Multivariate → data $(X_1, X_2 \cdots X_p)$, we consider live in a multivariate space, that is $p>2$ (it's a vector)

→ this variables $(X_1 \cdots X_p)$ are linked / dependent

analysis → MA is a collection of methods based on probabilistic models to understand, analyze and interpret the data

→ data reduction
discrimination
sorting
groups
dependence among variables

different goal

The model depend on the type of variables the type of the dependence, and the type of the problems.

社会P1：this class → unsupervised learning

include a lot of

classic statistics

you have data $(X_1 \cdots X_p) \in R^{p \times p}$,

you want fulfill some goals.

no label

---

you have data $(X_1 \cdots X_p) \in R^{p \times 1}$

and have some response variable

$(Y_1 \cdots Y_n) \in R$ or category

we want do $Y = f(X)$

have label

usually USL 在 SL 之前做下

∴

supervised learning

content P1

Background on multivariate distribution

① the data: sample mean, sample covariance

② the model: population versions, multivariate families of distributions.

③ densities, cumulative distribution functions

content P2

dependence modeling

① How we can describe the dependence between 2 variables

② Copulas: separating the marginals from the dependence

content P3

graphic model

① what's the structure of multivariate obs.

② graphs as a description of conditional independence relations.

③ High-dimensional data sets and regularization

content P4

Unsupervised learning

① Dimsion reduction data compression.

② Visualization of high-dimensional data

③ principal components analysis

④ clustering method

⑤ extend PCA to non-linear transformations (Autoders)

⑥ sample from the distribution of images)

Variational Autoencoder and GAN

VAE

Model and data perspective

Model perspective:

① the model is a random vector $X = (X_1 \cdots X_p)^T \in R^p$ following a certain distribution F_X describing the entire (infinite) population

② we can compute the statistics such as the population mean or population variance for X

Data perspective

① the data/sample $X_1 \cdots X_n \in R^p$ are thought of finite number n of observations from X

questions? range $n \in [0, p]$ right?

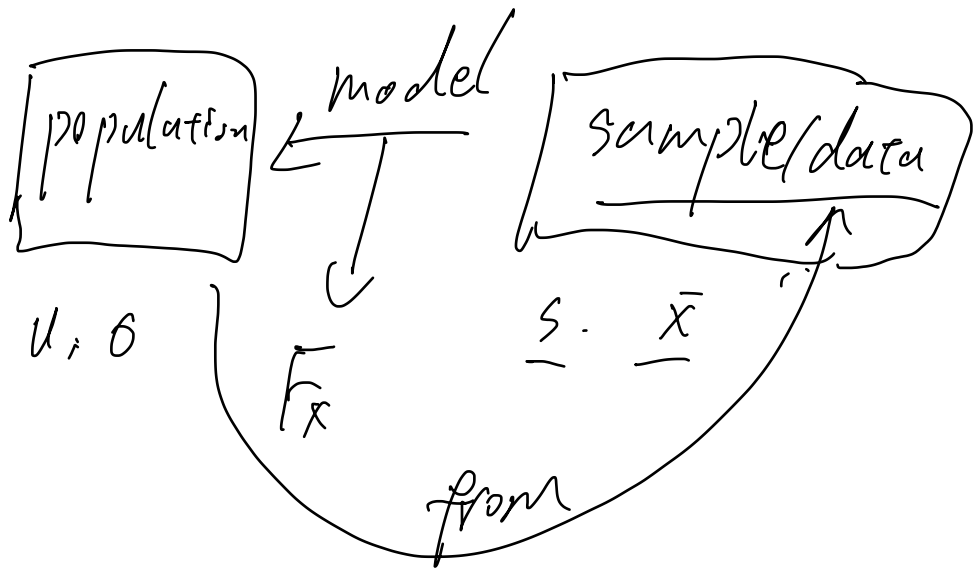
1-n 为 feature p 为维度 → nxp

unknown

this distribution is

② based on data, we can compute sample version of the mean or the variance. they are approximations of the population version.

model { vector distribution infinite population



the model: $p=2$, 2-dimensional

Let's $X = (X_1, X_2)^T$ be a bivariate normal distribution which is $N_2(\mu, \Sigma)$, with mean $(2,5)$ and covariance matrix: $\Sigma = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}$ ↳ model(的角度

the data:

we sumlate $n=50$ samples $x_1 \cdots x_n \in R^2$ from the normal distribution

一个x的取值. p.

we get $\bar{x} = (2.06, 5.07)$ $S = \begin{pmatrix} 0.98 & 0.71 \\ 0.71 & 1.11 \end{pmatrix}$

从样本对总体数据的估计

model(simulation → data

inference, but never know it exactly.

Some notation: data 个体 variable 变量

for a vector $x \in \mathbb{R}^p$ we write

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix} \qquad x^T = (x_1 \cdots x_p)$$

$u = \bar{x} = \sum_{i=1}^{n} x_i / n$ => unbiased if we get more data, $\bar{x} \to u$

mode 众数

sample 有限 ⇒

if $x, y \in \mathbb{R}^p$, $x^T y \in \mathbb{R}$ and $x y^T \in \mathbb{R}^{p \times p}$

$A \in \mathbb{R}^{m \times m}$ in boldface the diagonal matrix a's

$$\mathrm{diag}(a_1 \cdots a_m) = \begin{pmatrix} a_1 & 0 & \cdots & 0 \\ 0 & a_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & a_m \end{pmatrix} \in \mathbb{R}^{m \times m}$$

for n data $x_1 \cdots x_n \in \mathbb{R}^p$, we write

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & & & \vdots \\ \vdots & & & \\ x_{n1} & \cdots & & x_{np} \end{pmatrix} \in \mathbb{R}^{n \times p} = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}^T$$

obs n 数行数

→ feature p 数列数

population

Let $X=(X_1 \cdots X_p)^T \sim F$ be a random p-vetor with join distribution function:

$$F_X(y) = F(y) = F(y_1, y_2 \cdots y_p) = P(X_1 \le y_1, \cdots X_p \le y_p) \in R^p$$

the univariate marginal distribution of X_j, $j=1 \cdots p$ is

$$F_j(y_j) = P(\infty \cdots X_j \le y_j, \cdots \infty)$$

similarly, we get higher dimensional marginals by

$$F_{jk}(y_j, y_k) = P(X_j \le y_j, X_k \le y_k)$$

F describes the margins and the dependence structure of the population

if F is differentable (we assume), the join density is the derivative:

