250C Multivariate Biostatistics Review

Tomoki Okuno

Summer 2023

Contents

1 Matrix Normal

- Let $Z \sim MN(0, I_n, I_p)$. Show $X = AZB + C \sim MN(C, AA', B'B)$.

  Solution: First consider $Y = AZ$. Since $Z = A^{-1}Y$,

$$f(Y) = f_Z(A^{-1}Y)\left|\frac{\partial Z}{\partial Y}\right| = (2\pi)^{-np/2}|A^{-1}|^p \exp\left[-\frac{1}{2}\operatorname{tr}(Y'(A^{-1})'A^{-1}Y)\right]$$
$$= (2\pi)^{-np/2}|AA'|^{-p/2} \exp\left[-\frac{1}{2}\operatorname{tr}(Y'(AA')^{-1}Y)\right],$$

  leading to $Y \sim MN(0, AA', I_p)$. Second consider $X = YB + C \Rightarrow Y = (X - C)B^{-1}$, then

$$f(X) = f_Y((X - C)B^{-1})\left|\frac{\partial Y}{\partial X}\right|$$
$$= (2\pi)^{-np/2}|AA'|^{-p/2}|B'B|^{-n/2} \exp\left\{-\frac{1}{2}\operatorname{tr}[(B'B)^{-1}(X - C)'(AA')^{-1}(X - C)]\right\}.$$

- Let $\Omega = AA'$ and $\Sigma = B'B$. Show $\operatorname{Cov}(x_{ij}, x_{kl}) = \omega_{ik}\sigma_{jl}$.

  The covariance of $\operatorname{Vec}(\mathbf{X}) = (\mathbf{X}_1', \ldots, \mathbf{X}_n')'$, where $\mathbf{X}_i \in \mathbb{R}^p$, can be expressed as

$$\operatorname{Cov}(\operatorname{Vec}(\mathbf{X})) = \operatorname{Cov}\begin{pmatrix}\mathbf{X}_1\\ \mathbf{X}_2\\ \vdots\\ \mathbf{X}_n\end{pmatrix} = \mathbf{\Omega} \otimes \mathbf{\Sigma} = \begin{pmatrix}\omega_{11}\mathbf{\Sigma} & \cdots & \omega_{1n}\mathbf{\Sigma}\\ \omega_{21}\mathbf{\Sigma} & \cdots & \omega_{2n}\mathbf{\Sigma}\\ \vdots & \ddots & \vdots\\ \omega_{n1}\mathbf{\Sigma} & \cdots & \omega_{nn}\mathbf{\Sigma}\end{pmatrix} \quad \cdots \quad (*)$$

  Hence, $\operatorname{Cov}(\mathbf{X}_i) = \omega_{ii}\mathbf{\Sigma}$ and then $\operatorname{Var}(x_{ij}) = \operatorname{Cov}(\mathbf{X}_i)_{jj} = \omega_{ii}\sigma_{jj}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. Similarly, $\operatorname{Cov}(x_{ij}, x_{kl}) = \operatorname{Cov}(\mathbf{X}_i, \mathbf{X}_k)_{jl} = (\omega_{ik}\mathbf{\Sigma})_{jl} = \omega_{ik}\sigma_{jl}$ for $i, k = 1, \ldots, n$ and $j, l = 1, \ldots, p$.

2 Gamma/Inverse Gamma distribution

- X follows a **gamma** distribution with shape a and **rate** b parameters, denoted $\text{Gamma}(a, b)$, if

$$f(x) = \frac{b^a}{\Gamma(a)}x^{a-1}\exp(-bx), \quad x > 0,\ a, b > 0$$

  with

$$E(X) = \frac{a}{b}, \quad \operatorname{var}(X) = \frac{a}{b^2}$$

  A χ^2_k random variable corresponds to $\text{Gamma}(k/2, 1/2)$.

- X has a **inverse gamma** distribution with shape a and **scale** b, denoted $\text{InvGamma}(a, b)$, if

$$f(x) = \frac{b^a}{\Gamma(a)}x^{-a-1}\exp\left(-\frac{b}{x}\right), \quad x > 0,\ a, b > 0$$

  with

$$E(X) = \frac{b}{a-1},\ a > 1, \quad \operatorname{var}(X) = \frac{b^2}{(a-1)^2(a-2)},\ a > 2.$$

  If $Y \sim \text{Gamma}(a, b)$, then $X = Y^{-1} \sim \text{InvGamma}(a, b)$.

3 Wishart/Inverse Wishart distribution

- Let $S \sim W_p(n, \Sigma)$ with $\Sigma \succ O$ and $n > p$. S has a (non-singular) **Wishart** distribution with n degrees of freedom if the joint density of the $p(p+1)/2$

$$p(S \mid n, \Sigma) = c|\Sigma|^{-n/2}|S|^{(n-p-1)/2} \exp\left[-\frac{1}{2}\operatorname{tr}(\Sigma^{-1}S)\right]$$

with $E(S) = n\Sigma$. When $p = 1$, this is Gamma$(n/2, 1/(2\sigma^2))$. Further $\sigma^2 = 1$, this is χ^2_n with mean n.

- D has a **Inverse Wishart** distribution with n degrees of freedom, denoted $IW(v, S)$:

$$p(D \mid \nu, S) = c|S|^{v/2}|D|^{-(\nu+p+1)/2} \exp\left[-\frac{1}{2}\operatorname{tr}(SD^{-1})\right]$$

with $E(D) = \frac{S}{\nu-p-1}$ for $\nu > p+1$. When $p = 1$, this is InvGamma$(\nu/2, S/2)$ with mean $S/(\nu-2)$.

- Let $S_1 \perp\!\!\!\perp S_2$ with $S_i \sim W_p(n_i, \Sigma)$, $i = 1, 2$. Show that $S_1 + S_2 \sim W_p(n_1 + n_2, \Sigma)$.

  Solution: We can suppose that

$$\begin{aligned} X = (X_1, \ldots, X_{n_1})' &\sim MN(O, I_{n_1}, \Sigma), \\ Y = (Y_1, \ldots, Y_{n_2})' &\sim MN(O, I_{n_2}, \Sigma), \end{aligned}$$

  Let $Z = (X' \mid Y')' \sim MN(O, I_{n_1+n_2}, \Sigma)$, then

$$S_1 + S_2 = X'X + Y'Y = (X'\ Y')\begin{pmatrix} X \\ Y \end{pmatrix} = Z'Z \sim W_p(n_1 + n_2, \Sigma).$$

4 Hypothesis testing for general hypothesis

- Consider a standard multivariate linear model: $Y = XB + E \sim MN(XB, I_n, \Sigma)$, where $\Sigma : p \times p$.
- Wish to test $H_0 : \underbrace{C}_{g\times q}\underbrace{B}_{q\times p} = \underbrace{D}_{g\times p}$ vs. $H_1 : H_0^c$.
- Then LRT statistic and the null asymptotic distribution are given by

$$\lambda = \left(\frac{|\widehat{\Sigma}_0|}{|\widehat{\Sigma}|}\right)^{-n/2} \quad \Rightarrow \quad -2\log\lambda = n(\log|\widehat{\Sigma}_0| - \log|\widehat{\Sigma}|) \xrightarrow{D} \chi^2_r,$$

  where $r = \dim(\Theta_1) - \dim(\Theta_0)$. The LRT test rejects H_0 if $-2\log\lambda > \chi^2_{r,\alpha}$

  – Useful result: $\max_{\Sigma>0} |\Sigma|^{-n/2} e^{-\frac{1}{2}\operatorname{tr}(\Sigma^{-1}W)} = |\widehat{\Sigma}|^{-n/2} e^{-np/2}$.

