14

MULTIVARIATE STATISTICAL METHODS

14.1 Multivariate data and distribution

14.1.1 Sample statistics

notations:

- $\mathbf{1}$ is the vector of all 1.
- J is a square matrix with all 1.

The most basic sample statistics are sample mean, sample covariance, and sample correlation, as we introduce in the following.

Let $\boldsymbol{X}$ be the data matrix such that $\boldsymbol{X} = [X_1, X_2, \ldots, X_n]^T, X_i \in \mathbb{R}^p$. It follows that

- (sample mean)
$$\bar{X} \triangleq \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}X^T J$$
- (sample covariance)
$$S \triangleq \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T = \frac{1}{n}X^T(I - \frac{1}{n}J)X.$$
Because the demean operation can be realized by multiplying matrix $I - \frac{1}{n}J$, which is an orthogonal projector, we can also write the sample covariance by
$$S = \frac{1}{n-1}X^T(I - \frac{1}{n}J)X.$$
- (sample correlation)
$$R = D^{-1/2}SD^{-1/2},$$
where $D = diag(S)$.

Applying affine transformation to sample statistics can still have relative simple forms.

Lemma 14.1.1 (affine transformation of sample statistics). *Let $Y = AX$ and $Z = BX$.*

- $$\bar{Y} = A\bar{X}.$$
- $$S_Y = AS_XA^T.$$

- $$S_{Y,Z} = AS_XB^T.$$

Proof. (1)

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\sum_{i=1}^{n} AX_i = A\bar{X}.$$

(2)

$$S_Y = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})(Y_i - \bar{Y})^T = AS_XA^T.$$

(3)

$$S_{Y,Z} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})(Z_i - \bar{Z})^T = AS_XB^T.$$

□

14.1.2 Multivariate Gaussian distribution

The most basic joint distribution used to model multiple random variables are multivariate Gaussian/normal distribution. This section, we briefly review the basic properties of multivariate normal distribution, for more detailed discussion,see subsection 12.1.9.

Definition 14.1.1 (multivariate Gaussian/normal distribution). *A random vector is said be multivariate Gaussian/normal random variable if it's pdf is multivariate Gaussian/normal distribution, whose support is $\mathbb{R}^n$ and its pdf is*

$$\rho(x;\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}|det\Sigma|^{1/2}}\exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))$$

with mean $\mu \in \mathbb{R}^n$ and covariance matrix $\Sigma \in \mathbb{R}^{n\times n}$.

Example 14.1.1 (bivariate Gaussian distribution). Let $f(x,y)$ be the density of a bivariate Gaussian distribution $MN(\mu,\Sigma)$, where

$$\mu = \begin{Bmatrix} \mu_X \\ \mu_Y \end{Bmatrix}, \Sigma = \begin{Bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{Bmatrix}.$$

Then,

$$f(x,y) = \frac{\exp(-\frac{1}{2(1-\rho^2)}}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - 2\frac{\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}]).$$

Theorem 14.1.1 (affine transformation for Multivariate normal distribution). *[1, p. 183] Let X be a n dimensional random vector with $MN(\mu, \Sigma)$ distribution. Let $Y = AX + b, A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m$. Then Y is an m dimensional random vector having a $MN(A\mu + b, A\Sigma A^T)$ distribution.*

Proof. Use moment generating function to prove. Let $Y = AX + b$, then from Lemma 11.6.5

$$M_Y(t) = e^{t^T b} M_X(A^T t) = e^{t^T(A\mu+b)+\frac{1}{2}t^T A\Sigma A^T t}$$

which suggesting $Y \sim MN(A\mu + b, A\Sigma A^T)$ □

Lemma 14.1.2 (orthonormal transformation maintains independence). *Let X be a n dimensional random vector with $MN(0, I)$. If C is an orthonormal matrix, then $Y = CX$ has distribution $MN(0, I)$. That is, orthonormal transformation will preserve independence.*

Proof. $Cov(Y) = C^T IC = I$. □

Lemma 14.1.3 (marginal distribution). *The multivariate Gaussian distribution $\rho(x; \mu, \Sigma)$ on $\mathbb{R}^n$ has marginal distribution on $\mathbb{R}^k, k \leq n$ given as $\rho(x_1; \mu_1, \Sigma_{11}), x_1 \in \mathbb{R}^k$ where we decompose*

$$\mu = [\mu_1, \mu_2]^T, \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

Proof. Use above Theorem 14.1.1. Let

$$A = \begin{bmatrix} I & 0 \end{bmatrix}$$

Then $X_1 = AX$. □

Theorem 14.1.2 (conditional distribution). *The multivariate Gaussian distribution $\rho(x;\mu,\Sigma)$ on $\mathbb{R}^n$ has marginal distribution on $\mathbb{R}^k, k \leq n$ given as*

$$\frac{\rho(x_1,x_2)}{\rho(x_2)} = \rho(x_1;\mu_1+\Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2),\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$

where we decompose

$$\mu = [\mu_1^T,\mu_2^T]^T, \Sigma = \begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix},$$

with $\mu_1 \in \mathbb{R}^k, \mu_2 \in \mathbb{R}^{n-k}$.

Proof. See link □

Lemma 14.1.4. *Let a p dimensional random vector $X \sim MN(\mu,\Sigma)$ with Σ being nonsingular. Then*

- *$(X-\mu)^T\Sigma^{-1}(X-\mu)$ is distributed as χ_p^2, where χ_p^2 denote the chi-square distribution with p degrees of freedom.*
- *The $MN(\mu,\Sigma)$ distribution assigns probability $1-\alpha$ to the solid ellipsoid $\{x : (x-\mu)^T\Sigma^{-1}(x-\mu) \leq F(1-\alpha)\}$, where $F(x)$ denote the cdf for chi-square distribution with p degrees of freedom.*

Proof. Straight forward. Note that $(X-\mu)^T\Sigma^{-1}(X-\mu)$ is the sum of squares of independent Gaussian standard random variables. □

14.1.3 Estimation methods

14.1.3.1 *Maximum likelihood estimation*

Lemma 14.1.5 (likelihood function). *Assume $x_1, x_2, ..., x_n, \in x_i \in \mathbb{R}^p$ are independent random samples drawn from a multivariate normal distribution (μ,Σ). Then likelihood function is given by*

$$L(\mu,\Sigma) = \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}}\exp(-Tr(\Sigma^{-1}(\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T))).$$

Proof.

$$\begin{aligned} L(\mu,\Sigma) &= \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}} \exp(-((\sum_{i=1}^{n}(x_i-\mu)\Sigma^{-1}(x_i-\mu)^T))) \\ &= \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}} \exp(-Tr(\Sigma^{-1}(\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T))) \end{aligned}$$

Note that we use $Tr(x^TAx) = Tr(Axx^T)$ [Lemma A.8.8]. □

Lemma 14.1.6 (maximum likelihood estimator). *[2, p. 172] Let $X_1, X_2, ..., X_n$ be a random sample from a multivariate normal distribution with mean μ and covariance Σ. Then,*

$$\hat{\mu} = \bar{X}, \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})(X_i-\bar{X})^T = \frac{n-1}{n}S,$$

*are the **maximum likelihood estimator** of μ and Σ. That is*

$$\begin{aligned} (\hat{\mu},\hat{\Sigma}) &= \arg\max L(\mu,\Sigma) \\ &= \arg\max \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}} \exp(-Tr(\Sigma^{-1}(\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T))). \end{aligned}$$

Proof. (1) Note that

$$\begin{aligned} \sum_{i=1}^{n}(X_i-\mu)(X_i-\mu)^T &= \sum_{i=1}^{n}(X_i-\bar{X}+\bar{X}-\mu)(X_i-\bar{X}+\bar{X}-\mu)^T \\ &= \sum_{i=1}^{n}(X_i-\bar{X})(X_i-\bar{X})^T + \sum_{i=1}^{n}(\bar{X}-\mu)(\bar{X}-\mu)^T \\ &= \sum_{i=1}^{n}(X_i-\bar{X})(X_i-\bar{X})^T + n(\bar{X}-\mu)(\bar{X}-\mu)^T \end{aligned}$$

Note that we use

$$\sum_{i=1}^{n}(X_i-\bar{X})(\bar{X}-\mu)^T = (\sum_{i=1}^{n}X_i - n\bar{X})(\bar{X}-\mu)^T = 0$$

to eliminate the cross terms. Then, for the exponent in L, we have

$$Tr(\Sigma^{-1}(\sum_{i=1}^{n}(X_i-\mu)(X_i-\mu)^T)) = Tr(\Sigma^{-1}(\sum_{i=1}^{n}(X_i-\bar{X})(X_i-\bar{X})^T)) + n(\bar{X}-\mu)\Sigma^{-1}(\bar{X}-\mu).$$

It is easy to see that when $\mu = \bar{X}$, L is maximized **given any positive definite matrix** Σ. (2) Note that after replacing μ with $\bar{X}$, we have

$$-\ln L \approx \frac{n}{2}\ln|\Sigma| + \frac{1}{2}Tr(\Sigma^{-1}(\sum_{i=1}^{n}(x_i - \bar{X})(x_i - \bar{X})^T))$$

Taking the derivative w.r.t. Σ^{-1}, using

$$\frac{\partial}{\partial A}\ln|A| = A^{-T}, \frac{\partial}{\partial A}Tr(AB) = \frac{\partial}{\partial A}Tr(BA) = B^T,$$

we obtain

$$\frac{\partial}{\partial \Sigma^{-1}}\ln L = \frac{N}{2}\Sigma - \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{X})(x_i - \bar{X})^T).$$

Set the derivative to zero and we will get the results. □

14.1.3.2 *Weighted estimation*

Maximum likelihood method is sensitive to outliers. A more robust estimation method can be achieved by assigning smaller weight to samples with large deviation to the estimated mean[3, p. 97]. A typical iterative algorithm is given by algorithm 21.

Algorithm 21: Iterative reweighed estimation for multivariate normal distribution

Input: Sample set consists of $X_1, X_2, ..., X_N$

Initial estimate $\hat{\mu}(1) = \bar{X}, \hat{\Sigma}^{(1)} = S$.

Set $k = 1$.

repeat

For $i = 1, 2, ..., N$, set

$$D_i^2 = (X_i - \hat{\mu}^{(1)})^T [\hat{\Sigma}^{(1)}]^{-1} (X_i - \hat{\mu}).$$

Update location estimation using

$$\hat{\mu}^{(k+1)} = \frac{\sum_{i=1}^N w_1(D_i) X_i}{\sum_{i=1}^N w_1(D_i)},$$

where w_1 is a weight function.

Update dispersion matrix estimation using

$$\hat{\Sigma}^{(k+1)} = \frac{1}{N-1} \sum_{i=1}^N w_2(D_i^2)(X_i - \hat{\mu}^{(k+1)})(X_i - \hat{\mu}^{(k+1)})^T$$

where w_2 is a weight function.

set $k = k + 1$

until *terminal condition is met*;

Output: $\hat{\mu}, \hat{\Sigma}$

Additionally, we have a robust estimation procedure of correlation via Kendall's tau[3, p. 97]. From Lemma 14.4.22, the linear correlation is connected to the Kendall's tau via

$$\rho_\tau(X_1, X_2) = \frac{2}{\pi} \arcsin \rho.$$

The quantity $\rho_\tau(X_1, X_2)$ can be calculated via Definition 14.4.8.

14.2 Principal component analysis (PCA)

14.2.1 Statistical fundamentals of PCA

14.2.1.1 *PCA for random vectors*

Given a **zero mean** D-dimensional real-valued random vector $X = (X_1, X_2, ..., X_D)$ with covariance matrix Σ_X. PCA aims to find $d, d \ll D$ orthonormal vectors $u_i, u_i \in \mathbb{R}^D, i = 1, 2, ..., d$ such that the linearly transformed random variables $Y_i = u_i^T \Sigma_X, ...Y_D$, also known as **principal components**, has the maximum variances. Intuitively, the goal is to construct a new set of random variables or principal components, as a linear combinations of $X_1, ..., X_D$, that capture the most significant variations. Such new principal components can be further used in classification, regression and optimal control.

Mathematically, we can solve $u_1, ..., u_d$ by a series of optimization problems given by

$$
\begin{aligned}
u_1 &= \arg\max_u u^T \Sigma_X u, \ s.t. \ u^T u = 1 \\
u_2 &= \arg\max_u u^T \Sigma_X u, \ s.t. \ u^T u = 1, u^T u_1 = 0 \\
&\dots \\
u_d &= \arg\max_u u^T \Sigma_X u, \ s.t. \ u^T u = 1, u^T u_1 = 0, u^T u_2, ..., u^T u_{d-1} = 0
\end{aligned}
$$

It turns out that the optimization problems are indeed the top d eigenvectors of Σ_X, as we show in the following theorem.

Theorem 14.2.1 (principal components are top eigenvectors). *The principal components vectors $u_1, ..., u_d$ are given by the top d eigenvectors of Σ_X.*

Proof. (1) Use Reyleigh quotient theorem [Theorem 4.8.4] the top eigenvector of Σ_X maximize $u^T \Sigma_X u$ under the constraint $u^T u = 1$. (2) Use quadratic form maximization theorem [Theorem 4.12.4], we know that $u_1, ..., u_d$ are indeed the top eigenvectors. □

There are a number of critical properties regarding principal components.

Theorem 14.2.2 (principal components property).

- *Principal components $Y_i = u_i^T X, i = 1, 2, ..., d$ are uncorrelated to each other; that is*

$$cov(Y_i, Y_j) = u_i^T \Sigma_X u_j = 0, i \neq j;$$

- *Total **variance** preserved in $Y_1, ..., Y_d$ is*

$$Var[Y_1] + \cdots + Var[Y_d] = \lambda_1 + \cdots + \lambda_d,$$

where the total variance of $X_1, X_2, ..., X_D$ is

$$Var[X_1] + \cdots + Var[X_D] = \lambda_1 + \cdots + \lambda_D.$$

Proof. (1) Since u_i are eigenvector, such that

$$cov(y_i, y_j) = u_i^T \Sigma_X u_j = \lambda u_i^T u_j = 0, i \neq j.$$

(2)

$$Var[Y_1] + \cdots + Var[Y_d] = \sum_{i=1}^{d} u_i^T \Sigma_X u_i = \sum_{i=1}^{d} \lambda_i u_i^T u_i = \sum_{i=1}^{d} \lambda_i.$$

Use the property that trace of a matrix equals the sum of all its eigenvalues [Theorem 4.7.3], we have

$$Var[X_1] + \cdots + Var[X_D] = Tr(\Sigma_X) = \lambda_1 + \cdots + \lambda_D.$$

□

14.2.1.2 *Sample principal components*

In practice, we are given a set of sample point $x_1, ..., x_n, x_i \in \mathbb{R}^D$, and our goal is to find low-dimensional projected sample points $y_1, ..., y_n, y_i \in \mathbb{R}^d, , d \ll D$ via $y_i = U^T x_i, U \in \mathbb{R}^{D \times d}$ such that the variations of in $\{x_1, ..., x_n\}$ are maximally preserved [Figure 29.1.2]. The column vectors $u_1, ..., u_d \in \mathbb{R}^D$ are known as **principal component directions or principal components**, and the transformed sample point (a vector) y_i is known as **principal component scores**, with k component given by $u_k^T x_i$.

The goal of preserving sample variation can be formulated as the following optimization problem. Given a set of multi-dimensional sample point $x_1, ..., x_N \in \mathbb{R}^D$ with sample covariance matrix S. The d sample principal components are d unit vectors $u_i, u_i \in \mathbb{R}^D, i = 1, 2, ..., D, U = [u_1, ..., u_D]$ satisfying

$$\begin{aligned}
u_1 &= \arg\max_u u^T S u, \ s.t. \ u^T u = 1 \\
u_2 &= \arg\max_u u^T S u, \ s.t. \ u^T u = 1, u^T u_1 = 0 \\
&\cdots \\
u_d &= \arg\max_u u^T S u, \ s.t. \ u^T u = 1, u^T u_2 = 0, ..., u^T u_{d-1} = 0
\end{aligned}$$

Figure 14.2.1: Principal components for 2D samples.

Similar to Theorem 14.2.2, principal components $u_1, ..., u_d$ are top d eigenvectors of S defined by

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})(x_i - \overline{x})^T.$$

In addition, there are a number of critical properties regarding principal components and principal component scores.

Theorem 14.2.3. *Let $x_1, x_2, ..., x_N \in \mathbb{R}^D$ be a set of random samples. Let U of the matrix whose columns are the top d principal components of sample covariance matrix S. Let $\lambda_1, ..., \lambda_d$ be the top d eigenvalues. The principal component scores are given by $y_i = U^T x_i$. It follows that*

- *The sample covariance matrix Σ_y of $y_1, ..., y_N, y_i \in \mathbb{R}^d$ are diagonal. The diagonal terms are given by λ_i.*
- *The total variance of principal component scores is*

$$\sum_{i=1}^{N}(y_i - \overline{y})^T(y_i - \overline{y}) = (N-1)(\lambda_1 + \lambda_2 + \cdots \lambda_d),$$

where total variance of the original sample is

$$\Delta^2 = \sum_{i=1}^{N} (x_i - \overline{x})^T (x_i - \overline{x}) = (N-1)(\lambda_1 + \lambda_2 + \cdots \lambda_D).$$

Proof. (1) First note that

$$\begin{aligned}
\frac{1}{N-1} \sum_{i=1}^{N} (y_i - \overline{y})(y_i - \overline{y})^T &= \sum_{i=1}^{N} (U^T x_i - U^T \overline{x})(U^T x_i - U^T \overline{x})^T \\
&= \frac{1}{N-1} \sum_{i=1}^{N} U(x_i - \overline{x})^T (x_i - \overline{x})^T U^T \\
&= USU^T
\end{aligned}$$

To see the off diagonal terms are zero, we have $e_i^T USU^T e_j = u_i^T S u_j = \lambda_j u_i^T u_j = 0, i \neq j$. To see the diagonal terms, we have To see the off diagonal terms are zero, we have $e_i^T USU^T e_i = u_i^T S u_i = \lambda_i u_i^T u_i = \lambda_i$. To see the diagonal terms, we have

(2)

$$\begin{aligned}
&\sum_{i=1}^{N} (y_i - \overline{y})^T (y_i - \overline{y}) \\
&= \sum_{i=1}^{N} (U^T x_i - V^T \overline{x})^T (U^T x_i - U^T \overline{x}) \\
&= \sum_{i=1}^{N} (x_i - \overline{x})^T UU^T (x_i - \overline{x}) \\
&= \sum_{i=1}^{N} Tr((x_i - \overline{x})^T UU^T (x_i - \overline{x})) \\
&= \sum_{i=1}^{N} Tr(UU^T (x_i - \overline{x})(x_i - \overline{x})^T) \\
&= Tr(UU^T \sum_{i=1}^{N} (x_i - \overline{x})(x_i - \overline{x})^T) \\
&= (N-1) Tr(UU^T S) \\
&= Tr(U^T U \Lambda U^T U) \\
&= Tr(\Lambda) \\
&= \lambda_1 + \lambda_2 + \cdots \lambda_j
\end{aligned}$$

where we use matrix trace cyclic property [Lemma A.8.8]. □

Remark 14.2.1 (rank deficiency for high dimensional input data). When the input data $x_i \in \mathbb{R}^D$ is high dimensional such that $D \gg N$, the scattering matrix

$$S^2 = \sum_{i=1}^{N} (x_i - \overline{x})(x_i - \overline{x})^T,$$

will have rank at most $N - 1$ (when columns are linearly independent) or smaller (when there are linearly dependent columns).

14.2.2 Geometric fundamentals of PCA

14.2.2.1 *Optimization approach*

Consider a set of points $X = \{x_1, x_2, ..., x_N\}, x_i \in \mathbb{R}^D$ and assume they are given by

$$x_i = \mu + U_d y_i + \epsilon_i,$$

where $\mu \in \mathbb{R}^D, U_d \in \mathbb{R}^{D \times d}$ is a matrix with independent columns, $y_i \in \mathbb{R}^d$ is the linear combination coefficients, and $\epsilon_i \in \mathbb{R}^D$ is the additional noise.

The geometric picture of such representation is that data points are approximately lying on a low-dimensional affine space, characterized by shift μ and subspace basis U_d. The goal principal component analysis is to find $\mu, U_d, \{y_i\}$, when d is given, such that the sum of squared errors is minimized.

Remark 14.2.2 (redundancy in the representation). [4, p. 19]

- They are redundancy in the above representation because of the arbitrariness in the choice of μ and U. For example, $x_i = \mu + U_d y_i = (\mu + U_d y_0) + U_d(y_i - y_0)$. We can remove this translational ambiguity by requiring $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$; therefore we are effectively dealing with demeaned data.
- Another ambiguity is due to the arbitrariness in the choice of basis spanning the subspace. We can remove this ambiguity by enforcing orthonormality in the columns of U_d.

The principal component problem can be solved by the following optimization framework.

Lemma 14.2.1 (geometric PCA in the optimization framework). *Consider a set of points $X = \{x_1, x_2, ..., x_N\}, x_i \in \mathbb{R}^D$ and further assume they are demeaned such that $\frac{1}{N}\sum_{i=1}^{N} x_i = 0$.*

- *Then finding d orthonormal basis (i.e. U_d) and $y_i, i = 1, ..., N$ that minimizes the sum of squared errors is given by*

$$\min_{U_d, \{y_i\}} \|X - U_d Y\|_F^2 = \sum_{i=1}^{N} \|x_i - U_d y_i\|^2, s.t., U_d^T U_d = I_d.$$

- *The optimization problem can be alternatively formulated as*

$$\min_{U_d} \left\|X - U_d U_d^T X\right\|_F^2, s.t., U_d^T U_d = I_d.$$

- *The necessary condition for $y_i, i = 1, ..., N$ to achieve optimality is*

$$\hat{y}_i = U_d^T x_i.$$

Proof. The Lagrangian function is given by

$$L = \sum_{i=1}^{N} \|x_i - U_d y_i\|^2 + Tr((I_d - U^T U)\Lambda),$$

where $\Lambda \in \mathbb{R}^{d \times d}$ is the matrix of Lagrange multipliers. Then we have

$$\frac{\partial L}{\partial y_i} = 0 \implies -2U_d^T(x_i - U_d y_i) = 0 \implies y_i = U_d^T x_i,$$

where we use the fact that $U_d^T U_d = I_d$. □

Theorem 14.2.4 (PCA via SVD). *[4, p. 21]Let $X = [x_1, x_2, ..., x_N] \in \mathbb{R}^{D \times N}$ be the data matrix. Let $X = U\Sigma V^T$ be the singular value decomposition (SVD) of X. Then for any given $d < D$, a solution to PCA is the first d columns of$U$, given as $U_d = [u_1, u_2, ...u_d]$ and $\{y_i\}$ is the top $d \times N$ sub matrix $\Sigma_d V_d^T$ of the matrix ΣV^T (each column of length d is one y_i and in y_i each row is scaled in $\sqrt{\lambda_i}$).*

Proof.

$$\begin{aligned}
\left\|X - UU^T X\right\|_F^2 &= Tr((X - UU^T X)^T (X - UU^T X)) \\
&= Tr(X^T X) - Tr(XUU^T X) - Tr(UU^T X^T X) + Tr(X^T UU^T UU^T X) \\
&= Tr(X^T X) - 2Tr(XUU^T X) + Tr(X^T UU^T X) \\
&= Tr(X^T X) - Tr(X^T UU^T X)
\end{aligned}$$

Note that $Tr(XUU^T X) = Tr(U^T XX^T U) = \sum_{i=1}^{d} u_i^T XX^T u_i$. From Rayleigh quotient theorem [Theorem 4.8.4], we know that maximum of $Tr(U^T XX^T U$ is attained when u_i are the top d eigenvectors. □

Remark 14.2.3 (pitfalls for statistical approach and geometrical approach). Let X be the data matrix $X \in \mathbb{R}^{p\times N}$. In statistical approach, we calculate principal components by eigen-decomposition or SVD from sample covariance matrix of X, in which X **is demeaned**.