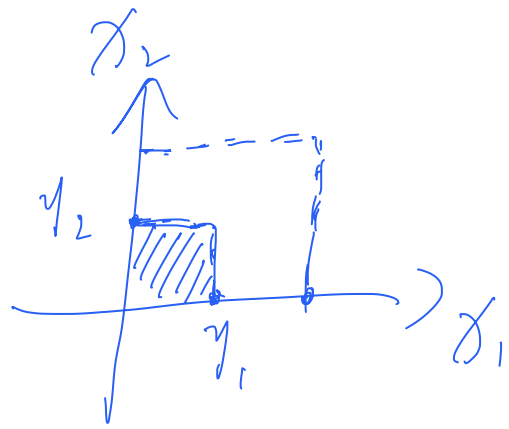
$$f(y) = f(y_1 \cdots y_p) = \frac{\partial^p F(y)}{\partial y_1 \cdots \partial y_p}$$

or equivalently:

$$F(y) = \int_{-\infty}^{y_1} \cdots \int_{-\infty}^{y_p} f(u_1 \cdots u_p)\, d_1 \cdots d u_p$$

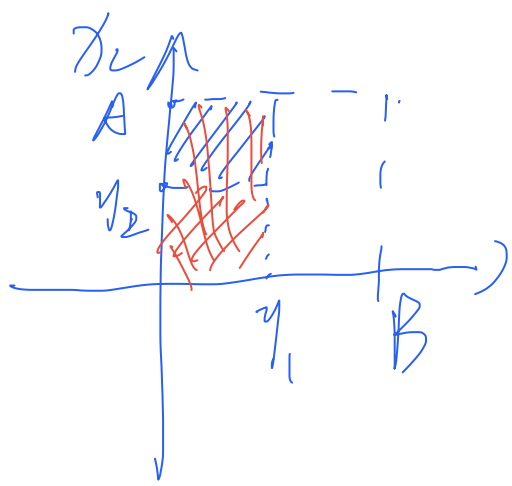
本质上：F[上页画0处]作为 distribution，可以替换为任意 distribution：

举例：uniform distribution



$$F_\gamma(y_1, y_2) = P(\gamma_1 \le y_1, \gamma_2 \le y_2)$$

"///" is the P



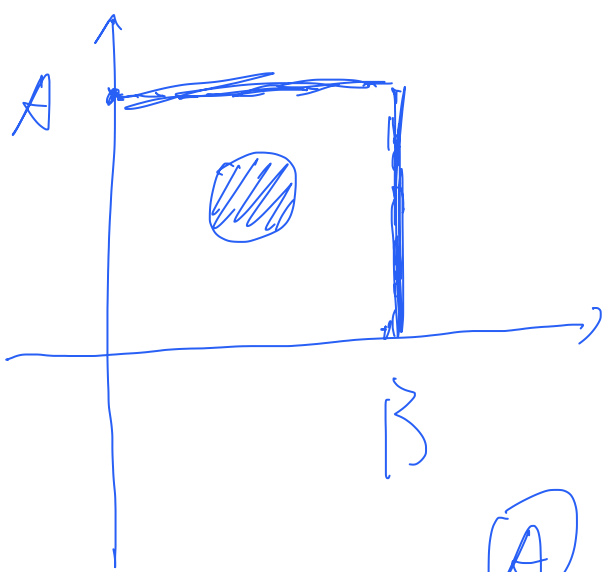
从图可推导得 $\Longrightarrow$ $P(\gamma_1 \le y_1, y_2 \le \gamma_2)$

A is the range $[c, A)$

$\Downarrow$ 得 F

$$F(y_1) - F(y_1, y_2)$$

"///" — "▨"



$\Rightarrow$ if we want to know the P(区) which is not limit to rectangles

Ⓐ one way we try to is 用 rectangles 渐近.

the joint density existence $\Longleftrightarrow$ the marginal densities

$$f_j(y_j) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(u_1 \cdots u_{j-1}\, u_{j+1} \cdots u_p)\, du_1 \cdots du_{j-1} du_{j+1} \cdots du_p$$

the density function integrates to 1

$$F(\infty, \infty) = 1$$

① PS

Sample quantity: we don't know distribution we need to estimate it

Let $X_1 \cdots X_n \in R^p$ be the n samples of X

You should see $X_1 \cdots X_n$ as random realizations of X

the emprical distribution function

③ $\hat{F}(y) = \frac{1}{n} \sum_{i=1}^{n} 1\{X_{i1} \le y_1, \cdots X_{ip} \le y_p\}$ ② PS

is a non-parametric estimtor of F

there is no "empirical" sample-equivalent for the density function without further assumptions. why?

example:

Let $X = (X_1, X_2)^T$ be a bivariate random variable with density function: $f(y_1, y_2) = \begin{cases} r(y_1^{\alpha} + y_2^{\beta}), & 0 < y_1, y_2 < 1 \\ 0, & \text{otherwise} \end{cases}$ $\alpha, \beta > 0$

$\int_0^1 \int_0^1 r(y_1^{\alpha} + y_2^{\beta})\, dy_1 dy_2 = 1 \Longrightarrow$ get the r

$0 < y_1, y_2 < 1$

$= \int_0^1 \int_0^1 r y_1^{\alpha} dy_1 dy_2 + \int_0^1 \int_0^1 r y_2^{\alpha} dy_1 dy_2$

we get: $r = \left[(\alpha+1)^{-1} + (\beta+1)^{-1}\right]^{-1}$

$= \int_0^1 r y_1^{\alpha} dy_1 \int_0^1 dy_2 + \int_0^1 r y_2^{\beta} dy_2 \int_0^1 dy_1$

$= \left[\frac{r}{\alpha+1} y_1^{\alpha+1}\right]_0^1 y_{2,0}^{\,1} + \left[\frac{r}{\beta+1} y_2^{\beta+1}\right]_0^1 y_{1,0}^{\,1}$

