The Multivariate Normal and the Choleski and Eigen Decompositions

Rob McCulloch

1. Introduction

A square matrix $A = [a_{ij}]$ is *symmetric* if $a_{ij} = a_{ji}$.

A square, symmetric matrix is *positive definite* (pd) if

$$x'Ax > 0 \ \ \forall x.$$

Our basic example is a covariance matrix.

If $X = (X_1, X_2, \ldots, X_p)'$ is a random (column) vector with $E(X) = \mu = (\mu_1, \mu_2, \ldots, \mu_p)'$, then the covariance of X is

$$\Sigma = E((X - \mu)(X - \mu)') = [E((X_i - \mu_i)(X_j - \mu_j)]$$

is symmetric.

Since

$$Var(a'X) = a'\Sigma a$$

Σ is positive definited unless some linear combination of the X_i has 0 variance.

Let's review two basic matrix decompositions for symmetric pd matrices and use them to review basic properties of the multivariate normal distribution.

We'll look at:

(i):

The eigen decomposition.

(i):

The Choleski decomposition.

Later we will also look at the Singular Value Decomposition.

2. Change of Variable

To develop the normal distribution based on matrix decompositions, we will need the change of variable formulas, univariate and multivariate.

Let's review these.

Let Θ be a random variable with density $p(\theta)$.

In Bayesian statistics, θ is often used for the parameter of the model so that $p(\theta)$ is the prior distribution.

The general Bayesian model consists of:

$$f(y \mid \theta), \;\; p(\theta).$$

Rather than think in terms of the parameter θ we may want to consider a 1-1 reparmetrization

$$\gamma = g(\theta),$$

where g is 1 to 1.

What is $p(\gamma)$??

Univariate change of variable

$\theta \in \mathbb{R}$

density $p_\theta(\theta)$

$\gamma = g(\theta)$, $\theta = h(\gamma)$ $(h = g^{-1})$

$p_\gamma(\gamma) = p_\theta(h(\gamma)) \, |h'(\gamma)|$

or $p(\gamma) = p(\theta(\gamma)) \, |\theta'(\gamma)|$

or $p(\gamma) = p(\theta(\gamma)) \left| \frac{d\theta}{d\gamma} \right|$

Univariate change of variable

A simple way to see it.

$$Y = g(\Theta) \quad , \quad \Theta = h(Y)$$

Let's assume h is monotonic increasing

$$h'(Y) > 0 \quad \forall\, Y$$

$$f_Y(Y_0) = F'(Y_0)$$

$$\begin{aligned}
F(Y_0) &= P\{Y \le Y_0\} \\
&= P\{h(Y) \le h(Y_0)\} \\
&= P\{\Theta \le h(Y_0)\} \\
&= F_\Theta(h(Y_0))
\end{aligned}$$

$$\begin{aligned}
F'(Y_0) &= F'_\Theta(h(Y_0))\, h'(Y_0) \\
&= f_\Theta(h(Y_0))\, h'(Y_0)
\end{aligned}$$

Example

Suppose $Y \sim \text{Bernoulli}(\theta)$ and $p(\theta) = 1$.

That is, we have the uniform prior on $\theta \in (0, 1)$.

Suppose we want to work with the odds-ratio, instead of the probability.

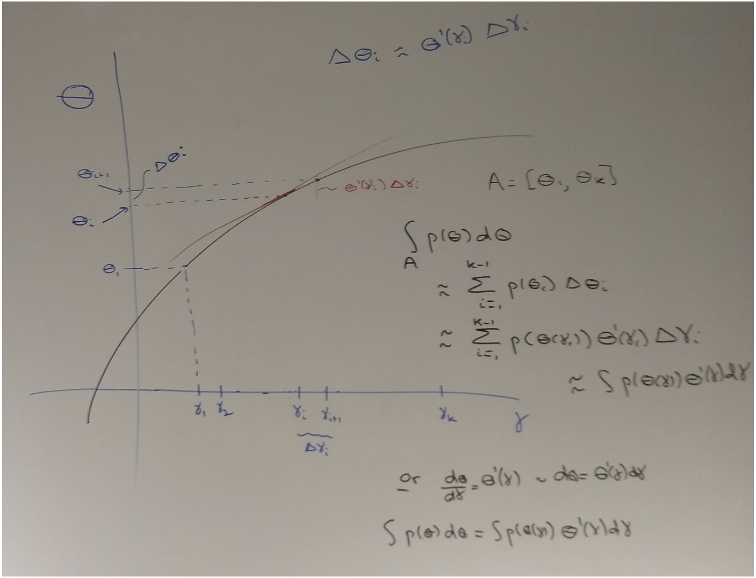
$$\gamma = \frac{1-\theta}{\theta} \in \mathbb{R}^+$$

$$\theta = \frac{1}{1+\gamma}$$

$$\frac{d\theta}{d\gamma} = -(1+\gamma)^{-2}$$

$$p(\gamma) = \frac{1}{(1+\gamma)^2}$$

Calculus intuition for univariate change of variable



Example, linear

Suppose $X \sim p(x|\alpha)$, where α is a "shape" parameter.

