

- **Goals:** Identify or estimate $\boldsymbol{F}_t$, $t \in \{1, \dots, n\}$, then model the distribution/dynamics of the (lower-dimensional) factors (instead of $\boldsymbol{X}_t$, $t \in \{1, \dots, n\}$).
- Factor models imply that $\Sigma := \mathrm{Cov}[\boldsymbol{X}] = B\Omega B^\top + \Upsilon$.
- With $B^* = B\Omega^{1/2}$ and $\boldsymbol{F}^* = \Omega^{-1/2}(\boldsymbol{F} - \mathbb{E}[\boldsymbol{F}])$, we have

$$\boldsymbol{X} = \boldsymbol{\mu} + B^* \boldsymbol{F}^* + \boldsymbol{\varepsilon},$$

  where $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{X}]$. We have $\Sigma = B^*(B^*)^\top + \Upsilon$. Conversely, if $\mathrm{Cov}[\boldsymbol{X}] = BB^\top + \Upsilon$ for some $B \in \mathbb{R}^{d \times p}$ with $\mathrm{rank}(B) = p < d$ and diagonal matrix Υ, then $\boldsymbol{X}$ has a factor-model representation for a p-dimensional $\boldsymbol{F}$ and d-dimensional $\boldsymbol{\varepsilon}$.

Example 6.28 (One-factor/equicorrelation model)

Let $\mathbb{E}[\boldsymbol{X}] = \mathbf{0}$, $\Sigma = \mathrm{Cov}[\boldsymbol{X}] = \rho J_d + (1-\rho) I_d$ ($J_d = (1) \in \mathbb{R}^{d\times d}$).

- Then $\Sigma = BB^\top + \Upsilon$ for $B = \sqrt{\rho}\mathbf{1}$ and $\Upsilon = (1-\rho)I_d$.
- Any Y with $\mathbb{E}Y = 0$, $\mathrm{Var}\,Y = 1$ independent of $\boldsymbol{X}$ leads to the *factor decomposition* of $\boldsymbol{X}$
$$F = \frac{\sqrt{\rho}}{1+\rho(d-1)}\sum_{j=1}^d X_j + \sqrt{\frac{1-\rho}{1+\rho(d-1)}}Y, \quad \varepsilon_j = X_j - \sqrt{\rho}F.$$
We have $\mathbb{E}[F] = 0$, $\mathrm{Var}[F] = 1$, so $\boldsymbol{X} = \mathbf{0} + BF + \boldsymbol{\varepsilon} = \sqrt{\rho}\mathbf{1}F + \boldsymbol{\varepsilon}$.
- The requirements of Definition 6.27 are fulfilled since $\mathrm{Cov}[F, \varepsilon_j] = 0$, $\mathrm{Cov}[\varepsilon_j, \varepsilon_k] = 0$ for all $j \neq k$.
- $\mathrm{Var}[\bar{X}_n] = \mathrm{Var}[\sqrt{\rho}F + \bar{\varepsilon}_d] = \rho + \frac{1-\rho}{d} \underset{(d\to\infty)}{\to} \rho$ (systematic factor matters!)
- If $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma)$, take $Y \sim \mathrm{N}(0,1)$ (then F is also normal). One typically writes this (one-factor) equicorrelation model as $\boldsymbol{X} = \sqrt{\rho}F + \sqrt{1-\rho}\boldsymbol{Z}$, where $F, Z_1, \dots, Z_d \overset{\text{ind.}}{\sim} \mathrm{N}(0,1)$.

6.4.2 Statistical estimation strategies

Consider $\boldsymbol{X}_t = \boldsymbol{a} + B\boldsymbol{F}_t + \boldsymbol{\varepsilon}_t$, $t \in \{1, \dots, n\}$. Three types of factor model are commonly used:

1) ***Macroeconomic factor models:*** Here we assume that $\boldsymbol{F}_t$ is observable, $t \in \{1, \dots, n\}$. Fitting $B, \boldsymbol{a}$ is accomplished by time series regression (see later).

2) ***Fundamental factor models:*** Here we assume that the matrix of factor loadings B is known but the factors $\boldsymbol{F}_t$ are unobserved (and have to be estimated from $\boldsymbol{X}_t$, $t \in \{1, \dots, n\}$, using cross-sectional regression at each t).

3) ***Fundamental factor models:*** Here we assume that neither the factors $\boldsymbol{F}_t$ nor the factor loadings B are observed (both have to be estimated from $\boldsymbol{X}_t$, $t \in \{1, \dots, n\}$). The factors can be found with principal component analysis (see later).

6.4.3 Estimating macroeconomic factor models

There are two equivalent approaches.

Univariate regression

- Consider the (univariate) *time series regression* model
$$X_{t,j} = a_j + \boldsymbol{b}_j^\top \boldsymbol{F}_t + \varepsilon_{t,j}, \quad t \in \{1, \dots, n\}.$$
- To justify the use of the ordinary least-squares (OLS) method to derive statistical properties of the method it is usually assumed that, conditional on the factors, the errors $\varepsilon_{1,j}, \dots, \varepsilon_{n,j}$ form a white noise process (i.e., are identically distributed and serially uncorrelated).
- $\hat{a}_j$ estimates a_j, $\hat{\boldsymbol{b}}_j$ estimates the jth row of B.

