MATH 556: MATHEMATICAL STATISTICS I
HIERARCHICAL MODELS: VARIANCE COMPONENTS

Consider the three-level hierarchical model:

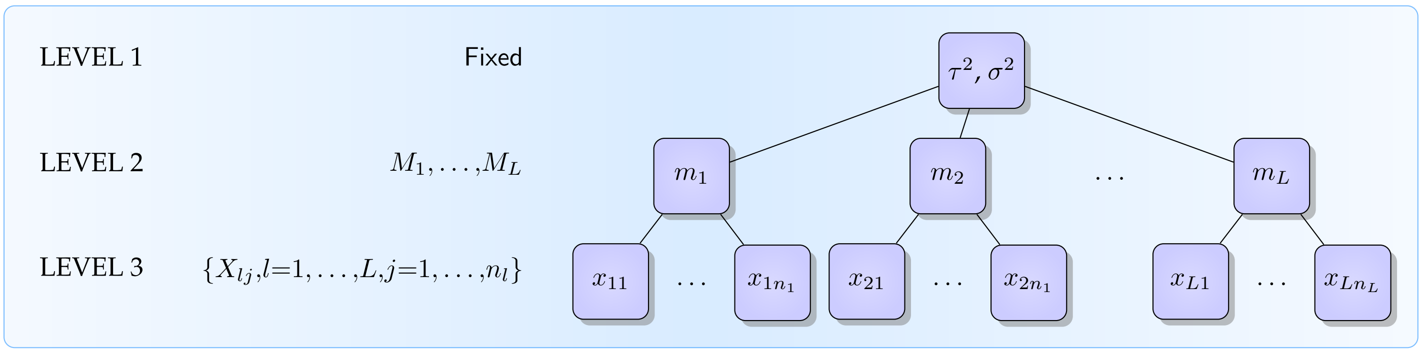
LEVEL 3: $\tau^2, \sigma^2 > 0$, fixed parameters;

LEVEL 2: $M_1, \ldots, M_L \sim Normal(0, \tau^2)$ independent;

LEVEL 1: For $l = 1, \ldots, L$

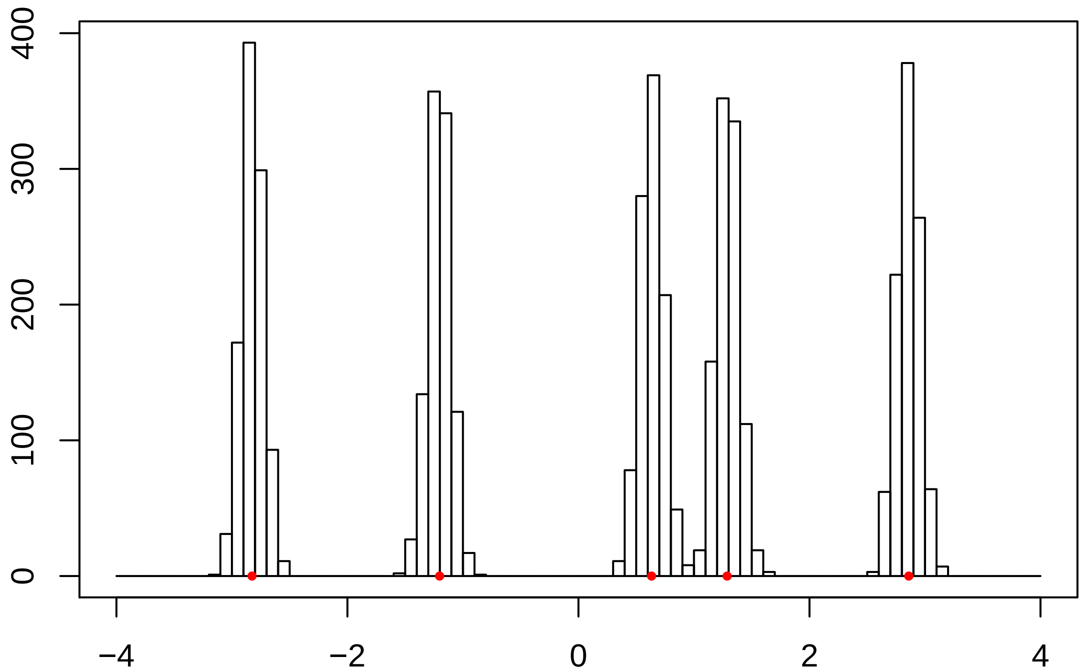
$$X_{l1}, \ldots, X_{ln_l} | M_l = m_l \sim Normal(m_l, \sigma^2)$$

where all the X_{lj} are conditionally independent given $M_1, \ldots, M_L$.