$= \frac{r}{\alpha+1} + \frac{r}{\beta+1} = 1$

只要满足即可估计 F

$\alpha = 1, \beta = 2, r = 6/5$

$F = \int_0^{y_1} \int_0^{y_2} \frac{6}{5}(u_1 + u_2^2)\, du_1 du_2$

$F = \frac{6}{5} y_1 y_2 \left[\frac{1}{2} y_1 + \frac{1}{3} y_2^2\right]$

$F(X_1 < \frac{1}{2}, X_2 > \frac{1}{2}) = \frac{1}{4}$ (求 Probabilities 的统计量

for $\hat{F}$ 类似的:

$\hat{P}(X_1 < \frac{1}{2}, X_2 > \frac{1}{2}) = \hat{F}(\frac{1}{2}, 1) - \hat{F}(\frac{1}{2}, \frac{1}{2}) = \frac{1}{n} \sum_{i=1}^{n} 1\{X_{i1} \le \frac{1}{2}, X_{i2} > \frac{1}{2}\}$

population quantity:

Let $X = (X_1 \cdots X_p)^T$ be a random vector

we can define the expectation of $g(x)$, where g $R^p \to R^m$ is some function

$$E(g(x)) = \int_{R^p} g(u)\, dF(u) = \int_{R^p} g(u) f(u)\, du \in R^m$$

sample quantity

for n samples $x_1 \cdots x_n \in R^p$ of X, we define the sample version of $E(g(x))$ as

$$\hat{E}(g(x)) = \int_{R^p} g(u)\, d\hat{F}(u) = \frac{1}{n} \sum_{i=1}^{n} g(x_{i1} \cdots x_{ip})$$

(期望就是 $\int_{-\infty}^{\infty} f(x)\, dx$)

For a random vector $X \in R^p$ and $Y \in R^q$ matrix $A \in R^{m \times p}$ $B \in R^{m \times q}$ and constant vector $b \in R^m$

$$Z = \underset{m \times p \cdot p \times 1}{AX} + \underset{m \times q \cdot q \times 1}{BY} + \underset{m \times 1}{b} \in R^m$$

$E(Z) = A \cdot E(X) + B \cdot E(Y) + b$

the trace is an operator on square matrices that returns the sum of the diagnoal elements. 对角元素和

Let $A = \{a_{kl}\}_{k,l=1}^{m}$ be $m \times m$ matrix, then

$$Tr(A) = \sum_{i=1}^{m} a_{ii}$$

the trace is linear and circular (with compatible matrix)

可乘的前提得 AB BA

$$Tr(A+B) = Tr(A) + Tr(B), \quad Tr(ABC) = Tr(CAB)$$

For a random vector $X \in R^p$ and a determinstic matrix 确定性矩阵

$A \in R^{p \times p}$

$$\underset{1 \times p\ p \times p\ p \times 1}{E(X^T A X)} = E[Tr(X^T A X)] = E[Tr(AXX^T)] = Tr[A \cdot E(XX^T)]$$

how can I prove it? $X^T A X$ is a scalar, $Tr(\text{scalar}) = \text{scalar}$

by linearity $E[Tr(C)] = Tr[E(C)]$

population quantity

the variance-covariance matrix of X is the matrix of variances and covariances between the p components

$$\Sigma = cov(X) = \{\sigma_{kl}\}_{k,l=1}^{p} = E[(X - E(X))(X - E(X))^T]$$

$$\sigma_{kl} = E[(X_k - E(X_k)) \cdot (X_l - E(X_l))], \quad \sigma_k^2 = \sigma_{kk}$$

the correlation matrix is the matrix of correlations obtained through normalization.

$$\rho_{kl} = \frac{\sigma_{kl}}{\sigma_k \cdot \sigma_l}, \quad \rho_{kk} = 1$$

$$\mathbf{T} = cor(X) = \{\rho_{kl}\}_{k,l=1}^{p} = \Delta^{-1/2} \Sigma \Delta^{-1/2}, \text{ with } diag(\sigma_1^2 \cdots \sigma_p^2)$$

recall by cauchy-schwarz inequality we can get $\rho_{kl} \in [-1, 1]$

$|\langle x, y \rangle|^2 \le \langle x, x \rangle \cdot \langle y, y \rangle$

$|\langle x, y \rangle| \le \|x\| \cdot \|y\|$

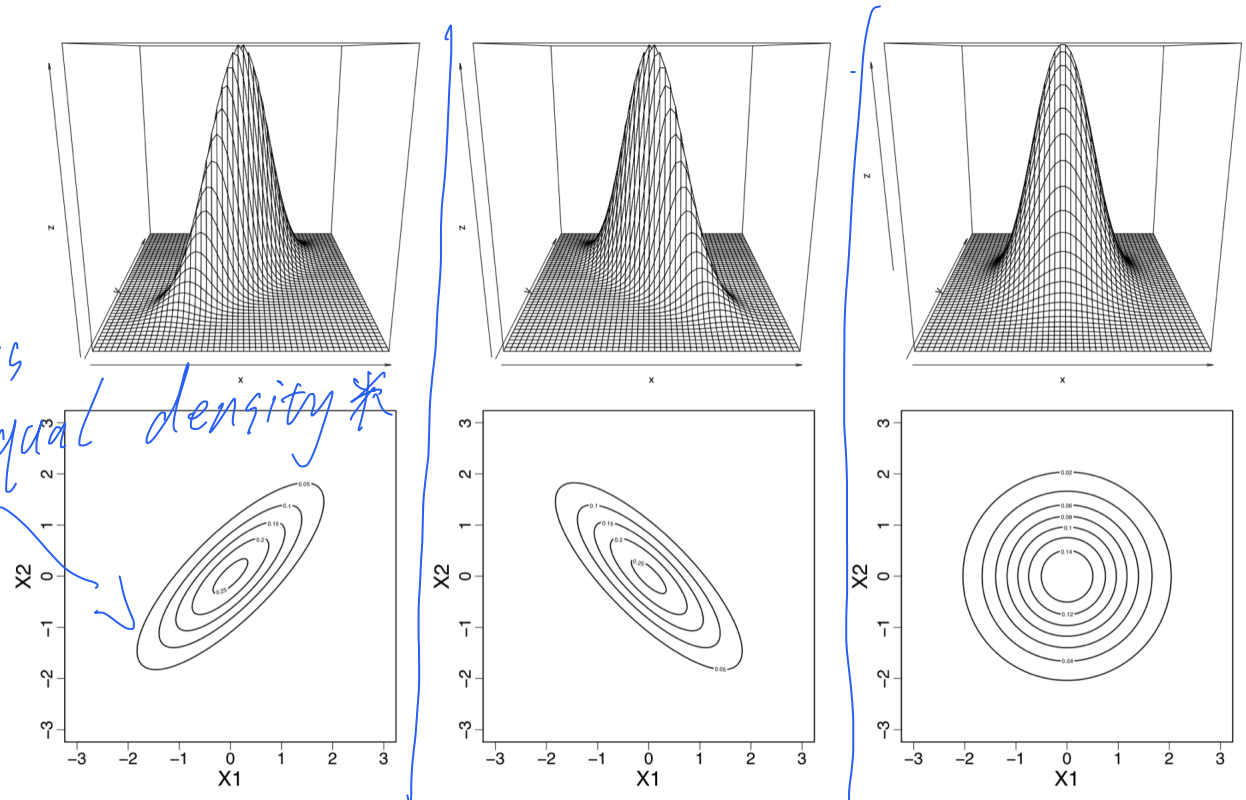
sample quantity.