- Instead of the LRT test, we often work with

$$\begin{aligned} \lambda^{2/n} = \frac{|\widehat{\Sigma}|}{|\widehat{\Sigma}_0|} = \frac{|(Y - X\widehat{B})'(Y - X\widehat{B})|}{|(Y - X\widehat{B}_0)'(Y - X\widehat{B}_0)|} &= \frac{|(Y - X\widehat{B})'(Y - X\widehat{B})|}{|(Y - X\widehat{B})'(Y - X\widehat{B}) + (\widehat{B} - \widehat{B}_0)'X'X(\widehat{B} - \widehat{B}_0)|} \\ = \frac{|E|}{|E_0|} = \frac{|E|}{|E+H|} &= \frac{|Y'QY|}{|Y'QY + (C\widehat{B} - D)'[C(X'X)C']^{-1}(C\widehat{B} - D)|}, \end{aligned}$$

  where (i) $E = Y'QY \sim W_p(n - q,\ \Sigma)$, (ii) $H = E_0 - E \sim W_p(g,\ \Sigma)$, and (iii) $E \perp H$

  Proof: (i) Let $P = X(X'X)^{-1}X'$ with rank of q. Since $PY = X\widehat{B}$ and P and Q are and idempotent,

$$E = (Y - X\widehat{B})'(Y - X\widehat{B}) = Y'(I_n - P)'(I_n - P)Y = Y'QQY = Y'QY.$$

By the spectral decomposition,

$$P = U \begin{pmatrix} I_q & O \\ O & O \end{pmatrix} U' \quad \Rightarrow \quad Q = I_n - P = U \begin{pmatrix} O & O \\ O & I_{n-q} \end{pmatrix} U' = U_1 U_1',$$

where U is an $n \times n$ orthogonal matrix and U_1 has $n-q$ orthogonal columns ($U_1'U_1 = I_{n-q}$). Thus,

$$\begin{aligned} E = Y'QY &= (Y - X\widehat{B})'Q(Y - X\widehat{B}) \quad \because QX = O,\ X'Q = O \\ &= (Y - X\widehat{B})'U_1U_1'(Y - X\widehat{B}) \\ &\sim W_p(n-q, \Sigma), \end{aligned}$$

as $U_1'(Y - X\widehat{B}) \sim MN(0, U_1'U_1 = I_{n-q}, \Sigma)$.

(ii) We have the same restriction,

$$\widehat{B}_0 = \widehat{B} - (X'X)^{-1}C'[C(X'X)^{-1}C']^{-1}(C\widehat{B} - D).$$

Using this, we get

$$\begin{aligned} H &= E_0 - E \\ &= (Y - X\widehat{B}_0)'(Y - X\widehat{B}_0) - (Y - X\widehat{B})'(Y - X\widehat{B}) \\ &= [Y - X\widehat{B} + X(\widehat{B} - \widehat{B}_0)]'[Y - X\widehat{B} + X(\widehat{B} - \widehat{B}_0)] - (Y - X\widehat{B})'(Y - X\widehat{B}) \\ &= Y'QX(\widehat{B} - \widehat{B}_0) + (\widehat{B} - \widehat{B}_0)'X'QY + (\widehat{B} - \widehat{B}_0)'X'X(\widehat{B} - \widehat{B}_0) \\ &= (\widehat{B} - \widehat{B}_0)'X'X(\widehat{B} - \widehat{B}_0) \\ &= (C\widehat{B} - D)'[C(X'X)^{-1}C']^{-1}(C\widehat{B} - D). \end{aligned}$$

Further, since $(X'X)^{-1} \succ O$ and C is of full row rank, i.e., $C'a = 0 \Rightarrow a = 0$,

$$a'C(X'X)^{-1}C'a = (C'a)'(X'X)^{-1}C'a > 0,\ \forall a \neq 0,$$

meaning that $C(X'X)^{-1}C$ is positive definite. Thus, we can write

$$H = (C\widehat{B} - D)'[C(X'X)^{-1}C']^{-1/2}[C(X'X)^{-1}C']^{-1/2}(C\widehat{B} - D) = W'W \in \mathbb{R}^{p \times p},$$

where $W = [C(X'X)^{-1}C']^{-1/2}(C\widehat{B} - D) \in \mathbb{R}^{g \times p}$. Here we have

$$\widehat{B} = (X'X)^{-1}X'Y \sim MN(B, (X'X)^{-1}, \Sigma).$$

It follows that

$$C\widehat{B} - D \sim MN(0, C(X'X)^{-1}C', \Sigma) \quad \text{under } H_0$$

and then $W = [C(X'X)^{-1}C']^{-1/2}(C\widehat{B} - D) \sim MN(0, I_g, \Sigma)$. so that

$$H = W'W \sim W_p(g, \Sigma).$$

(iii) Suppose $Y = (Y_1, \ldots, Y_p)$, where $Y_j \sim N_n(XB_j, I_n)$, $j = 1, \ldots, p$. Then since $X'Q = X'(I_n - P) = O$, $X'Y_i \perp\!\!\!\perp QY_i$ and so $X'Y \perp\!\!\!\perp QY$ by Theorem 2.5 in the Seber textbook (Craig's theorem). Hence,

$$\begin{aligned} X'Y \perp\!\!\!\perp QY &\Rightarrow (X'X)^{-1}X'Y \perp\!\!\!\perp (QY)'(QY) \\ &\Rightarrow \widehat{B} \perp\!\!\!\perp E \quad \because (QY)'(QY) = Y'QY = E \\ &\Rightarrow H \perp\!\!\!\perp E \end{aligned}$$

since we see that H is a function of $\widehat{B}$ by (ii).

- If $n - q > g$ Wilk's lambda statistic and the null distribution are

$$\lambda^{2/n} = \frac{|E|}{|E+H|} \quad \sim \quad U(p,\ n-q,\ g).$$

- Note that the Wilk's distribution is invariant under changes of the scale parameters of E and H.
- If the null hypothesis is a form of

$$H_0 : \underbrace{C}_{g\times q}\underbrace{B}_{q\times p}\underbrace{M}_{p\times m} = \underbrace{D}_{g\times m},$$

then simply consider the transformation

$$W = YM = XBM + EM := X\Gamma + F,$$

where $F = EM \sim MN(0, I_n, M'\Sigma M)$ and $W \sim MN(X\Gamma, I_n, M'\Sigma M)$. Thus, H_0 can be simplified to a form of $H_0 : C\Gamma = D$, so that the previous argument can apply to this.

- (Midterm) Extend the univariate F test (all coefficients other than the intercept are 0 in a standard 1-dim regression framework) to the multivariate case. Derive the form of the test, and define its null distribution and rejection region. You may assume the first column of X is identically equal to 1_n.

Solution: The null hypothesis is a form of $CBM = 0$, where $C = (0\ I_{q-1}) : (q-1)\times q$ and $M = I_p$. Hence, $E \sim W_p(n-q, \Sigma)$ and $H \sim W_p(q-1, \Sigma)$ so that the Wilk's test statistic follows $U(p, n-q, q-1)$ as $g = \text{rank}(C) = q-1$ under H_0, i.e.,

$$\lambda^{n/2} = \frac{|E|}{|E+H|} \sim_{H_0} U(p, n-q, q-1).$$

- **Union-Intersection Principle**. Consider $H_0 : \theta \in \cap_{\gamma\in\Gamma} H_{0\gamma}$ vs. $H_1 : H_0^c$ with

$$H_{0\gamma} : \theta \in \Theta_\gamma, \quad \text{vs.} \quad H_{1\gamma} : \theta \in \Theta_\gamma^c\ (= H_{0\gamma}^c).$$

If the rejection region for the test of $H_{0\gamma}$ is $\{y : -2\log\lambda_\gamma(y) > c\}$, then the rejection region for H_0 is

$$R = \cup_{\gamma\in\Gamma}\{y : -2\log\lambda_\gamma(y) > c\} = \{y : \max_\gamma[-2\log\lambda_\gamma(y)] > c\}.$$

- **Intersection-Union Principle**. Consider $H_0 : \theta \in \cup_{\gamma\in\Gamma} H_{0\gamma}$ vs. $H_1 : H_0^c$ with

$$H_{0\gamma} : \theta \in \Theta_\gamma, \quad \text{vs.} \quad H_{1\gamma} : \theta \in \Theta_\gamma^c\ (= H_{0\gamma}^c).$$

If the rejection region for the test of $H_{0\gamma}$ is $\{y : -2\log\lambda_\gamma(y) > c\}$, then the rejection region for H_0 is

$$R = \cap_{\gamma\in\Gamma}\{y : -2\log\lambda_\gamma(y) > c\} = \{y : \min_\gamma[-2\log\lambda_\gamma(y)] > c\}.$$

5 Full conditionals

- The full conditional posterior distribution for a given node (parameter) is proportional to the product of the probability model for that node and its parent node, i.e.,

$$\text{Full conditional} \ \propto\ \text{Likelihood / Prior} \ \times\ \text{Distribution for its node}$$

- Recall that normal posteriors are defined as Posterior $\propto$ Likelihood $\times$ Prior.