Multivariate regression

- Here, construct large matrices:

$$X = \underbrace{\begin{pmatrix} \boldsymbol{X}_1^\top \\ \vdots \\ \boldsymbol{X}_n^\top \end{pmatrix}}_{n\times d}, \quad F = \underbrace{\begin{pmatrix} 1 & \boldsymbol{F}_1^\top \\ \vdots & \vdots \\ 1 & \boldsymbol{F}_n^\top \end{pmatrix}}_{n\times(p+1)}, \quad \tilde{B} = \underbrace{\begin{pmatrix} \boldsymbol{a}^\top \\ B^\top \end{pmatrix}}_{(p+1)\times d}, \quad E = \underbrace{\begin{pmatrix} \boldsymbol{\varepsilon}_1^\top \\ \vdots \\ \boldsymbol{\varepsilon}_n^\top \end{pmatrix}}_{n\times d}.$$

  This model can be expressed by $X = F\tilde{B} + E$ (estimate $\tilde{B}$).

- Assume the unobserved $\boldsymbol{\varepsilon}_1, \dots, \boldsymbol{\varepsilon}_n$ form a white noise process. Then, conditional on $\boldsymbol{F}_1, \dots, \boldsymbol{F}_n$, we have a multivariate linear regression, see, e.g., Mardia et al. (1979), with estimator $\hat{\tilde{B}} = (F^\top F)^{-1} F^\top X$.

- Now examine the condititions of Definition 6.27: Do the errors vectors $\boldsymbol{\varepsilon}_t$ come from a distribution with diagonal covariance matrix, and are they uncorrelated with the factors?

- Consider the sample correlation matrix of $\hat{E} = X - F\hat{\hat{B}}$ (model residual matrix; hopefully shows that there is little correlation in the errors) and take the diagonal elements as an estimator $\hat{\Upsilon}$ of Υ.

6.4.4 Estimating fundamental factor models

- Consider the cross-sectional regression model $\boldsymbol{X}_t = B\boldsymbol{F}_t + \boldsymbol{\varepsilon}_t$ (B known; $\boldsymbol{F}_t$ to be estimated; $\mathrm{Cov}[\boldsymbol{\varepsilon}] = \Upsilon$); note that $\boldsymbol{a}$ can be absorbed into $\boldsymbol{F}_t$. To obtain precision in estimating $\boldsymbol{F}_t$, we need $d \gg p$.
- First estimate $\boldsymbol{F}_t$ via OLS by $\hat{\boldsymbol{F}}_t^{\mathrm{OLS}} = (B^\top B)^{-1} B^\top \boldsymbol{X}_t$. This is the best linear unbiased estimator if the $\boldsymbol{\varepsilon}$ is homoskedastic. However, it is possible to obtain linear unbiased estimates with a smaller covariance matrix via generalized least squares (GLS).
- To this end, estimate Υ by $\hat{\Upsilon}$ via the diagonal of the sample covariance matrix of the residuals $\hat{\boldsymbol{\varepsilon}}_t = \boldsymbol{X}_t - B\hat{\boldsymbol{F}}_t^{\mathrm{OLS}}$, $t \in \{1, \dots, n\}$.
- Then estimate $\boldsymbol{F}_t$ via $\hat{\boldsymbol{F}}_t = (B^\top \Upsilon^{-1} B)^{-1} B^\top \Upsilon^{-1} \boldsymbol{X}_t$.

6.4.5 Principal component analysis

- **Goal:** Reduce the dimensionality of highly correlated data by finding a small number of uncorrelated linear combinations which account for most of the variance in the data; this can be used for finding factors.
- **Key:** Any symmetric A admits a *spectral decomposition*

$$A = \Gamma \Lambda \Gamma^\top,$$

where

1) $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_d)$ is the diagonal matrix of eigenvalues of A which, w.l.o.g., are ordered so that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$; and
2) Γ is an orthogonal matrix whose columns are eigenvectors of A standardized to have length 1.

- Let $\Sigma = \Gamma \Lambda \Gamma^\top$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$ (positive semidefiniteness $\Rightarrow$ all eigenvalues ≥ 0) and $\boldsymbol{Y} = \Gamma^\top(\boldsymbol{X} - \boldsymbol{\mu})$ (the so-called *principal component transform*). The jth component $Y_j = \boldsymbol{\gamma}_j^\top(\boldsymbol{X} - \boldsymbol{\mu})$ is the *jth principal component of $\boldsymbol{X}$* (where $\boldsymbol{\gamma}_j$ is the jth column of Γ).