$$S = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T \in R^{p \times p} \qquad r_{kl} = \frac{s_{kl}}{s_k s_l}$$

why not n, because it's an unbiased estimator of Σ bessel's correction

$R = D^{-1/2} S D^{-1/2}$ with $D = diag(s_1^2 \cdots s_p^2)$

PS:① bivariate normal distribution



points is equal density 来

这是 different density function

上面的图可以理解为: $z = f_{x,y}(x, y)$

z 为 probability

下图为其在 x-y 平面上的投影

其 $U = \begin{pmatrix} a \\ b \end{pmatrix}$ 改变的是下图的位置

$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$ σ_{11} σ_{12} 改变在 x or y 的跨度

σ_{21} σ_{12} 是相关小量

来怎么理解: $z = 0 = f_{(x,y)}(x, y) = \dfrac{1}{(2\pi)^{n/2} \cdot |\Sigma|^{1/2}} e^{[-\frac{1}{2}(x-u)\Sigma^{-1}(x-u)]}$

其 $n = len(x)$

$= \text{当 } z = 0$

$x = (x_1, x_2, x_3 \cdots)$

covariance = V = density value

(2) PS: $F(\gamma) = P(\chi_1 \le \gamma_1, \cdots \chi_p \le \gamma_p)$

① $= E[\text{Indicator function}]$

$= E[I(\chi_1 \le \gamma_1 \cdots \chi_p \le \gamma_p)]$

cove: prob is expectation of indicator function

$I\{A\} = \begin{cases} 1 & \text{if it's true} \\ 0 & \text{if it's false} \end{cases}$

∴ if we want estimate the $F(\gamma)$.

we need to estimate the expectation of the indicator function.

So how to estimate the expectation of the indicator function? => use monte-carlo

$$\hat{F}(\gamma) = \frac{1}{n} \cdot \sum_{i=1}^{n} 1 \cdot \{\chi_{i1} \le \gamma_1, \cdots \chi_{ip} \le \gamma_p\} \quad (3)$$

expectation: [prob x value] 对应 Indicator function.

PS: it's non-parametric estimator

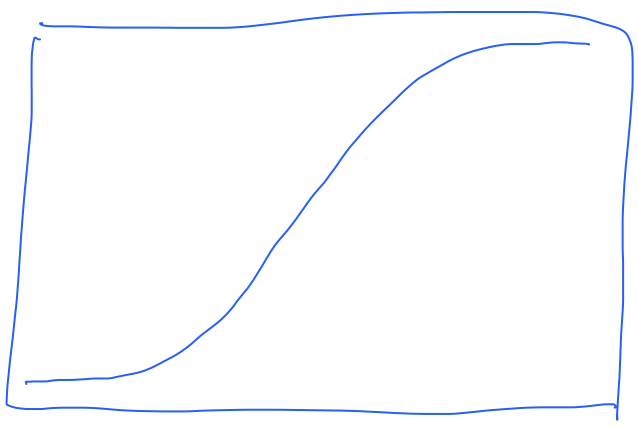
good estimator:

it means it coverage to the true CDF

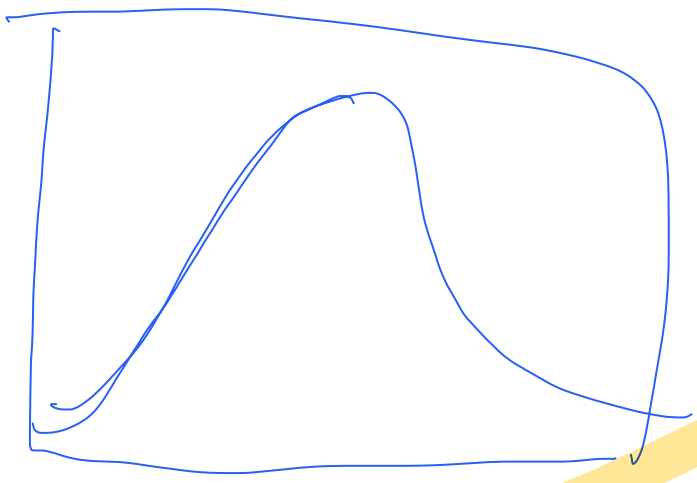
$\hat{F}(x) \to F(x)$

move data

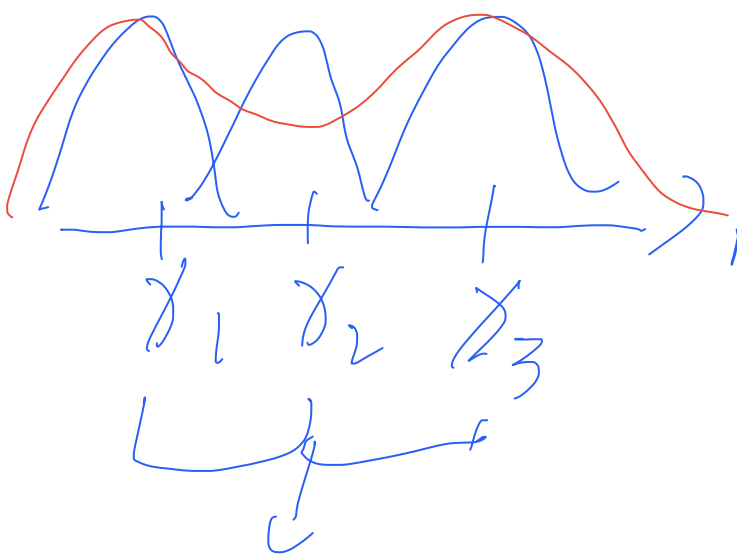
CDF of normal distribution



density function of normal distribution



这里只是估计了 CDF
没有估计 density
如何应对? way1: 连续下、求导.
way2: kernel function estimate



need some assumption

how to choose

red : kernel density

non-parameter

but it not empirical

$\gamma_1 \cdots \sim$ normal distribution

a covariance matrix $\Sigma = cov(X)$ is symmetric and positive-semidefinite, that is,