6 Estimating functions and sandwich variance

- The need to specify a full probability model for the data is undesirable. Here, we give a framework within which the asymptotic properties of a broad range of estimation recipes may be evaluated.
- Let $Y \in \mathbb{R}^n$ represent n observations from a distribution indexed by a p-dimensional parameter θ, with $\text{cov}(Y_i, Y_j \mid \theta) = 0$ $(i \neq j)$.
- Assume that $Y_1, \ldots Y_n$ are iid. Suppose that $\widehat{\theta}_n$ is a solution to the estimating function

$$G_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} G(\theta, Y_i) = 0, \quad \text{i.e.,} \quad G_n(\widehat{\theta}_n) = 0.$$

  Then $\widehat{\theta}_n \xrightarrow{P} \theta$ (consistency) and

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{D} N_p(0, A^{-1}B(A')^{-1}) \quad \Rightarrow \quad \text{Var}(\widehat{\theta}_n) \approx \frac{A^{-1}B(A^T)^{-1}}{n} \text{ (sandwich form)},$$

  where

$$A = E\left[\frac{\partial}{\partial \theta^T} G(\theta, Y)\right], \quad B = E\left[G(\theta, Y)G(\theta, Y)^T\right] = \text{var}[G(\theta, Y)].$$

  Again, Y_i's are iid. We can derive this by expanding $G_n(\widehat{\theta}_n)$ in a Taylor series around the true θ.
- Empirically, the sandwich variance estimator can be obtained as

$$\widehat{\text{Var}}(\widehat{\theta}_n) \approx \frac{\widehat{A}^{-1}\widehat{B}(\widehat{A}')^{-1}}{n}$$

  by evaluating A and B empirically, by Method of Moment,

$$\widehat{A} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \theta^T} G(\widehat{\theta}, Y_i), \quad \widehat{B} = \frac{1}{n}\sum_{i=1}^{n} G(\widehat{\theta}, Y_i)G(\widehat{\theta}, Y_i)^T.$$

- If Y's are independent but *not* identically distributed, then

$$[A_n^{-1}B_n(A_n')^{-1}]^{-1/2}(\widehat{\theta}_n - \theta) \xrightarrow{D} N(0, I_p) \quad \Rightarrow \quad \text{Var}(\widehat{\theta}_n) \approx A_n^{-1}B_n(A_n')^{-1},$$

  where

$$A_n = E\left[\frac{\partial}{\partial \theta^T} G_n(\theta)\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[\frac{\partial}{\partial \theta^T} G(\theta, Y_i)\right],$$

$$B_n = E\left[G_n(\theta)G_n(\theta)^T\right] = \text{var}[G_n(\theta)] = \frac{1}{n^2}\sum_{i=1}^{n} \text{var}[G(\theta, Y_i)].$$

- Consider the situation in which we View the score function as an estimating function as follows

$$G_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} S(\theta, Y_i) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \log p(Y_i \mid \theta)}{\partial \theta},$$

  which implies $G_n(\widehat{\theta}_{\text{MLE}}) = 0$.

7 Quasi-likelihood (more complicated case)

- We describe an estimating function that is based upon the mean and variance of the data only. Suppose we specify the first two moments of the data as

$$\begin{aligned} E(Y_i \mid \beta) &= \mu_i(\beta) \\ \text{Var}(Y_i \mid \beta) &= V_i(\alpha, \beta), \ \alpha > 0, \\ \text{Cov}(Y_i, Y_j \mid \beta) &= 0, \ i \neq j \end{aligned}$$

  where $\beta \in \mathbb{R}^p$, and α is an $r \times 1$ vector of parameters that appear only in the variance model.

- Another expression using vector is

$$\begin{aligned} E(\mathbf{Y} \mid \beta) &= \boldsymbol{\mu}(\beta) : n \times 1 \\ \text{Var}(\mathbf{Y} \mid \beta) &= \text{diag}(V_1, \ldots, V_n) \end{aligned}$$

- Let $\widehat{\alpha}$ be a consistent estimator of α, which would be

$$\widehat{\alpha} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)}.$$

  The estimator $\widehat{\beta}$ that satisfies the estimating equation

$$\begin{aligned} 0 = G(\beta, \widehat{\alpha}) &= \mathbf{D}(\beta)' \mathbf{V}(\widehat{\alpha}, \beta)^{-1} [\mathbf{Y} - \boldsymbol{\mu}(\beta)] \\ &= \sum_{i=1}^{n} \left[\frac{\partial \mu_i}{\partial \beta} \right] V_i(\widehat{\alpha}, \beta)^{-1} [Y_i - \mu_i(\beta)] \end{aligned}$$

  where $\mathbf{D}$ is the $n \times p$ matrix of derivatives with elements $\partial \mu_i / \partial \beta_j$.

- Suppose $\text{Var}(Y) = V$ (assumption is true), then we have asymptotic distribution

$$(\widehat{D}^T \widehat{V}^{-1} \widehat{D})^{1/2} (\widehat{\beta} - \beta) \xrightarrow{D} N_p(0, I_p), \quad \Rightarrow \quad \text{Var}(\widehat{\beta}) = (\widehat{D}^T \widehat{V}^{-1} \widehat{D})^{-1}$$

- The word "quasi" refers to the fact that the resulting sampling model may not correspond to a specific distribution. Thus, no probabilistic prediction is usually warranted.
- Consistency of $\widehat{\beta}$ depends only on specifying correctly the mean structure ($E[G(\alpha, \beta)] = 0 \Rightarrow E(Y) = \mu(\beta)$) and consistency of $\widehat{\alpha}$.
- Asymptotically appropriate variances (SEs) depend on the correct specification of the mean/variance.

8 Quasi-likelihood + Sandwich estimation

- Assume $E(Y_i) = \mu_i$, $\text{var}(Y_i) = V_i(\alpha, \mu_i(\beta))$ and $\text{cov}(Y_i, Y_j) = 0$ $(i \neq j)$ as *working* covariate model.
- Take the quasi-score function as an estimating function and in this case, the variance of $\widehat{\beta}$ is

$$\text{Var}(\widehat{\beta}) = (D' V^{-1} D)^{-1} (D' V^{-1} \, \text{Var}(Y) V^{-1} D)(D' V^{-1} D)^{-1}.$$

- The model-based variance estimator is obtained by substituting $\text{Var}(Y) = V$ to get $(D' V^{-1} D)^{-1}$.
- The robust (naive) variance estimator is given by

$$\widehat{\text{Var}}(\widehat{\beta}) = (D' V^{-1} D)^{-1} (D' V^{-1} \, \text{diag}(RR') V^{-1} D)(D' V^{-1} D)^{-1},$$

  where $R = (R_1, \ldots, R_n)'$ is the $n \times 1$ vector of (unstandardized) residuals with elements $R_i = Y_i - \mu_i(\widehat{\beta})$ so that $\text{diag}(RR') = \text{diag}\left([Y_1 - \mu_1(\widehat{\beta})]^2, \ldots, [Y_n - \mu_n(\widehat{\beta})]^2 \right)$.