In the geometrical approach, if we directly perform SVD on X without demeaning X, we will get different results.

14.2.2.2 *Properties*

Theorem 14.2.5 (representation in the principal component space). *Let $X = [x_1, x_2, ..., x_N] \in \mathbb{R}^{D\times N}$ be the data matrix. Let $X = U\Sigma V^T$ be the singular value decomposition (SVD) of X.*

- *Then the coordinate vector of x_i in the basis of $U \in \mathbb{R}^{D\times p}$ is given by*

$$y_i = U^T x_i, y_i \in \mathbb{R}^p.$$

- *The matrix $P = UU^T$ is an orthonormal projection matrix that projects a vector in $\mathbb{R}^D$ to a subspace $S \subseteq \mathbb{R}^D$, where S has the basis U. We have recovery representation*

$$Uy_i = UU^T x_i = x_i.$$

- *If X has D independent columns, then $U \in \mathbb{R}^{D\times D}$, then $P = I$*

Proof. (2) Since $X = U\Lambda V^T$, then $x_i \in S$. (x_i can be written as a linear combination of columns vectors of U, and the coefficients are the i column of the matrix ΛV^T). Therefore $x_i = Px_i$. □

Theorem 14.2.6 (PCA distance and inner product preserving properties). *Let $X = [x_1, x_2, ..., x_N] \in \mathbb{R}^{D\times N}$ be the data matrix. Let $X = U\Sigma V^T$ be the singular value decomposition (SVD) of X. It follows that*

- *(distance preservation) For each data point x_i, let $y_i = U^T x_i$ be the (new) coordinates projected on the eigen-basis.*

$$\left\|y_i - y_j\right\|^2 = \left\|x_i - x_j\right\|^2.$$

- *Let U_d be the matrix whose columns are the top d eigenvectors, and let $y_i = U_d^T x_i$. Denote $P_d = U_d U_d^T$. Then*

$$\left\|y_i - y_j\right\|^2 = \left\|P_d(x_i - x_j)\right\|^2 \leq \left\|P_d(x_i - x_j)\right\|^2.$$

- *(beat d rank approximation to inner product matrix) Let $y_i = U_d^T x_i$. Denote $Y = [y_1, ..., y_N]$. Then $Y^T Y$ is the best d rank approximation to $X^T X$ Moreover, $\left\|Y^T Y\right\|_F^2 = \sum_i = 1^d \sigma_i^2$, $\left\|X^T X\right\|_F^2 = \sum_i = 1^D \sigma_i^2$*

Proof. (1)

$$\begin{aligned}\left\|y_i - y_j\right\|^2 &= (y_i - y_j)^T (y_i - y_j) \\ &= (U^T x_i - U^T x_j)^T (U^T x_i - U^T x_j) \\ &= (x_i - x_j)^T U U^T (x_i - x_j) \\ &= (x_i - x_j)^T (x_i - x_j)\end{aligned}$$

where we used the factor that $x_i, x_j, x_i - x_j$ are all lying inside the subspace S spanned by U, and the orthogonal projection matrix to S is $P = UU^T$. Therefore $P(x_i - x_j) = (x_i - x_j)$. (2) Similar to (1). We then use matrix norm inequality [Theorem 4.13.2] that gives

$$\left\|A(x_i - x_j)\right\|^2 \leq \|A\|^2 \left\|x_i - x_j\right\|^2 = \sigma_1(P_d) \left\|x_i - x_j\right\|^2 = \left\|x_i - x_j\right\|^2,$$

where we use the fact that largest eigenvalue of orthogonal projector matrix P_d is 1. (3) Note that $Y^T Y = X^T U_d U_d^T X = V_d^T \Sigma_d^2 V_d$, and $X^T X$ has SVD $X^T X = V^T \Sigma^2 V$. Therefore, $Y^T Y$ is the best d rank approximation to $X^T X$.

The conclusion that $\left\|Y^T Y\right\|_F^2 = \sum_i = 1^d \sigma_i^2$ is from the relationship between SVD and Frobenius norm [Theorem 4.9.2].

□

Remark 14.2.4. It is not possible to derive a lower bound for the ration $\frac{\left\|y_i - y_j\right\|^2}{\left\|x_i - x_j\right\|^2}$ because if $x_i - x_j$ lies in the null space of P_d, then the ratio can be zero.

14.2.3 Probabilistic PCA

In the previous sections, we have addressed two approaches to PCA, one is statistical approach that seeks principal components that preserve variation and another is geometry

approach that view principal components as the orthogonal basis of a low-dimensional affine subspace which the data primarily lie in. This section, we introduce probabilistic PCA [5], which aims to characterize data structure from probabilistic data generation perspective.

Probabilistic PCA offers several flexibility in practical use of PCA, including

- Multiple probabilistic PCA models can be combined as a probabilistic mixture.
- Maximum-likelihood estimates can be computed for elements associated with principal components.
- Probabilistic PCA can handle missing data or incomplete data in a more formal probabilistic framework, rather than using ad-hoc methods.
- Probabilistic PCA can be used to generate new data samples.

Suppose we are given N data points $\{x_1, ..., x_N\}, x_i \in \mathbb{R}^D$, **probabilistic PCA** specifies a data generation model for X, given by

- $X \sim WY + \mu + \epsilon$, where $W \in \mathbb{R}^{D\times d}$, $Y \in MN(0, I_d), \epsilon \sim MN(0, \sigma^2 I), \mu \in \mathbb{R}^n$.
- $X \sim MN(\mu, C_Y), C_Y = WW^T + \sigma^2 I$

Usually, we call $Y = (Y_1, ..., Y_d)$ **latent factors**.

Remark 14.2.5 (interpretation).

- Probabilistic PCA can be viewed a latent factor model with X being the latent variable.
- Normal PCA is a limiting case of probabilistic PCA, taken as the limit as the covariance of the noise becomes infinitesimally small ($\sigma^2 \to 0$).

One goal in probabilistic PCA is to estimate parameters W, μ, σ.

Lemma 14.2.2 (likelihood function). *Assume $x_1, x_2, ..., x_n, \in x_i \in \mathbb{R}^D$ are independent random samples drawn from a multivariate normal distribution (μ, Σ). Then likelihood function is given by*

$$L(\mu, \Sigma) = \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}} \exp(-Tr(\Sigma^{-1}(\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T))).$$

Proof.

$$\begin{aligned} L(\mu, \Sigma) &= \frac{1}{(2\pi)^{nD/2}|\Sigma|^{n/2}} \exp(-((\sum_{i=1}^{n}(x_i - \mu)\Sigma^{-1}(x_i - \mu)^T))) \\ &= \frac{1}{(2\pi)^{nD/2}|\Sigma|^{n/2}} \exp(-Tr(\Sigma^{-1}(\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T))) \end{aligned}$$

Note that we use $Tr(x^T Ax) = Tr(Axx^T)$ [Lemma A.8.8]. □

Similar to Lemma 14.1.6 and follow **[??]**, we can show

$$\hat{W} = U_d \left(\Lambda_d - \sigma^2 \mathbf{I}\right)^{1/2} R$$

where U_d consists the top d eigenvectors of S, and the top d corresponding eigenvalues are in the diagonal matrix Λ_q, and R is an arbitrary $d \times d$ orthogonal rotation matrix.

And

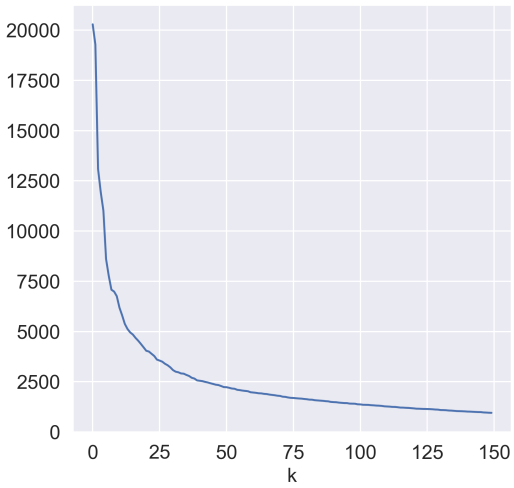
$$\hat{\sigma}^2 = \frac{1}{D-d} \sum_{j=d+1}^{D} \lambda_j.$$

14.2.4 Applications

14.2.4.1 *Eigenfaces and eigendigits*

We first apply PCA to the face image from Labeled Faces in the Wild to extract top 'eigenfaces', which are faces can be used to synthesize the majority of human faces by linear combination. We perform PCA on around one thousand randomly selected face images. Each image has 50×37 pixels. The PCA results are in Figure 14.2.2. Note that the first several components capture the major variations of image data. On the other hand, Components associated with smaller singular values mostly captures noise.

(a) Raw face image examples.

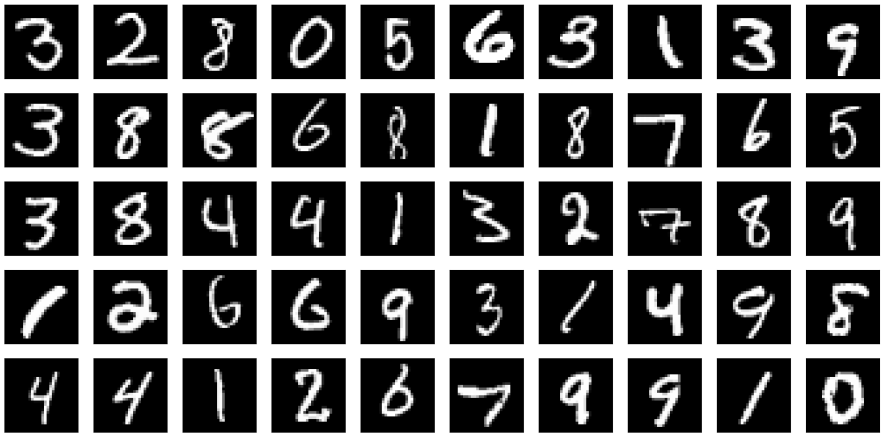


(b) Singular value spectrum.

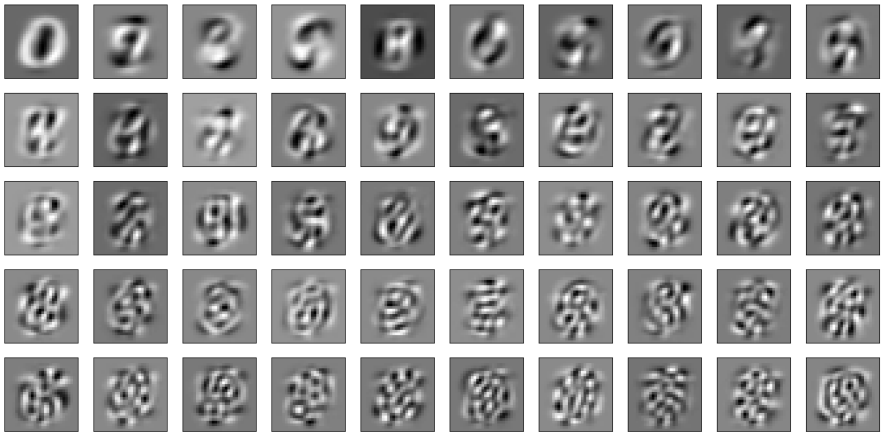
(c) Top eigenfaces.

Figure 14.2.2: PCA eigenface analysis.

Similar PCA on the MNIST data set is showed in Figure 14.2.3.



(a) Raw digit image examples.



(b) Top 50 eigen-digits.

Figure 14.2.3: PCA eigen-digit analysis for MNIST dataset.

14.2.4.2 *Interest rate curve dynamics modeling*

PCA method has found important application in financial industry. One example is to modeling the interest rate curve dynamics. It is common knowledge that in borrowing and lending business, the interest rate we borrow or lend is not a constant independent of

the length of loan. Instead, the interest rate is a function of life of the loan, which takes into account the economy, credit risk and other factors.

At every date t, the interest rates can be represented by a curve, as showed in Figure 14.2.4, whose x is swap tenor approximating the meaning the loan life and y axis is the borrowing or lending rate.

At different t, we observe different curve shape. Usually, the interest rate curve is represented by a vector of points on the curve, denoted by $y \in \mathbb{R}^N$, where N is the number of different tenors.

One brute force way is to model the curve dynamics is to model $y_1, ...y_N$ altogether, leading to a N variable model like

$$y_i(t + dt) = y_i(t) + \sigma_i z_i, i = 1, ..., N,$$

where z_i is zero mean random shock, and $z_1, ..., z_N$ has certain joint distribution to capture the correlation structure in $y_1, ..., y_N$. The full model is usually impractical because the difficulties in accurately estimating covariance structure of $y_1, ..., y_N$ when N is large.

Another low-dimensional modeling approach is to exploit that the high correlation nature among $y_1, ..., y_N$ and seek some model like

$$y_i(t + dt) = y_i(t) + \sum_{j=1}^{k} b_i^{(j)} z_j, i = 1, ..., N,$$

where $z_1, ..., z_k$ are k zero mean random shock **with zero correlation**, and $b^{(1)}, ..., b^{(k)} \in \mathbb{R}^N$ are top k eigenvectors obtained for historical covariance matrix of $y_i(t + dt) - y_i(t)$. Using PCA can lead to much more robust estimation that capture the covariance structure.

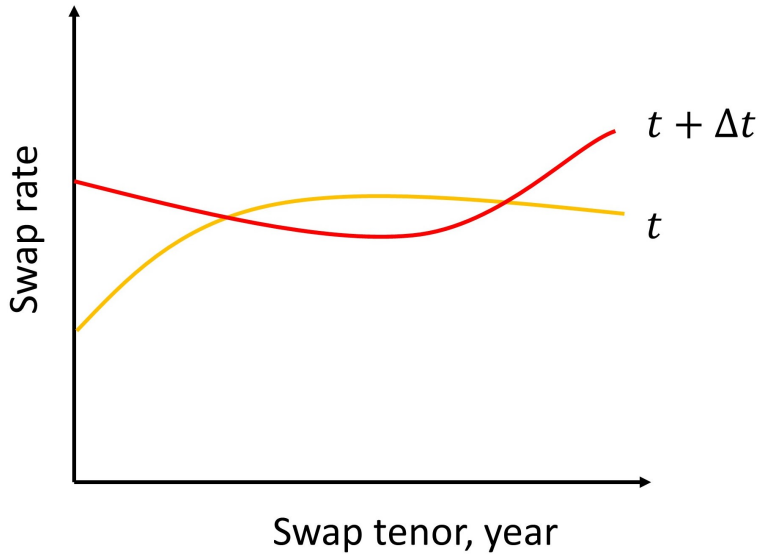


Figure 14.2.4: Demonstration of interest rate curve dynamics.

As a demonstration, we now analyze **daily interest rates** with maturities of 1 year, 2 years, 3 years, 4 years, 5 years, 7 years, 10 years, and 30 years observing between 2000 and 2011. Table 14.2.1 shows the PCA factors and the associated eigenvectors. Figure 14.2.5 plots the first three dominating PCA factors. The first eigenvector/factor has deeper interpretations, they represent the parallel shift mode, the steepening model, and the bending mode that dominating curve changing dynamics.

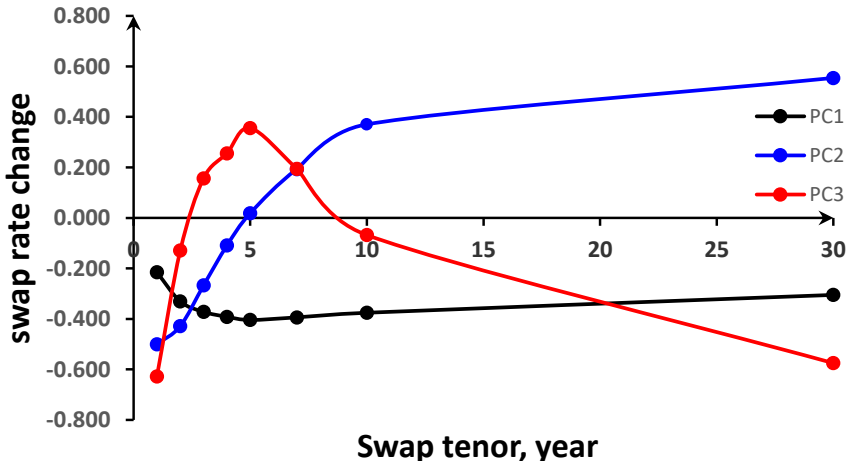


Figure 14.2.5: Demonstration of first three dominating PCA factor in the swap rate curve daily change.

Table 14.2.1: Eigenvectors and eigenvalues for swap rate daily change

(a) Eigenvectors for swap rate daily change

	PC1	PC2	PC3	···	PC8
1Y	0.216	-0.501	0.627	···	-0.034
2Y	0.331	-0.429	0.129	···	0.236
3Y	0.372	-0.267	-0.157	···	-0.564
4Y	0.392	-0.110	-0.256	···	0.512
5Y	0.404	0.019	-0.355	···	-0.327
7Y	0.394	0.194	-0.195	···	0.422
10Y	0.376	0.371	0.068	···	-0.279
30Y	0.305	0.554	0.575	···	0.032

(b) Eigenvalues for swap rate daily change

	PC1	**PC2**	**PC3**	···	**PC8**
eigenvalue	4.77	2.08	1.29	···	-0.034

14.3 Canonical correlation analysis

14.3.1 Basics

CCA looks for linear combinations of variables in the two groups that are highly correlated with each other. Consider an example that we want to study the factors affecting student's academic performance. Let X be the random vector of activities from N students. Each one quantifies the time spents on sports, electronics, reading, and homework. Let Y be the random vector of academic performance from N students. Each one quantifies the performance on reading, math, and language. In CCA, we seek vectors a and b such that $Corr(a^T X, b^T Y)$ is maximized. The results are useful in the following ways:

- Examine the overall relationship between the two set of variables. Suppose $\max Corr(a^T X, b^T Y)$ is low, then it suggests that the two set of variables are statistically irrelevant, i.e., activities are not affecting academic performance.
- Examine the coefficients a and b to understand relationships between individual components. Suppose the result is $0.6 * Reading + 0.9 * Math + 0.7 * Logic$ is highly correlated with $0.5 * Sports - 2 * Electronics + 0.8 * Reading + 5 * Homework$, then we can understand that spending too much time in electronics will negatively affect academic performance and spending more time on homework will have positive impact.
- Seek low dimensional representations. Following previous point, we can view $0.6 * Reading + 0.9 * Math + 0.7 * Logic$ as a new quantity that characterize academic performance and $0.5 * Sports - 2 * Electronics + 0.8 * Reading + 5 * Homework$ as a new quantity characterizing activities.

More formally, given two column vector of random variables $X = (X_1, X_2, ..., X_p)$ and $Y = (Y_1, Y_2, ..., Y_q)$ with finite second moments. **Canonical correlation analysis** seeks vector $a \in \mathbb{R}^m$ and $b \in \mathbb{R}^n$ such that the random variable $a^T X$ and $b^T Y$ has the maximum correlation. More formally, we want to solve

$$\max_{a \in \mathbb{R}^p, b \in \mathbb{R}^q} corr(a^T X, b^T Y) = \max_{a \in \mathbb{R}^p, b \in \mathbb{R}^q} \frac{a^T \Sigma_{XY} b}{\sqrt{a^T \Sigma_{XX} a} \sqrt{b^T \Sigma_{XX} b}}.$$

To understand the maximization problem better, we first examine its equivalent optimization problems before we proceed to its solution.

Lemma 14.3.1 (equivalent optimization problems). *Let $c = \Sigma_{XX}^{1/2} a, d = \Sigma_{YY}^{1/2} b$, then the original optimization problem can be written by*

$$\max_{c \in \mathbb{R}^p, d \in \mathbb{R}^q} \frac{c^T \Sigma_{XX}^{-1/2} \Sigma_{XY} \sigma_{YY}^{-1/2} d}{\sqrt{c^T c}\sqrt{d^T d}}.$$

We can equivalently write the unconstrained optimization as

$$\max_{c \in \mathbb{R}^p, d \in \mathbb{R}^q} c^T \Sigma_{XX}^{-1/2} \Sigma_{XY} \sigma_{YY}^{-1/2} d$$
$$\text{subject to } c^T c = 1, d^T d = 1$$

Proof. Straight forward. □

Theorem 14.3.1 (solution to canonical correlation problem).