$$X^T \Sigma X \geq 0, \quad 0, \text{ for all } X \in \mathbb{R}^p$$

if the above holds strictly with > 0, then Σ is positive definite

if Σ is positive definite and Σ^{-1} with { invertible / positive-d } we can have a decomposition $\Sigma^{1/2}$ satisfying

$$\Sigma^{1/2}(\Sigma^{1/2})^T = \Sigma$$

for a matrix $A \in \mathbb{R}^{m \times p}$, a constant vector a vector $b \in \mathbb{R}^m$, and a random vector $X \in \mathbb{R}^p$

$$cov(b + aX) = E[A \cdot (X - E(X))(X - E(X))^T \cdot A^T] = A\Sigma A^T \in \mathbb{R}^{m \times m}$$

the covariance notion also generalizes to two vectors

Let $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ be random vectors, then their covariance matrix is $p \times q$.

$$cov(X;Y) = E[(X - E(X))(Y - E(Y))^T] = [cov(X_i, Y_j)]_{i,j=1}^{p,q}$$

note that $\Sigma_{xy} = \Sigma_{yx}^T$

population quantity

Let $X \in \mathbb{R}^p$ be a random vector with $\Sigma = cov(X)$ and $\Delta = diag(\sigma_1^2, \dots \sigma_p^2)$ and correlation matrix =

$$cor(X) = P = \Delta^{-1/2} \Sigma \Delta^{-1/2}$$

then the random vector: satisfies:

$$Y = \Delta^{-1/2}(X - E(X)) \Rightarrow E(Y) = 0, \quad cov(Y) = P \quad (iii)$$

for for standardization, consider:

$$Y = \Sigma^{-1/2}(X - E(X)) \quad (iv)$$

then $E(Y) = 0$ and $cov(Y) = I_p = diag(1, \dots 1)$

for (iii) $cov(Y) = cov(\Delta^{-1/2}X - \Delta^{-1/2}E(X)) = cov(\Delta^{-1/2}X)$

$$= E[\Delta^{-1/2}(X - E(X))(X - E(X))^T(\Delta^{-1/2})^T]$$

$$= \Delta^{-1/2} \cdot \Sigma \cdot (\Delta^{-1/2})^T = \Delta^{-1/2} \cdot \Sigma \cdot \Delta^{-1/2} = P$$

$\Delta = diag(\sigma_1^2 \dots \sigma_p^2) = \begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_p^2 \end{bmatrix} \Rightarrow (\Delta^{-1/2}) = (\Delta^{-1/2})^T$

for (iv): $cov(Y) = cov(\Sigma^{-1/2}X - \Sigma^{-1/2}E(X)) = cov(\Sigma^{-1/2} \cdot X)$

$$= E[\Sigma^{-1/2} \cdot (X - E(X))(X - E(X))^T(\Sigma^{-1/2})^T]$$

$$= \Sigma^{-1/2} \cdot \Sigma \cdot (\Sigma^{-1/2})^T = \Sigma^{-1} \cdot \Sigma = I$$

sample quantity:

the analogous sample standardization for the data $X_1 \cdots X_n$ are

$$y_i = D^{-1/2}(x_i - \bar{x}) \qquad y_i = S^{-1/2}(x_i - \bar{x})$$

S is the sample covariance

D is the $diag(s_1^2 \cdots s_p^2)$

Independence:

By definite: the p components $X_1 \cdots X_p$ of $X \sim F$ are mutually independent

$$F(y) = \prod_{j=1}^{p} F_j(y_i), \quad \forall y \in \mathbb{R}^p$$

$$\Updownarrow$$

$$f(y) = \prod_{j=1}^{p} f_j(y_j), \quad \forall y \in \mathbb{R}^p$$

if two random variables are independent, then the $E(xy)$

$$E(xy) = E(x) \cdot E(y)$$

consequently, if $X = (X_1 \cdots X_p)^T$ mutually independent,

$$\sigma_{kk} = cov(X_k; X_k) = E[(X_k - E(X_k))(X_k - E(X_k))] = 0$$

$$cov(X) = diag(\sigma_1^2 \cdots \sigma_p^2)$$

Let $X=(X_1,X_2)^T$ be a bivariate normal random variables with mean = 0 and covariance $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$

the density function:

$$f(x_1,x_2) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{\{-\frac{1}{2(1-\rho^2)}(x_1^2 - 2\rho x_1 x_2 + x_2^2)\}}$$

原形式: (补充)

$$f(x,y) = \frac{1}{2\pi \sigma_x \sigma_y} \cdot e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-u_x}{\sigma_x}\right)^2 - 2\rho \cdot \frac{x-u_x}{\sigma_x} \cdot \frac{y-u_y}{\sigma_y} + \left(\frac{y-u_y}{\sigma_y}\right)^2\right]}$$

$$u = \begin{bmatrix} u_x \\ u_y \end{bmatrix} \quad \Sigma = \begin{pmatrix} \sigma_x^2 & \sigma_x\sigma_y \cdot \rho \\ \sigma_x\sigma_y\rho & \sigma_y^2 \end{pmatrix}$$

$$f_1(x_1) = \int_{-\infty}^{+\infty} f(x_1,x_2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x_1^2}{2}} \quad \text{same to } x_2$$

if $\rho = 0 \Rightarrow X_1, X_2$ independent ?