- This robust estimator is consistent for the variance of $\widehat{\beta}$ under correct specification of the mean and with uncorrelated data.
- There are two things to bear in mind when one considers the use of the sandwich technique.
  1. Unless the sample size is sufficiently large, the sandwich estimator may be highly unstable. The model-based estimators may be preferable for small- to medium-sized n.
  2. The second consideration is that if the assumed mean–variance model is correct, then a model-based estimator is more efficient.

9 GEE (Generalized Estimating Equations)

- GEE = multivariate (correlated within individual) + Quasi-likelihood + Sandwich.
- Assume the marginal mean model $E(Y_i) = \mu_i(\beta)$, $i = 1, \ldots, m$, and consider the $n_i \times n_i$ working correlation matrix

$$\mathrm{Var}(Y_i) = W_i(\alpha) = \Delta_i^{1/2} R_i(\alpha) \Delta_i^{1/2} \quad \text{with} \quad \mathrm{Cov}(Y_i, Y_j) = 0 \ (i \neq j),$$

  where α is usually unknown, so that observations on different individuals are assumed uncorrelated, and $\Delta_i^{1/2} = \mathrm{diag}(V_{i1}, \ldots, V_{in_i})$.
- Then we obtain the estimating equation:

$$G(\alpha, \beta) = \mathbf{D}'\mathbf{W}(\alpha)^{-1}(\mathbf{Y} - \boldsymbol{\mu}(\beta)) = \sum_{i=1}^{m} D_i' W_i(\alpha)^{-1}(Y_i - \mu_i(\beta)) = 0, \quad D_i = \frac{\partial \mu_i}{\partial \beta}.$$

- Let $\widehat{\beta}$ be a solution to the function such that $G(\alpha, \widehat{\beta}_m) = 0$, then

$$\widehat{\beta}_m \xrightarrow{P} \beta,$$
$$(A_m^{-1} B_m A_m^{-1})^{-1/2} (\widehat{\beta}_m - \beta) \xrightarrow{D} N(0, I_p),$$

  where

$$\begin{aligned}
A_m &= E\left[\frac{\partial G(\alpha, \beta)}{\partial \beta}\right] = -\sum_{i=1}^{m} D_i' W_i(\alpha)^{-1} D_i, \\
B_m &= E[G(\alpha, \beta) G(\alpha, \beta)'] \\
&= E\left\{ \left[\sum_{i=1}^{m} D_i' W_i(\alpha)^{-1}(Y_i - \mu_i(\beta))\right] \left[\sum_{i=1}^{m} (Y_i - \mu_i(\beta))' W_i(\alpha)^{-1} D_i\right] \right\} \\
&= E\left\{ \left[\sum_{i=1}^{m} D_i' W_i(\alpha)^{-1}[Y_i - \mu_i(\beta)][Y_i - \mu_i(\beta)]' W_i(\alpha)^{-1} D_i\right] \right\} \\
&= \sum_{i=1}^{m} D_i' W_i(\alpha)^{-1} E\left\{[Y_i - \mu_i(\beta)][Y_i - \mu_i(\beta)]'\right\} W_i(\alpha)^{-1} D_i \\
&= \sum_{i=1}^{m} D_i' W_i(\alpha)^{-1} \mathrm{Var}(Y_i) W_i(\alpha)^{-1} D_i
\end{aligned}$$

  leading to

$$\mathrm{Var}(\widehat{\beta}) = \left(\sum_{i=1}^{m} D_i' W_i(\alpha)^{-1} D_i\right)^{-1} \left(\sum_{i=1}^{m} D_i' W_i(\alpha)^{-1} \mathrm{Var}(Y_i) W_i(\alpha)^{-1} D_i\right)^{-1} \left(\sum_{i=1}^{m} D_i' W_i(\alpha)^{-1} D_i\right)^{-1}.$$

- If $\text{Var}(Y_i) = W_i(\alpha)$, i.e., the assumed variance-covariance matrix is true, the model-based (naive) variance estimator is given by

$$\widehat{\text{Var}}(\widehat{\beta}) = \left(\sum_{i=1}^{m} D_i' \widehat{W}_i^{-1} D_i\right)^{-1}$$

where $\widehat{W}_i = W_i(\widehat{\alpha})$.

- Alternatively, $\text{Var}(Y_i)$ is estimated as $\widehat{\text{Var}}(Y_i) = R_i R_i'$, where $R_i = Y_i - \mu_i(\widehat{\beta})$, so that we obtain

$$\widehat{\text{Var}}(\widehat{\beta}) = \left(\sum_{i=1}^{m} D_i' \widehat{W}_i^{-1} D_i\right)^{-1} \left(\sum_{i=1}^{m} D_i' \widehat{W}_i^{-1} R_i R_i' \widehat{W}_i^{-1} D_i\right)^{-1} \left(\sum_{i=1}^{m} D_i' \widehat{W}_i^{-1} D_i\right)^{-1},$$

which is called a robust (empirical) variance estimator. *Empirical* is a more appropriate description since the form can be highly unstable for small m.

- Note that we cannot estimate $\text{Var}(Y)$ when there is dependence between individuals.
- If we consider longitudinal linear models, i.e., $E(Y_i) = X_i\beta$, then $D_i = X_i$.
- For inference, we may use the asymptotic distribution

$$\widehat{\text{Var}}(\widehat{\beta})^{-1/2}(\widehat{\beta} - \beta) \xrightarrow{D} N_p(0, I).$$

Consistency of $\widehat{\beta}$ depends only on specifying correctly the mean structure and consistency of $\widehat{\alpha}$.

- No subject-level inference is guaranteed, i.e., we cannot obtain $\widehat{Y}_1, \ldots, \widehat{Y}_n$.
- For the choice of "working" correlation structure, if we choose a simple structure, there are few elements in α to estimate such as independence and exchangeability, but there is a potential loss of efficiency. On the other hand, a more complex structure may provide greater efficiency but more instability in estimation of α. This choice depends on the sample size, with relatively sparse data encouraging the use of a simple correlation structure.
- Hypothesis testing with GEE. Consider $H_0 : L\beta = 0$ for $L : r \times p$, $r \leq p$.
  - LRTs are not available because no likelihood (sampling distribution) is assumed, unlike GLMM.
  - Wald statistics (less than optimal for binary data)

$$W = (L\widehat{\beta} - L\beta)'(L\widehat{\text{Var}}(\widehat{\beta})L')(L\widehat{\beta} - L\beta)$$

so that under H_0

$$W = (L\widehat{\beta})'(L\widehat{\text{Var}}(\widehat{\beta})L')^{-1}(L\widehat{\beta}) \xrightarrow{D} \chi^2_r.$$

  - Quasi-score statistics (less than optimal for binary data). Under H_0,

$$S = G_n(\widetilde{\beta})(\widehat{\text{Var}}(\widetilde{\beta}))^{-1}G_n(\widetilde{\beta}) \xrightarrow{D} \chi^2_r$$

with $\widetilde{\beta}$ being the GEE estimate of β under H_0.

10 Estimating equation in Poisson (log-linear) GEE

- Consider Poisson GEE with $\mathbf{Y}_i : n_i \times 1$ for $i = 1, \ldots, m$, $\log \mu_{ij} = \mathbf{X}_{ij}'\beta$ and

$$E(Y_{ij}) = \mu_{ij}(\beta) = \exp(\mathbf{X}_{ij}'\beta),$$
$$V(Y_{ij}) = \alpha\mu_{ij}(\beta) = V_{ij},$$

where $\mathbf{X}_{ij} : p \times 1$ and $\beta : p \times 1$. Let $\boldsymbol{\mu}_i(\beta) = (\mu_{i1}, \ldots, \mu_{in_i})'$. Then

$$\mathbf{D}_{ij} = \frac{\partial \mu_{ij}}{\partial \beta} = \exp(\mathbf{X}_{ij}'\beta)\mathbf{X}_{ij} = \mu_{ij}\mathbf{X}_{ij} : p \times 1$$

so that $\mathbf{D}_i = \begin{pmatrix} \mu_{i1}\mathbf{X}_{i1} & \cdots & \mu_{in_i}\mathbf{X}_{in_i} \end{pmatrix}' : n_i \times p$.