Let $Y = a + bX$.

$X = \frac{Y-a}{b}$. $\frac{dx}{dy} = \frac{1}{b}$.

$$f(y \mid a, b, \alpha) = p(\frac{y-a}{b})\,\frac{1}{b}$$

example:

```
Details:

     If ‘scale’ is omitted, it assumes the default value of ‘1’.

     The Gamma distribution with parameters ‘shape’ = a and ‘scale’ = s
     has density

                  f(x)= 1/(s^a Gamma(a)) x^(a-1) e^-(x/s)

     for x >= 0, a > 0 and s > 0.  (Here Gamma(a) is the function
     implemented by R’s ‘gamma()’ and defined in its help.  Note that a
     = 0 corresponds to the trivial distribution with all mass at point
     0.)
```

$$f(x|a) = \frac{1}{\Gamma(a)} x^{a-1} \, e^{-x}.$$

$$Y = sX, \;\; X = Y/s, \;\; dx/dy = 1/s.$$

$$f(y|s,a) = \frac{1}{\Gamma(a)} \, (y/s)^{a-1} \, e^{-y/s} \, (1/s)$$

Multivariate Change of Variable

$$\theta \in \mathbb{R}^k$$

$$\gamma = g(\theta) \qquad \theta = h(\gamma)$$

$$h'(\gamma) = \left[\frac{\partial \theta_i}{\partial \gamma_j} \right] \quad k \times k$$

$$p(\gamma) = p(h(\gamma)) \, |h'(\gamma)| \, +$$

density of γ

density of θ

absolute value of the determinant of $h'(\gamma)$

Example, Linear, $R^k \Rightarrow R^k$

$Z = (Z_1, Z_2, \ldots, Z_k)'$.

$\mu \in R^k$, A, $k \times k$, invertible.

$$y = \mu + Az, \;\; z = A^{-1}(y - \mu), \;\; \frac{dz}{dy} = A^{-1}.$$

$$f(y) = f_z(A^{-1}(y - \mu))\,|A^{-1}|.$$

3. Orthogonal Matrices and Rotation

A matrix $p \times p$ matrix P is orthogonal if

$$P'P = PP' = I$$

where I is the identity matrix.

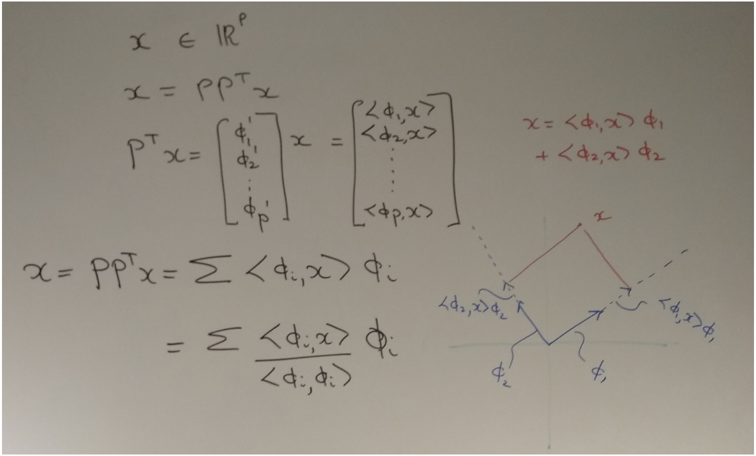
This means all the rows and columns have euclidean length 1 and all the rows are orthogonal to each other and all the columns are orthogonal to each other.

$$P = [\phi_1, \phi_2, \dots \phi_p]$$

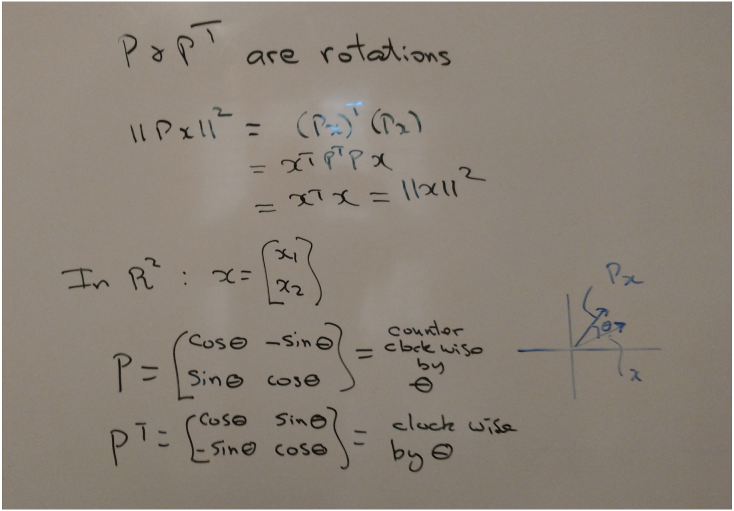
$$I = P^T P = \begin{bmatrix} \phi_1' \\ \phi_2' \\ \vdots \\ \phi_p' \end{bmatrix} [\phi_1, \phi_2, \dots \phi_p] = [\langle \phi_i, \phi_j \rangle]$$

$$\Rightarrow \langle \phi_i, \phi_j \rangle = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

The columns of P (or the rows) form an orthonormal basis for R^p.