In the following plot, we have $L = 5$, with $n_l = 1000$ for $l = 1, \ldots, L$, with $\tau^2 = 2^2$ and $\sigma^2 = 0.1^2$.

```
set.seed(23984)
L<-5
nvec<-rep(1000,L)
tau<-2; sig<-0.1
M<-rnorm(L,0,tau)
mvec<-rep(M,nvec)
X<-rnorm(sum(nvec),mvec,sig)
par(mar=c(3,3,2,1))
hist(X,breaks=seq(-4,4,by=0.1),main='');box()
points(M,rep(0,L),pch=19,col='red',cex=0.5)
```



In the histogram,

- the red dots indicate the position of the sampled $m_1, \dots, m_5$;
- the histograms represent the sampled X_{lj} for $l = 1, \dots, 5$ and $j = 1, \dots, 1000$.

We can implement the same model with the variables having bivariate Normal distributions: for example

$$\mathbf{M}_l = \begin{bmatrix} M_{l1} \\ M_{l2} \end{bmatrix} \sim Normal_2(\mathbf{0}, \mathbf{V})$$

for $l = 1, \dots, L$, independently, with

$$\mathbf{V} = \begin{bmatrix} 2 & 3 \\ 3 & 4 \end{bmatrix}$$

and, for $j = 1, \dots, n_l$

$$\mathbf{X}_{lj} | \mathbf{M}_l = \mathbf{m}_l \sim Normal_2(\mathbf{m}_l, \Sigma)$$

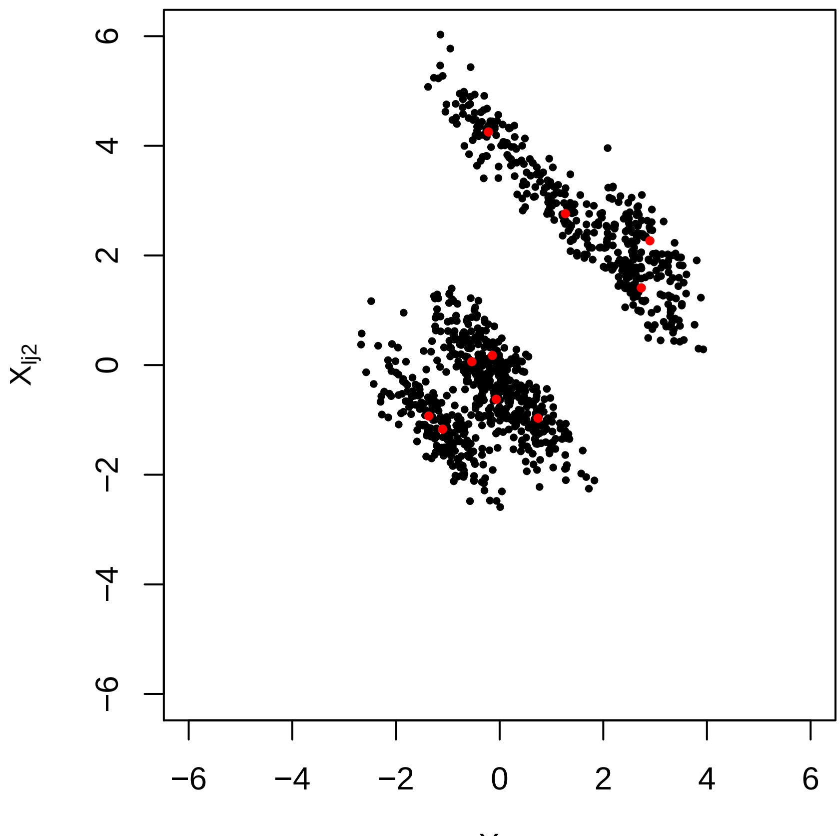
conditionally independent, with the factorization of Σ as