- *The optimizer of the transformed optimization problem is given by*
  - *c is eigenvector of $\Sigma_{XX}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{YX} \Sigma_{XX}^{-1/2}$ associated with the largest eigenvalue.*
  - *d is proportional to $\Sigma_{YY}^{-1/2} \Sigma_{YX} \Sigma_{XX}^{-1/2} c$*

  Or reciprocally,
  - *d is eigenvector of $\Sigma_{YY}^{-1/2} \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XY} \Sigma_{YY}^{-1/2}$ associated with the largest eigenvalue.*
  - *c is proportional to $\Sigma_{XX}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1/2} d$*
- *The optimizer of the original optimizer problem is*
  - *a is eigenvector of $\Sigma_{XX}^{-1} \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{YX}$ associated with the largest eigenvalue.*
  - *b is proportional to $\Sigma_{YY}^{-1} \Sigma_{YX} c$*

  Or reciprocally,
  - *b is eigenvector of $\Sigma_{YY}^{-1} \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XY}$ associated with the largest eigenvalue.*
  - *a is proportional to $\Sigma_{XX}^{-1} \Sigma_{XY} b$*

Proof. Using Cauchy-Schwartz inequality [Theorem 11.9.4],

$$\begin{aligned}&(c^T \Sigma_{XX}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1/2}) d \\ \leq &(c^T \Sigma_{XX}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{YX})^{1/2} (d^T d)^{1/2}\end{aligned}$$

Therefore,

$$\rho \leq \frac{(c^T \Sigma_{XX}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{YX})^{1/2}}{(c^T c)^{1/2}}$$

Base on Rayleigh quotient [Theorem 4.8.4], ρ will take the maximum value when c is eigenvector of $\Sigma_{XX}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{YX} \Sigma_{XX}^{-1/2}$ associated with the largest eigenvalue.

□

Corollary 14.3.1.1 (solution when components are standardized). *Suppose $\Sigma_{XX} = I$ and $\Sigma_{YY} = I$. Then the solution to the following optimization problem is given by*

$$\max_{a\in\mathbb{R}^m, b\in\mathbb{R}^n} corr(a^T X, b^T Y) = \max_{a\in\mathbb{R}^m, b\in\mathbb{R}^n} \frac{a^T\Sigma_{XY}b}{\sqrt{a^T\Sigma_{XX}a}\sqrt{b^T\Sigma_{XX}b}}.$$

- *a is eigenvector of $\Sigma_{XY}\Sigma_{YX}$ associated with the largest eigenvalue.*
- *b is eigenvector of $\Sigma_{YX}\Sigma_{XY}$ associated with the largest eigenvalue.*

In practice, we collect n random samples. Let $X \in \mathbb{R}^{n\times p}, Y \in \mathbb{R}^{n\times q}$ be the **centered** data matrix. Then the optimization will be maximizing sample correlations, which is given by

$$\begin{aligned}&\max_{c\in\mathbb{R}^p, d\in\mathbb{R}^q} c^T(X^TX)^{-1/2}(X^TY)(Y^TY)^{-1/2}d\\ &subject\ to c^Tc = 1, d^Td = 1\end{aligned}$$

14.3.2 Sparse CCA

In some applications such as bioinfomatics [6], we often encounter the situation where $\min(p, q) \gg n$. This leads to the case where $(X^TX)^{-1}$ and $(Y^TY)^{-1}$ do not exist. As a result, we need to use the original optimization form. Further, we might want to induce sparsity in a, b. Eventually, our optimization problem becomes

$$\begin{aligned}&\max_{a\in\mathbb{R}^p, b\in\mathbb{R}^q} a^TX^TYb\\ &subject\ to\ a^TX^TXa = 1, b^TY^TYb = 1,\\ &\|a\|_1 \le c_1, \|b\|_1 \le c_2.\end{aligned}$$

We can relax the constraint to yield a convex optimization given by

$$\begin{aligned}&\max_{a\in\mathbb{R}^p, b\in\mathbb{R}^q} a^TX^TYb\\ &subject\ to\ a^TX^TXa \le 1, b^TY^TYb \le 1,\\ &\|a\|_1 \le c_1, \|b\|_1 \le c_2.\end{aligned}$$

Note that this optimization problem is bi-convex. That is, when we fix a, the optimization is convex with respect to b, and vice versa.

This optimization is derived from the well-known as penalized matrix factorization problem[7, 8], and have a general iterative algorithm below [algorithm 22].

Algorithm 22: Alternating sparse canonical correlation analysis algorithm

Input: Data matrix X, Y
Initialize a to have $a^T XXa = 1$.
repeat
  Solve $b = \arg\max_b a^T X^T Yb$, subject to $b^T Y^T Yb \leq 1, \|b\|_1 \leq c_2$.
  Solve $a = \arg\max_a a^T X^T Yb$, subject to $a^T X^T Xa \leq 1, \|a\|_1 \leq c_1$.
until *terminal condition is met*;
Output: a, b

14.4 Copulas and dependence modeling

14.4.1 Definitions and properties

Definition 14.4.1. *A **copulas** $C : [0,1]^d \to [0,1]$ is a multivariate CDF of a $d-$dimensional random vector on the unit cube whose univariate marginal distributions are all $U(0,1)$.*

*The **density of a copulas**, denoted by c, is given by,*

$$c(u_1, u_2, ..., u_d) = \frac{\partial^d}{\partial u_1 \partial u_2 ... \partial u_d} C(u_1, u_2, ..., u_d).$$

Note 14.4.1 (marginal distribution function from joint distribution function)**.**

$$F_1(y_1) = \lim_{y_2,...,y_d \to \infty} F(y_1, y_2, ..., y_d)$$

Lemma 14.4.1 (basic properties of Copulas).

- $$C(u_1, ..., u_{i-1}, 0, u_{i+1}, ..., u_d) = 0$$
- $$C(1, ..., 1, u, 1, ..., 1) = u$$
- *C is non-decreasing in each of the d variables.*
- *(marginalization property) Let $C_{1:d}(u_1, u_2, ..., u_d)$ be copula(i.e., joint cdf) associated with uniform random variables $U_1, U_2, ..., U_d$, then*

$$C_{1:k}(u_1, u_2, ..., u_k) \triangleq C_{1:d}(u_1, u_2, ..., u_k, 1, ..., 1), k < d,$$

is the marginal cdf associated with uniform random variables $U_1, U_2, ..., U_k$.

Proof. Directly from the property of multivariate cdf. □

Corollary 14.4.0.1. *Let $C : [0,1] \times [0,1] \to [0,1]$ be a bivariate copula. Then*

$$C(0,u) = C(u,0) = 0, C(1,u) = C(u,1) = u$$

and

$$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0, \forall 0 \leq u_1 \leq u_2 \leq 1, 0 \leq v_1 \leq v_2 \leq 1.$$

Lemma 14.4.2 (probability integral transform for a random variable). *Let X be a random variable with support $\mathbb{R}$ and let $F_X : \mathbb{R} \to [0,1]$ be its cdf.*

It follows that

- *The new random variable $Y = F_X(X)$ has an uniform distribution.*
- *Let U be a uniform random variable on $[0,1]$, then*

$$X = F_X^{-1}(U).$$

- *Let ϕ be the cdf of a standard normal, then the random variable defined by*

$$Z = \phi^{-1}(F_X(X)).$$

is a Gaussian random variable.

Proof. (1)Note that

$$P(Y \leq y) = Pr(F_X(X) \leq y) = Pr(X \leq F_X^{-1}(y)) = F_X[F_X^{-1}(y)] = y.$$

(2) Note that $Pr(X < x) = Pr(F_X^{-1}(U) < x) = Pr(U < F_X(x)) = F_X(x)$. where we use the fact that $Pr(U < y) = y, \forall y \in [0,1]$. (3) use (1) and (2). □

Example 14.4.1. If X has an exponential distribution with unit mean, then

$$F_X(x) = 1 - \exp(-x).$$

It follows that the random variable Y, defined as

$$Y = 1 - \exp(-X)$$

has a uniform distribution.

Lemma 14.4.3 (probability transform for a random vector and its properties). *Consider a random vector $(X_1, X_2, ..., X_d)$. Suppose its marginal cdfs $F_i(x) = P(X_i \leq x)$ are continuous functions. Then the random vector*

$$(Y_1, Y_2, ..., Y_d) \triangleq (F_1(X_1), F_2(X_2), ..., F_d(X_d))$$

has uniformly distributed marginals. Moreover, the joint cdf of $(Y_1, Y_2, ..., Y_d)$ *is the copulas associated with the joint cdf of* $(X_1, X_2, ..., X_d)$.

Proof. (1) Note that from the definition of marginal cdf, we have

$$P(Y_1 < y_1, Y_2 < \infty, ..., Y_d < \infty) = F_{Y_1, Y_2, ..., Y_d}(y_1, \infty, ..., \infty) = F_{Y_1}(y_1).$$

$F_{Y1}(y_1)$ is a uniform distributed cdf, as showed in Lemma 14.4.2. (2) (a)To show the copula is the same we can use monotone transformation invariance Lemma 14.4.7 since F_i is increasing;(b)

$$\begin{aligned} &Pr(F_1(x_1) < u_1, ..., F_d(x_d) < u_d) \\ =&Pr(x_1 < F_1^{-1}(u_1), ..., x_d < F_d^{-1}(u_d)) \\ =&F_X(F_1^{-1}(u_1), ..., F_d^{-1}(u_d)) \end{aligned}$$

That is, the joint cdf of $(Y_1, Y_2, ..., Y_d)$is the copulas associated with the joint cdf of $(X_1, X_2, ..., X_d)$ [Theorem 14.4.2]. □

Lemma 14.4.4 (Frechet-Hoeffding bounding). *[3, p. 189]*

$$W(u_1, u_2, ..., u_d) \leq C(u_1, u_2, ..., u_d) \leq M(u_1, ..., u_d)$$

where

$$W(u_1, u_2, ..., u_d) = \max(1 - d + \sum_{i=1}^{d} u_i, 0)$$

and

$$M(u_1, u_2, ..., u_d) = \min(u_1, u_2, ..., u_d)$$

In particular, for $d = 2$, *we have*

$$\max(u + v - 1, 0) \leq C(u, v) \leq \min(u, v).$$

Proof. (1)

(2) For the lower bound, we have

$$\begin{aligned} C(u_1, ..., u_d) &= Pr(\cap_{1\leq i\leq d}\{U_i \leq u_i\}) \\ &= 1 - Pr(\cup_{1\leq i\leq d}\{U_i > u_i\}) \\ &= 1 - \sum_{i=1}^{d} Pr(U_i > u_i) \\ &= 1 - \sum_{i=1}^{d}(1 - u_i) \\ &= 1 - d + \sum_{i=1}^{d} u_i \end{aligned}$$

where we use Demorgan's law [Lemma 1.1.2] and union bound. □

Corollary 14.4.0.2. *For a multivariate joint cdf F with margins $F_1, F_2, ..., F_d$, we have*

$$\max(1 - d + \sum_{i=1}^{d} F_i(x_i), 0) \leq F(x_1, ..., x_d) \leq \min(F_1(x_1), ..., F_d(x_d))$$

Proof. Note that

$$W(F_1(x_1), ..., F_d(x_d)) \leq C(F_1(x_1), ..., F_d(x_d)) \leq M(F_1(x_1), ..., F_d(x_d)),$$

and

$$C(F_1(x_1), ..., F_d(x_d)) = F(x_1, ..., x_d).$$

□

Example 14.4.2 (Gaussian copula).

$$C_\rho(u, v) = \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \frac{1}{2\pi(1-\rho^2)^{1/2}} \exp(-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)})dxdy$$

where ρ is the linear correlation coefficient.

Example 14.4.3 (Student's t copula).

$$C_{\rho,\nu}(u, v) = \int_{-\infty}^{t^{-1}(u)} \int_{-\infty}^{t^{-1}(v)} \frac{1}{2\pi(1-\rho^2)^{1/2}} \exp(1 + \frac{x^2 - 2\rho xy + y^2}{\nu(1-\rho^2)})^{-(\nu+2)/2}dxdy$$

where ρ is the linear correlation coefficient.

14.4.2 Copulas and distributions

Note 14.4.2. In this section we define inverse function for a cdf as

$$F^{-1}(t) = \inf\{x \in \mathbb{R} : F(x) \geq t\}.$$

14.4.2.1 *Fundamentals*

Definition 14.4.2 (copula associated with a distribution). *[9, p. 3] Let $F(x_1, x_2, ..., x_n)$ be a cdf of a random vector $(X_1, X_2, ..., X_n)$ with support $\mathbb{R}^n$. Let $F_1, F_2, ..., F_n$ be the corresponding marginal cdf. The copula $C(u_1, u_2, ..., u_n)$ associated with F is such that*

$$F(x_1, x_2, ..., x_n) = C(F_1(x_1), F_2(x_2), ..., F_n(x_n)).$$

Theorem 14.4.1 (construct a multivariate cdf from a copulas and margins). *Consider d random variables $X_1, X_2, ..., X_d$ with $F_1, F_2, ..., F_d$ being the univariate cdf. Let C be a d dimensional copulas, then we can construct a d-dimensional multivariate cdf as*

$$F(x_1, x_2, ..., x_n) = C(F_1(x_1), F_2(x_2), ..., F_d(x_d))$$

such that the marginal cdf is given by $F_1, F_2, ..., F_d$.

Proof. (1)Use the definition of marginal cdf and Lemma 14.4.1, we have

$$\begin{aligned} F_1(x_1) &\triangleq F(x_1, \infty, ..., \infty) \\ &= C(F_1(x_1), F_2(\infty), ..., F_n(\infty)) \\ &= F_1(x_1). \end{aligned}$$

□

Remark 14.4.1 (back out the copula from constructed cdf). Note that we construct a new cdf F from a given copula $C(u_1, u_2, ..., u_d), u_i \in [0, 1]$ via

$$F(x_1, x_2, ..., x_n) = C(F_1(x_1), F_2(x_2), ..., F_d(x_d)).$$

On the other hand, Theorem 14.4.2 gives a way to back out the original copula from joint cdf via

$$\begin{aligned} C_n(u_1, u_2, ..., u_d) &= F(F_1^{-1}(u_1), F_2^{-1}(u_2), ..., F_d^{-1}(u_d)) \\ &= C_o(F_1F_1^{-1}(u_1), F_2F_2^{-1}(u_2), ..., F_dF_d^{-1}(u_d)) \\ &= C_o(F_1F_1^{-1}(u_1), F_2F_2^{-1}(u_2), ..., F_dF_d^{-1}(u_d)) \\ &= C_o(u_1, u_2, ..., u_d) \end{aligned}$$

where $u_1, u_2, ..., u_d \in [0, 1]$ and F is given by

$$F(x_1, x_2, ..., x_n) = C_o(F_1(x_1), F_2(x_2), ..., F_d(x_d)).$$

Note 14.4.3 (caution!).

- It is **incorrect** that univariate margins and correlation matrix will allow the construction of the multivariate joint cdf.
- It is **incorrect** that univariate Gaussian margins and correlation matrix will allow the construction of the multivariate joint cdf.
- It is **only correct** that, **if we know the joint distribution is multivariate Gaussian**, univariate Gaussian margins and correlation matrix will allow the construction of the multivariate joint cdf.

Theorem 14.4.2 (construct a copulas for a joint distribution; any continuous multivariate cdf has a copula). *Consider a random vector $(X_1, X_2, ..., X_d)$ with continuous cdf F and marginal cdf $F_1, ..., F_d$. The copula for this random vector $(X_1, X_2, ..., X_d)$ is defined as*

$$C(u_1, u_2, ..., u_d) = F(F_1^{-1}(u_1), F_2^{-1}(u_2), ..., F_d^{-1}(u_d))$$

where $u_1, u_2, ..., u_d \in [0, 1]$.

Proof. Based on the definition of a copula associated with a cdf, we want to show that $F(x_1, x_2, ..., x_d) = C(F_1(x_1), F_2(x_2), ..., F_d(x_d))$. We have

$$\begin{aligned} C(u_1, u_2, ..., u_d) &= F(F_1^{-1}(u_1), F_2^{-1}(u_2), ..., F_d^{-1}(u_d)) \\ \implies C(F_1(u_1), F_2(u_2), ..., F_d(u_d)) &= F(F_1^{-1}(F_1(u_1)), F_2^{-1}(F_2(u_2)), ..., F_d^{-1}(F_d(u_d))) \\ &= F(F_1^{-1}(F_1(u_1)), F_2^{-1}(F_2(u_2)), ..., F_d^{-1}(F_d(u_d))) \\ &= F(u_1, u_2, ..., u_n). \end{aligned}$$

□

Lemma 14.4.5 (copula of a uniform distribution). *For d-dimensional uniform cdf F, its copulas $C = F$.*

Proof. Note that

$$C(u_1, u_2, ..., u_d) = F(F_1^{-1}(u_1), F_2^{-1}(u_2), ..., F_d^{-1}(u_d)) = F(u_1, u_2, ..., u_d).$$

since for all marginals, $F_i(x) = x$. □

Remark 14.4.2. Given a copula, we can obtain different multivariate cdf by selecting different marginal distribution.

Theorem 14.4.3 (Sklar's theorem, construct joint cdf and pdf from copula and margins). *For every multivariate cumulative distribution function*

$$H(x_1, ..., x_d) \triangleq P(X_1 \leq x_1, ..., X_d \leq x_d).$$

of a random vector $(X_1, X_2, ..., X_d)$, there exists a copula C such that

$$H(x_1, ..., x_d) = C(F_1(x_1), ..., F_d(x_d)).$$

If H has a density h, then

$$h(x_1, ..., x_d) = c(F_1(x_1), ..., F_d(x_d)) \cdot f_1(x_1) \cdots f_d(x_d),$$

or equivalently,

$$c(x_1, ..., x_d) = \frac{h(F_1(x_1), ..., F_d(x_d))}{f_1(x_1) \cdots f_d(x_d)},$$

where $f_1, f_2, ..., f_d$ are marginal density functions.

Proof. (1) directly from Theorem 14.4.2.(2) Use chain rule. Note that

$$h = \frac{\partial^d H}{\partial x_1 \partial x_2 ... \partial x_d},$$

and

$$\frac{\partial H}{\partial x_1} = \frac{\partial C}{\partial x_1} = \frac{\partial C}{\partial F_1} \frac{\partial F_1}{\partial x_1}.$$

Therefore,

$$h = \frac{\partial^d H}{\partial x_1 \partial x_2 ... \partial x_d} = \frac{\partial^d C}{\partial F_1 \partial F_2 ... \partial F_d} \frac{\partial F_1}{\partial x_1} \frac{\partial F_2}{\partial x_2} ... \frac{\partial F_d}{\partial x_d} = c(F_1, F_2, ..., F_d) f_1 \cdot f_2 \cdots f_d.$$

□

Remark 14.4.3. The marginal distribution and the copulas can be modeled and estimated separately and independently.

Example 14.4.4 (joint density with Gaussian copula). Let $c(u_1, u_2, ..., u_d)$ be a Gaussian copula given by

$$c(u_1, u_2, ..., u_d) = \frac{1}{\sqrt{det(R)}} \exp(-\frac{1}{2} \begin{bmatrix} \phi^{-1}(u_1) \\ \vdots \\ \phi^{-1}(u_d) \end{bmatrix}^T \cdot (R^{-1} - I) \begin{bmatrix} \phi^{-1}(u_1) \\ \vdots \\ \phi^{-1}(u_d) \end{bmatrix}),$$

where ϕ is the cdf for a standard normal variable.

- If we have Gaussian margins all given by ϕ, then we have the joint pdf given by

$$\begin{aligned} & h(x_1, x_2, ..., x_d) \\ & = c(\phi(x_1), \phi(x_2), ..., \phi(x_d))\phi(x_1)\phi(x_2)\cdots\phi(x_d) \\ & = \frac{1}{\sqrt{det(R)}} \exp(-\frac{1}{2} \begin{bmatrix} \phi^{-1}(\phi(x_1)) \\ \vdots \\ \phi^{-1}(\phi(x_d)) \end{bmatrix}^T \cdot (R^{-1} - I) \begin{bmatrix} \phi^{-1}(\phi(x_1)) \\ \vdots \\ \phi^{-1}(\phi(x_d)) \end{bmatrix}) \frac{1}{(2\pi)^{d/2}} \exp(-\frac{1}{2} \sum_{i=1}^{d} x_i^2) \\ & = \frac{1}{\sqrt{det(R)}} \exp(-\frac{1}{2} x^T R^{-1} x) \end{aligned}$$

- Let $\phi(x)$ be the standard Gaussian cdf, then a Gaussian random variable $N(m, \sigma^2)$ has pdf given by $\phi(\frac{x-m}{\sigma})$.

Then,

$$
\begin{aligned}
& h(x_1, ..., x_d) \\
& = c(F_1(x_1), ..., F_d(x_d)) \cdot f_1(x_1) \cdots f_d(x_d) \\
& = \frac{1}{\sqrt{det(R)}} \exp(-\frac{1}{2} \begin{bmatrix} \phi^{-1}(\phi(x_1 - m_1/\sigma_1)) \\ \vdots \\ \phi^{-1}(\phi(x_d - m_d/\sigma_d)) \end{bmatrix}^T \\
& \cdot (R^{-1} - I) \begin{bmatrix} \phi^{-1}(\phi(x_1 - m_1/\sigma_1)) \\ \vdots \\ \phi^{-1}(\phi(x_d - m_d/\sigma_d)) \end{bmatrix}) \\
& \frac{1}{(2\pi)^{d/2}} \exp(\begin{bmatrix} (x_1 - m_1/\sigma_1)) \\ \vdots \\ (x_d - m_d/\sigma_d)) \end{bmatrix}^T \cdot \begin{bmatrix} (x_1 - m_1/\sigma_1)) \\ \vdots \\ (x_d - m_d/\sigma_d)) \end{bmatrix}) \\
& = \frac{1}{\sqrt{det(R)}} \exp(-\frac{1}{2} \begin{bmatrix} (x_1 - m_1/\sigma_1) \\ \vdots \\ (x_d - m_d/\sigma_d) \end{bmatrix}^T \cdot (R^{-1}) \begin{bmatrix} (x_1 - m_1/\sigma_1) \\ \vdots \\ (x_d - m_d/\sigma_d) \end{bmatrix}) \frac{1}{(2\pi)^{d/2}} \\
& = \frac{1}{\sqrt{det(R)}(2\pi)^{d/2}} \exp(-\frac{1}{2} \begin{bmatrix} (x_1 - m_1/\sigma_1) \\ \vdots \\ (x_d - m_d/\sigma_d) \end{bmatrix}^T \cdot (\Sigma^{-1}) \begin{bmatrix} (x_1 - m_1/\sigma_1) \\ \vdots \\ (x_d - m_d/\sigma_d) \end{bmatrix})
\end{aligned}
$$

where $\Sigma^{-1} = D^{-1}R^{-1}D^{-1}$, $D = diag(\sigma_1, \sigma_2, ..., \sigma_d)$.