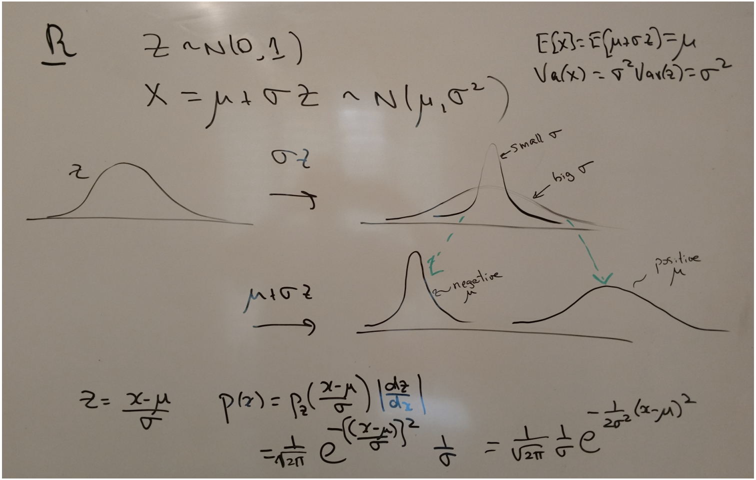


P may be viewed as a rotation.

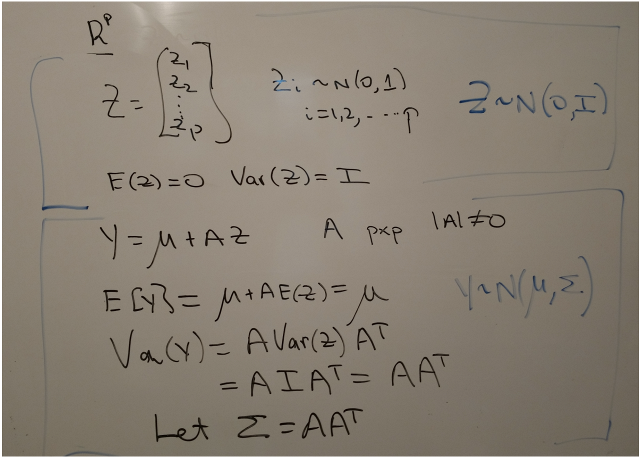


4. Multivariate Normal

In the univariate normal case it is useful to think a general $Y \sim N(\mu, \sigma^2)$ as a linear function of a standard normal:



What about the multivariate normal? Can we express it as a linear function of a "standard normal"?



The multivariate normal density from the change of variable and $Y = \mu + AZ$:

$$f(z) = f(z_1, z_2, \ldots z_p)$$

$$= \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} z_i^2} = (2\pi)^{-\frac{p}{2}} e^{-\frac{1}{2} \|z\|^2}$$

$$f(y) = (2\pi)^{-\frac{p}{2}} e^{-\frac{1}{2}\|A^{-1}(y-\mu)\|^2} |A^{-1}|$$

$$\|A^{-1}(y-\mu)\|^2 = (y-\mu)^T (A^{-1})^T A^{-1} (y-\mu)$$

$$= (y-\mu)^T \{AA^T\}^{-1} (y-\mu)$$

$$= (y-\mu)^T \Sigma^{-1} (y-\mu)$$

$$|\Sigma|^{-\frac{1}{2}} = |AA^T|^{-\frac{1}{2}} = (|A|^2)^{-\frac{1}{2}} = |A^{-1}|$$

$$\boxed{f(y) = (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(y-\mu)^T \Sigma^{-1} (y-\mu)}}$$

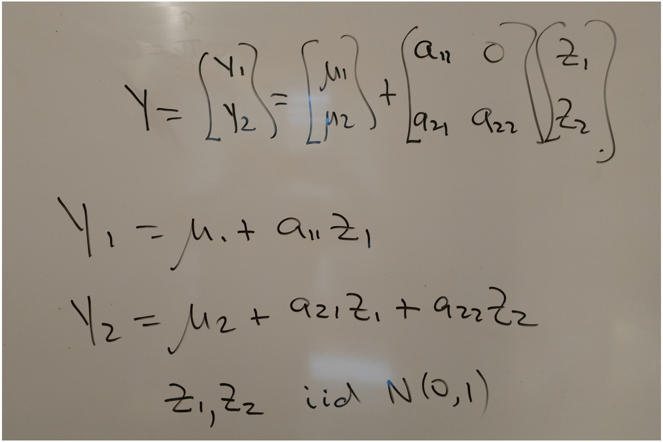
But, can we choose A is such a way that it tell us a nice story about how the Z_i are combined to create a dependent structure embodied in a given Σ?