$$\Sigma = \begin{bmatrix} 0.50 & 0.00 \\ 0.00 & 0.60 \end{bmatrix} \begin{bmatrix} 1.0 & -0.8 \\ -0.8 & 1.0 \end{bmatrix} \begin{bmatrix} 0.50 & 0.00 \\ 0.00 & 0.60 \end{bmatrix} = \begin{bmatrix} 0.25 & -0.24 \\ -0.24 & 0.36 \end{bmatrix}.$$

In the following plot, we have $L = 10$, with $n_l = 100$ for $l = 1, \dots, L$

```
set.seed(23984)
library(mvnfast)
L<-10
nvec<-rep(100,L)
V<-matrix(c(4,3,3,4),2,2)
Sigma<-diag(c(0.5,0.60)) %*% matrix(c(1,-0.8,-0.8,1),2,2) %*% diag(c(0.5,0.60))
M<-rmvn(L,rep(0,2),V)
X<-numeric(length=2)
for(l in 1:L){
    Z<-rmvn(nvec[l],M[l,],Sigma)
    X<-rbind(X,Z)
}
```

```r
par(mar=c(3,4,1,1))
plot(X,pch=19,cex=0.4,xlim=range(-6,6),ylim=range(-6,6),
        xlab=expression(X[lj1]),ylab=expression(X[lj2]))
points(M,pch=19,col='red',cex=0.5)
```



For each l, the X_{lj} values for $j = 1, \ldots, n_l$ are conditionally independent given $M_l = m_l$, but are not (unconditionally) independent. For the univariate case, the marginal distribution of $\mathbf{X}_l = (X_{l1}, \ldots, X_{ln_l})^\top$ is computed as

$$\begin{aligned} f_{X_{l1},\ldots,X_{ln_l}}(x_{l1}, \ldots, x_{ln_l}; \tau^2, \sigma^2) &= \int_{-\infty}^{\infty} \prod_{j=1}^{n_l} f_{X_{lj}|M_l}(x_{lj}|m_l; \sigma^2) f_{M_l}(m_l; \tau^2)\, dm_l \\ &= \int_{-\infty}^{\infty} \prod_{j=1}^{n_l} \left\{ \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left\{-\frac{1}{2\sigma^2}(x_{lj} - m_l)^2\right\}\right\} \left(\frac{1}{2\pi\tau^2}\right)^{1/2} \exp\left\{-\frac{1}{2\tau^2} m_l^2\right\} dm_l \\ &= \left(\frac{1}{2\pi}\right)^{(n_l+1)/2} \left(\frac{1}{\sigma^2}\right)^{n_l/2} \left(\frac{1}{\tau^2}\right)^{1/2} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma^2}\sum_{j=1}^{n}(x_{lj} - m_l)^2 + \frac{1}{\tau^2} m_l^2\right]\right\} dm_l. \end{aligned}$$

Completing the square gives

$$\frac{1}{\sigma^2}\sum_{j=1}^{n}(x_{lj} - m_l)^2 + \frac{1}{\tau^2} m_l^2 = \frac{1}{\sigma^2}\sum_{j=1}^{n_l}(x_{lj} - \overline{x}_l)^2 + \left(\frac{n_l}{\sigma^2} + \frac{1}{\tau^2}\right)\left(m_l - \frac{n_l\overline{x}/\sigma^2}{n/\sigma^2 + 1/\tau^2}\right)^2 + \frac{n/\sigma^2}{n/\sigma^2 + 1/\tau^2}\overline{x}_l^2$$

and thus integrating out m_l yields

$$f_{X_{l1},\ldots,X_{ln_l}}(x_{l1},\ldots,x_{ln_l};\tau^2,\sigma^2) = \left(\frac{1}{2\pi}\right)^{n_l/2}\left(\frac{1}{\sigma^2}\right)^{n_l/2}\left(\frac{1}{\tau^2}\right)^{1/2}\left(\frac{n_l}{\sigma^2}+\frac{1}{\tau^2}\right)^{-1/2}$$
$$\exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma^2}\sum_{j=1}^{n_l}(x_{lj}-\overline{x}_l)^2 + \frac{n/\sigma^2}{n/\sigma^2+1/\tau^2}\overline{x}_l^2\right]\right\}$$