- Assume working independence:

$$\mathbf{W}_i = \boldsymbol{\Delta}_i^{1/2}\boldsymbol{\Delta}_i^{1/2} = \boldsymbol{\Delta} = \alpha \operatorname{diag}(V_{i1}, \ldots, V_{in_i}) = \alpha \operatorname{diag}(\mu_{i1}, \ldots, \mu_{in_i}) : n_i \times n_i$$

- Hence, the estimating function is given by

$$G(\alpha, \beta) = \sum_{i=1}^{m} \mathbf{D}_i'\mathbf{W}_i^{-1}[\mathbf{Y}_i - \boldsymbol{\mu}_i(\beta)] = \frac{1}{\alpha}\sum_{i=1}^{m} \mathbf{X}_i'[\mathbf{Y}_i - \boldsymbol{\mu}_i(\beta)],$$

which is equivalent to the score function, $\ell'(\beta)$.

11 Estimating equation in Binomial (logit) GEE

- Consider Binomial (logit) GEE with $\mathbf{Y}_i : n_i \times 1$ for $i = 1, \ldots, m$, $\log \frac{\mu_{ij}}{1-\mu_{ij}} = \mathbf{X}_{ij}'\beta$ and

$$E(Y_{ij}) = n_{ij}\mu_{ij}(\beta) = n_{ij}\frac{\exp(\mathbf{X}_{ij}'\beta)}{1 + \exp(\mathbf{X}_{ij}'\beta)},$$
$$V(Y_{ij}) = \alpha\mu_{ij}(1 - \mu_{ij}) = V_{ij},$$

where $\mathbf{X}_{ij} : p \times 1$ and $\beta : p \times 1$. Let $\boldsymbol{\mu}_i(\beta) = (\mu_{i1}, \ldots, \mu_{in_i})'$. Then

$$\mathbf{D}_{ij} = \frac{\partial \mu_{ij}}{\partial \beta} = \frac{\exp(\mathbf{X}_{ij}'\beta)}{[1 + \exp(\mathbf{X}_{ij}'\beta)]^2}\mathbf{X}_{ij} = \mu_{ij}(1 - \mu_{ij})\mathbf{X}_{ij} : p \times 1$$

so that $\mathbf{D}_i = \begin{pmatrix} \mu_{i1}(1-\mu_{i1})\mathbf{X}_{i1} & \cdots & \mu_{in_i}(1-\mu_{in_i})\mathbf{X}_{in_i} \end{pmatrix}' : n_i \times p$

- Assume working independence. Then

$$\mathbf{W}_i = \boldsymbol{\Delta}_i^{1/2}\boldsymbol{\Delta}_i^{1/2} = \boldsymbol{\Delta} = \alpha \operatorname{diag}(V_{i1}, \ldots, V_{in_i}) = \alpha \operatorname{diag}(\mu_{i1}(1-\mu_{i1}), \ldots, \mu_{in_i}(1-\mu_{in_i})) : n_i \times n_i$$

- Hence, the estimating function is given by

$$G(\alpha, \beta) = \sum_{i=1}^{m} \mathbf{D}_i'\mathbf{W}_i^{-1}[\mathbf{Y}_i - \boldsymbol{\mu}_i(\beta)] = \frac{1}{\alpha}\sum_{i=1}^{m} \mathbf{X}_i'[\mathbf{Y}_i - \boldsymbol{\mu}_i(\beta)],$$

which is exactly the same as Poisson GEE and is equivalent to the score function, $\ell'(\beta)$.

12 GLMM Inference

- A GLMM is defined as follows:

$$p(y_{ij} \mid \theta_{ij}) = \exp\left[\frac{y_{ij}\theta_{ij} - d(\theta_{ij})}{a(\alpha)} + c(y_{ij}, \alpha)\right]$$

with $\eta_{ij} = g(\mu_{ij}) = x_{ij}'\beta + z_{ij}b_i$ and $b_i \overset{iid}{\sim} N(0, D(\alpha))$. Then we can show that

$$E(y_{ij} \mid \theta_{ij}) = \frac{\partial d(\theta_{ij})}{\partial \theta_{ij}} = d'(\theta_{ij}) = \mu_{ij}$$
$$\operatorname{var}(y_{ij} \mid \theta_{ij}) = a(\alpha)d''(\theta_{ij}) = a(\alpha)V(\mu_{ij})$$

- Two inference strategies
  1. **Marginal Likelihood**: Make specific assumptions about the distribution of the random effects b_i and carry out inference for the fixed effects (β) by marginalizing/integrating over b_i.
  2. **Conditional Likelihood**: Treat b_i as a nuisance parameter and estimate β and α conditioning on a sufficient statistic for b_i.
- Marginal Likelihood. Consider mixed effects model for $p(y_{ij} \mid \theta_{ij})$. Estimation may be based on

$$L(\beta, \alpha) = \prod_{i=1}^{n} \int p(y_{ij} \mid \beta, \alpha, b_i) p(b_i \mid \alpha) db_i,$$

  This integral cannot be expressed in closed form. In addition to numerically approximating this integral (e.g. via **Gauss–Hermite quadrature**), methods motivated by **Laplace approximation** have been proposed such as **PQL**, penalized quasi-likelihood.
- Example (binary longitudinal data). Consider $Y_{ij} \mid p_{ij} \sim \text{Bern}(p_{ij})$, with

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = x_{ij}'\beta + b_i \quad \Leftrightarrow \quad p_{ij} = \frac{\exp(x_{ij}'\beta + b_i)}{1+\exp(x_{ij}'\beta + b_i)}$$

  and $b_i \overset{iid}{\sim} N(0, \tau^2)$, where τ is known. Then we have

$$\begin{aligned} p(y_{ij} \mid p_{ij}) &= p_{ij}^{y_{ij}}(1-p_{ij})^{1-y_{ij}} \\ &= \exp[y_{ij}\log p_{ij} + (1-y_{ij})\log(1-p_{ij})] \\ &= \exp\left[y_{ij}\log\left(\frac{p_{ij}}{1-p_{ij}}\right) + \log(1-p_{ij})\right] \quad \text{Form of the exponential family} \\ &= \exp\left\{y_{ij}(x_{ij}'\beta + b_i) - \log[1+\exp(x_{ij}'\beta + b_i)]\right\} = p(y_{ij} \mid \beta, b_i) \end{aligned}$$

  and $p(b_i) \propto (\tau^2)^{-1/2}\exp[b_i^2/(2\tau^2)]$. Hence, β can be estimated based on the marginal likelihood

$$L(\beta) = \prod_{i=1}^{n} \int p(y_{ij} \mid \beta, b_i)\ p(b_i)\ db_i.$$

13 Interpretation of fixed effects in Poisson GLMM and GEE

- Poisson model in GLMM: $\log[E(Y_{ij} \mid b_i)] = \log(\mu_{ij} \mid b_i) = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + b_i$, where x_{ij1} is categorical (0 = placebo, 1 = treatment) and x_{ij2} is continuous.
  - $\exp(\beta_0)$: Rate (risk) of response for a typical individual under placebo ($x_{ij1} = 0$) and when $x_{ij2} = 0$.
  - $\exp(\beta_1)$: Ratio of the response rate (Relative risk of the response) for a typical individual under treatment to under placebo, adjusting for x_{ij2}.
  - $\exp(\beta_2)$: Ratio of the response rate (Relative risk of the response) between two typical individuals whose x_{ij2} values differ by one unit in the same treatment group.
  - β_0: Log rate (risk) of the response for a typical individual under placebo and when $x_{ij2} = 0$.
  - β_1: Log of the ratio of the response rate (Relative risk of the response) for a typical individual under treatment compared to under placebo, adjusting for x_{ij2}.
  - β_2: Log of the ratio of the response rate (Relative risk of the response) between two typical individuals whose x_{ij2} values differ by one unit in the same treatment group.
- Poisson model in GEE: $\log[E(Y_{ij})] = \log(\mu_{ij}) = \gamma_0 + \gamma_1 x_{ij1} + \gamma_2 x_{ij2}$.