- If we have pdf margins given by $f_1, f_2, ..., f_d$ and cdf margins given by $F_1, F_2, ..., F_d$, then we have the joint pdf given by

$$\begin{aligned} h(x_1, x_2, ..., x_d) &= c(F_1(x_1), F_2(x_2), ..., F_d(x_d)) f_1(x_1) f_2(x_2) \cdots f_d(x_d) \\ &= \frac{1}{\sqrt{det(R)}} \exp(-\frac{1}{2} \begin{bmatrix} \phi^{-1}(F_1(x_1)) \\ \vdots \\ \phi^{-1}(F_d(x_d)) \end{bmatrix}^T \\ &\cdot (R^{-1} - I) \begin{bmatrix} \phi^{-1}(F_1(x_1)) \\ \vdots \\ \phi^{-1}(F_d(x_d)) \end{bmatrix}) f_1(x_1) f_2(x_2) \cdots f_d(x_d) \end{aligned}$$

Lemma 14.4.6 (copula density for random vector with independent components). *If a random vector $(X_1, X_2, ..., X_d)$ has independent components, then its copula density is given by*

$$c = 1.$$

Proof. From Theorem 14.4.3, we know that

$$c(x_1, ..., x_d) = \frac{h(F_1(x_1), ..., F_d(x_d))}{f_1(x_1) \cdots f_d(x_d)} = \frac{f_1(x_1) \cdots f_d(x_d)}{f_1(x_1) \cdots f_d(x_d)} = 1.$$

□

Lemma 14.4.7 (copulas invariance under monotone transform). *Let C be the copulas associated with random vector $(X_1, X_2, ..., X_d)$ and its cdf F. It follows that*

- *Let $T_1, T_2, ..., T_d$ be increasing functions. Then C is also the copulas associated with random vector $(T_1(X_1), T_2(X_2), ..., T_n(X_n))$.*
- *Let $T_1, T_2, ..., T_d$ be monotone(either increasing or decreasing) functions. **If** $(X_1, X_2, ..., X_n)$ **have uniform distribution**, then C is also the copulas associated with random vector $(T_1(X_1), T_2(X_2), ..., T_n(X_n))$.*

Proof. (1)Denote $(Y_1, Y_2, ..., Y_n) = (T_1(X_1), T_2(X_2), ..., T_n(X_n))$. Note that $(Y_1, Y_2, ..., Y_n)$ has marginal

$$F_{Y_i}(y_i) = F_{X_i}(T_i^{-1}(y_1)).$$

and cdf

$$F_Y(y_1, ..., y_2) = F_X(T_1^{-1}(y_1), T_2^{-1}(y_2), ..., T_n^{-1}(y_n)).$$

Then

$$\begin{aligned} C_Y(u_1, u_2, ..., u_d) &= F_Y(F_{Y_1}^{-1}(u_1), F_{Y_2}^{-1}(u_2), ..., F_{Y_n}^{-1}(u_n)) \\ &= F_X(T_1^{-1}(F_{Y_1}^{-1}(u_1)), T_2^{-1}(F_{Y_2}^{-1}(u_2)), ..., T_n^{-1}(F_{Y_n}^{-1}(u_n))) \\ &= F_X(F_{X_1}^{-1}(u_1), F_{X_2}^{-1}(u_2), ..., F_{X_n}^{-1}(u_n)) \\ &= C_X(u_1, u_2, ..., u_d) \end{aligned}$$

where we use the relation $F_{Y_i} = F_{X_i} \circ T^{-1}$ such that

$$F_{X_1}^{-1} = (F_{Y_1} \circ T)^{-1} = T^{-1} \circ F_{Y_1}^{-1}.$$

Note that $(F_{Y_1} \circ T)$ is invertible only if T is increasing.

(2) We can replace $F_{X_i} = I$ in the above proof and get the result.

□

Remark 14.4.4 (caution!). This does not hold for decreasing transform.

14.4.2.2 *Survival copula*

Definition 14.4.3 (survival copula).

- *Consider a uniform random vector $(U_1, U_2, ..., U_d)$ with joint cdf/copula C. The **survival copula** associated with C is defined by*

$$C^s(u_1, u_2, ..., u_d) \triangleq Pr(U_1 > u_1, U_2 > u_2, ..., U_d > u_d).$$

- *We have relation*

$$1 - F(U_1 < u_1, U_2 < u_2, ..., U_d < u_d) = C^s(1 - F_1(u_1), 1 - F_2(u_2), ..., 1 - F_d(u_d)).$$

or equivalently

$$S(u_1, u_2, ..., u_d) = C^s(S_1(u_1), S_2(u_2), ..., S_d(u_d)),$$

where $S \triangleq 1 - F, S_i \triangleq 1 - F_i, i = 1, 2, ..., d$.

Lemma 14.4.8 (conversion between copula and survival copula).

- *Let $(U_1, U_2, ..., U_n)$ be uniform random variables. Let C be its joint cdf/copula, let C^s be its survival copula. Then*

$$C(u_1, u_2, ..., u_n) = 1 - C^s(1 - u_1, 1 - u_2, ..., 1 - u_n).$$

Or equivalently,

$$C^s(u_1, u_2, ..., u_n) = 1 - C(1 - u_1, 1 - u_2, ..., 1 - u_n).$$

- *Let $(X_1, X_2, ..., X_n)$ be random variables with margins $F_1, F_2, ..., F_n$ and copula C. Then the joint cdf*

$$\begin{aligned} F_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) &= C(F_1(x_1), F_2(x_2), ..., F_n(x_n)) \\ &= 1 - C^s(1 - F_1(x_1), 1 - F_2(x_x), ..., 1 - F_n(x_n)) \end{aligned}$$

Or equivalently,

$$\begin{aligned} F^s_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) &\triangleq Pr(X_1 > x_1, X_2 > x_2, ..., X_n > x_n) \\ &= C(F_1(x_1), F_2(x_2), ..., F_n(x_n)) \\ &= 1 - C^s(1 - F_1(x_1), 1 - F_2(x_x), ..., 1 - F_n(x_n)) \end{aligned}$$

$$C^s(u_1, u_2, ..., u_n) = 1 - C(1 - u_1, 1 - u_2, ..., 1 - u_n).$$

14.4.2.3 *Partial differential and conditional distribution*

Theorem 14.4.4 (partial differential of copula gives conditional distribution of uniform random variables). *[10, p. 16]*

- *(bivariate case)Let $C(u, v)$ be a copula, i.e., the joint cdf associated with uniform random variables (U, V). Assume $C(u, v)$ can be differentiated. Then*

$$\frac{\partial}{\partial v} C(u, v)$$

is the conditional cdf given by

$$\frac{\partial}{\partial u} C(v, v) = Pr(U < u | V = v).$$

- *(multivariate case, multiple conditioning on one)Let $C(u_1, u_2, ..., u_n)$ be a differentiable copula, i.e., the joint cdf associated with uniform random variables $(U_1, U_2, ..., U_n)$. Then*

$$\frac{\partial}{\partial u_k} C(u_1, u_2, ..., u_n)$$

is the conditional cdf given by

$$\frac{\partial}{\partial u_k} C(u_1, u_2, ..., u_n) = Pr(U_i \leq u_i, 1 \leq i \leq n, i \neq k | U_k = u_k).$$

- *(multivariate case, one conditioning on multiple)Let $C(u_1, u_2, ..., u_n)$ be a differentiable copula, i.e., the joint cdf associated with uniform random variables $(U_1, U_2, ..., U_n)$. Then the conditional cdf given by*

$$\frac{\frac{\partial^{n-1}}{\partial u_1 ... \partial u_{k-1} ... \partial u_n} C_{1:n}(u_1, ..., u_n)}{\frac{\partial^{k-1}}{\partial u_1 ... \partial u_{k-1}} C_{1:k-1}(u_1, ..., u_k - 1)} = Pr(U_k \leq u_k | U_i = u_i, 1 \leq i \leq n, i \neq k).$$

Proof. (1)

$$\begin{aligned} Pr(U \leq u | V = v) &= \lim_{h \to 0} Pr(U \leq u | v \leq V \leq v + h) \\ &= \lim_{h \to 0} \frac{C(u, v+h) - C(u, v)}{C_V(v+h) - C_V(v)} \\ &= \lim_{h \to 0} \frac{C(u, v+h) - C(u, v)}{h} \\ &= \lim_{h \to 0} \frac{C(u, v+h) - C(u, v)}{h} \qquad = \frac{\partial}{\partial u} C(u, v) \end{aligned}$$

where C_V is the marginal distribution of V given by [Lemma 14.4.1]

$$C_V(v) = C(1, v) = v.$$

(2)

$$\begin{aligned} &Pr(U_i \leq u_i, 1 \leq i \leq n, i \neq k | U_k) \\ &= \lim_{h \to 0} Pr(U_i \leq x_i, 1 \leq i \leq n, i \neq k | u_k \leq U_k \leq u_k + h) \\ &= \lim_{h \to 0} \frac{C(u_1, ..., u_k + h, ..., u_n) - C(u_1, ..., u_k, ..., u_n)}{C_k(x_k + h) - C_k(x_k)} \\ &= \lim_{h \to 0} \frac{C(u_1, ..., u_k + h, ..., u_n) - C(u_1, ..., u_k, ..., u_n)}{h} \\ &= \frac{\partial}{\partial u_k} C(u_1, ..., u_k, ..., u_n) \end{aligned}$$

where C_k is the marginal distribution of U_k given by Lemma 14.4.1

$$C_k(u_k) = C(1, ...1, , u_k, 1, ..., 1) = u_k.$$

(3) informally, we can think of the upper as

$$Pr(U_k \leq u_k, U_i = u_i, 1 \leq i \leq n, i \neq k);$$

and the lower as

$$Pr(U_i = u_i, 1 \leq i \leq n, i \neq k).$$

□

Lemma 14.4.9 (partial differential of copula gives conditional distribution for general random variables, bivariate distribution). *[10, p. 16] Let $C(u, v)$ be a copula, i.e., the joint cdf associated with uniform random variables (U, V). Assume $C(u, v)$ can be differentiated. Then*

$$\frac{\partial}{\partial u} C(u, v)$$

is the conditional cdf given by

$$\frac{\partial}{\partial u} C(u, v) = Pr(V < v | U = u).$$

Proof.

$$\begin{aligned}
Pr(X \leq x | Y = y) &= \lim_{h \to 0} Pr(X \leq x | y \leq Y \leq y + h) \\
&= \lim_{h \to 0} \frac{F_{XY}(x, y + h) - F_{XY}(x, y)}{F_Y(y + h) - F_Y(y)} \\
&= \lim_{h \to 0} \frac{C(F_X(x), F_Y(y + h)) - C(F_X(x), F_Y(y))}{F_Y(y + h) - F_Y(y)} \\
&= \lim_{h \to 0} \frac{C(F_X(x), F_Y(y) + \Delta(h)) - C(F_X(x), F_Y(y))}{\Delta(h)} \\
&= \frac{\partial}{\partial q} C(p, q)|_{p = F_X(x), q = F_Y(y)}
\end{aligned}$$

where $\Delta(h) = F_Y(y + h) - F_Y(y)$. □

Lemma 14.4.10 (partial differential gives conditional distribution, multiple conditioned on one). *[10, p. 19] Let $X_1, X_2, ..., X_n$ be real-valued random variables with corresponding copula C and continuous marginals $F_1, ..., F_n$. Then for any $k \in \{1, 2, ..., n\}$,*

$$Pr(X_i \leq x_i, 1 \leq i \leq n, i \neq k | X_k) = \frac{\partial}{\partial F_k(X_k)} C(F_1(x_1), ..., F_k(x_k), ..., F_n(x_n)),$$

for all $x_1, x_2, ..., x_n \in \mathbb{R}$.

Proof.

$$\begin{aligned}
&Pr(X_i \leq x_i, 1 \leq i \leq n, i \neq k | X_k) \\
&= \lim_{h \to 0} Pr(X_i \leq x_i, 1 \leq i \leq n, i \neq k | x_k \leq X_k \leq x_k + h) \\
&= \lim_{h \to 0} \frac{F_{X_1:X_n}(x_1, ..., x_k + h, ..., x_n) - F_{X_1:X_n}(x_1, ..., x_k, ..., x_n)}{F_k(x_k + h) - F_k(x_k)} \\
&= \lim_{h \to 0} \frac{C(F_1(x_1), ..., F_k(x_k + h), ..., F_n(x_n)) - C(F_1(x_1), ..., F_k(x_k), ..., F_n(x_n))}{F_k(x_k + h) - F_k(x_k)} \\
&= \lim_{h \to 0} \frac{C(F_1(x_1), ..., F_k(x_k) + \Delta(h), ..., F_n(x_n)) - C(F_1(x_1), ..., F_k(x_k), ..., F_n(x_n))}{\Delta(h)} \\
&= \frac{\partial}{\partial F_k(X_k)} C(F_1(x_1), ..., F_k(x_k), ..., F_n(x_n))
\end{aligned}$$

where $\Delta(h) = F_k(x_k + h) - F_k(x_k)$. □

Lemma 14.4.11 (partial differential gives conditional distribution, one conditioned on multiple). *[10, p. 20] Let $X_1, X_2, ..., X_n$ be real-valued random variables with corresponding copula C and continuous marginals $F_1, ..., F_n$. Then for any $k \in \{1, 2, ..., n\}$,*

$$Pr(X_i \leq x_i, 1 \leq i \leq n, i \neq k | X_k) = \frac{\partial}{\partial F_k(X_k)} C(F_1(x_1), ..., F_k(x_k), ..., F_n(x_n)),$$

for all $x_1, x_2, ..., x_n \in \mathbb{R}$.

Proof.

$$\begin{aligned}
Pr(X \leq x|Y = y) &= \lim_{h\to 0} Pr(X \leq x|y \leq Y \leq y+h) \\
&= \lim_{h\to 0} \frac{F_{XY}(x, y+h) - F_{XY}(x, y)}{F_Y(y+h) - F_Y(y)} \\
&= \lim_{h\to 0} \frac{C(F_X(x), F_Y(y+h)) - C(F_X(x), F_Y(y))}{F_Y(y+h) - F_Y(y)} \\
&= \lim_{h\to 0} \frac{C(F_X(x), F_Y(y) + \Delta(h)) - C(F_X(x), F_Y(y))}{\Delta(h)} \\
&= \frac{\partial}{\partial q} C(p, q)|_{p=F_X(x), q=F_Y(y)}
\end{aligned}$$

where $\Delta(h) = F_Y(y+h) - F_Y(y)$. □

14.4.3 Common copula functions

14.4.3.1 *Gaussian copula*

Definition 14.4.4 (Gaussian copula). *A Gaussian copula characterized by correlation matrix $R \in [-1, 1]^{d\times d}$ is given by*

$$C(u_1, u_2, ..., u_d; R) = \Phi(\phi^{-1}(u_1), \phi^{-1}(u_2), ..., \phi^{-1}(u_d)),$$

where $u_1, u_2, ..., u_d \in [0, 1]$, Φ is the cdf for a multivariate normal distribution with zero mean and covariance matrix R, ϕ is the cdf for a standard normal variable.

The copula density function is given by

$$c(u_1, u_2, ..., u_d) = \frac{1}{\sqrt{det(R)}} \exp(-\frac{1}{2} \begin{bmatrix} \phi^{-1}(u_1) \\ \vdots \\ \phi^{-1}(u_d) \end{bmatrix}^T \cdot (R^{-1} - I) \begin{bmatrix} \phi^{-1}(u_1) \\ \vdots \\ \phi^{-1}(u_d) \end{bmatrix}).$$

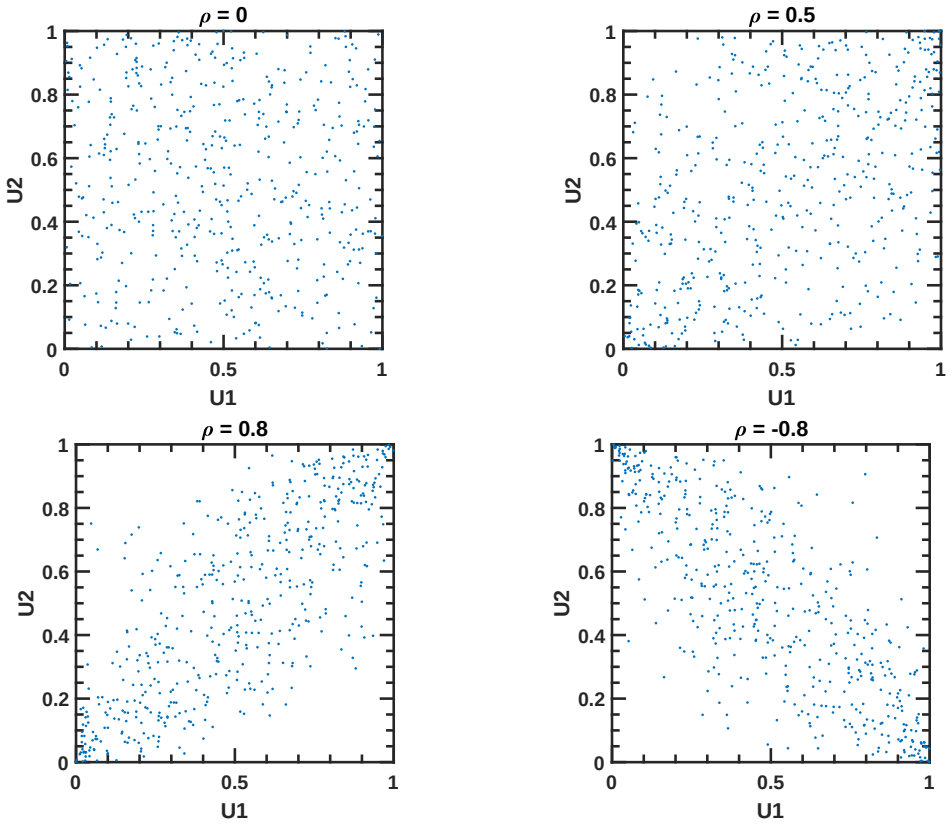


Figure 14.4.1: Gaussian copula with different correlations.

Remark 14.4.5 (derivation of copula density function). From Theorem 14.4.3, we know that

$$c(x_1, ..., x_d) = \frac{h(F_1(x_1), ..., F_d(x_d))}{f_1(x_1) \cdots f_d(x_d)}.$$

Note that

$$h(x) = \frac{1}{(2\pi)^{d/2}|detR|^{1/2}} \exp(-\frac{1}{2}(x)^T R^{-1}(x)), x \in \mathbb{R}^d,$$

and

$$f_1 \cdot f_2 \cdots f_d = \frac{1}{(2\pi)^{d/2}} \exp(-\frac{1}{2}(x)^T(x)).$$

Lemma 14.4.12 (the copula associated with multivariate Gaussian distribution is Gaussian copula).

- *If F is a multivariate normal distribution $MN(0, R)$, where R is correlation matrix and $\Sigma = R$(that is, the margins are standard normal such that covariance matrix is*

equal to correlation matrix); then the copula associated with F is the Gaussian copula with correlation matrix R

- *If F is a multivariate normal distribution $MN(0, \Sigma)$, then the copula associated with F is the Gaussian copula with correlation matrix $R = D^{-1/2}\Sigma D^{-1/2}, D = diag(\Sigma)$.*
- *If F is a multivariate normal distribution $MN(\mu, \Sigma)$, then the copula associated with F is the Gaussian copula with correlation matrix $R = D^{-1/2}\Sigma D^{-1/2}, D = diag(\Sigma)$.*

Proof. (1)Straight forward [Theorem 14.4.2]. (2)(3) Let Φ denote the joint cdf $MN(\mu, \Sigma)$ and ϕ_i the marginal cdf $N(\mu_i, \sigma_i^2)$. Let Φ^S denote the joint cdf $MN(\mu, R)$ and ϕ_i^S the marginal cdf $N(0, 1)$.

Then,

$$\begin{aligned}\Phi(u_1, u_2, ..., u_d) &= \Phi^S(\frac{u_1 - \mu_1}{\sigma_1^2}, \frac{u_2 - \mu_2}{\sigma_2^2}, ..., \frac{u_d - \mu_d}{\sigma_d})\\ \phi_i(u_i) &= \phi_i^S(\frac{u_i - \mu_i}{\sigma_i})\\ \phi_i^{-1}(u_1) &= \sigma_i(\phi_i^S)^{-1} + \mu_i\end{aligned}$$

Therefore,

$$\begin{aligned}&\Phi(\phi_i^{-1}(u_1), \phi_2^{-1}(u_2), ..., \phi_2^{-1}(u_d))\\ =&\Phi^S((\phi_i^{-1}(u_1) - \mu_1)/\sigma_1, (\phi_i^{-1}(u_2) - \mu_2)/\sigma_2, ..., (\phi_i^{-1}(u_d) - \mu_d)/\sigma_d)\\ =&\Phi^S((\phi_1^S)^{-1}, (\phi_2^S)^{-1}, ..., (\phi_d^S)^{-1})\end{aligned}$$

□

Lemma 14.4.13 (construct a multivariate cdf from Gaussian copulas and margins). *Consider d random variables $X_1, X_2, ..., X_d$ with $F_1, F_2, ..., F_d$ being the univariate cdf. Let C be a d dimensional Gaussian copulas given by*

$$C(u_1, u_2, ..., u_d; R) = \Phi(\phi^{-1}(u_1), \phi^{-1}(u_2), ..., \phi^{-1}(u_d)),$$

where $u_1, u_2, ..., u_d \in [0, 1]$, Φ is the cdf for a multivariate normal distribution with zero mean and covariance matrix R, ϕ is the cdf for a standard normal variable.

Then the joint distribution for $X_1, X_2, ..., X_d$ is given by

$$F(x_1, x_2, ..., x_n) = \Phi(\phi^{-1}(F_1(x_1)), \phi^{-1}(F_2(x_2)), ..., \phi^{-1}(F_d(x_d)));$$

and the joint density function is given by

$$f(x) = \frac{1}{(2\pi)^{d/2}|detR|^{1/2}} \exp(-\frac{1}{2}(x)^T R^{-1}(x)) f_1(x_1) f_2(x) \cdots f_d(x_d),$$

where

$$x = (\phi^{-1}(F_1(x_1)), \phi^{-1}(F_2(x_2)), ..., \phi^{-1}(F_d(x_d))).$$

and $f_1, f_2, ..., f_d$ *are the marginal densities of* $X_1, X_2, ..., X_n$.

Proof. (1)See Theorem 14.4.1. (2)See Theorem 14.4.3. □

Example 14.4.5 (Gaussian margin with Gaussian copula will give multivariate Gaussian). Note that from Theorem 14.4.3, we know that the joint cdf can be constructed from margins $f_1, f_2, ..., f_d$ and $c(u_1, u_2, ..., u_d)$ via

$$h(x_1, ..., x_d) = c(F_1(x_1), ..., F_d(x_d)) \cdot f_1(x_1) \cdots f_d(x_d).$$

Let $\phi(x)$ be the standard Gaussian cdf, then a Gaussian random variable $N(m, \sigma^2)$ has pdf given by $\phi(\frac{x-m}{\sigma})$.