This joint pdf does not factorize into a product of functions of the individual x_{lj} values, and hence the random variables are not indepenent. We can compute the distribution more concisely using mgfs and iterated expectation. We have for the multivariate mgf

$$\begin{aligned}
M_{\mathbf{X}_l}(\mathbf{t}) = \mathbb{E}_{\mathbf{X}}\left[\exp\{\mathbf{t}^\top\mathbf{X}\}\right] &= \mathbb{E}_{M_l}\left[\mathbb{E}_{\mathbf{X}|M_l}\left[\exp\{\mathbf{t}^\top\mathbf{X}\}\middle|M_l\right]\right] && \text{by iterated expectation}\\
&= \mathbb{E}_{M_l}\left[\mathbb{E}_{\mathbf{X}|M_l}\left[\exp\left\{\sum_{j=1}^{n_l} t_j X_{lj}\right\}\middle|M_l\right]\right] && \text{expanding the inner product}\\
&= \mathbb{E}_{M_l}\left[\prod_{j=1}^{n_l}\exp\left\{M_l t_j + \frac{t_j^2\sigma^2}{2}\right\}\right] && \text{using the Normal mgf for } X_{lj}\\
&= \exp\left\{\frac{(\mathbf{t}^\top\mathbf{t})\sigma^2}{2}\right\}\mathbb{E}_{M_l}\left[\exp\left\{M_l(\mathbf{1}^\top\mathbf{t})\right\}\right]\\
&= \exp\left\{\frac{(\mathbf{t}^\top\mathbf{t})\sigma^2}{2}\right\}\exp\left\{\frac{(\mathbf{1}^\top\mathbf{t})\tau^2}{2}\right\} && \text{using the Normal mgf for } M_l\\
&= \exp\left\{\frac{\mathbf{t}^\top\mathbf{V}\mathbf{t}}{2}\right\}
\end{aligned}$$

where, by inspection, we have that

$$\mathbf{V} = \sigma^2\mathbf{I}_{n_l} + \tau^2\mathbf{1}\mathbf{1}^\top$$

where, for the (j,k)th element, we have

$$[\mathbf{V}]_{jk} = \left\{\begin{array}{cc} \sigma^2+\tau^2 & j=k \\ \tau^2 & j\neq k \end{array}\right.$$

Thus we can conclude that $\mathbf{X}_l \sim Normal_{n_l}(\mathbf{0},\mathbf{V})$.

We can verify this by direct calculation: we have that

$$\mathbb{E}_{X_{lj}}[X_{lj}] = \mathbb{E}_{M_l}[\mathbb{E}_{X_{lj}|M_l}[X_{lj}|M_l]] = \mathbb{E}_{M_l}[M_l] = 0.$$

and

$$\mathbb{E}_{X_{lj}}[X_{lj}^2] = \mathbb{E}_{M_l}[\mathbb{E}_{X_{lj}|M_l}[X_{lj}^2|M_l]] = \mathbb{E}_{M_l}[M_l^2+\tau^2] = \sigma^2+\tau^2.$$

so $\text{Var}_{X_{lj}}[X_{lj}] = \sigma^2+\tau^2$. Finally, for $j\neq k$,

$$\mathbb{E}_{X_{lj},X_{lk}}[X_{lj}X_{lk}] = \mathbb{E}_{M_l}[\mathbb{E}_{X_{lj},X_{lk}|M_l}[X_{lj}X_{lk}|M_l]] = \mathbb{E}_{M_l}[M_l^2] = \tau^2$$

as X_{lj} and X_{lk} are conditionally independent given M_l, each with mean M_l. We therefore conclude that

$$\text{Corr}_{X_{lj},X_{lk}}[X_{lj},X_{lk}] = \frac{\text{Cov}_{X_{lj},X_{lk}}[X_{lj},X_{lk}]}{\text{Var}_{X_{lj}}[X_{lj}]} = \frac{\tau^2}{\sigma^2+\tau^2}.$$