- $\exp(\gamma_0)$: Expected rate (risk) of response under placebo and when $x_{ij2} = 0$ over the population.
- $\exp(\gamma_1)$: Ratio of the expected response rate (Relative risk of the response) in the treatment group compared to the placebo group, adjusting for x_{ij2}.
- $\exp(\gamma_2)$: Ratio of the expected response rate (Relative risk of the response) over two populations of individuals who differ in x_{ij2} by one unit in the same treatment group.
- γ_0, γ_1, γ_2: Log of the ...

14 Interpretation of fixed effects in Logistic GLMM and GEE

- Logistic model in GLMM: $\text{logit}[E(Y_{ij} \mid b_i)] = \text{logit}(\mu_{ij} \mid b_i) = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + b_i$, where x_{ij1} is categorical (0 = placebo, 1 = treatment) and x_{ij2} is continuous.
- Logistic model in GEE: $\text{logit}[E(Y_{ij})] = \text{logit}(\mu_{ij}) = \gamma_0 + \gamma_1 x_{ij1} + \gamma_2 x_{ij2}$.
- Both interpretations can be obtained by changing rate/risk to **odds** (intercept term) and ratio of rate/relative risk to **odds ratio** (covariates) for the interpretation in the Poisson GLMM and GEE.

15 Comparison of repression coefficients for GLMM vs GEE

- Regression coefficients β in GLMM are interpreted conditioning on $b_i = 0$ (typical individual).
- **Gaussian** model (LMM): $Y_i \mid b_i \sim N_p(X_i\beta + Z_i b_i, \sigma^2_\epsilon I_p)$, where $\beta \in \mathbb{R}^q$, $b_i \overset{iid}{\sim} N_k(0, D)$
  - Conditional mean $E(Y_i \mid b_i) = X_i\beta + Z_i b_i$.
  - Marginal mean $E(Y_i) = E[E(Y_i \mid b_i)] = E(X_i\beta + Z_i b_i) = X_i\beta$.
  - $E(Y_i \mid b_i = 0) = X_i\beta = E(Y_i)$, i.e., the mean for a typical individual equals the marginal mean.
  - $E(Y_i) = X_i\gamma$ in GEE, meaning that $\gamma = \beta$ (same as regression coefficients in LMM).
- **Poisson** model: $Y_{ij} \mid \beta, b_i \sim \text{Poisson}(\mu_{ij})$ with $\log(\mu_{ij}) = x'_{ij}\beta + z'_{ij}b_i$ and $b_i \overset{iid}{\sim} N_k(0, D)$ with $D \succ O$.
  - $E(Y_{ij} \mid b_i) = \text{var}(Y_{ij} \mid b_i) = \exp(x'_{ij}\beta + z'_{ij}b_i)$.
  - $E(Y_{ij}) = E[E(Y_{ij} \mid b_i)] = E[\exp(x'_{ij}\beta + z'_{ij}b_i)] = \exp(x'_{ij}\beta + z'_{ij}Dz_{ij}/2)$.
  - $E(Y_{ij} \mid b_i = 0) = \exp(x'_{ij}\beta) < E(Y_{ij})$.
  - $E(Y_{ij}) = \exp(x'_{ij}\gamma)$ in **GEE**. So $x'_{ij}\gamma = x'_{ij}\beta + z'_{ij}Dz_{ij}/2 > x'_{ij}\beta \;\Rightarrow\; |\gamma| > |\beta|$ holds. That is, unlike a probit model, γ is **inflated** compared to the marginal β in GLMM.
- **Probit** model: $Pr(Y_{ij} = 1 \mid b_i) = Pr(Y^*_{ij} > 0 \mid b_i) = \Phi(x'_{ij}\beta + z_{ij}b_i + \epsilon_{ij})$ or

$$\begin{aligned} Y^*_{ij} \mid b_i &\sim N(x'_{ij}\beta + z'_{ij}b_i,\ \sigma^2) \\ Y_{ij} \mid Y^*_{ij} &= I\{Y^*_{ij} > 0\}, \end{aligned}$$

  where $\epsilon_{ij} \sim N(0, \sigma^2)$ and $b_i \overset{iid}{\sim} N_q(0, D(\alpha))$.
  - $E(Y_{ij} \mid b_i) = Pr(Y^*_{ij} > 0 \mid b_i) = Pr\left(\dfrac{Y^*_{ij} - x'_{ij}\beta - z_{ij}b_i}{\sigma} > \dfrac{-(x'_{ij}\beta + z_{ij}b_i)}{\sigma}\right) = \Phi\left(\dfrac{x'_{ij}\beta + z_{ij}b_i}{\sigma}\right)$
  - $E(Y_{ij}) = Pr(Y^*_{ij} > 0) = Pr\left(\dfrac{Y^*_{ij} - z'_{ij}\beta}{\sqrt{z'_{ij}Dz_{ij} + \sigma^2}} > \dfrac{-x'_{ij}\beta}{\sqrt{z'_{ij}Dz_{ij} + \sigma^2}}\right) = \Phi\left(\dfrac{x'_{ij}\beta}{\sqrt{z'_{ij}Dz_{ij} + \sigma^2}}\right)$
  - $E(Y_{ij} \mid b_i = 0) = \Phi\left(\dfrac{x'_{ij}\beta}{\sigma}\right) > E(Y_{ij})$, i.e., the mean for a typical individual is larger than the marginal mean.

– $E(Y_i) = \Phi\left(\frac{x'_{ij}\gamma}{\sigma}\right)$ in **GEE**, implying that γ in GEE is **attenuated** compared to the marginal β in the probit GLMM since

$$\frac{|\beta|}{\sqrt{z'_{ij}Dz_{ij}+\sigma^2}} = \frac{|\gamma|}{\sigma} \quad \Rightarrow \quad |\gamma| = \frac{\sigma}{\sqrt{z'_{ij}Dz_{ij}+\sigma^2}}|\beta| < |\beta|$$

- **Logit** model: $Y_{ij} \mid \beta, b_i \sim \text{Bernoulli}(p_{ij})$ with $\log[p_{ij}/(1-p_{ij})] = x'_{ij}\beta + z_{ij}b_i$ and $b_i \overset{iid}{\sim} N_q(0, D(\alpha))$.
  - $E(Y_{ij} \mid \mathbf{b}_i) = \dfrac{\exp(\mathbf{x}_{ij}\beta + \mathbf{z}_{ij}\mathbf{b}_i)}{1+\exp(\mathbf{x}_{ij}\beta + \mathbf{z}_{ij}\mathbf{b}_i)}$.
  - $E(Y_{ij}) = E[E(Y_{ij} \mid \mathbf{b}_i)] = E\left[\dfrac{\exp(\mathbf{x}_{ij}\beta + \mathbf{z}_{ij}\mathbf{b}_i)}{1+\exp(\mathbf{x}_{ij}\beta + \mathbf{z}_{ij}\mathbf{b}_i)}\right] \approx \dfrac{\exp(\mathbf{x}_{ij}\gamma)}{1+\exp(\mathbf{x}_{ij}\gamma)}$, $|\gamma| = \dfrac{|\beta|}{|c^2\mathbf{D}\mathbf{z}_{ij}\mathbf{z}'_{ij} + \mathbf{I}_q|^{q/2}}$.
  - $E(Y_{ij} \mid \mathbf{b}_i = \mathbf{0}) = \dfrac{\exp(\mathbf{x}_{ij}\beta)}{1+\exp(\mathbf{x}_{ij}\beta)} > E(\mathbf{Y})$ since $|\beta| > |\gamma|$.
  - If $q = 1$, $\mathbf{z}_{ij} = 1$ and $D = \sigma_0^2$, then $|c^2\mathbf{D}\mathbf{z}_{ij}\mathbf{z}'_{ij} + \mathbf{I}_q|^{q/2} = \sqrt{c\sigma_0^2 + 1}$.
  - $E(Y_{ij}) = \dfrac{\exp(\mathbf{x}_{ij}\gamma)}{1+\exp(\mathbf{x}_{ij}\gamma)}$ in GEE, which is **attenuated** compared to the marginal model in GLMM.