Given Σ, there is more than one way to choose A such that $\Sigma = AA^T$!!!!

Choleski Decomposition

Given symmetric, positive definite Σ we can always write $\Sigma = AA^T$ where A is lower triangular.

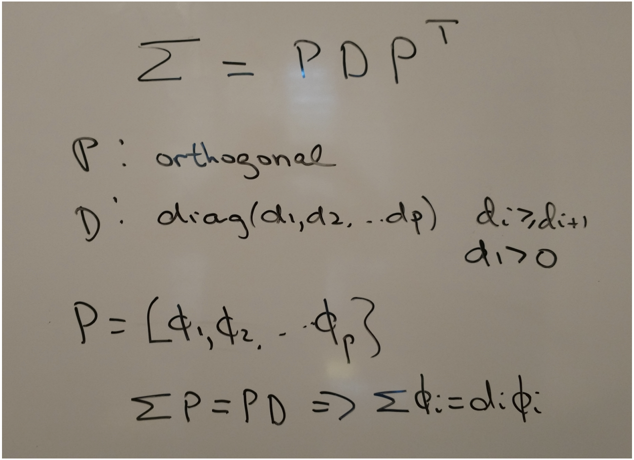
In R^2 we have:



Eigen Decomposition

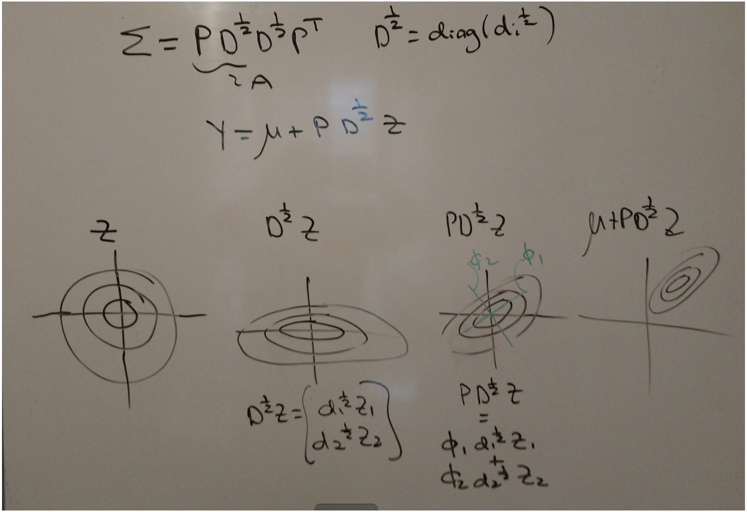
Also called the spectral decomposition.

We can always write a symmetric positive definite $\Sigma = PDP^T$.



The columns of P are the eigen vectors of Σ and the diagonal elements are the corresponding eigen values.

The geometric picture is:



?worth a thousand words?

Note:

A symmetric, pd.

$$A = PDP'$$

(i)

$$|A| = |P|^2 \, |D| = |D| = \prod d_{ii}$$

(ii)

$$tr(A) = tr(DP'P) = tr(D) = \sum d_{ii}$$

Note:

$$A = PDP'$$

$D^{\frac{1}{2}} = diag(d_{ii}^{\frac{1}{2}})$.

$$A = PD^{\frac{1}{2}}D^{\frac{1}{2}}P' = PD^{\frac{1}{2}}P'PD^{\frac{1}{2}}P'$$

Let $A^{\frac{1}{2}} = PD^{\frac{1}{2}}P'$.

So,

$$A = A^{\frac{1}{2}}A^{\frac{1}{2}}.$$

$A^{\frac{1}{2}}$ is called the symmetric pd square root of A.

5. The Choleski Decomposition

Not only is the Choleski decomposition very powerful, you can figure out basic things about it very simply!!

Simple and powerful, my favorite!!

Computing the Choleski:

Choleski:

A symmetric, positive definite $\rightarrow \exists$ lower triangular L such that

$$A = LL'$$

To compute L, you can recursively solve the system of equations give by

$$LL' = A$$

The simple 2×2 case:

$$L = \begin{bmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{bmatrix} \qquad A = \begin{bmatrix} a_{11} & a_{21} \\ a_{21} & a_{22} \end{bmatrix}$$

$LL^T = A$:

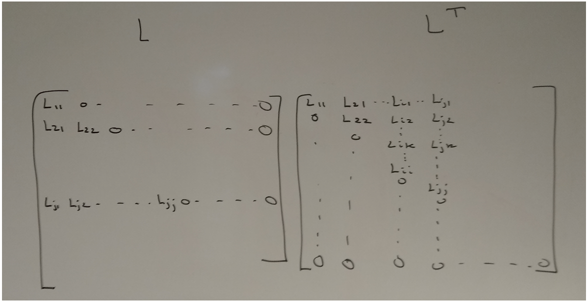
$$LL^T = \begin{bmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{bmatrix} \begin{bmatrix} L_{11} & L_{21} \\ 0 & L_{22} \end{bmatrix} = \begin{bmatrix} L_{11}^2 & L_{11}L_{21} \\ L_{21}L_{11} & L_{21}^2 + L_{22}^2 \end{bmatrix}$$

$$L_{11} = \sqrt{a_{11}}$$

$$L_{21} = a_{21} / L_{11}$$

$$L_{22} = (a_{22} - L_{21}^2)^{\frac{1}{2}}$$

In general we have:

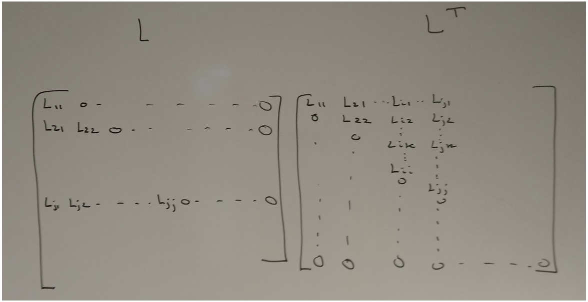


Notice that the top 2×2 corner is just like the simple 2×2 case!

After that we can do solve for L by interating over the rows, and doing each row by iterating over the columns.
Assume we know all the rows of L for rows $1, 2, \ldots, (j-1)$.

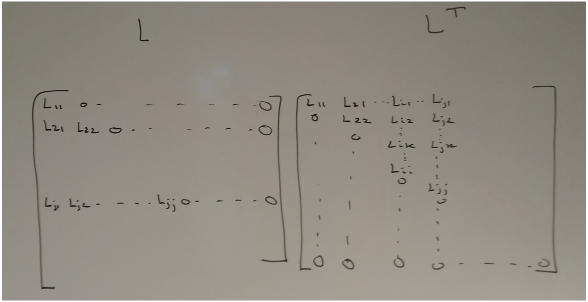
j^{th} row of L times first column of L':

$$L_{j1}\, L_{11} = a_{j1} \rightarrow L_{j1} = a_{j1}/L_{11}.$$



j^{th} row of L times second column of L':

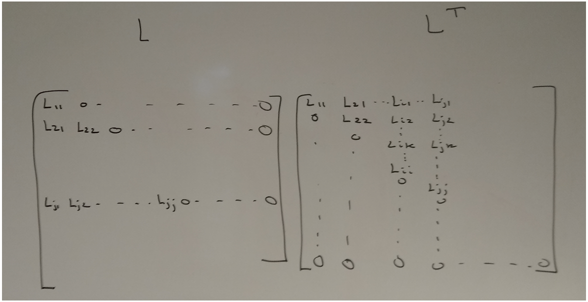
$$L_{j1}L_{21} + L_{j2}L_{22} = a_{j2} \rightarrow L_{j2} = (a_{j2} - L_{j1}L_{21})/L_{22}$$



j^{th} row of L times i^{th} column of L', $(j > i)$:

$$\sum_{k=1}^{i} L_{jk} L_{ik} = a_{ji} \rightarrow L_{ji} = (a_{ji} - \sum_{k=1}^{(i-1)} L_{jk} L_{ik}) / L_{ii}$$

and, finally,



j^{th} row of L times j^{th} column of L', $(j > i)$:

$$\sum_{k=1}^{j} L_{jk}^2 = a_{jj} \to L_{jj} = (a_{jj} - \sum_{k=1}^{(j-1)} L_{jk}^2)^{1/2}$$

Other basic properties:

(i)

For L lt (lower triangular), L^{-1} is lt and fast to compute.

(ii) The system

$$Lx = b$$

is quickly recursively solved.

(iii)

If A is symmetric, pd, then the system

$$Ax = b$$

can be solved by

$$A = LL' \to LL'x = b \to L'x = L^{-1}b$$

Let $y = L^{-1}b$ and solve for y using $Ly = b$.

Then solve for x using $L'x = y$.

As previously noted:

Solve $y \approx Xb$ for b.

solve: $X^T X b = X^T y$

$X = QR$: $X^T X = R^T R$ $X^T y = R^T Q^T y$

solve: $R^T R b = R^T Q^T y$

solve: $R b = Q^T y$

See Murphy, page section 7.5.2.

QR is $O(np^2)$.

6. More on the Multivariate Normal

We'll use the Choleski decomposition to derive fundamental properties of the multivariate normal distribution.

- (a) The marginal from a multivariate normal.
- (b) For normals, uncorrelated $\Rightarrow$ independent.
- (c) The conditional from a multivariate normal.
- (d) Linear of normal is normal.

We partition a normal vector into X and Y.

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix} \right)$$

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

$$\Sigma_{xy} = E\left((X - \mu_x)(Y - \mu_y)^T \right)$$

We take the choleski root of Σ.

$$\Sigma = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}$$

$\Sigma = L L^T$ $\quad$ L lower triangular

$$L = \begin{bmatrix} L_1 & 0 \\ A & L_2 \end{bmatrix}$$ $\quad$ L_i lower tri

We have $(X, Y)'$ in terms of the choleski.
We have Σ in terms of the choleski.