证)

$$f_1(x_1) f_2(x_2) = \frac{1}{2\pi} e^{-\frac{1}{2}(x_1^2 + x_2^2)} = T.$$

if $T = f(x_1, x_2) \Rightarrow \rho = 0$.

summary:

$$\begin{cases} X_1, X_2 \sim N \\ \rho = 0 \end{cases} \Rightarrow X_1, X_2 \text{ independent}$$

$X_1\ X_2$ indep $\Rightarrow \rho = 0$

conditional distribution:

Let $X \in R^p$, $Y \in R^q$ be random vectors, and Let $(X^T, Y^T)^T$ be the joint vector with density $f(x,y)$. Then the conditional distribution of X given the $Y=y$

$$f_{X|Y=y}(x) = \frac{f_{(x,y)}(x,y)}{f_y(y)}$$

if X, Y independent, $f_x(x) = f_y(y)$

for the bivariate normal distribution our numerical example with $X = X_1$, $Y = X_2$.

$$f_{X|Y=y}(x) = \frac{1}{2\pi\sqrt{1-\rho^2}} \cdot e^{\{-\frac{(x^2-\rho y)}{2(1-\rho^2)}\}}$$

therefore $(X|Y=y) \sim N(\rho y, 1-\rho^2)$

we also can define the $E(X|Y=y)$:

$$E(X|Y=y) = \int_x x \cdot f_{X|Y}(X|Y=y)\, dx$$

more generally: $E(g(x)|Y=y) = \int_x g(x) \cdot f_{X|Y}(X|Y=y)\, dx$

the conditional expectation can be seen as a random variable: $E(X|Y)$ with the important property

$$E(X) = E[E(X|Y)]$$

similarly:

$$cov(X|Y=y) = E[(X - E(X|Y))(X - E(X|Y))^T]$$

if X and Y ~~indepent~~ indenpent.

$E(X|Y) = E(X)$

$cov(X|Y) = cov(X)$

$F_{X|Y}(X|Y) = F_X(x)$,

$f_{X|Y}(X|Y) = f_X(x)$

$\Rightarrow$ both is sufficient condition

Module 02:
multivariate normal and other distributions

---

the univariate uniform distribution: 均匀分布(别看到1个distribution就是正态分布)

Let $S=(a,b)$ be interval, then the uniform distribution $U \sim U(a,b)$ is a uniformly distributed random variable if $f(u) = \frac{1}{b-a}$, for $u \in (a,b)$

prob density function

a important example: $S=(0,1) \Rightarrow$ standard uniform distribution

recall that any random variable X with continous cumulative distribution function F satisfies:

$F(x) \sim U(0,1)$, since:

$$P(F(x) \le u) = P(x \le F^{-1}(u)) = F(F^{-1}(u)) = u, \forall u \in (0,1)$$

we can thus generate the samples of X using a standard uniform random variable $U \sim U(0,1)$, by setting $X = F^{-1}(u)$,

$$P(X \le x) = P(F^{-1}(u) \le x) = F(x)$$

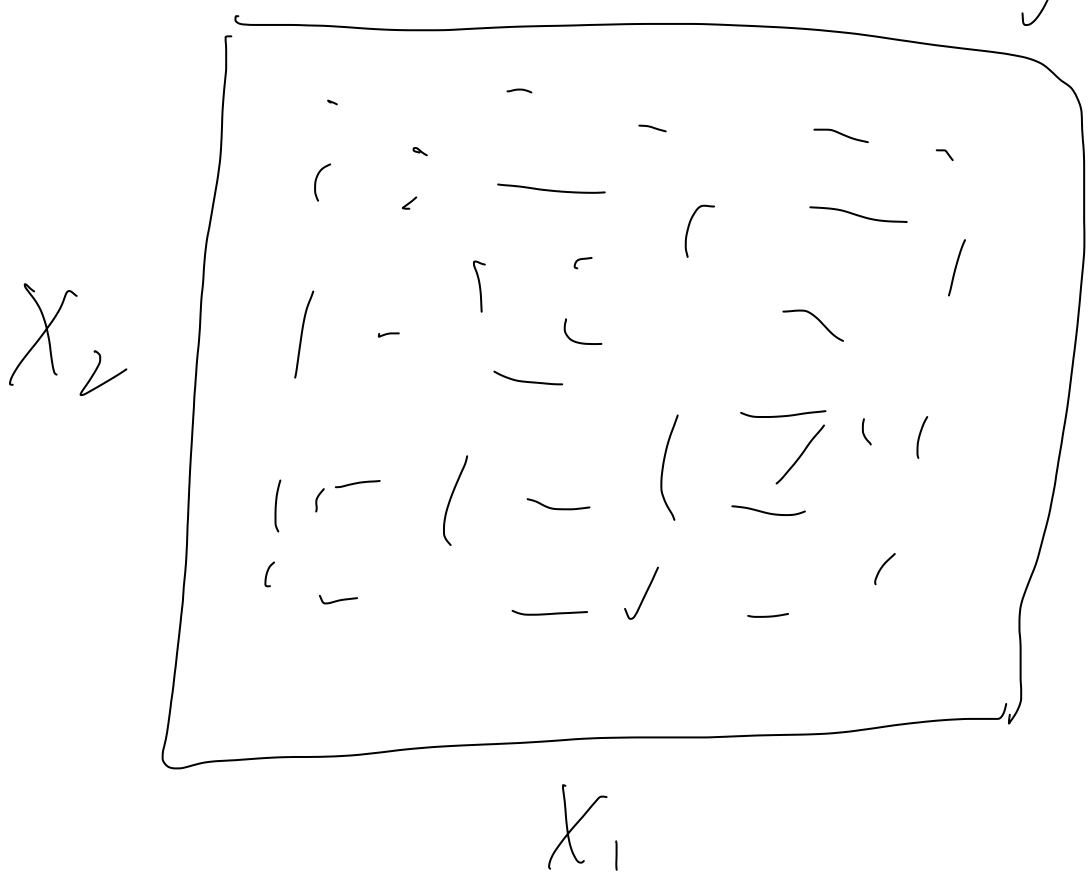
In general, we can construct a multivariate distribution for $X = (X_1 \cdots X_p)^T$ by letting $X_1 \cdots X_p$ be independent univariate random variables. The density is:

$$f(x_1, \cdots x_p) = f_{x_1}(x_1) \cdot f_{x_2}(x_2) \cdots f_{x_p}(x_p)$$

Let X_j, $j = 1 \cdots p$ be random variables uniformly distributed on interval $S_j = (a_j, b_j)$, $j = 1 \cdots p$. Then $X = (X_1 \cdots X_p)^T$ has a multivariate uniform distribution with density

$$f(x_1 \cdots x_p) = \frac{1}{\prod_{j=i}^{p} (b_j - a_j)}, \quad x_j \in S_j$$

by definition the marginals X_j are mutually independent and all uniformly distributed