16 Marginal variance, and covariance in GLMM

- **Gaussian** model:

$$\begin{aligned}
&\text{var}(Y_i) = E_{b_i}[\text{var}(Y_i \mid b_i)] + \text{var}_{b_i}[E(Y_i \mid b_i)] = \sigma_\epsilon^2 I_p + Z_i D Z_i' \quad \text{or} \quad \text{var}(Y_{ij}) = \sigma_\epsilon^2 + z'_{ij}Dz_{ij},\\
&\text{cov}(Y_{ij}, Y_{ik}) = \text{cov}_{b_i}[E(Y_{ij} \mid b_i), E(Y_{ik} \mid b_i)] = \text{cov}_{b_i}(z'_{ij}b_i, z'_{ij}b_i) = z'_{ij}Dz_{ik},\\
&\text{cov}(Y_{ij}, Y_{i'j}) = 0, \ i \neq i'
\end{aligned}$$

- **Poisson** model:

$$\begin{aligned}
\text{var}(Y_{ij}) &= E_{b_i}[\text{var}(Y_{ij} \mid b_i)] + \text{var}_{b_i}[E(Y_{ij} \mid b_i)]\\
&= E_{b_i}[\exp(x'_{ij}\beta + z'_{ij}b_i)] + E[\exp(2x'_{ij}\beta + 2z'_{ij}b_i)] - E_{b_i}[\exp(x'_{ij}\beta + z'_{ij}b_i)]^2\\
&= \exp(x'_{ij}\beta + z'_{ij}Dz_{ij}/2) + \exp(2x'_{ij}\beta + 2z'_{ij}Dz_{ij}) - \exp(2x'_{ij}\beta + z'_{ij}Dz_{ij})\\
&= E(Y_{ij}) + E(Y_{ij})^2[\exp(z'_{ij}Dz_{ij}) - 1]\\
&= E(Y_{ij})[1 + E(Y_{ij})\kappa_D],\\
\text{cov}(Y_{ij}, Y_{ik}) &= \text{cov}_{b_i}[E(Y_{ij} \mid b_i), E(Y_{ik} \mid b_i)]\\
&= \text{cov}[E_{b_i}[\exp(x'_{ij}\beta + z'_{ij}b_i)], E_{b_i}[\exp(x'_{ik}\beta + z'_{ik}b_i)]]\\
&= E\{\exp[(x'_{ij} + x'_{ik})\beta + (z'_{ij} + z'_{ik})b_i]\} - E(Y_{ij})E(Y_{ik})\\
&= \exp[(x'_{ij} + x'_{ik})\beta + (z'_{ij} + z'_{ik})D(z_{ij} + z_{ik})/2] - E(Y_{ij})E(Y_{ik})\\
&= E(Y_{ij})E(Y_{ik})\exp[z'_{ik}Dz_{ij}] - E(Y_{ij})E(Y_{ik})\\
&= E(Y_{ij})E(Y_{ik})[\exp(z'_{ik}Dz_{ij}) - 1]\\
\text{cov}(Y_{ij}, Y_{i'j}) &= 0, \ i \neq i'
\end{aligned}$$

- **Logit** model. The marginal inference is possible, but one needs to do a little work: using the approximation shown in the last section, or calculate the required integrals using a **Monte Carlo** estimate. By the previous section, the *approximate* marginal mean with only random intercept is given by

$$\mu^*_{ij} = \frac{\exp(\mathbf{x}_{ij}\gamma)}{1+\exp(\mathbf{x}_{ij}\gamma)}$$

so that

$$\begin{aligned}
\text{var}(Y_{ij}) &= E_{b_i}[\text{var}(Y_{ij} \mid b_i)] + \text{var}_{b_i}[E(Y_{ij} \mid b_i)] \\
&= E_{b_i}[\mu_{ij}(1-\mu_{ij})] + E(\mu_{ij}^2) - E(\mu_{ij})^2 \\
&= E[\mu_{ij}(1-\mu_{ij})] \\
&= \mu_{ij}^*(1-\mu_{ij}^*) \\
\text{cov}(Y_{ij}, Y_{ik}) &= \text{cov}_{b_i}[E(Y_{ij} \mid b_i), E(Y_{ik} \mid b_i)] \\
&= E\left[\left(\frac{\exp(\mathbf{x}_{ij}\beta + b_i)}{1+\exp(\mathbf{x}_{ij}\beta + b_i)}\right)\left(\frac{\exp(\mathbf{x}_{ik}\beta + b_i)}{1+\exp(\mathbf{x}_{ik}\beta + b_i)}\right)\right] - \mu_{ij}^*\mu_{ik}^*.
\end{aligned}$$

We see that the marginal covariance is not constant and not of easily interpretable form.

17 Bayesian Inference for LMM

- Classical LMM: $Y_i = X_i\beta + Z_i b_i + \epsilon_i$, where $b_i \overset{iid}{\sim} N(0, D) \perp\!\!\!\perp \epsilon_i \sim N(0, \sigma_\epsilon^2 I_{m_i})$.
- We have a three-stage hierarchical prior
  1. Likelihood: $p(y_i \mid \beta, b_i, \sigma_\epsilon^2)$, $i = 1, 2, \ldots m_i$.
  2. Random effects prior: $p(b_i \mid D)$, $i = 1, 2, \ldots m_i$.
  3. Hyperprior: $\pi(\beta, D, \sigma_\epsilon^2) \overset{e.g.}{=} p(D)\ p(\sigma_\epsilon^2)\ p(\beta \mid \sigma_\epsilon^2)$ with
     - $\sigma_\epsilon^2 \sim \text{InvGamma}(a, b) \ \Leftrightarrow\ (\sigma_\epsilon^2)^{-1} \sim \text{Gamma}(a, b)$
     - $\beta \mid \sigma_\epsilon^2 \sim N(\beta_0, (\sigma^2/\delta)(X'X)^{-1})$
     - $D \sim IW_r(\nu, S)$

     Note that hyperpriors (hyperparameters) are sensitive.
- Inference is based on the posterior distribution

$$p(\beta,\ \mathbf{b},\ D,\ \sigma_\epsilon^2 \mid \mathbf{Y}).$$

- Simulation-based inference is often implemented via MCMC.
- An efficient Gibbs sequence simulates iteratively from the following distributions.
  - $p(\beta, \mathbf{b} \mid \mathbf{Y},\ D,\ \sigma_\epsilon^2) = p(\mathbf{b} \mid \mathbf{Y},\ \beta,\ D,\ \sigma_\epsilon^2)\ p(\beta \mid \mathbf{Y},\ D,\ \sigma_\epsilon^2)$
  - $p(\sigma_\epsilon^2 \mid \mathbf{Y},\ \beta,\ \mathbf{b})$
  - $p(D \mid \mathbf{Y},\ \beta,\ \mathbf{b},\ \sigma_\epsilon^2)$.