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} + \begin{bmatrix} L_1 & 0 \\ A & L_2 \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix}$$

$$X = \mu_x + L_1 Z_1$$

$$Y = \mu_Y + A Z_1 + L_2 Z_2$$

$$LL^T = \begin{bmatrix} L_1 & 0 \\ A & L_2 \end{bmatrix} \begin{bmatrix} L_1^T & A^T \\ 0 & L_2^T \end{bmatrix} = \begin{bmatrix} L_1 L_1^T & L_1 A^T \\ A L_1^T & A A^T + L_2 L_2^T \end{bmatrix}$$

$$= \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}$$

(a) Marginal of X, (b) uncorrelated implies independence.

(a) $X = \mu_x + L_1 Z$

$\Rightarrow X \sim N(\mu_x, L_1 L_1^T) = N(\mu_x, \Sigma_{xx})$

(b) $\Sigma_{xy} = 0 \Rightarrow A = 0$

$X = \mu_x + L_1 Z_1$

$Y = \mu_Y + L_2 Z_2$

$Z_1 \perp\!\!\!\perp Z_2 \Rightarrow X \perp\!\!\!\perp Y$

(c) $Y \mid X = x$.

(c)

$$Y - \mu_Y = A Z_1 + L_2 Z_2$$

Z_1 and X are 1-1

$$Y - \mu_Y = A L_1^{-1} (x - \mu_X) + L_2 Z_2$$

$$Z_2 \perp\!\!\!\perp X$$

$$Y \mid X = x \sim N\left(\mu_Y + A L_1^{-1}(x - \mu_X),\ L_2 L_2^T\right)$$

Solve for x coefficients in terms of Σ.

$$A L_1^T = \Sigma_{yx}$$

$$A L_1^{-1} L_1 L_1^T = \Sigma_{yx}$$

$$A L_1^{-1} \Sigma_{xx} = \Sigma_{yx}$$

$$A L_1^{-1} = \Sigma_{yx} \Sigma_{xx}^{-1}$$

$Y \mid X = x.$

$$Y = \mu_y + \Sigma_{yx}\Sigma_{xx}^{-1}(X - \mu_x) + E$$

$$E = L_2 Z_2 \perp\!\!\!\perp X$$

$$\Sigma_{yy} = \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xx}\Sigma_{xx}^{-1}\Sigma_{xy} + \mathrm{Var}(E)$$

$$\mathrm{Var}(E) = \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}$$

$$\boxed{Y \mid X = x \sim N\left(\mu_y + \Sigma_{yx}\Sigma_{xx}^{-1}(x - \mu_x),\ \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}\right)}$$

An important special case:

$$\begin{bmatrix} Y \\ X \end{bmatrix} \qquad \begin{array}{l} Y \in \mathbb{R} \\ X \in \mathbb{R}^p \end{array}$$

$$\begin{bmatrix} Y \\ X \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_Y \\ \mu_X \end{bmatrix}, \begin{bmatrix} \sigma_Y^2 & \overset{1 \times p}{\sigma_{YX}} \\ \underset{p \times 1}{\sigma_{XY}} & \underset{p \times p}{\Sigma_{XX}} \end{bmatrix} \right).$$

$$\sigma_{YX} = E\left((Y - \mu_Y)(X - \mu_X)^T \right) \; ; \; \sigma_{XY} = \sigma_{YX}^T$$

$$Y \mid X = x \sim N\left(\mu_Y + \sigma_{YX} \Sigma_{XX}^{-1} (x - \mu_x), \right.$$
$$\left. \sigma_Y^2 - \sigma_{YX} \Sigma_{XX}^{-1} \sigma_{XY} \right)$$

So, if (X, Y) are multivariate normal, $X \in R^p$, $Y \in R$, then,

$$Y \mid X = x \sim N\left(\mu_Y + \sigma_{YX} \Sigma_{XX}^{-1} (x - \mu_X),\right.$$
$$\left.\sigma_Y^2 - \sigma_{YX} \Sigma_{XX}^{-1} \sigma_{XY}\right)$$

$$\beta = \Sigma_{XX}^{-1} \sigma_{XY}$$

$$\sigma^2 = \sigma_Y^2 - \sigma_{YX} \Sigma_{XX}^{-1} \sigma_{XY}$$

$$Y \mid X = x \sim N\left(\mu_Y + (x - \mu_X)^T \beta, \sigma^2\right)$$

So that the conditional distribution of $Y \mid X = x$ has the form of the standard multiple regression model with iid homoscedastic normal errors.