Then,

$$
\begin{aligned}
&h(x_1, ..., x_d) \\
&= c(F_1(x_1), ..., F_d(x_d)) \cdot f_1(x_1) \cdots f_d(x_d) \\
&= \frac{1}{\sqrt{det(R)}} \exp(-\frac{1}{2} \begin{bmatrix} \phi^{-1}(\phi(x_1 - m_1/\sigma_1)) \\ \vdots \\ \phi^{-1}(\phi(x_d - m_d/\sigma_d)) \end{bmatrix}^T \\
&\cdot (R^{-1} - I) \begin{bmatrix} \phi^{-1}(\phi(x_1 - m_1/\sigma_1)) \\ \vdots \\ \phi^{-1}(\phi(x_d - m_d/\sigma_d)) \end{bmatrix}) \\
&\frac{1}{(2\pi)^{d/2}} \exp(\begin{bmatrix} (x_1 - m_1/\sigma_1)) \\ \vdots \\ (x_d - m_d/\sigma_d)) \end{bmatrix}^T \cdot \begin{bmatrix} (x_1 - m_1/\sigma_1)) \\ \vdots \\ (x_d - m_d/\sigma_d)) \end{bmatrix})) \\
&= \frac{1}{\sqrt{det(R)}} \exp(-\frac{1}{2} \begin{bmatrix} (x_1 - m_1/\sigma_1) \\ \vdots \\ (x_d - m_d/\sigma_d) \end{bmatrix}^T \cdot (R^{-1}) \begin{bmatrix} (x_1 - m_1/\sigma_1) \\ \vdots \\ (x_d - m_d/\sigma_d) \end{bmatrix}) \frac{1}{(2\pi)^{d/2}} \\
&= \frac{1}{\sqrt{det(R)}(2\pi)^{d/2}} \exp(-\frac{1}{2} \begin{bmatrix} (x_1 - m_1/\sigma_1) \\ \vdots \\ (x_d - m_d/\sigma_d) \end{bmatrix}^T \cdot (\Sigma^{-1}) \begin{bmatrix} (x_1 - m_1/\sigma_1) \\ \vdots \\ (x_d - m_d/\sigma_d) \end{bmatrix})
\end{aligned}
$$

where $\Sigma^{-1} = D^{-1}R^{-1}D^{-1}$, $D = diag(\sigma_1, \sigma_2, ..., \sigma_d)$.

Example 14.4.6 (bivariate Gaussian copula).

$$C_\rho(u, v) = \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \frac{1}{2\pi(1-\rho^2)^{1/2}} \exp(-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}) dxdy$$

where ρ is the linear correlation coefficient.

14.4.3.2 *t copula*

Definition 14.4.5 (*t* copula). *[11, p. 419] A t-copula characterized by correlation matrix $R \in [-1,1]^{n\times n}$ and degree of freedom parameter v is the copula associated with the multivariate Student's t probability distribution*

$$C(u_1, u_2, ..., u_n; \rho, v) = T_n(T_v^{-1}(u_1), T_v^{-1}(u_2), ..., T_v^{-1}(u_n); R, v).$$

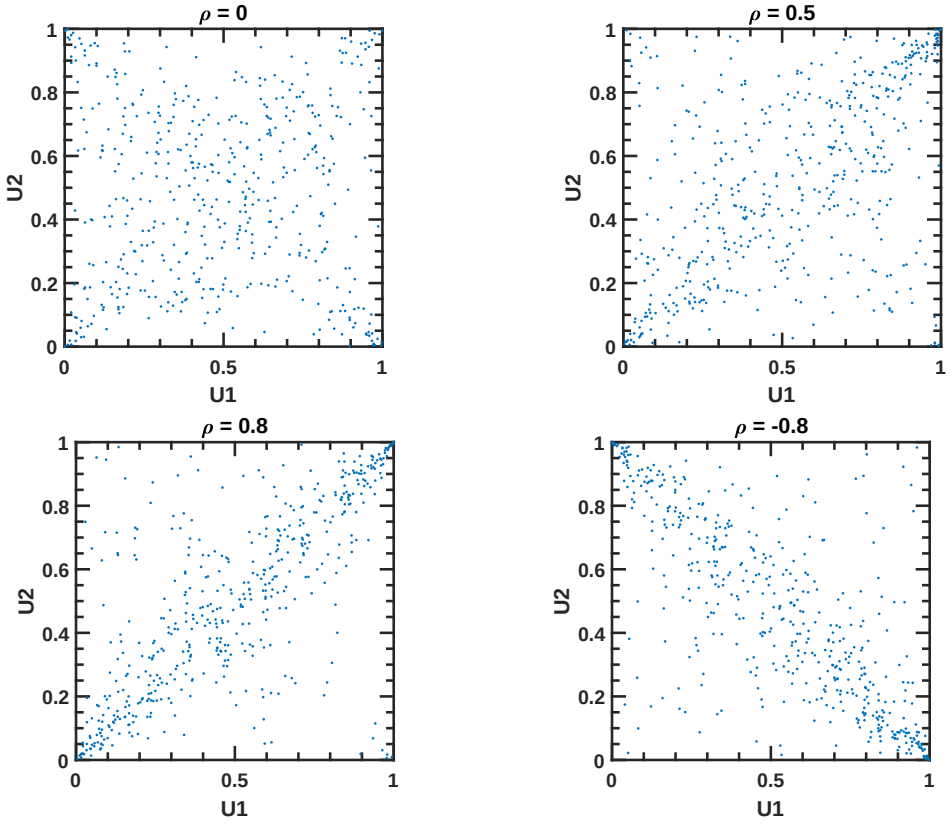


Figure 14.4.2: Student T copula with different correlations.

Example 14.4.7 (bivariate t copula).

$$C(u_1, u_2; \rho, v) = \int_{-\infty}^{T_v^{-1}(u_1)} \int_{-\infty}^{T_v^{-1}(u_2)} \frac{1}{2\pi(1-\rho^2)^{1/2}}(1 + \frac{x_1^2 + x_2^2 - 2\rho x_1 x_2}{v(1-\rho^2)})dx_1 dx_2$$

where ρ is the linear correlation coefficient.

14.4.3.3 *Common copula functions: other copula*

Definition 14.4.6 (product copula, co-monotonicity copula). *[12, p. 187][3, p. 190]*

- ***(product copula, independence copula)****The d-dimensional product copula is given by*
$$C(u_1, u_2, ..., u_d) = u_1 u_2 \cdots u_d, u_i \in [0,1], \forall i.$$
The product copula corresponds to independence and it can be viewed as the cdf of $(U_1, ..., U_d)$, where $U_1, ..., U_d$ are independent uniform random variables.
- ***(co-monotonicity copula)*** *The joint cdf of the random vector $(U_1, U_2, ..., U_d)$, where U is a uniform random variable is called a co-monotonicity copula, which characterizes perfect positive correlation. It is given by*
$$C(u_1, u_2, ..., u_d) = \min(u_1, u_2, ..., u_d).$$
- ***two dimensional counter-monotonicity copula*** *is defined as the joint cdf of $(U, 1-U)$.Therefore,*
$$\begin{aligned} C(u_1, u_2) \triangleq & Pr(U \leq u_1, (1-U) \leq u_2) \\ = & Pr(1-u_2 \leq U \leq u_1) \\ = & \max\{u_1 + u_2 - 1, 0\} \end{aligned}$$

Remark 14.4.6 (interpretation). [12, p. 185]

- The co-monotonicity copula is the joint cdf of $\boldsymbol{U} = (U, U, ..., U)$;that is, $\boldsymbol{U}$ contains d copies of $U(0,1)$. The co-monotonicity copula is the upper bound of all copula functions [Lemma 14.4.4].
- The two dimensional counter-monotonicity copula is the lower bound of all copula functions.

14.4.4 Dependence and copula

14.4.4.1 *Linear correlations*

Example 14.4.8. Consider discrete-valued random variable V_1 and V_2 given by:

- V_1 equally take three different values $-1, 0, +1$.
- If $V_1 = -1$ *or* $+1$, $V_2 = 1$. If $V_1 = 0$, then $V_2 = 0$.

It is clearly that $E[V_1V_2] = 0, E[V_1] = 0 \implies Cov[V_1, V_2] = 0$; however, it is clearly that V_1 and V_2 are uncorrelated but they are dependent since

$$Pr(V_2|V_1 = v) \neq Pr(V_2).$$

where

$$\begin{aligned} Pr(V_2 = 1) &= Pr(V_2 = 1|V_1 = 1)Pr(V_1 = 1) + Pr(V_2 = 1|V_1 = 1)Pr(V_1 = 1) \\ &= 1 \times 1/3 + 1 \times 1/3 = 2/3; \end{aligned}$$

and

$$Pr(V_2 = 0) = Pr(V_2 = 0|V_1 = 0)Pr(V_1 = 0) = 1 \times 1/3 = 1/3.$$

Definition 14.4.7 (linear correlation, Pearson correlation). *[3, p. 202] Given two random variables X and Y. The linear correlation is defined by*

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var[X]Var[Y]}}.$$

Note 14.4.4 (characteristics of linear correlations).

- sensitive to outliers.
- measure the 'average dependence' between X and Y.
- invariant under strictly **increasing linear transformation**.
- May be misleading when multivariate distribution is not elliptical.

Lemma 14.4.14 (invariance of correlation under affine transformation).

- *Let random variables X and Y have correlation $\rho(X, Y)$. Then*

$$\rho(aX + b, cY + d) = \rho(X, Y), \forall a, c > 0, b, d \in \mathbb{R}.$$

- *Let random variables X and Y have correlation $\rho(X, Y)$. Then*

$$\rho(aX + b, cY + d) = \frac{cd}{|cd|}\rho(X, Y), \forall a, b, b, d \in \mathbb{R}.$$

Proof. (1)

$$\rho(aX + b, cY + d) \triangleq \frac{Cov(aX + b, cY + d)}{\sqrt{Var[aX + b]Var[cY + d]}} = \frac{aCov(X, Y)d}{\sqrt{a^2Var[X]Var[Y]c^2}} == \rho(X, Y).$$

(2) straight forward. □

Remark 14.4.7 (generally not invariant under nonlinear transformation). Let $T : \mathbb{R} \to \mathbb{R}$ be a nonlinear strictly increasing function. Then generally

$$\rho(X, Y) \neq \rho(T(X), T(Y)).$$

Lemma 14.4.15 (perfect linear correlation and linear function relationship). *[3, p. 202] Let X and Y be two random variable defined on the same probability space. It follows that*

- *$\rho(X, Y) = 1$ implies $Y = \alpha + \beta X$ almost surely for some $\alpha, \beta \in \mathbb{R}, \beta > 0$.*
- *$\rho(X, Y) = -1$ implies $Y = \alpha + \beta X$ almost surely for some $\alpha, \beta \in \mathbb{R}, \beta < 0$.*
- *Conversely, if $Y = \alpha + \beta X$, then $\rho(X, Y) = sign(\beta)$.*

Proof. (1)(2) Note that the equality holds only when the equality holds in Cauchy inequality [Theorem 11.9.4].Then, X and Y must have linear dependence almost everywhere. (3) Use the affine transformation invariance property of correlation [Lemma 14.4.14], □

14.4.4.2 *Rank correlations*

Definition 14.4.8 (Kendall's tau for observations).

- *Let $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ be a set of joint observations from two random variables X and Y.*
- *A pair of observations (x_i, y_i) and (x_j, y_j) are **concordant** if both $x_i > x_j$ and $y_i > y_j$ or if both $x_i < x_j$ and $y_i < y_j$;*
- *They are **discordant**, if $x_i > x_j$ and $y_i < y_j$ or if $x_i < x_j$ and $y_i > y_j$.*
- *If $x_i = x_j, y_i = y_j$, the pair is neither concordant nor discordant.*
- *The **Kendall τ coefficient** is defined as*

$$\rho_\tau = \frac{num\ concordant\ pairs - num\ disconcordant\ pairs}{n(n-1)/2}.$$

Note that $n(n-1)/2$ is the total number of pairs to compare.

Definition 14.4.9 (Kendall's tau for random variables). *[3, p. 207] Let X_1 and X_2 be two random variables. Then the Kendall's tau is given by*

$$\rho_\tau = E[sign((X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2))],$$

where $(\tilde{X}_1, \tilde{X}_2)$ is a independent copy of (X_1, X_2); that is,$(\tilde{X}_1, \tilde{X}_2)$ has the same cdf of (X_1, X_2), but they are statistically independent.

Lemma 14.4.16 (Kendall's tau from copula). *[3, p. 207] The C be the copula associated with the joint cdf of (X, Y), then*

$$\begin{aligned}\rho_\tau(X,Y) &= 4\int_0^1\int_0^1 C(u,v)dC(u,v) - 1\\ &= 4E[C(U,V)] - 1\end{aligned}$$

Proof. From the definition

$$\begin{aligned}\rho_\tau &= E[sign((X_1-\tilde{X}_1)(X_2-\tilde{X}_2))]\\ &= E[\mathbf{1}((X_1-\tilde{X}_1)(X_2-\tilde{X}_2)>0)] - E[\mathbf{1}((X_1-\tilde{X}_1)(X_2-\tilde{X}_2)<0)]\\ &= Pr((X_1-\tilde{X}_1)(X_2-\tilde{X}_2)>0) - Pr((X_1-\tilde{X}_1)(X_2-\tilde{X}_2)<0)\\ &= 2Pr((X_1-\tilde{X}_1)(X_2-\tilde{X}_2)>0) - 1\\ &= 2Pr((X_1-\tilde{X}_1)>0,(X_2-\tilde{X}_2)>0) + 2Pr((X_1-\tilde{X}_1)<0,(X_2-\tilde{X}_2)<0) - 1\\ &= 4Pr((X_1-\tilde{X}_1)<0,(X_2-\tilde{X}_2)<0) - 1\\ &= 4\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Pr((X_1<s_1, X_2<s_2) f_{\tilde{X}_1,\tilde{X}_2}(s_1,s_2)ds_1ds_2 - 1\\ &= 4\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F(s_1,s_2)dF(s_1,s_2) - 1\\ &= 4\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} C(F_1(s_1),F_2(s_2))dC(F_1(s_1),F_2(s_2)) - 1\\ &= 4\int_0^1\int_0^1 C(u_1,u_2)dC(u_1,u_2) - 1\end{aligned}$$

□

Remark 14.4.8 (Kendall's tau is independent of the marginal cdf). Note that Kendall's tau only depends on the correlation structure characterized by the copula; it is independent of the marginal cdf.

Lemma 14.4.17 (Hoffding formula for covariance). *[3, p. 204] If (X_1, X_2) has joint cdf F and marginal cdf F_1 and F_2, then*

- $$Cov(X_1,X_2) = \frac{1}{2}E[(X_1-\tilde{X}_1)(X_2-\tilde{X}_2)],$$

where $(\tilde{X}_1, \tilde{X}_2)$ is a independent copy of (X_1, X_2); that is,$(\tilde{X}_1, \tilde{X}_2)$ has the same cdf of (X_1, X_2), but they are statistically independent.

- $$Cov(X_1, X_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(F(x_1, x_2) - F_1(x_1)F_2(x_2))dx_1dx_2.$$

Proof. (1) Directly expand the rhs. Note that $E[X_1\tilde{X}_2] = E[X_1]E[\tilde{X}_2]$ due to independence. (2) A useful identity is for any $a \in \mathbb{R}, b \in R$, we have

$$(a - b) = \int_{-\infty}^{\infty} H(x - b) - H(x - a)dx.$$

We have

$$\begin{aligned}
&E[(X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2)] \\
=&E[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(H(s_1 - X_1) - H(s_1 - \tilde{X}_1))(H(s_2 - X_2) - H(s_2 - \tilde{X}_2))ds_1ds_2] \\
=&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}E[(H(s_1 - X_1) - H(s_1 - \tilde{X}_1))(H(s_2 - X_2) - H(s_2 - \tilde{X}_2))]ds_1ds_2 \\
=&2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}Pr(X_1 \le s_1, X_2 \le s_2) - Pr(X_1 \le s_1)Pr(X_2 \le s_2)ds_1ds_2 \\
=&2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(F(s_1, s_2) - F_1(s_1)F_2(s_2))ds_1ds_2
\end{aligned}$$

□

Definition 14.4.10 (Spearman's rho). *[3, p. 207] Let X_1 and X_2 be two random variables with marginal cdf F_1 and F_2. **Spearman's rho** is defined as*

$$\rho_S(X, Y) = \rho(F_1(X_1), F_2(X_2)).$$

In other words, Spearman's rho is the linear correlation of the transform random variables.

Definition 14.4.11 (Spearman's rho for observations).

- *Let $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ be a set of joint observations from two random variables X and Y.*
- *Let $R(x_i)$ denote the rank of x_i among $x_1, x_2, ..., x_n$, where $R(x_i)$ will take value from 1 to n. Similarly we denote $R(y_i)$ as the rank of y_i.*

- *The **Spearman's ρ coefficient** is defined as*

$$\rho_S = \frac{\frac{1}{n}\sum_{i=1}^{n}(R(x_i)-\bar{R}(x))(R(y_i)-\bar{R}(y))}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(R(x_i)-\bar{R}(x))^2\frac{1}{n}\sum_{i=1}^{n}(R(y_i)-\bar{R}(y))^2}},$$

  where $\bar{R}(y) = \frac{1}{n}\sum_{i=1}^{n} R(y_i)$.
- *It can be showed that*

$$\rho_S = 1 - \frac{\sum_{i=1}^{n}(R(x_i)-R(y_i))^2}{n^3-n}.$$

Lemma 14.4.18 (properties of Spearman's rho for observations). *Let $(x_1,y_1),(x_2,y_2),...,(x_n,y_n)$ be a set of joint observations from two random variables X and Y. It follows that*

- $$n\bar{R}(y) = \sum_{i=1}^{n} R(y_i) = \frac{1}{2}n(n+1).$$
- $$\sum_{i=1}^{n} R(y_i)^2 = \sum_{i=1}^{n} R(x_i)^2 = \frac{n(n+1)(2n+1)}{6}.$$
- $$\rho_S = 1 - \frac{\sum_{i=1}^{n}(R(x_i)-R(y_i))^2}{n^3-n}.$$

Proof. (1)straight forward.(2) this is the sum of squares from 1 to n. (3) Note that

$$\sum_{i=1}^{n}(R(x_i)-\bar{R}(x))^2 = \sum_{i=1}^{n} R(x_i)^2 - (\sum_{i=1}^{n} R(x_i))^2 = \frac{n(n+1)(2n+1)}{6} - (\frac{1}{2}n(n+1))^2.$$

and

$$\sum(R(x_i)-R(y_i))^2 = \sum R(x_i)^2 + \sum R(y_i)^2 - 2\sum R(x_i)R(y_i)$$

implies

$$2\sum R(x_i)R(y_i) = \sum R(x_i)^2 + \sum R(y_i)^2 - \sum(R(x_i)-R(y_i))^2.$$

□

Lemma 14.4.19 (Spearman's rho from copula). *[3, p. 207]*

$$\rho_S(X,Y) = 12\int_0^1\int_0^1 C(u,v)dudv - 3$$

Proof. From definition and use Lemma 14.4.17, we have

$$\begin{aligned}
\rho_S(X,Y) &= \rho(F_1(X_1), F_2(X_2)) \\
&= Cov(F_1(X_1), F_2(X_2))/(\sqrt{Var[F_1[X_1]]Var[F_2[X_2]]}) \\
&= 12Cov(F_1(X_1), F_2(X_2)) \\
&= 12\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \hat{F}(F_1(x_1), F_2(x_2)) - \hat{F}_1(F_1(x_1))\hat{F}_2(F_2(x_2))dx_1dx_2 \\
&= 12\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \hat{F}(F_1(x_1), F_2(x_2)) - x_1x_2dx_1dx_2 \\
&= 12(\int_0^1\int_0^1 C(u,v)dudv - \frac{1}{4})
\end{aligned}$$

where we use the property $F_1(X_1), F_2(X_2)$ are uniform random variable with variance $1/12$, and the joint cdf of $(F_1(X_1), F_2(X_2))$ is the copula associated with joint cdf of (X_1, X_2)(Lemma 14.4.3). □

Lemma 14.4.20 (Spearman's rho for monotonic relation). *Let X be a random variable and let Y be a monotonic function of X, denoted by $Y = f(X)$. It follows that*

- *If f is a monotonically increasing function, the $\rho_S(X,Y) = 1$.*
- *If f is a monotonically decreasing function, the $\rho_S(X,Y) = -1$.*

Proof. (1)Consider the Spearman's rho for observations [Definition 14.4.11]. For each sample (x_i, y_i), each component has the same rank. (2)For each sample (x_i, y_i), each component has ranks satisfying

$$R(x_i) - \bar{R}(x_i) = -(R(y_i) - \bar{R}(y_i)).$$

□

Lemma 14.4.21 (first quadrant probability of bivariate Gaussian distribution). *[3, p. 215] Let $(X_1, X-2)$ be a random vector with joint multivariate Gaussian distribution $MN(0, \Sigma)$. Let $\rho = \rho(X_1, X_2)$. Then*

$$Pr(X_1 > 0, X_2 > 0) = \frac{1}{4} + \frac{\arcsin \rho}{2\pi}.$$

Proof. See reference. □

Lemma 14.4.22 (rank correlation for Gaussian copula). *[3, p. 215] Let (X_1, X_2) be a bivariate random vector with Gaussian copula characterized by correlation coefficient ρ and continuous margins. Then*

- $$\rho_\tau(X_1, X_2) = \frac{2}{\pi} \arcsin \rho$$
- $$\rho_S(X_1, X_2) = \frac{6}{\pi} \arcsin \frac{1}{2}\rho$$

Proof. (1) Note that Kendall's tau only depends on the copula; therefore we can assume (X_1, X_2) has bivariate normal distribution($MN(0, 2\Sigma)$, correlation ρ). From Lemma 14.4.16,

$$\begin{aligned}
\rho_\tau &= 4Pr((X_1 - \tilde{X}_1) > 0, (X_2 - \tilde{X}_2) > 0) - 1 \\
&= 4Pr(Y_1 > 0, Y_2 > 0) - 1 \\
&= 4(\frac{1}{4} + \frac{\arcsin \rho}{2\pi}) - 1 \\
&= \frac{2}{\pi} \arcsin \rho
\end{aligned}$$

where $(\tilde{X}_1, \tilde{X}_2)$ is the independent copy of (X_1, X_2), and $Y_1 = X_1 - \tilde{X}_2 \sim MN(0, 2\Sigma)(Y_1$ has the same correlation ρ), we use Lemma 14.4.21. (2) From Lemma 14.4.19, we have

$$\begin{aligned}
\rho_S(X_1, X_2) &= 12 \int_0^1 \int_0^1 C(u, v) du dv - 3 \\
&= 12 \int_0^1 \int_0^1 \Phi(\phi^{-1}(u), \phi^{-1}(v)) du dv - 3 \\
&= 12 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Phi(s_1, s_2) d\phi(s_1) d\phi(s_2) - 3 \\
&= 12 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Phi(s_1, s_2) f(s_1) f(s_2) ds_1 ds_2 - 3 \\
&= 12 Pr(X_1 - S_1 < 0, X_2 - S_2 < 0) - 3 \\
&= 12 Pr(Y_1 < 0, Y_2 < 0) - 3 \\
&= 12(\frac{1}{4} + \frac{\arcsin \rho/2}{2\pi}) - 3 \\
&= \frac{6}{\pi} \arcsin \frac{1}{2}\rho
\end{aligned}$$

where $(\tilde{X}_1, \tilde{X}_2)$ is the independent copy of (X_1, X_2), and $Y_1 = X_1 - S_1 \sim MN(0, \Sigma + I_2)(Y_1$ has the correlation $\rho/2$), we use Lemma 14.4.21. □

Remark 14.4.9 (applications in robust correlation estimation for multivariate Gaussian random variables). Note that multivariate Gaussian distribution has Gaussian copula [Lemma 14.4.12]. Therefore, we can estimate ρ_τ, ρ_S first(which is robust) and then convert them to linear correlation coefficients.