- We have $\mathbb{E}\boldsymbol{Y} = \boldsymbol{0}$ and $\mathrm{Cov}[\boldsymbol{Y}] = \Gamma^\top\Sigma\Gamma = \Gamma^\top\Gamma\Lambda\Gamma^\top\Gamma = \Lambda$, so the principal componenents are uncorrelated with $\mathrm{Var}[Y_j] = \lambda_j$, $j \in \{1, \dots, d\}$. The principal components are thus ordered by variance (from largest to smallest).
- One can show:
  - The first principal component is that standardized linear combination of $\boldsymbol{X}$ which has maximal variance among all such combinations, i.e., $\mathrm{Var}(\boldsymbol{\gamma}_1^\top\boldsymbol{X}) = \max\{\mathrm{Var}(\boldsymbol{a}^\top\boldsymbol{X}) : \boldsymbol{a}^\top\boldsymbol{a} = 1\}$.
  - For $j \in \{2, \dots, d\}$, the jth principal component is that standardized linear combination of $\boldsymbol{X}$ which has maximal variance among all such linear combinations which are orthogonal to (and hence uncorrelated with) the first $j-1$-many linear combinations.
- $\sum_{j=1}^d \mathrm{Var}(Y_j) = \sum_{j=1}^d \lambda_j = \mathrm{trace}(\Sigma) = \sum_{j=1}^d \mathrm{Var}(X_j)$, so we can interpret $\sum_{j=1}^k \lambda_j / \sum_{j=1}^d \lambda_j$ as the fraction of total variance explained by the first k principal components.

Principal components as factors

- Inverting the principal component transform $\boldsymbol{Y} = \Gamma^\top(\boldsymbol{X} - \boldsymbol{\mu})$, we have

$$\boldsymbol{X} = \boldsymbol{\mu} + \Gamma\boldsymbol{Y} = \boldsymbol{\mu} + \Gamma_1\boldsymbol{Y}_1 + \Gamma_2\boldsymbol{Y}_2 =: \boldsymbol{\mu} + \Gamma_1\boldsymbol{Y}_1 + \boldsymbol{\varepsilon}$$

  where $\boldsymbol{Y}_1 \in \mathbb{R}^k$ contains the first k principal components. This is reminiscent of the basic factor model.

- Although $\varepsilon_1, \ldots, \varepsilon_d$ will tend to have small variances, the assumptions of the factor model are generally violated (since they need not have a diagonal covariance matrix and need not be uncorrelated with $\boldsymbol{Y}_1$). Nevertheless, principal components are often interpreted as factors.

Sample principal components

- Assume $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ with identical distribution, unknown mean vector $\boldsymbol{\mu}$ and covariance matrix Σ with the spectral decomposition $\Sigma = \Gamma\Lambda\Gamma^\top$ as before.

- Estimate $\boldsymbol{\mu}$ by $\bar{\boldsymbol{X}}$ and Σ by $S_x = \frac{1}{n}\sum_{t=1}^n(\boldsymbol{X}_t - \bar{\boldsymbol{X}})(\boldsymbol{X}_t - \bar{\boldsymbol{X}})^\top$.

- Apply the spectral decomposition to S_x to get $S_x = GLG^\top$, where G is the eigenvector matrix and $L = \operatorname{diag}(l_1, \ldots, l_d)$ is the diagonal matrix consisting of ordered eigenvalues.
- Define the "sample principle component transforms" $\boldsymbol{Y}_t = G^\top(\boldsymbol{X}_t - \bar{\boldsymbol{X}})$, $t \in \{1, \ldots, n\}$. The jth component $Y_{t,j} = \boldsymbol{g}_j^\top(\boldsymbol{X}_t - \bar{\boldsymbol{X}})$ is the *jth sample principal component at time t* ($\boldsymbol{g}_j$ is the jth column of G).
- The rotated vectors $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ have sample covariance matrix L:

$$S_y = \frac{1}{n}\sum_{t=1}^{n}(\boldsymbol{Y}_t - \bar{\boldsymbol{Y}})(\boldsymbol{Y}_t - \bar{\boldsymbol{Y}})^\top = \frac{1}{n}\sum_{t=1}^{n}\boldsymbol{Y}_t\boldsymbol{Y}_t^\top$$
$$= \frac{1}{n}\sum_{t=1}^{n}G^\top(\boldsymbol{X}_t - \bar{\boldsymbol{X}})(\boldsymbol{X}_t - \bar{\boldsymbol{X}})^\top G = G^\top S_x G = L.$$

Thus the rotated vectors show no correlation between components and the components are ordered by their sample variances, from largest to smallest.

- Now use G and $\boldsymbol{Y}_t$ to calibrate an approximate factor model. We assume our data are realisations from the model

$$\boldsymbol{X}_t = \bar{\boldsymbol{X}} + G_1 \boldsymbol{F}_t + \boldsymbol{\varepsilon}_t, \quad t \in \{1, \dots, n\},$$

  where G_1 consists of the first k columns of G and $\boldsymbol{F}_t = (Y_{t,1}, \dots, Y_{t,k})$, $t \in \{1, \dots, n\}$.

- In practice, the errors $\boldsymbol{\varepsilon}_t$ do not have a diagonal covariance matrix and are not uncorrelated with $\boldsymbol{F}_t$. Nevertheless the method is a popular approach to constructing time series of statistically explanatory factors from multivariate time series of risk-factor changes.