④

Example

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

$$X, Y \sim N(0,1)$$

$$Y = \rho X + E \qquad E \sim N(0, 1-\rho^2)$$

Note $\mu_X, \mu_Y = 0$

$$E((Y - BX)X^T) = 0$$

$$\Rightarrow \quad \Sigma_{YX} = B\,\Sigma_{XX}$$

$$B = \Sigma_{YX} \Sigma_{XX}^{-1}$$

$$\Rightarrow \quad Y = BX + E \qquad E \perp\!\!\!\perp X .$$

(d)

$$Y \sim N(\mu, \Sigma)$$

$$\Rightarrow \quad a + BY \sim N(a + B\mu,\ B\Sigma B')$$

$$\begin{aligned}
X &= a + BY \\
&= a + B[\mu + AZ] \quad ; \quad AA^T = \Sigma \\
&= a + B\mu + BAZ \quad ; \quad (BA)(BA^T) = BAA^TB^T \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad = B\Sigma B^T \\
&\sim N(a + B\mu,\ B\Sigma B^T)
\end{aligned}$$

7. Simulating a Multivariate Normal

Suppose we wish to draw $Y \sim N(\mu, \Sigma)$.

Let $Z = (Z_1, Z_2, \ldots, Z_p)'$, $Z_j \sim N(0,1)$, *iid*.

Then let,

$$Y = \mu + A\,Z$$

where,

$$\Sigma = AA'$$

If A is cholesky, multiplication $A\,Z$ is fast.

8. Likelihood, Sufficiency, and MLE

Let's use our spectral decomposition to learn about the multivariate normal likelihood.

Let,

$$X_i \sim N_p(\mu, \Sigma), \quad iid, \quad i = 1, 2, \ldots, n.$$

$$X_i = (X_{i1}, X_{i2}, \ldots, X_{ip})'$$

Recall that for a parametric model,

$$f(y \mid \theta), \quad \theta \in \theta,$$

given data, y, the maximum likelihood estimator is obtained by finding the θ that makes what you have seen most likely:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, f(y \mid \theta)$$

In practice we often maximize the log of the likelihood or minimize the negative of the log likelihood.

Example:

Bernoulli MLE

$Y_i \sim \text{Bern}(p) \quad Y_i \in \{0,1\}$

$$p(y_1, y_2, \cdots y_n | p) = \prod_{i=1}^{n} p^{y_i} (1-p)^{1-y_i}$$

$$= p^k (1-p)^{n-k} \qquad k = \#(Y_i = 1)$$

$$\log p = k \log p + (n-k) \log(1-p)$$

FOC: $\dfrac{k}{p} - \dfrac{(n-k)}{1-p} = 0 \Rightarrow (n-k)p = k(1-p)$

$$\Rightarrow p = \frac{k}{n}$$

FOC: “first order condition”, $f' = 0$.
So, *the observed sample frequency is the MLE!*

In our problem we will observe $X_i = x_i$ for $X_i \sim N_p(\mu, \Sigma), \;\; iid, \;\; i = 1, 2, \ldots, n.$

note:

x a p dimensional column vector. A $p \times p$.

$$x'Ax = tr(x'Ax) = tr(Axx'),$$

where tr is the trace.

$$p(x_1, x_2, \ldots x_n) =$$

$$\prod_{i=1}^{n} (2\pi)^{-p/2} |\Sigma|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) \right\}$$

$$= (2\pi)^{-\frac{np}{2}} |\Sigma|^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2} \sum_i (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) \right\}$$

$$= (2\pi)^{-\frac{np}{2}} |\Sigma|^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2} \sum_i \mathrm{tr}\left(\Sigma^{-1} (x_i - \mu)(x_i - \mu)^T \right) \right]$$

$$= (2\pi)^{-\frac{np}{2}} |\Sigma|^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2} \mathrm{tr}\left(\Sigma^{-1} \sum_i (x_i - \mu)(x_i - \mu)^T \right) \right]$$

Note: $\bar{x} = \frac{1}{n} \sum x_i$; $\sum (x_i - \bar{x})$

$= \sum x_i - n\bar{x}$

$= n\bar{x} - n\bar{x} = 0.$

$$\sum (x_i - \mu)(x_i - \mu)^T$$

$$= \sum ((x_i - \bar{x}) - (\mu - \bar{x}))(\quad)^T$$

$$= \sum (x_i - \bar{x})(x_i - \bar{x})^T + n(\mu - \bar{x})(\mu - \bar{x})^T$$

$$+ \underbrace{\text{cross terms}}_{=0}$$

Cross terms:

$$\sum (\mu - \bar{x})(x_i - \bar{x})^T = (\mu - \bar{x}) \sum (x_i - \bar{x})^T = 0$$

$$A = \sum_i (x_i - \bar{x})(x_i - \bar{x})'$$

$$\mathrm{tr}\left(\Sigma^{-1} \sum_i (x_i - \mu)(x_i - \mu)^T \right]$$

$$= \mathrm{tr}\left(\Sigma^{-1} \left(A + n (\bar{x} - \mu)(\bar{x} - \mu)^T \right) \right)$$

$$= \mathrm{tr}\, \Sigma^{-1} A + n\, \mathrm{tr}\, \Sigma^{-1} (\bar{x} - \mu)(\bar{x} - \mu)^T$$

$$= \mathrm{tr}\, \Sigma^{-1} A + n\, (\bar{x} - \mu)^T \Sigma^{-1} (\bar{x} - \mu)$$

Sufficiency:

Give data, functions of the data are *sufficient* is they are all we need to compute the likelihood.