14.4.4.3 *Tail dependence*

Definition 14.4.12 (tail dependence). *Let X and Y be random variables with marginal cdf F_X and F_Y.*

- *The **coefficient of upper tail dependence** of X and Y is*

$$\begin{aligned}
\lambda_u(X, Y) \triangleq &= \lim_{\alpha \to 1} Pr(F_Y(Y) > \alpha | F_X(X) > \alpha) \\
&= \lim_{\alpha \to 1} Pr(Y > F_Y^{-1}(\alpha) | X > F_X^{-1}(\alpha)).
\end{aligned}$$

- *The **coefficient of lower tail dependence** of X and Y is*

$$\begin{aligned}\lambda_l(X,Y) \triangleq &= \lim_{\alpha\to 0} Pr(F_Y(Y) \le \alpha | F_X(X) \le \alpha) \\ &= \lim_{\alpha\to 0} Pr(Y > F_Y^{-1}(\alpha) | X > F_X^{-1}(\alpha)).\end{aligned}$$

Tail dependence is the probability of observing a large(small) Y given that X is large(small). If $\lambda_u > 0 (\lambda_l > 0)$, then we say (X,Y) has an upper(lower) tail dependence.

Lemma 14.4.23 (tail dependence from copula).

- $$\lambda_l = \lim_{q\to 0^+} \frac{Pr(F_Y(Y) \le q, F_X(X) \le q)}{Pr(F_X(X) \le q)} = \lim_{q\to 0^+} \frac{C(q,q)}{q}.$$
- $$\lambda_u = \lim_{q\to 1^-} \frac{Pr(F_Y(Y) > q, F_X(X) > q)}{Pr(F_X(X) > q)} = \lim_{q\to 1^-} \frac{C^s(q,q)}{1-q},$$

  where C^s is the survival copula.

Remark 14.4.10 (tail dependence is independent of the marginal cdf).

- Note that tail dependence only depends on the correlation structure characterized by the copula; it is independent of the marginal cdf.
- The existence of tail will depend on margins.

Lemma 14.4.24 (tail independence of Gaussian copula). *Let (X,Y) be a bivariate random vector with Gaussian copula characterized by correlation coefficient ρ and continuous margins. Then*

$$\lambda_u = \lambda_l = 2 \lim_{x\to\infty} \Phi(x\sqrt{1-\rho}/\sqrt{1+\rho}) = 0.$$

Proof. From definition, we have

$$\begin{aligned}
\lambda_l &= \lim_{q\to 0^+} \frac{Pr(F_Y(Y) \le q, F_X(X) \le q)}{Pr(F_X(X) \le q)} \\
&= \lim_{q\to 0^+} \frac{Pr(Y \le f^{-1}(q), X \le \phi^{-1}(q))}{Pr(X \le f^{-1}(q))} \\
&= \lim_{q\to -\infty} \frac{Pr(Y \le q, X \le q)}{Pr(X \le q)} \\
&= \lim_{q\to -\infty} \frac{\int_{-\infty}^{q}\int_{-\infty}^{q} f(x,y)dxdy}{\int_{-\infty}^{q} f(x)dx} \\
&= \lim_{q\to -\infty} \frac{\int_{-\infty}^{q} f(x,q)dx}{f(q)} + \frac{\int_{-\infty}^{q} f(q,y)dy}{f(q)} \\
&= 2\lim_{q\to -\infty} \frac{\int_{-\infty}^{q} f(x,q)dx}{f(q)} \\
&= 2\lim_{q\to -\infty} f(x|y=q)dx
\end{aligned}$$

where $f(x,y)$ is the density of (X,Y), $f(x)$ is the marginal density, and we use L'hospital rule in the derivation. Note that $X|y = q \sim N(\rho q, 1-\rho^2)$ [Theorem 14.1.2]; therefore,

$$\lim_{q\to -\infty} f(x|y=q)dx = \Phi(\frac{q-\rho q}{\sqrt{1-\rho^2}}).$$

□

14.4.5 Estimating copula function

14.4.5.1 *Empirical copula method*

Definition 14.4.13 (empirical copula). *[11, p. 424] Suppose we have observation of n iid d dimensional random vectors,*

$$X^i = (X_1^{(i)}, X_2^{(i)}, ..., X_d^{(i)}), i = 1, 2, ..., n,$$

We can construct its empirical copula associated with the joint distribution of X via the following procedures:

- *(construct empirical marginal cdf)*

$$\hat{F}_k(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_k^{(i)} \le x), k = 1, 2, ..., d$$

- *(construct transformed uniform sample)*

$$(\hat{U}_1^{(i)}, ..., \hat{U}_d^{(i)}) = (\hat{F}_1(X_1^i), ..., \hat{F}_d(X_d^{(i)})), i = 1, ..., n$$

- *(construct empirical copula)*

$$\hat{C}(u_1, u_2, ..., u_d) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(\hat{U}_1^{(i)} \leq u_1, ..., \hat{U}_1^{(i)} \leq u_d)$$

Remark 14.4.11. The nature of copula is the cdf of the transformed uniform random vector $(F_1(X_1), F_2(X_2), ..., F_n(X_n))$ [Lemma 14.4.3].

14.4.5.2 *Maximum likelihood method*

Lemma 14.4.25 (maximum likelihood function and two-stage estimation method). *[11, p. 429] Suppose we have observation of n iid d dimensional random vectors,*

$$X^i = (X_1^{(i)}, X_2^{(i)}, ..., X_d^{(i)}), i = 1, 2, ..., n.$$

Assume the joint cdf for X is given by

$$F(x_1, x_2, ..., x_n) = C(F_1(x_1; \theta_1), ..., F_d(x_d; \theta_d)),$$

such that we have two sets of parameters given by

- $(\theta_1, ..., \theta_d)$ *for univariate distribution function* $F_1, F_2, ..., F_d$*;*
- θ_c *for the copula function* $C(u_1, ..., u_d)$*.*

The maximum log likelihood function is given by

$$\begin{aligned} & l(\theta_1, ..., \theta_d, \theta_c) \\ &= \sum_{i=1}^{n} \ln c(F_1(x_1^{(j)}; \hat{\theta}_1), ..., F_1(x_d^{(j)}; \hat{\theta}_d) + \sum_{i=1}^{n} \sum_{j=1}^{d} \end{aligned}$$

The first stage is to estimate univariate parameter $\theta_1, ..., \theta_d$ via

$$\hat{\theta}_i = \arg\max \sum_{j=1}^{N} \ln f_i(x_i^{(j)}; \theta_i).$$

The second stage is to estimate the copula parameters θ_c with the estimated univariate parameters $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_d$ fixed via

$$\hat{\theta}_c = \arg\max \sum_{j=1}^{N} \ln c(F_1(x_1^{(j)}; \hat{\theta}_1), ..., F_1(x_d^{(j)}; \hat{\theta}_d); \theta_c).$$

14.4.6 Applications of copula

14.4.6.1 *Generating correlated uniform random number*

Methodology 14.4.1 (conditional method for generating bivariate uniform random number with arbitrary joint distribution). *[10, p. 96] Suppose that we want to generate a pair of random variables with marginal uniform $U(0,1)$ and joint cdf given by copula $C(u,v)$.*

We use the following procedures:

- *Generate U and T independently from $U(0,1)$;*
- *Set $V = C_u^{-1}(T)$, where $C_u = \frac{\partial}{\partial u} C(u,v)$.*
- *The desired pair is (U,V).*

Proof. Note that [Theorem 14.4.4]

$$C_u = \frac{\partial}{\partial u} C(u,v) = Pr(V < v | U = u),$$

which is the conditional cdf. Then we use inverse transformation method to get V. □

Methodology 14.4.2 (conditional method for generating bivariate uniform random number with arbitrary joint distribution). *[10, p. 96]*

Suppose that we want to generate a set of n random variables with marginal uniform $U(0,1)$ and joint cdf given by copula $C_n(u_1, u_2, ..., u_n)$.

Further denote $C_{1:k}(u_1, ..., u_k)$ be the copula of $(U_1, U_2, ..., U_k), 2 \leq k \leq n$ and set $C_1(u_1) = u_1$.

We use the following procedures:

- *Simulate u_1 from $U(0,1)$.*
- *Simulate u_2 from the conditional distribution function $C_2(u_2|u_1)$.*
- *Simulate u_3 from the conditional distribution $C_3(u_3|u_1, u_2)$.*
- $\cdots$
- *Simulate u_n from the conditional distribution $C_n(u_n|u_1, u_2, ..., u_{n-1})$.*

where

$$\begin{aligned} C_k(u_k|u_1,...,u_{k-1}) &\triangleq Pr(U_k \le u_k | U_1 = u_1,...,U_{k-1} = u_{k-1}) \\ &= \frac{\frac{\partial^{k-1}}{\partial u_1 ... \partial u_{k-1}} C_{1:k}(u_1,...,u_k)}{\frac{\partial^{k-1}}{\partial u_1 ... \partial u_{k-1}} C_{1:k-1}(u_1,...,u_k - 1)} \end{aligned}$$

Proof. See Theorem 14.4.4 for how partial derivative is associated with conditional distribution. □

Example 14.4.9 (generate correlated uniform random variables with Gaussian copula correlation structure). Given a Gaussian copula C characterized by correlation matrix Σ. We can generate random vector $(X_1, X_2, ..., X_n)$ with uniform distribution and Gaussian copula correlation structure using the following procedures:

- First generate $(Y_1, Y_2, ..., Y_n) \sim MN(0, \Sigma)$.
- Then transform $X_1 = \phi(Y_1), X_2 = \phi(Y_2), ..., X_n = \phi(Y_n)$, where ϕ is standard normal cdf.

To understand the mechanism, note that Y_1, Y_2 alone is standard normal variable. Therefore $\phi(Y_1)$ and $\phi(Y_2)$ are uniform random variables [Lemma 14.4.2]. To show C is the copula associated with the cdf F, we have

$$\begin{aligned} F(x_1, x_2) &= Pr(X_1 \le x_1, X_2 \le x_2) \\ &= Pr(\phi(Y_1) \le x_1, \phi(Y_2) \le x_2) \\ &= Pr(\phi(Y_1) \le x_1, \phi(Y_2) \le x_2) \\ &= Pr(Y_1 \le \phi^{-1}(x_1), Y_2 \le \phi^{-1}(x_2)) \\ &= \Phi(\phi^{-1}(x_1), \phi^{-1}(x_2)) \\ &= C(x_1, x_2) \end{aligned}$$

where $C(u_1, u_2) = \Phi(\phi^{-1}(u_1), \phi^{-1}(u_2))$,$\Phi$ is the cdf for a multivariate normal distribution with zero mean and covariance matrix Σ, ϕ is the cdf for a standard normal variable. Note that the copula of a uniform distribution is itself [Lemma 14.4.5].

Example 14.4.10 (generate correlated uniform random variables with t copula correlation structure). [11, p. 420] Suppose we have n marginal distribution $F_i : \mathbb{R} \to [0,1], i = 1, 2, ..., n$. Given a T copula C characterized by correlation matrix Σ and degree of freedom v. We can generate samples of random vector $(X_1, X_2, ..., X_n)$ characterized by margins $F_1, F_2, ..., F_n$ and copula C using the following procedures:

- Generate $Z_1, Z_2, ..., Z_n$ as iid $N(0,1)$, and let $Z = (Z_1, Z_2, ..., Z_n)$.

- Generate a random $W \sim \chi^2(n)$ independent of Z.
- Return $X = \sqrt{\frac{v}{W}}CZ$, where C is the Cholesky decomposition of Σ such that $\Sigma = CC^T$.
- Return $U_i = T_v(X_i), i = 1, 2, ..., n$, where T_v is the univariate student t distribution with v degrees of freedom.

14.4.6.2 *Generating general correlated random number*

Methodology 14.4.3 (generate pair correlated random variables with Gaussian copula correlation structure). *Suppose we have two univariate marginal distribution $F_1 : \mathbb{R} \to [0,1]$ and $F_2 : \mathbb{R} \to [0,1]$. Given a Gaussian copula C characterized by correlation matrix Σ. We can generate samples of random vector (X_1, X_2) characterized by margins F_1, F_2 and copula C using the following procedures:*

- *First generate $(Y_1, Y_2) \sim MN(0, \Sigma)$.*
- *Then transform $X_1 = F_1^{-1}(\phi(Y_1)), X_2 = F_2^{-1}(\phi(Y_2))$, where ϕ is standard normal cdf.*
- *The random vector (X_1, X_2) has marginal distribution (F_1, F_2) and cdf*

$$F = C(F_1, F_2).$$

Proof. (1)Note that Y_1, Y_2 alone is standard normal variable. Therefore $\phi(Y_1)$ and $\phi(Y_2)$ are uniform random variables [Lemma 14.4.2]. Therefore X_1 and X_2 have marginal distribution (F_1, F_2). To show C is the copula associated with the cdf F, we have

$$\begin{aligned} F(x_1, x_2) &= Pr(X_1 \le x_1, X_2 \le x_2) \\ &= Pr(F_1^{-1}(\phi(Y_1)) \le x_1, F_2^{-1}(\phi(Y_2)) \le x_2) \\ &= Pr(\phi(Y_1) \le F_1(x_1), \phi(Y_2) \le F_2(x_2)) \\ &= Pr(Y_1 \le \phi^{-1}(F_1(x_1)), Y_2 \le \phi^{-1}(F_2(x_2))) \\ &= \Phi(\phi^{-1}(F_1(x_1)), \phi^{-1}(F_2(x_2))) \\ &= C(F_1(x_1), F_2(x_2)) \end{aligned}$$

where $C(u_1, u_2) = \Phi(\phi^{-1}(u_1), \phi^{-1}(u_2))$,$\Phi$ is the cdf for a multivariate normal distribution with zero mean and covariance matrix R, ϕ is the cdf for a standard normal variable. (2) Alternatively, we can use Methodology 14.4.6. □

Methodology 14.4.4 (generate correlated random variables with Gaussian copula correlation structure). *Suppose we have n marginal distribution $F_i : \mathbb{R} \to [0,1], i = 1, 2, ..., n$. Given a Gaussian copula C characterized by correlation matrix Σ. We can generate*

samples of random vector $(X_1, X_2, ..., X_n)$ *characterized by margins* $F_1, F_2, ..., F_n$ *and copula* C *using the following procedures:*

- *First generate* $(Y_1, Y_2, ..., Y_n) \sim MN(0, \Sigma)$.
- *Then transform* $X_1 = F_1^{-1}(\phi(Y_1)), X_2 = F_2^{-1}(\phi(Y_2)), ..., X_n = F_n^{-1}(\phi(Y_n))$,, *where* ϕ *is standard normal cdf.*
- *The random vector* $(X_1, X_2, ..., X_n)$ *has marginal distribution* $(F_1, F_2, ..., F_n)$ *and cdf*

$$F = C(F_1, F_2, ..., F_n).$$

Example 14.4.11 (generation of dependent default time). Let $T_1, T_2, ..., T_n$ denote the random default time for n loan borrowers. Assume the hazard curve for each borrower is given by $h_i(t), t \geq 0$ such that the marginal cdf of default time is given by

$$F_i(t) = Pr(T_i \leq t) = 1 - \exp(-\int_0^t h(s)ds).$$

Further assume the copula associated with the joint cdf is Gaussian copula with correlation matrix R.

Consider the following random number generating process

- Simulate $Y_1, Y_2, ..., Y_n$ from $MN(0, R)$.
- Obtain $T_1, T_2, ..., T_n$ using $T_i = F_i^{-1}(\phi(Y_i))$, where ϕ is the cdf of a standard normal.

Then samples of $T_1, T_2, ..., T_n$ will follow the joint cdf F.

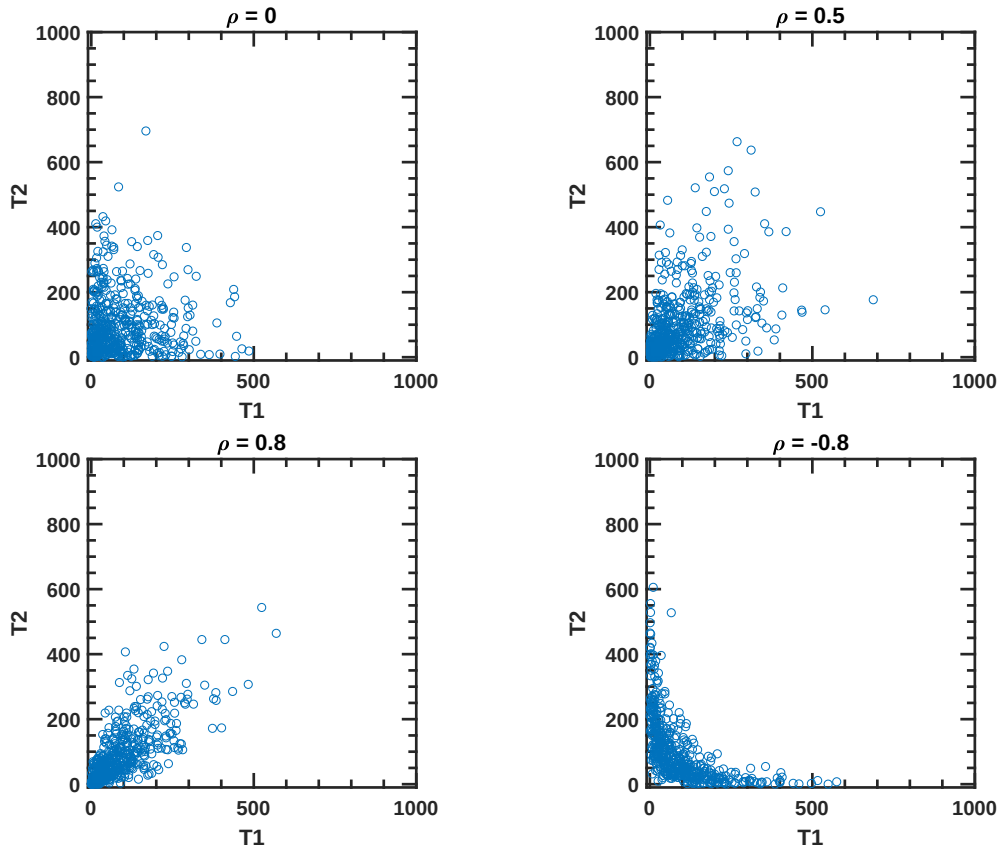


Figure 14.4.3: Generated correlated default time via Gaussian copula with different correlations. The hazard rate for both parties is $h(t) = 0.01$.

Methodology 14.4.5 (generating correlated random number with t copula). *Suppose we have n marginal distribution $F_i : \mathbb{R} \to [0,1], i = 1,2,...,n$. Given a T copula C characterized by correlation matrix Σ and degree of freedom v. We can generate samples of random vector $(X_1, X_2, ..., X_n)$ characterized by margins $F_1, F_2, ..., F_n$ and copula C using the following procedures:*

- *Generate $Z_1, Z_2, ..., Z_n$ as iid $N(0,1)$, and let $Z = (Z_1, Z_2, ..., Z_n)$.*
- *Generate a random $W \sim \chi^2(n)$ independent of Z.*
- *Return $X = \sqrt{\frac{v}{W}}CZ$, where C is the Cholesky decomposition of Σ such that $\Sigma = CC^T$.*
- *Set $U_i = T_v(X_i), i = 1,2,...,n$, where T_v is the univariate student t distribution with v degrees of freedom.*
- *Return $Y_i = F_i^{-1}(U_i), i = 1,2,...,n$.*

Example 14.4.12 (copula method for pricing basket option). Consider a financial basket option with payoff at future time T given by

$$V_T = max(\sum_{i=1}^{n} \frac{1}{n} S_T^{(i)} - K, 0),$$

where $S_T^{(i)}, i = 1, ..., n$ is the stochastic stock price i at time T, K is a constant.

To evaluate V_T via Monte Carlo method, we can generate samples of $S_T^{(i)}, i = 1, ..., n$ and then evaluate the expectation.

We make the following assumptions:

- We can construct the implied cdf F_i for each $S_T^{(i)}, i = 1, 2, ..., n$ by estimating historical data.
- The joint distribution of $S_T^{(1)}, S_T^{(2)}, ..., S_T^{(n)}$ has Gaussian copula with correlation matrix Σ. Note that the correlation matrix can be estimated from historical data.

Then we can use the following simulation method to generate one sample

- First generate $(Y_1, Y_2, ..., Y_n) \sim MN(0, \Sigma)$.
- Return $S_T^{(1)} = F_1^{-1}(\phi(Y_1)), S_T^{(2)} = F_2^{-1}(\phi(Y_2)), ..., S_T^{(n)} = F_n^{-1}(\phi(Y_n))$,, where ϕ is standard normal cdf.

We can similarly generate samples with t copula.