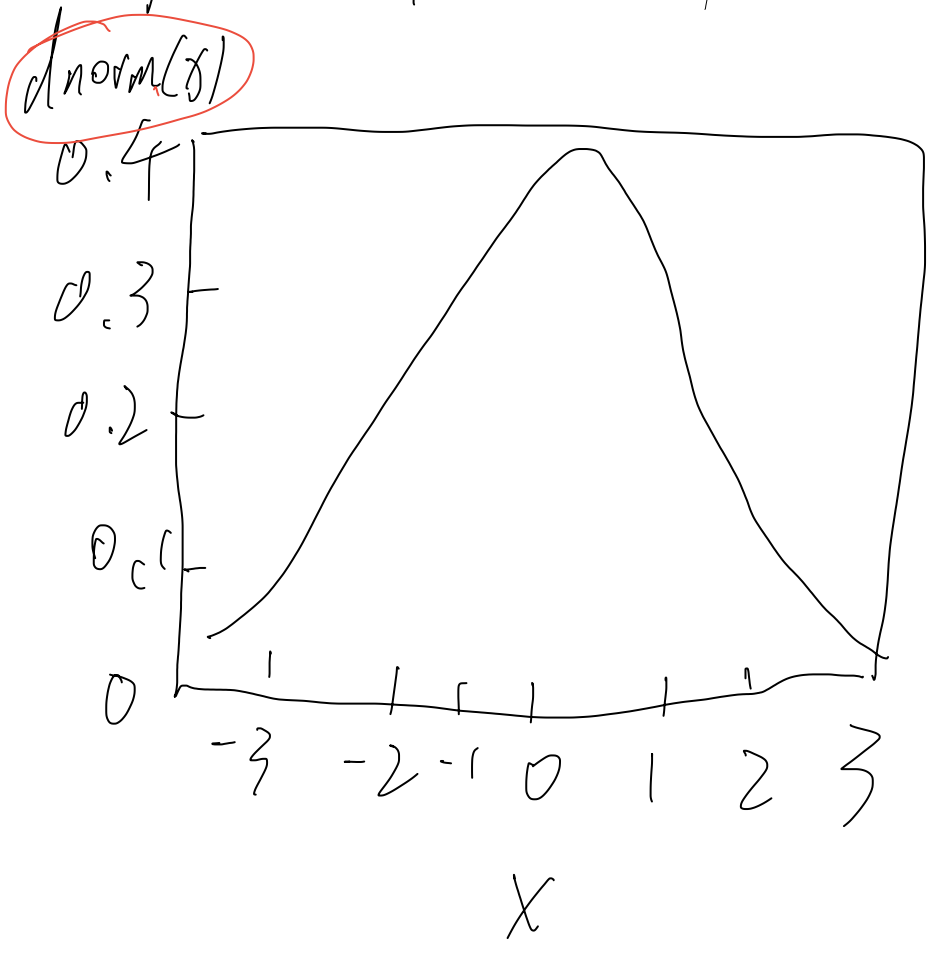
the multivariate normal distribution

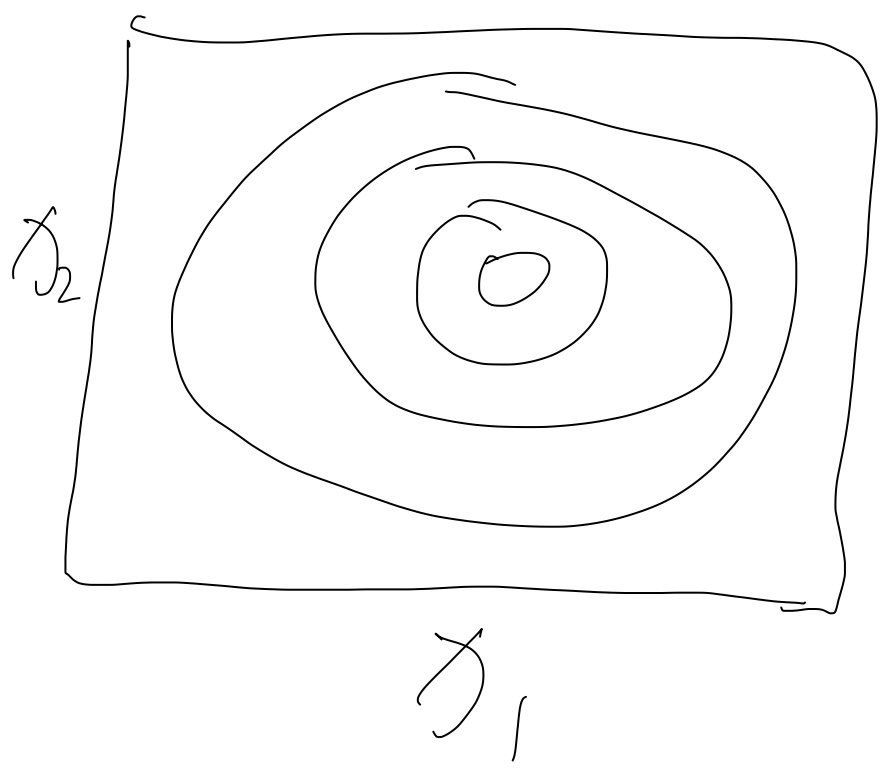
recall the $X \sim N(u, \sigma^2)$ has a univariate normal distribution if its density is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-u)^2}{2\sigma^2}\right\} \qquad u = E(x)$$
$$\sigma^2 = \mathrm{Var}(x)$$

the standard multivariate normal distribution follows by the same "independent construction" as before, namely let $X_1 \cdots X_p$ be standard normal $N(0,1)$, then $X = (X_1 \cdots X_p)^T$ has density

$$f(x) = f_{x_1}(x_1) \cdots f_{x_p}(x_p) = \frac{1}{\sqrt{(2\pi)^p}} \exp\left\{-\frac{1}{2} x^T x\right\}, \quad x \in \mathbb{R}^p$$





Density of transformed random vectors.