Clearly, for iid MVN data,

$$\bar{x} \;\; and \;\; A$$

are sufficient.

$p + \frac{p(p+1)}{2}$ quantities instead of the $n\,p$ data.

What is A?

The k, j element of A is:

$$A_{jk} = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

The sample covariance between X_j and X_k is

$$s_{jk} = \frac{A_{jk}}{(n-1)}$$

The sample variance of X_j is

$$s_{jj} = \frac{A_{jj}}{(n-1)}$$

MLE:

$$L \propto |\Sigma|^{-n/2} \exp(tr(-\tfrac{1}{2}\Sigma^{-1}A)) \exp(-\tfrac{n}{2}(\bar{x}-\mu)'\Sigma^{-1}(\bar{x}-\mu))$$

Clearly, for any Σ, maximum over μ is attained at

$$\hat{\mu} = \bar{x}$$

Notation: $etr(A) = exp(tr(A))$.

$$L(\hat{\mu}, \Sigma) \propto |\Sigma|^{-\frac{n}{2}} \operatorname{etr}\left[-\frac{1}{2}\Sigma^{-1}A\right]$$

$A = T T'$; $\Gamma = T' \Sigma^{-1} T$

$$|\Gamma| = \frac{|T|^2}{|\Sigma|}$$

$$L(\hat{\mu}, \Gamma) \propto |\Gamma|^{\frac{n}{2}} \operatorname{etr}\left[-\frac{n}{2}\frac{\Gamma}{n}\right]$$

γ_i : roots of $\frac{\Gamma}{n}$

$$L \propto \prod_i \gamma_i^{\frac{n}{2}} e^{-\frac{n}{2}\gamma_i}$$

Let $a = \frac{n}{2}$

$$\max_{\gamma} \gamma^{a} e^{-a\gamma}$$

$$\max_{\gamma} a \log(\gamma) - a\gamma$$

FOC: first order condition
set derivative equall to 0

$$\frac{a}{\gamma} = a \Rightarrow \gamma^{*} = 1.$$

So $\frac{\hat{\Gamma}}{n} = P I P^T = I$

$$\hat{\Gamma} = nI$$

Now we simply solve for $\hat{\Sigma}$

$$\hat{\Gamma} = T'\hat{\Sigma}^{-1}T = nI$$

$$\hat{\Sigma}^{-1} = n(T')^{-1}T^{-1} = n(TT')^{-1}$$
$$= nA^{-1}$$

$$\hat{\Sigma} = \frac{A}{n}$$

9. Checking for Normality

Suppose $Y \sim N(\mu, \Sigma)$.

$$\Sigma = PD^{\frac{1}{2}}D^{\frac{1}{2}}P'.$$

$$\Sigma^{-1} = PD^{-\frac{1}{2}}D^{-\frac{1}{2}}P'.$$

Then

$$Y = \mu + PD^{\frac{1}{2}}Z, \;\; Z \sim N(0, I).$$

So,

$$Z = D^{-\frac{1}{2}}P'(Y - \mu).$$

$$Z'Z = (Y - \mu)'PD^{-\frac{1}{2}}D^{-\frac{1}{2}}P'(Y - \mu) = (Y - \mu)'\Sigma^{-1}(Y - \mu).$$

So,

$$(Y - \mu)'\Sigma^{-1}(Y - \mu) = Z'Z = \sum Z_i^2 \sim \chi^2_p.$$

So if $Y_i \sim N(\mu, \Sigma)$, *iid*, then

$$D_i = (Y_i - \hat{\mu})' \hat{\Sigma}^{-1} (Y_i - \hat{\mu}) \approx \chi^2_p, \ \ iid$$

So you can check to see if the D_i look right.

I usually use a qqplot.

10. Weighted Regression

Review Linear Case

Usual: $Y = X\beta + \varepsilon \quad \varepsilon \sim N(0, \sigma^2 I)$

Consider: $Y = X\beta + \varepsilon \quad \varepsilon \sim N(0, \Sigma)$

$$\Sigma = LL^T \quad \text{cholesky: } L \text{ is } \triangle \text{ lower triangular.}$$

$$\Sigma^{-1} = (L^T)^{-1} L^{-1} = (L^{-1})^T L^{-1}$$

$$\tilde{\varepsilon} = L^{-1}\varepsilon \; ; \quad E(\tilde{\varepsilon}) = 0 \quad Var(\tilde{\varepsilon}) = L^{-1}[LL^T](L^{-1})^T$$

$$= (L^{-1}L)(L^T(L^T)^{-1}) = I.$$

$$L^{-1}Y = L^{-1}X\beta + L^{-1}\varepsilon$$

$$\tilde{Y} = \tilde{X}\beta + \tilde{\varepsilon} \quad \tilde{\varepsilon} \sim N(0, I)$$

$$\hat{\beta} = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T\tilde{Y} = (X^T(L^{-1})^T L^{-1}X)^{-1} X^T(L^{-1})^T L^{-1}Y$$

$$= (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y$$