Methodology 14.4.6 (transform correlated uniform distribution to arbitrary distribution with same copula structure). *Let $(U_1, U_2, ..., U_n)$ be a uniform random vector with cdf F(or equivalently copula$C$ [Lemma 14.4.5]. Let $F_1, F_2, ..., F_n$ be the marginal cdf of a target cdf F^{target}. It follows that the random vector $(F_1^{-1}(U_1), F_2^{-1}(U_2), ..., F_n^{-1}(U_2))$ has marginal cdf $F_1, F_2, ..., F_n$ and copula structure C; that is, the cdf of $(F_1^{-1}(U_1), F_2^{-1}(U_2), ..., F_n^{-1}(U_2))$ is F^{target}.*

Proof. Let $U_i \sim U(0, 1)$, then

$$Pr(F_i^{-1}(u_i) < x_i) = Pr(u_i < F_i(x_i)) = F_i(x_i),$$

which implies that $F_i^{-1}(U_i)$ has marginal F_i.

The show the cdf of $(F_1^{-1}(U_1), F_2^{-1}(U_2), ..., F_n^{-1}(U_2))$ has the same copula as $(U_1, U_2, ..., U_n)$, we use the monotone transform invariance property of copula [Lemma 14.4.7]. □

14.4.6.3 *Multivariate distribution approximation with Gaussian copula*

Theorem 14.4.5. *Given random variable $X_1, X_2, ..., X_n$ with margins $F_{X_1}, F_{X_2}, ..., F_{X_n}$ and pair-correlation matrix R. We can construct a multivariate distribution such that recovers the margins and correlation via*

$$F(x_1, x_2, ..., x_n) = \Phi(\phi^{-1}(F_1(x_1)), \phi^{-1}(F_2(x_2)), ..., \phi^{-1}(F_n(x_n))),$$

Φ is the cdf for a multivariate normal distribution with zero mean and covariance matrix R, ϕ is the cdf for a standard normal variable.

Proof. From the theorem of constructing multivariate distribution from margins [Theorem 14.4.1], we have

$$F(x_1, x_2, ..., x_n) = C(F_1(x_1), F_2(x_2), ..., F_d(x_d)).$$

Note that the Gaussian copula with correlation matrix R [Definition 14.4.4] is given by

$$C(u_1, u_2, ..., u_d; R) = \Phi(\phi^{-1}(u_1), \phi^{-1}(u_2), ..., \phi^{-1}(u_d)),$$

where $u_1, u_2, ..., u_d \in [0, 1]$, Φ is the cdf for a multivariate normal distribution with zero mean and covariance matrix R, ϕ is the cdf for a standard normal variable. □

Remark 14.4.12 (implications).

- To fully determine the multivariate distribution, we usually need all the margins and all cross-term moments.
- With only margins and correlations given, we can construct a multivariate Gaussian distribution as an approximation.

14.5 Covariance structure and factor analysis

14.5.1 The orthogonal factor model

14.5.1.1 *Motivation and factor models*

In some real world applications, we often face the challenges of modeling the dependence for a large set of random variables. Directly modeling of the joint distribution is usually intractable. In section 14.4, we introduce Copula method as a parametric method to model dependence. In the section, we introduce factor model, which aims model the correlation structure by connecting them to a few random variables, called latent factors[2, p. 482].

Consider a p dimensional observable random vector X and a $m, m \ll p$ dimensional random vector F, called common/latent factors. Let $E[X] = \mu, \mu \in \mathbb{R}^p, Cov(X) = \Sigma, \Sigma \in \mathbb{R}^{p \times p}$. In the **orthogonal factor model**, **we assume** X **is related to** F via

$$X - \mu = LF + \epsilon,$$

where

- $L \in \mathbb{R}^{p \times m}$ is called matrix of **factor loadings**;
- ϵ is a p dimensional random vector such that $E[\epsilon] = 0_{p \times 1}, Cov(\epsilon) = \psi, Cov(\epsilon)_{ij} = \psi_i, i = j, Cov(\epsilon)_{ij} = 0, i \neq j$.
- $E[F] = 0_{m \times 1}, Cov(F) = I_{m \times m}$.
- F and ϵ are independent such that $Cov(\epsilon, F) = 0$.

orthogonal factor model

Remark 14.5.1 (parameter summary in factor models). The original covariance structure $\Sigma \in \mathbb{R}^{p \times p}$ has $p(p+1)/2$ parameters. In the new covariance structure resulted from factor model, we have in total $p(m+1)$ parameters:

- mp factor loadings in the matrix L.
- p specific variance in the matrix ψ.

14.5.1.2 *Covariance structure implied by factor model*

An important application of factor models is to serve as a low dimensional approximation to high-dimensional covariance matrix of the random vector $X_1, ..., X_p$. In the factor model, we use a L matrix of pm elements and ψ matrix of p elements to approximate/reproduce the original covariance matrix Σ of X, containing $p(p+1)/2$ elements.

Lemma 14.5.1 (covariance structure implied by factor model). *[2, p. 483] In the factor model, the covariance of X is given by*

$$E[(X-\mu)(X-\mu)^T] = LL^T + \psi.$$

In addition, the covariance matrix between X and F is given by

$$Cov(X,F) \triangleq E[(X-\mu)F^T] = L.$$

Proof. (1)

$$\begin{aligned} E[(X-\mu)(X-\mu)^T] &= E[(LF+\epsilon)(LF+\epsilon)^T] \\ &= E[LFF^TL^T + \epsilon(LF)^T + LF\epsilon^T + \epsilon\epsilon^T] \\ &= LIL^T + 0 + 0 + \psi. \end{aligned}$$

(2)

$$\begin{aligned} E[(X-\mu)F^T] &= E[(LF+\epsilon)F^T] \\ &= E[LFF^T + \epsilon F^T] \\ &= LI + 0 = L. \end{aligned}$$

□

Remark 14.5.2 (non-uniqueness of factor model). **[anderson2009introduction]**[2, pp. 488, 504] Let Σ, L, F, ψ be the results of factorization such that

$$X - \mu = LF + \epsilon,$$

and

$$\Sigma = LL^T + \psi.$$

Let $U \in \mathbb{R}^{m\times m}$ be a orthonormal matrix such that $UU^T = I$. Then the factors and the associated factor loadings can also take form $\hat{L} = LU, \hat{F} = U^TF$ also satisfy

$$X - \mu = \hat{L}\hat{F} + \epsilon,$$

and

$$\Sigma = \hat{L}\hat{L}^T + \psi.$$

To understand this, note that

$$\hat{L}\hat{F} = LUU^TF = LUU^TL^T = LL^T$$

and

$$\hat{L}\hat{L}^T = LU(LU)^T = LUU^TL^T = LL^T.$$

Example 14.5.1. Consider a one factor model given by

$$V_i = a_i Y + \sqrt{1 - a_i^2} Z_i, i = 1, 2, ..., n,$$

where $(F, Z_1, Z_2, ..., Z_n)$ are independent standard normal variables.

Then the covariance structure implied by the factor model is given by

$$\begin{aligned} Cov &= \begin{bmatrix} a_1 \\ a_2 \\ \cdots \\ a_n \end{bmatrix} [a_1, a_2, ..., a_n] + \begin{bmatrix} 1 - a_1^2 & & & \\ & 1 - a_2^2 & & \\ & & \ddots & \\ & & & 1 - a_n^2 \end{bmatrix} \\ &= \begin{bmatrix} a_1^2 & a_1 a_2 & \cdots & a_1 a_n \\ a_2 a_1 & a_2 a_2 & \cdots & a_2 a_n \\ \vdots & \cdots & \ddots & \vdots \\ a_n a_1 & a_n a_2 & \cdots & a_n a_n \end{bmatrix} \begin{bmatrix} 1 - a_1^2 & & & \\ & 1 - a_2^2 & & \\ & & \ddots & \\ & & & 1 - a_n^2 \end{bmatrix} \\ &= \begin{bmatrix} 1 & a_1 a_2 & \cdots & a_1 a_n \\ a_2 a_1 & 1 & \cdots & a_2 a_n \\ \vdots & \cdots & \ddots & \vdots \\ a_n a_1 & a_n a_2 & \cdots & 1 \end{bmatrix} \end{aligned}$$

which is also a valid correlation matrix.

14.5.2 Parameter estimation

14.5.2.1 *Data collection and preparation*

In the factor model, we need to estimate L and Ψ. We will go through two methods PCA and maximum likelihood estimation. The first step is to prepare data.

- Suppose each data vector x_i has p components. Then we can form a $n \times p$ data matrix X.

- We can transform the data matrix to sample covariance matrix [subsection 14.1.1] via
$$S = \frac{1}{n-1} X^T (I - \frac{1}{n} J) X.$$
- The sample covariance matrix and sample correlation matrix will then be used to build factor models.

14.5.2.2 *PCA method*

Lemma 14.5.2 (PCA method for estimation of factor loadings). *[2, p. 488] Let $S \in \mathbb{R}^{p \times p}$ be the sample covariance matrix of data matrix X. Let S adopt eigendecomposition*

$$S = \sum_{i=1}^{p} \lambda_i u_i u_i^T,$$

where $\lambda_1 \leq \lambda_2 \cdots \leq \lambda_p$.

It follows that

- *Then an estimation of column i of factor loading matrix L is given by*
$$\hat{L}_i = \sqrt{\lambda_i} u_i$$
such that
$$\hat{L}\hat{L}^T = \sum_{i=1}^{m} \lambda_i u_i u_i^T.$$
- *An estimation of ψ is given by*
$$\hat{\psi} = diag(S - \hat{L}\hat{L}^T).$$
- *The factors $F_1, F_2, ..., F_m$ are assumed to be independent standard normal random variables estiamted to have samplings given by*

Proof. Note that
$$\hat{L}_i \hat{L}_j^T = 0, \forall i \neq j.$$
Then
$$LL^T = \sum_{i=1}^{m} \hat{L}_i \hat{L}_i^T = \sum_{i=1}^{m} \lambda_i u_i u_i^T.$$
□

Remark 14.5.3 (how to determine the number of factors)**.** We can empirically determine the number of factors based on the following considerations:

- m is chosen to explain most of the variance.
- m is chosen such that the number of parameters $p(m+1)$ is smaller than the number of parameters $p(p+1)/2$ in the full covariance matrix.

14.5.2.3 *Maximum likelihood method*

Using the Maximum Likelihood Estimation Method we must assume that the data are independently sampled from a multivariate normal distribution with mean vector μ and variance-covariance structure take the form of

$$\Sigma = LL^T + \psi,$$

where $L \in \mathbb{R}^{p\times m}$ is the factor loadings and ψ is the diagonal matrix of specific variances.

Methodology 14.5.1. *[2, p. 496] Suppose we have data vectors from n observations $x_1, x_2, ..., x_n \in \mathbb{R}^p$. The maximum likelihood estimation involves estimating the mean μ, the matrix of factor loadings, and the specific variance matrix ψ.*

The likelihood function is given by

$$L(\mu,\Sigma) = (2\pi)^{-\frac{np}{2}}|\Sigma|^{-\frac{n}{2}}\exp(-\frac{1}{2}Tr(\Sigma^{-1}(\sum_{j=1}^{n}(x_j-\mu)(x_j-\mu)^T)))$$

where $\Sigma = LL^T + \psi$. and the log likelihood function is given by

$$l(\mu,\Sigma) \triangleq \ln L = \frac{np}{2}\ln(2\pi) - \frac{n}{2}\ln|\Sigma| - \frac{1}{2}Tr(\Sigma^{-1}(\sum_{j=1}^{n}(x_j-\mu)(x_j-\mu)^T)).$$

14.5.3 Factor score estimation

Methodology 14.5.2 (weighted least square method to find factor score). *Suppose we have data vectors from n observations $x_1, x_2, ..., x_n \in \mathbb{R}^p$. Further suppose that we have already had the factor model given by*

$$X = \mu + Lf + \epsilon,$$

with $\mu \in \mathbb{R}^p, L \in \mathbb{R}^{p\times m}$ being given. Then the factor scores $f_1, f_2, ..., f_n \in \mathbb{R}^m$ associated with data vector can be estimated by minimizing the following optimization problem

$$\min_{f_i \in \mathbb{R}^m} (x_i - \mu - Lf_i)^T \psi^{-1} (x_i - \mu - Lf_i).$$

Furthermore, the minimizer is given by [see generalized least square Theorem 15.1.12]

$$\hat{f}_i = (L^T \psi^{-1} L)^{-1} L^T \psi^{-1} (x_i - \mu).$$

14.5.4 Application I: Joint default modeling

14.5.4.1 *Single factor model*

We consider a financial application of factor modeling. Consider n parties that will default before the next period t with unconditional probability p_i (i.e., a Bernoulli random variable). Because default behavior of multiple parties can be quite correlated, particularly during periods of financial crisis, we also like to capture their **joint default behavior** for risk management purpose.

Define a new proxy **standard random variable** X_i by

$$X_i = a_i F + \sqrt{1 - a_i^2} Z_i, i = 1, 2, ..., n,$$

where F is a common factor (such as GDP) affecting defaults for all parties and Z_i is a noise term affecting only party i. F and Z_i are independent standard normal variables.

It follows that

- $$Pr(X_i \text{ will default}) \triangleq p_i = Pr(X_i < \phi^{-1}(p_i)),$$
  where ϕ is the cdf of the standard normal variable.
- The **conditional default probability** of firm i before t conditioning on the observation of F is given by

$$Pr(X_i \text{ will default} | F = f) \triangleq Pr(T_i < t | F = f) = \phi\left(\frac{\phi^{-1}(p_i) - a_i f}{\sqrt{1 - a_i^2}}\right),$$

To see this, note that $X_i \sim N(0, 1)$. Therefore,

$$Pr(X_i < \phi^{-1}(p_i)) = \phi(\phi^{-1}(p_i)) = p_i.$$

For the second point,

$$\begin{aligned} Pr(X_i < \phi^{-1}(p_i)|F = f) &= Pr(a_i f + \sqrt{1 - a_i^2} Z_i < \phi^{-1}(p_i)) \\ &= Pr(Z_i < \frac{\phi^{-1}(p_i) - a_i f}{\sqrt{1 - a_i^2}}) \end{aligned}$$

We can see that when a_i is large, all parties tend to default together when an event $f < 0$ occurs. On the other hand, when $a_i = 0$, all parties will default independently.

The single factor approach can be extended to capture the **joint default time modeling**. Let $T_1, T_2, ..., T_n$ denote the random default time for n parties. Assume marginal cdf of default time is given by $Q_i(t)$. Define a new proxy random variable $X_i = \phi^{-1}(Q_i(T_i)), i = 1, 2, ..., n$, and **assume** X_i can be modeled by

$$X_i = a_i F + \sqrt{1 - a_i^2} Z_i, i = 1, 2, ..., n,$$

where F is a common factor affecting defaults for all firms and Z_i is a factor affecting only firm i. F and Z_i are independent standard normal variables. It follows that the **conditional default probability** of firm i before t conditioning on the observation of F is given by

$$Q_i(t|F = f) \triangleq Pr(T_i < t|F = f) = \phi(\frac{\phi^{-1}(Q_i(t)) - a_i f}{\sqrt{1 - a_i^2}}),$$

where ϕ is the standard normal cdf.

To see this, we have

$$\begin{aligned} Pr(T_i < t) &= Pr(T_i < t) \\ &= Pr(Q_i(T_i) < Q_i(t)) \\ &= Pr(\phi^{-1}(Q_i(T_i)) < \phi^{-1}(Q_i(t))) \\ &= Pr(X_i < \phi^{-1}(Q_i(t))) \\ &= Pr(a_i F + \sqrt{1 - a_i^2} Z_i < \phi^{-1}(Q_i(t))) \\ &= Pr(Z_i < \frac{\phi^{-1}(Q_i(t)) - a_i F}{\sqrt{1 - a_i^2}}) \\ \implies Q_i(T|F = f) \triangleq Pr(T_i < t|F) &= Pr(Z_i < \frac{\phi^{-1}(Q_i(T)) - a_i F}{\sqrt{1 - a_i^2}}|F = f) \\ &= \phi(\frac{\phi^{-1}(Q_i(t)) - a_i f}{\sqrt{1 - a_i^2}}) \end{aligned}$$

Remark 14.5.4 (the correlation structure for the Gaussian copula implied by the factor model)**.** Consider a one factor model given by

$$V_i = a_i Y + \sqrt{1 - a_i^2} Z_i, i = 1, 2, ..., n,$$

where $(F, Z_1, Z_2, ..., Z_n)$ are independent standard normal variables.

Then the correlation structure implied by the factor model is given by [Figure 14.5.1]

$$\begin{aligned} Cov &= \begin{bmatrix} a_1 \\ a_2 \\ \cdots \\ a_n \end{bmatrix} [a_1, a_2, ..., a_n] + \begin{bmatrix} 1 - a_1^2 & & & \\ & 1 - a_2^2 & & \\ & & \ddots & \\ & & & 1 - a_n^2 \end{bmatrix} \\ &= \begin{bmatrix} a_1^2 & a_1 a_2 & \cdots & a_1 a_n \\ a_2 a_1 & a_2 a_2 & \cdots & a_2 a_n \\ \vdots & \cdots & \ddots & \vdots \\ a_n a_1 & a_n a_2 & \cdots & a_n a_n \end{bmatrix} \begin{bmatrix} 1 - a_1^2 & & & \\ & 1 - a_2^2 & & \\ & & \ddots & \\ & & & 1 - a_n^2 \end{bmatrix} \\ &= \begin{bmatrix} 1 & a_1 a_2 & \cdots & a_1 a_n \\ a_2 a_1 & 1 & \cdots & a_2 a_n \\ \vdots & \cdots & \ddots & \vdots \\ a_n a_1 & a_n a_2 & \cdots & 1 \end{bmatrix} \end{aligned}$$

(a) Calculation of first default probability(thick solid black) from individual default probability of 10 reference names using Gaussian one factor model with $\rho = 0.5$.

(b) First default probability as a funciton of correlation of the underlying names.

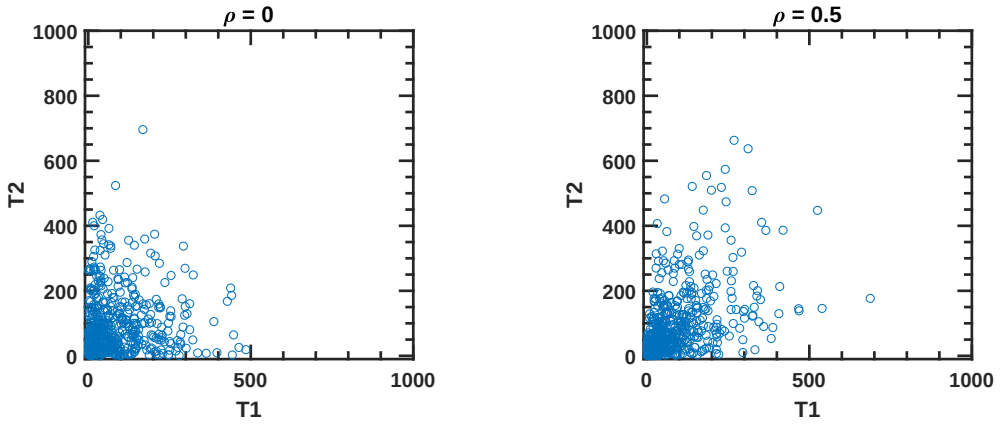


Figure 14.5.1: Correlation structure for the Gaussian copula implied by the factor model.

14.5.4.2 *Multiple factor model*

Finally, we can introduce multiple factors to give a richer representation of the correlation structure, as we make the following modification. Let $T_1, T_2, ..., T_n$ denote the random default time for n parties. Assume marginal cdf of default time is given by $Q_i(t)$. Define a new proxy random variable $X_i = \phi^{-1}(Q_i(T_i)), i = 1, 2, ..., n$, and **assume** X_i can be modeled by

$$X_i = \sum_{j=1}^{m} a_{ij} F_j + \sqrt{1 - \sum_{j=1}^{m} a_{ij}^2} Z_i, i = 1, 2, ..., n,$$

where $F_1, F_2, ..., F_m$ are common factors affecting defaults for all firms and Z_i is a factor affecting only firm i. $F_1, F_2, ..., F_m$ and Z_i are mutually independent standard normal variables.

The **conditional default probability** of firm i before t conditioning on the observation of $F_1, F_2, ..., F_m$ is given by

$$Q_i(t|F_1 = f_1, ..., F_m = f_m) \triangleq Pr(T_i < t|F_1 = f_1, ..., F_m = f_m) = \phi\left(\frac{\phi^{-1}(Q_i(t)) - \sum_{j=1}^{m} a_{ij} f_j}{\sqrt{1 - \sum_{j=1}^{m} a_{ij}^2}}\right),$$

where ϕ is the standard normal cdf.

14.5.5 Application II: factor models for stock return

14.5.5.1 *Overview*

In portfolio management and financial risk analytics, it is often desirable to identify a small set of risk factors that underlying the stochastic dynamics of hundreds and thousands of stocks.

Mathematically, let $r_1, ..., r_n$ be the stochastic return of n assets. Seeking risk factors can be simplified to a linear risk model given by

$$r_i(t) = a_i(t) + \sum_{j=1}^{m} b_{i,j}(t) f_j(t) + e_i(t), \forall i = 1, 2, ..., n$$

where $f_1, ..., f_m, m \ll n$ are stochastic risk factors, e_i is noise term specific to asset i satisfying $E[e_i] = 0, E[e_i e_j] = 0, \forall i \neq j$.

The applications of this linear factor model include

- Capture the correlation structure via fewer parameters.
- Managing risks from $r_1, ..., r_n$ is reduced to managing risks from a smaller set of risks $f_1, ..., f_n$.
- Understand the risk dynamics by studying these common factors.

These factors $f_1, ..., f_M$ can be external, such that GDP, inflation, etc. they can also be internal, such as linear combination of X_1, X_2, X_3 obtained via PCA or other statistical methods like data mining.

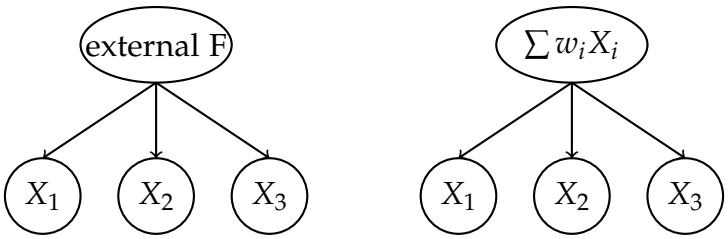


Figure 14.5.2: Factor model using (a) external factors or (b) internal factors.