Let X be a p random vector and $Y = H(X)$ for $H: R^p \to R^p$ a bijective (i.e: invertible) transformation

双射的 → (可逆的)

if f_X is the density of X, then Y has the density $f_Y(y) = f_X(H^{-1}(y)) \left| \det(J_{H^{-1}}(y)) \right|$ 要搞懂

where for a mapping $T = (T_1 \cdots T_p): R^p \to R^p$

$J_T(y)$ = is the Jacobian matrix at point $y \in R^p$

$$J_T(y) = \begin{pmatrix} \frac{\partial T_1}{\partial y_1} & \cdots & \frac{\partial T_1}{\partial y_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial T_p}{\partial y_1} & \cdots & \frac{\partial T_p}{\partial y_p} \end{pmatrix} \in R^{p \times p}$$

the matrix J_T is the best linear approximation of T at y

how to prove

example: matrix $A \in \mathbb{R}^{p\times p}$, $b \in \mathbb{R}^p$

the mapping $H(x) = Ax + b$ is affine (?)

$\Rightarrow$ $J_H(x) = A$

scaling and rotation:

further construction priciples for multivariate distributions are

scaling: Let $\Lambda = diag(\lambda_1 \dots \lambda_p)$ be a scaling matrix and define a new random vector $\Lambda^{1/2} Z$

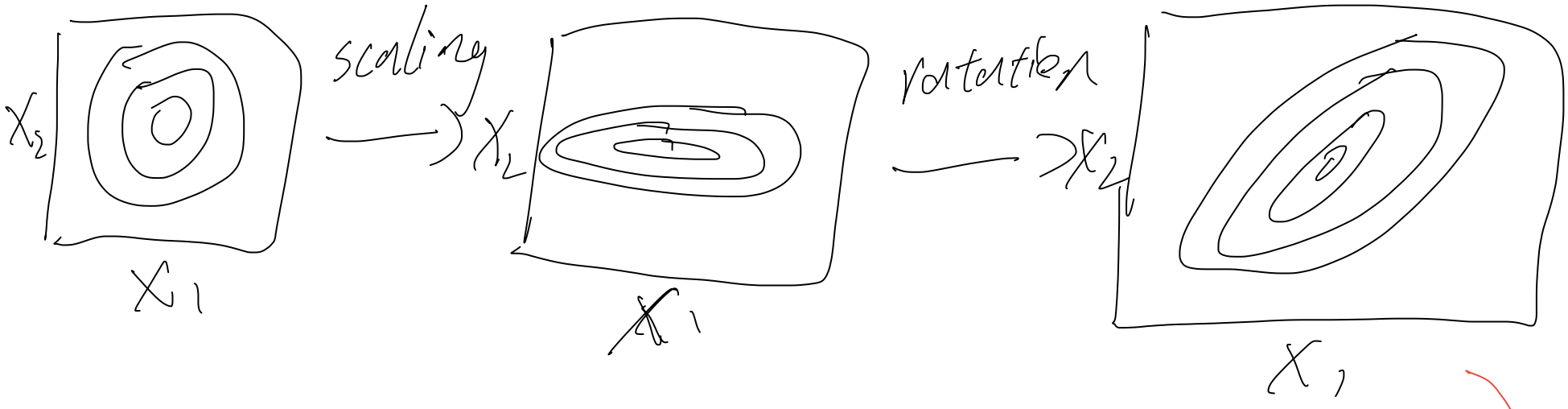
$z \in \mathbb{R}^p$ $f_{\Lambda^{1/2}Z}(z) = \frac{1}{\sqrt{(2\pi)^p det(\Lambda)}} exp\{-\frac{1}{2} z^T \Lambda^{-1} z\}$

PS $y = H(x) = \Lambda^{1/2} x$

$f_y(y) = f_x(H^{-1}(y)) \left| det(J_{H^{-1}}(y)) \right|$ — why?

Rotation: let V be a rotation matrix and define a new random vector $V\Lambda^{1/2}X$.

$f_{V\Lambda^{1/2}Z}(x) = \frac{1}{\sqrt{(2\pi)^p det(V\Lambda V^T)}} \cdot exp\{-\frac{1}{2} x^T (V\Lambda V^T)^{-1} x\}$



location shifting: for some $u \in \mathbb{R}^p$, let $H(x) = x + u$

$$f_{UV^{1/2}X}(x) = \frac{1}{\sqrt{(2\pi)^p \det(UV^T)}} \exp\left\{-\frac{1}{2}(x-u)^T (UV^T)^{-1}(x-u)\right\}$$

let Σ be a positive-definite $p \times p$-matrix and $u \in \mathbb{R}^p$. Then $X = (X_1 \cdots X_p)^T$ has a multivariate normal distribution $N_p(u, \Sigma)$ if its density satisfies

$$f(x) = \frac{1}{\sqrt{(2\pi)^p \det(\Sigma)}} \exp\left\{-\frac{1}{2}(x-u)^T \Sigma^{-1}(x-u)\right\}, \quad x \in \mathbb{R}^p$$

every positive-definite matrix Σ has eigendecomposition $\Sigma = V \Lambda V^T$ where $\Lambda = diag(\lambda_1 \cdots \lambda_p)$, λ is the eigenvalues of Σ, and V is a rotation matrix. Therefore, we get this represetation

$$X \sim u + V \Lambda^{1/2} N_p(0, I_p)$$

by the rules for expectation and covariance this gives: $E[X] = u$, $cov(X) = \Sigma$

Let $X \sim N_p(u, \Sigma)$

① the marginal X_j has a univariate normal distribution $N(u_j, \sigma_j^2)$

② Let A be a $m \times p$ matrix, $b \in R^m$, then $b + AX$ is multivariate normal $N_m(b + Au, A \Sigma A^T)$

in particular $\Sigma^{-1/2}(X - u) \sim N_p(0, I_p)$

where $\Sigma^{-1/2}$ is the choleski decomposition of Σ^{-1} satisfying $\Sigma^{-1/2} \cdot \Sigma^{-1/2} = \Sigma^{-1}$

3) X_j and X_k are independent if their covariance $\sigma_{jk} = 0$. That is the components of X are mutually independent if and only if

$$\Sigma = \text{diag}(\sigma_1^2 \cdots \sigma_p^2)$$

this equivalence between "independent" and "uncorrelated" if specific to the normal distribution.

the normal distribution is completely characterized by its mean vector μ. and covariance matrix Σ