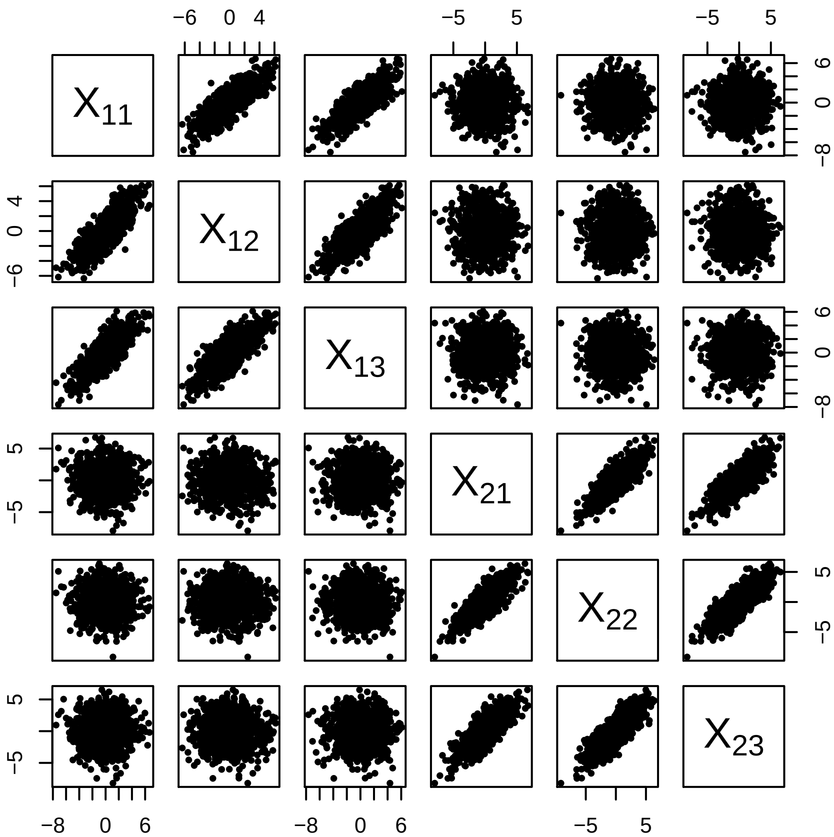
Note that

- this correlation is always positive;
- if $\tau^2 \longrightarrow 0$, the correlation converges to zero, and we have reverted to the iid $Normal(0, \sigma^2)$ case;
- as $\sigma^2 \longrightarrow 0$, the correlation converges to one.
- In this model, $l_1 \neq l_2$, the $X_{l_1 j_1}$ and $X_{l_2 j_2}$ are **unconditionally independent** for all j_1 and j_2.

In the following simulation, we have $L = 2$ and $n_1 = n_2 = 3$, with $\tau = 2$ and $\sigma = 1$, yielding a within-cluster correlation of

$$\frac{\tau^2}{\sigma^2 + \tau^2} = 0.8.$$

```
tau<-2;sig<-1
X<-t(replicate(1000,c(rnorm(3,rnorm(1,0,tau),sig),c(rnorm(3,rnorm(1,0,tau),sig)))))
par(mar=c(3,3,0,1))
pairs(X,cex=0.5,pch=19,labels=c(expression(X[11]),expression(X[12]),expression(X[13]),
                              expression(X[21]),expression(X[22]),expression(X[23])))
```




```
round(cor(X),4)

+          [,1]    [,2]    [,3]    [,4]   [,5]   [,6]
+ [1,] 1.0000  0.8106  0.8059  0.0099 0.0209 0.0394
+ [2,] 0.8106  1.0000  0.8075 -0.0003 0.0158 0.0190
+ [3,] 0.8059  0.8075  1.0000 -0.0013 0.0181 0.0281
+ [4,] 0.0099 -0.0003 -0.0013  1.0000 0.8191 0.8070
+ [5,] 0.0209  0.0158  0.0181  0.8191 1.0000 0.8203
+ [6,] 0.0394  0.0190  0.0281  0.8070 0.8203 1.0000
```