Example 14.5.2 (a macroeconomic factor model). [13, p. 635] Consider a factor model in which the returns of stocks are correlated with surprises in interest rates and surprises in GDP growth. For stock $i = 1, 2, ..., N$, the return is modeled by

$$R_i = a_i + b_{i1} F_{INT} + b_{i2} F_{GDP} + \epsilon_i,$$

where

- R_i is the return of stock i
- a_i is the expected return to stock i
- b_{i1} is the sensitivity of the return to stock i to interest rate surprise
- b_{i2} is the sensitivity of the return to stock i to GDP growth surprise
- F_{INT} is the surprise in interest rate
- F_{GDP} is the surprise in GDP
- ϵ_i an error term with a zero mean that represents the portion of the return to stock i not explained by the factor model

Note that we define **surprise** in general as the actual value minus the predicted(or exptected) value. For example,

$$actual\ inflation = \ predicted\ inflation\ +\ inflation\ surprise.$$

Example 14.5.3 (factor model for portfolio optimization). In the classical portfolio optimization framework, we require the mean and covariance matrix for the n assets. They can be related to the single-factor model parameters as:

$$E[r_i] = a_i + \sum_{j=1}^{m} b_{ij} E[f_j]$$

$$\sigma_i^2 = Var[r_i] = \sum_{j=1}^{m} b_{ij}^2 \sigma_{f_j}^2 + \sum_{k<j}^{m} 2 b_{ik} b_{ij} cov(f_k, f_j) + \sigma_{e_i}^2$$

$$\sigma_{ij} = Cov(r_i, r_j) = \sum_{k=1}^{m} \sum_{p=1}^{m} b_{ik} b_{jp} cov(f_k, f_p), i \neq j$$

Remark 14.5.5 (cross-panel parameter fitting). Consider a model of return

$$r = Xb + u,$$

where r is an N vector of excess returns, X is an N by K matrix of factor exposures, b is a K dimension vector of factor returns, and u is an N dimensional vector of specific returns, and we further assume the factor loading matrix X is given.

Then we can obtain factor vector b from the weighted least square minimization given by

$$\min_b (Xb - r)^T \Delta^{-1} (Xb - r).$$

The final result is from [Theorem 15.1.12]:

$$b = (X^T \Delta^{-1} X)^{-1} X^T \Delta^{-1} r.$$

14.5.5.2 *The Fama-French 3 factor model*

The section we study the arguably most famous factor model, known as the Fama-French 3 factor model[14, 15], for stock market returns. In the Fama-French 3 factor model, the asset i return is given by

$$r_i = E[r_i] + \beta_{i1}(R_m - E[R_m]) + \beta_{i2}(SMB - E[SMB]) + \beta_{i3}(HML - E[HML]) + \epsilon_i$$

where

- $R_m - r_f$ is the return on a market value-weighted index in excess of the one-month T-bill rate.
- SMB ='small [market capitalization] minus big' factor. Mathematically, SMB is the average **raw**[1] return on three small-cap portfolios minus the average return on three large-cap portfolio. When small stocks do well relative to large stocks this will be positive, and when they do worse than large stocks, this will be negative. SMB also refers to as **size premium**. Note that $E[SMB] \neq 0$.
- HML = 'high [book/price] minus low' factor. Mathematically, HML is the average **raw** return on two high book-to-market portfolios minus the average return on two low book-to-market portfolios. HML also refers to a **value premium**. $E[HML] \neq 0$.

Remark 14.5.6 (factor portfolios as proxy risk drivers). [16, p. 70]

- The fundamental reasons determines a stock's return is the company's management system, technology, the ability to adaptive, efficiency etc. However, these reasons/factors are hard to quantify and observe. Therefore, we use the performance of different companies as the proxy to these factors.
- The Fama-French model views the size and value factors as representing ('proxying for') a set of underlying risk factors. For example, small market-cap companies may be subject to risk factors such as less ready access to private and public credit markets and competitive disadvantages. High book-to-market may represent shares with depressed prices because of exposure to financial distress. The model views the return premiums to small size and value as compensation for bearing types of systematic risk.
- Fama and French create a portfolio designed to have returns that mimic the returns associated with the size effect and propose using the returns of this portfolio as a risk factor.

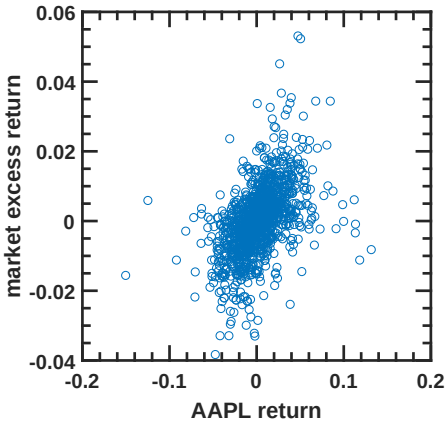
We first analyze the statical properties of the risk factors. Table 14.5.1 shows the monthly excess return statistics of Fama-french 3 factor portfolios. We can see that SMB and HML portfolios have positive premium, which agree with the size and value premium we discussed.

1 when say here raw return to emphasize that it is not excess return

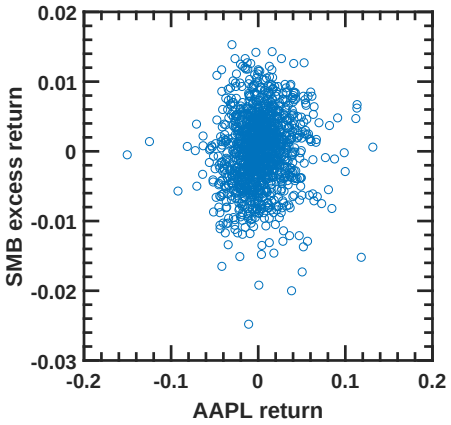
Table 14.5.1: statistics on Fama-French 3 factors from July 1963 to Dec. 1991.

factor name	**mean**	**std**	**correlation**		
			$r_M - r_f$	$SMB - r_f$	$HML - r_f$
$r_M - r_f$	0.43	4.54	1		
$SMB - r_f$	0.27	2.89	-0.38	1	
$HML - r_f$	0.40	2.54	0.34	-0.08	1

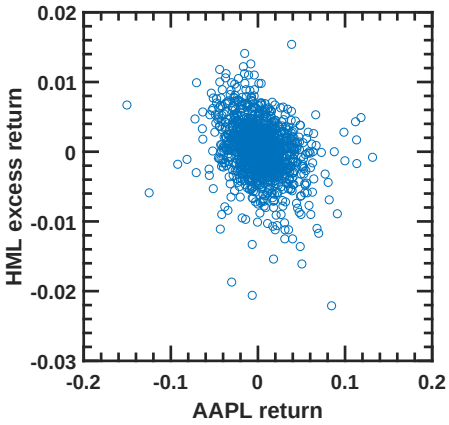
For each asset i, say AAPL stock, we can perform a linear regression based on observed returns and factor data to get coefficients $\beta_1, \beta_2, \beta_3$. Figure 14.5.3 shows the AAPL return data vs. factor realization in a historical period. Clearly, AAPL return is quite positively correlated with the factor R_M, and much less with SMB and HML. More detailed results can be found in Table 14.5.2.



(a) Scatter plot for AAPL daily return vs. market excess daily return from 2001-Oct to 2006-Oct.



(b) Scatter plot for AAPL daily return vs. SMB factor excess daily return from 2001-Oct to 2006-Oct.



(c) Scatter plot for AAPL daily return vs. HML factor excess daily return from 2001-Oct to 2006-Oct.

Figure 14.5.3: Scatter plot of AAPL return vs. market excess return, SMB excess return and HML excess return.

Table 14.5.2: AAPL stock return modeled by the Fama-French 3 factor model.

	estimate	**SE**	**tStat**	**pValue**
α	0.00186	0.00065	2.877	0.0041
β_{MKT}	1.2058	0.07012	17.196	1.4e-59
β_{SMB}	0.3753	0.12183	3.0806	0.00211
β_{HML}	-0.6583	0.17566	-3.7475	0.000187
R^2	0.264	adjusted R^2	0.263	

14.6 Graphical models

14.6.1 Fundamentals

Graphical models are an intuitive way of representing and visualizing the relationships between many random variables. A graph can help extract the conditional independence relationships among random variables. Thus we can answer questions like "Is A independent from B given that we know the value of C?"

Graphical models can be roughly divided into directed graphical models and undirected graphical models, Our focus is directed graphical models, known as Bayesian graphical model. More formally, a graph $G = (\mathcal{V}, \mathcal{E})$ consists of a set of nodes or vertices $\mathcal{V} = \{1, 2, ..., V\}$, and a set of edges $\mathcal{E} = \{(s, t) : s, t \in \mathcal{V}\}$. Nodes represent random variables, and edges representing assumed causal relationship and conditional independence between nodes or random variables. The connectivity between nodes can be represented by a matrix G, where $G(s, t) = 1$ if $(s, t) \in \mathcal{E}$. **Directed acyclic graph(DAG)** is a directed graph with no directed cycles. Related random variables of a random variable are classified as **parents**, **children**, **ancestors**, and **descendants**, which are defined by

- Parent of a node: $Pa(s) = \{t | G(t, s) = 1\}$
- Child of a node: $ch(s) = \{t | G(s, t) = 1\}$
- ancestors of a node: $anc(t)$ is the set of nodes s that has a directed path from s to t.
- descendants of a node: $anc(t)$ is the set of nodes s that has a directed path from t to s.

A directed graphical model $G = (\mathcal{V}, \mathcal{E})$ is a type of graphical model whose graph is a directed acyclic graph(DAG). In the graph, each node s is conditional independent of all its ancestor nodes except the parent nodes when conditioned on the parent nodes of s. Directed graphical model is also called **Bayes network**(the parameter as a random variable can be represented by a node), **brief network**, and **causal network**(because the directed arrows can be interpreted as causal relations.)

The first important application of graphical model is to allow decomposition and factorization of a complex joint distribution into simpler one.

We consider a naive decomposition of joint distributions.

Lemma 14.6.1 (joint distribution decomposition,chain rule). *Let P be the joint distribution on random variables $X_{1:N} \triangleq (X_1, X_2, ..., X_N)$, then we can decompose the joint distribution as*

$$P(X_{1:V}) = P(X_1)P(X_2|X_1)P(X_3|X_{1:2}...P(X_N|X_{1:N-1}) = \prod_{i=1}^{N} P(X_i|X_{1:i-1}).$$

This decomposition still holds if we permute the index of the X_i.

Proof. Apply the definition of conditional distribution $P(X, Y) = P(X)P(Y|X)$ from front to back. □

For any joint distribution, we can always represent it by a fully connected graphical model, as showed in Figure 14.6.1.

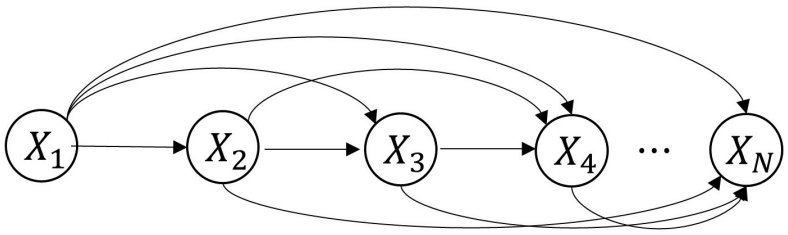


Figure 14.6.1: A fully connected graphical model representing the joint distribution $P(X_1, ..., X_N)$.

Graphical models can be used to represent the conditional independence among a set of random variable. Formally, we have

Definition 14.6.1 (set of conditional independence). *[17, p. 60][18, p. 324]Let P be a distribution over a set of random variables $\mathcal{X}$. Random variables X and Y are conditional independent given Z, denoted as $X \perp Y|Z$, if*

$$P(X, Y|Z) = P(X|Z)P(Y|Z)$$

the set of all such conditional independence is denoted as $I(P)$.

A graph model G is an independence map for a joint distribution P if

$$I(G) \subseteq I(P)$$

where $I(G)$ the set of all conditional independence assumptions encoded by G.

Therefore, a fully connected graph G [Figure 14.6.1] is the I-map for all distributions P defined over the same random variables. Note that for a fully connected graph G, $I(G) = \emptyset$.

The goal of graphical model is to capture all the conditional independence relationship, but not omit any, for a joint distribution P. In other words, we are seeking a graph model G that the **minimal I-map** of P.

A graphical model can lead to simplified factorization of a joint distribution.

Theorem 14.6.1 (factorization theorem). *If graph G is an I-map of P, then*

$$P(X_1...X_n) = \prod_i^n P(X_i|Pa(X_i))$$

Proof. by the chain rule lemma, we have

$$P(X_1...X_n) = \prod_i^n P(X_i|X_1...X_i))$$

□

Example 14.6.1. Consider the following graph models in Figure 14.6.2.

- For graphical model (A), we can have the following factorization

$$P(A, B, C, D, E) = P(A)P(B)P(C|A, B)P(D|B, C)P(E|C, D).$$

- Consider the graphical model (B). A data set of N points are generated iid from a Gaussian distribution with parameters μ and σ. The joint probability is given by

$$P(X_1, X_2, ..., X_N, \mu, \sigma) = P(\mu)P(\sigma)\prod_{n=1}^N P(X_n|\mu, \sigma).$$

- Consider the graphical model (C) representing a Markov chain. The joint probability can be decomposed by

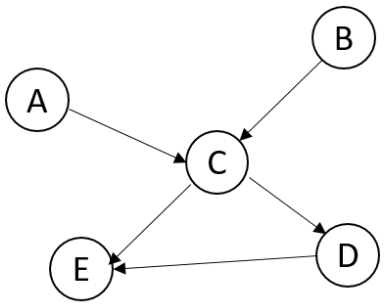
$$P(X_1, X_2, ..., X_N) = P(X_N|X_{N-1})P(X_{N-1}|X_{N-2})\cdots P(X_1).$$

- Consider the graphical model (D) representing a second-order Markov chain. The joint probability can be decomposed by

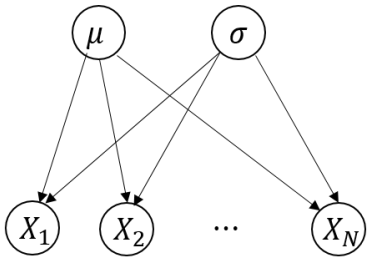
$$P(X_1, X_2, ..., X_N) = P(X_N|X_{N-1}, X_{N-2})P(X_{N-1}|X_{N-2}, X_{N-3}) \cdots P(X_3|X_1, X_2)P(X_1)P(X_2).$$

- Consider the graphical model (E) representing a hidden Markov chain. The joint probability can be decomposed by

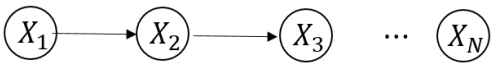
$$P(X_1, ..., X_N, Z_1, ..., Z_N) = P(Z_N|Z_{N-1})P(Z_{N-1}|Z_{N-2}) \cdots P(Z_1) \prod_{n=1}^{N} P(X_i|Z_i).$$



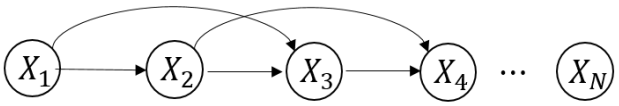
(a) Graphicial model (a).



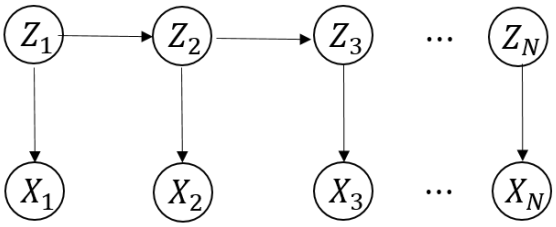
(b) Graphicial model (b). μ, σ are model parameters.



(c) Graphical model (c) for a Markov chain model.



(d) Graphical model (d) for a second-order Markov chain model.



(e) Graphical model (e) for a hidden Markov chain model.

Figure 14.6.2: Graphical model examples.

Example 14.6.2 (efficient inference via factorization). Consider a joint distribution on random variables A, B, C, D, E has the following factorization given by

$$P(A, B, C, D, E) = P(A)P(B)P(C|A, B)P(D|B, C)P(E|C, D).$$

To calculate

$$P(A|C = c) = \frac{P(A, C = c)}{P(C = c)},$$

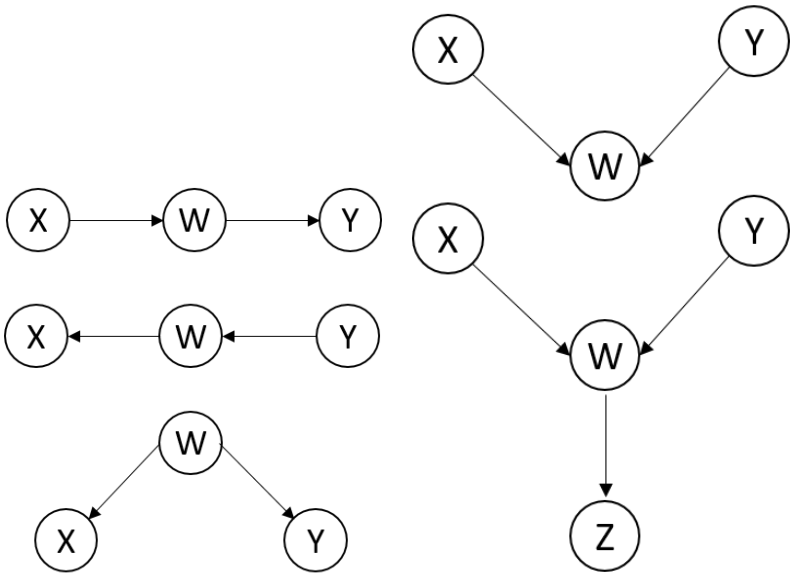
we have

$$\begin{aligned} P(A, C = c) &= \sum_{B,D,E} P(A)P(B)P(C = c|A, B)P(D|B, C = c)P(E|C = c, D) \\ &= \sum_{B} P(A)P(B)P(C = c|A, B) \sum_{D} P(D|B, C = c) \sum_{E} P(E|C = c, D) \\ &= \sum_{B} P(A)P(B)P(C = c|A, B) \end{aligned}$$

We now study how a graph model encodes conditional independence relationship. An important concept, d-separation [Figure 14.6.3], is introduced as follows.

Definition 14.6.2 (d-separation). *[18, p. 324] In a directed acyclic graphic model, a set of nodes $\mathcal{V}$ d-separates X from Y if **every undirected path** between X and Y is blocked by $\mathcal{V}$. A path is blocked by $\mathcal{V}$ if there is a node W on the path such that either*

- *W has converging arrows along the path $(\rightarrow W \leftarrow)$ and neither W nor its descendants are in $\mathcal{V}$.*
- *W does not have converging arrows along the path $(\rightarrow W \rightarrow)$ or $\leftarrow W \rightarrow$ and $W \in \mathcal{V}$.*



(a) X and Y are d-separated by W

(b) X and Y are not d-separated by W and Z.

Figure 14.6.3: d-separation examples.

Theorem 14.6.2 (d-sepration and conditional independence). *In a directed acyclic graphic model, X is conditional independent from Y given $\mathcal{V}$ if $\mathcal{V}$ d-separates X from Y.*

Remark 14.6.1 (correlation vs. causation)**.** The graphical modeling framework also help us distinguish correlation and causation

- causation will imply correlation.
- correlation will **not** imply causation.

For any two correlated events, A and B, the following relationships are possible:

- A causes B; (direct causation)
- B causes A; (reverse causation)
- A and B are **consequences of a common cause of C**, but do not cause each other; graphically, we have $A \leftarrow C \rightarrow B$;
- A causes B and B causes A (bidirectional or cyclic causation);
- A causes C which causes B (indirect causation); graphically, we have $A \rightarrow C \rightarrow B$;

14.7 Notes on Bibliography

For linear regression models, see [19][20]. For, linear models with *R* resources, see [21].

For multivariate statistical analysis, see [2][**anderson2009introduction**].

For copula, see [9][22][3][23].

BIBLIOGRAPHY

1. Hogg, R. V., McKean, J. & Craig, A. T. Introduction to Mathematical Statistics, 7 ed (2012).
2. Johnson, R. & Wichern, D. *Applied Multivariate Statistical Analysis* ISBN: 9780131877153 (Pearson Prentice Hall, 2007).
3. McNeil, A. J., Frey, R. & Embrechts, P. *Quantitative risk management: Concepts, techniques and tools* (Princeton university press, 2015).
4. Ma, Y. & Vidal, R. Generalized principal component analysis. *Unpublished Notes* (2002).
5. Tipping, M. E. & Bishop, C. M. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61,** 611–622 (1999).
6. Witten, D. M. & Tibshirani, R. J. Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical applications in genetics and molecular biology* **8,** 1–27 (2009).
7. Witten, D. M., Tibshirani, R. & Hastie, T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10,** 515–534 (2009).
8. Suo, X., Minden, V., Nelson, B., Tibshirani, R. & Saunders, M. Sparse canonical correlation analysis. *arXiv preprint arXiv:1705.10865* (2017).
9. Rüschendorf, L. *Mathematical Risk Analysis: Dependence, Risk Bounds, Optimal Allocations and Portfolios* ISBN: 9783642335907 (Springer Berlin Heidelberg, 2013).
10. Schmitz, V. *Copulas and stochastic processes* (Bibliothek der RWTH Aachen, 2003).
11. Roncalli, T. *Lecture Notes on Risk Management & Financial Regulation* (2016).
12. Ruppert, D. *Statistics and data analysis for financial engineering, 2ed* (Springer, 2015).
13. DeFusco, R. A., McLeavey, D. W., Pinto, J. E., Anson, M. J. & Runkle, D. E. *Quantitative investment analysis* (John Wiley & Sons, 2015).
14. Fama, E. F. & French, K. R. The cross-section of expected stock returns. *the Journal of Finance* **47,** 427–465 (1992).
15. Fama, E. F. & French, K. R. Common risk factors in the returns on stocks and bonds. *Journal of financial economics* **33,** 3–56 (1993).

16. Henry, E., Robinson, T. R., Stowe, J. D., *et al. Equity asset valuation* (John Wiley & Sons, 2010).

17. Koller, D. & Friedman, N. *Probabilistic graphical models: principles and techniques* (MIT press, 2009).

18. Murphy, K. P. *Machine learning: a probabilistic perspective* (MIT press, 2012).

19. Kutner, M., Nachtsheim, C. & Neter, J. *Applied Linear Regression Models* ISBN: 9780072955675 (McGraw-Hill Higher Education, 2003).

20. Seber, G. A. & Lee, A. J. *Linear regression analysis* (John Wiley & Sons, 2012).

21. Faraway, J. J. *Linear models with R* (CRC press, 2014).

22. Lindskog, F. *et al. Modelling dependence with copulas and applications to risk management* ().

23. Cherubini, U., Luciano, E. & Vecchiato, W. *Copula methods in finance* (John Wiley & Sons, 2004).