18 Bayesian Inference for GLMM

- Consider the sampling model of exponential families:

$$p(y_{ij}\theta_{ij}, \alpha) = \exp\left[\frac{y_{ij}\theta_{ij} - d(\theta_{ij})}{a(\alpha)} + c(y_{ij}, \alpha)\right],$$

  where usually $a(\alpha) = 1$.
- We have $E(y_{ij} \mid \theta_{ij}, \alpha) = d'(\theta_{ij}) = \mu_{ij}$ with $g(\mu_{ij}) = x'_{ij}\beta + z'_{ij}b_i$.
- Bayesian inference requires a joint prior: $\pi(\beta, \alpha, b_1, \ldots, b_n)$
- Inference is based on the posterior: $p(\beta, \alpha, b_1, \ldots, b_n \mid y_1, \ldots, y_n)$.

- Hierarchical models are used for Bayesian inference. If we assume $\theta_1, \ldots, \theta_n$ are exchangeable, within a hypothetical infinite sequence, then

$$p(\theta_1, \ldots, \theta_n) = \int \prod_{j=1}^{n} p(\theta_j \mid \phi)\pi(\phi)d\phi$$

hence a two-stage hierarchical prior is (1) $\theta_j \mid \phi \overset{iid}{\sim} p(\cdot \mid \phi)$ and (2) $\phi \sim \pi(\cdot)$.

19 Predictive Distribution

- Compared with the marginal likelihood, $p(y) = \int p(y \mid \theta)p(\theta)d\theta$, the **predictive distribution** is a likelihood of future data averaged over all parameters values supported by the posterior distribution $p(\theta \mid y)$, i.e.,

$$p(y' \mid y) = \int p(y' \mid \theta,\ y)p(\theta \mid y)\ d\theta$$

- In the context of hierarchical models, we can assume $p(y' \mid \theta,\ y) = p(y' \mid \theta)$.

$$p(y' \mid y) = \int p(y' \mid \theta)\ p(\theta \mid y)d\theta$$

- A useful way to calculate $p(y' \mid y)$ is to consider the joint posterior: $p(y', \theta \mid y)$.
- Monte Carlo inference for $p(y' \mid y)$ can then target the distribution above, by iteration through the following steps: (1) $\theta \sim p(\theta \mid y)$ and (2) $y' \sim p(y' \mid \theta)$.
- **Conditional Predictive Ordinates (CPO)** evaluates the likelihood of y_i given all other observations $y_{(i)}$, by averaging over all parameters supported by $p(\theta \mid y_{(i)})$, i.e.,

$$\text{CPO}_i = p(y_i \mid y_{(i)}) = \int p(y_i \mid \theta)\ p(\theta \mid y_{(i)})d\theta.$$

A Monte Carlo estimate of this can be obtained as follows.

$$\begin{aligned}[p(y_i \mid y_{(i)})]^{-1} &= \frac{p(y_{(i)})}{p(y)} = \int \frac{p(y_{(i)} \mid \theta)p(\theta)}{p(y)}d\theta \\ &= \int \frac{1}{p(y_i \mid \theta)}\frac{p(y \mid \theta)p(\theta)}{p(y)}d\theta = \int \frac{1}{p(y_i \mid \theta)}p(\theta \mid y)d\theta = E_{\theta|y}\left[\frac{1}{p(y_i \mid \theta)}\right]\end{aligned}$$

follows that after a Monte Carlo sample from $p(\theta \mid y)$ is obtained, we compute:

$$\widehat{\text{CPO}} = \left(\frac{1}{M}\sum_{i=1}^{M}\frac{1}{p(y_i \mid \theta^{(i)})}\right)^{-1}.$$

20 Difference LMM and GEE

- **GEE** has the fewest assumptions (only the mean/variance structures are specified) and is designed for population-level inference rather than individual-level.
  - GEE uses a "working" correlation matrix; "working" refers to the choice of a variance model that may not necessarily correspond to exactly the form we believe to be true but rather to be a choice that is statistically convenient.
  - Asymptotics are required for inference, so, GEE is less appealing when the number of individuals (m) is small.

- A sufficiently large sample size is required for both normality of the estimator and reliability of the sandwich variance estimator.
- The use of the sandwich variance estimator makes GEE the most dependable method in large sample situations. However, there can be losses in efficiency if we choose a working correlation matrix that is far from reality (not true).
- With GEE, it is not possible to make inference for individuals or incorporate prior information.

- **LMMs** are more flexible than GEE in terms of the questions that can be addressed with the data, but this flexibility leads to a greater number of assumptions. LMMs have two approaches: likelihood and Bayesian approaches.
  - For likelihood inference, as with GEE, we require the number of individuals m to be sufficiently large for asymptotic inference.
  - Prior information cannot be incorporated in a likelihood analysis.
  - For a small number of individuals. a Bayesian approach fully captures the uncertainty but inference is completely model-based.
  - With a small sample size, it is unlikely that we will be able to check the modeling assumptions.

21 PCA

- The method of principal components is used to find the linear combinations with large variance. The first PC is the normalized linear combination with maximum variance.
- In many exploratory studies, the number of variables under consideration is too large to handle. A way of reducing the number of variables of interest is to discard the linear combinations that have small variances and study only those with large variances.
- Background of PCA (HW4) Suppose $X : p$ follows a general multivariate T distribution. Show that the PCs have the same geometric interpretation as they would if X were normally distributed.

  Solution: Consider a multivariate normal distribution and T distribution with the same location and scale/dispersion parameters, μ and Σ. Let ν be the tail weight parameter (degrees of freedom) of the multivariate T distribution. Then the multivariate normal and T density are proportional to

$$|\Sigma|^{-1/2} \exp\left[-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right], \quad |\Sigma|^{-1/2}\left[1+\frac{1}{\nu}(x-\mu)'\Sigma^{-1}(x-\mu)\right]^{-(\nu+p)/2},$$

  respectively. In these cases, X is said to be elliptically distributed, i.e., the density contours are ellipses, as both densities of X are of the form: $g[(x-\mu)'\Sigma^{-1}(x-\mu)]$, where g is called the radial function.

  Further let Γ be an $p \times p$ orthogonal matrix s.t. $\Gamma'\Sigma\Gamma = \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p)$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. We have $\Sigma^{-1} = (\Gamma\Lambda\Gamma')^{-1} = \Gamma\Lambda^{-1}\Gamma'$ and the principal components can be defined as $Y = \Gamma'(X-\mu)$ so that we can write density contours as

$$(x-\mu)'\Sigma^{-1}(x-\mu) = (x-\mu)'\Gamma\Lambda^{-1}\Gamma'(x-\mu) = y'\Lambda^{-1}y = \sum_{i=1}^{p}\frac{y_i^2}{\lambda_i} := c^2,$$

  where c is constant. This means that the principal components have the same geometric interpretation between the two densities. Specifically, the direction cosine of the i-th principal axis is the eigenvector corresponding to λ_i, and the half-length of the i-th principal axis is $c\sqrt{\lambda_i}$ for both distributions.

- Population Principal Components: Let $X \in \mathbb{R}^p$ be a random vector with mean μ and variance Σ (may not be normal), and let Γ: $p \times p$ be an orthogonal matrix such that $\Gamma'\Sigma\Gamma = \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p)$ with $\lambda_1 \geq \ldots \geq \lambda_p$ and $\Gamma'\Gamma = \Gamma\Gamma' = I_p$. Then we can define p-principal components of X as

$$Y = \Gamma'(X-\mu),$$

which means linear combinations of X and the components of Γ are the standardized coefficients.

  - $E(Y) = 0$ and $\text{Cov}(Y) = \Gamma'\Sigma\Gamma = \Lambda$ (uncorrelated across principal components).
  - $\sum_i \text{Var}(Y_i) = \sum_i \lambda_i = \text{tr}(\Sigma)$ and $\prod_i \text{Var}(Y_i) = \prod_i \lambda_i = |\Sigma|$.

- Theorem: *No* standardized linear combination of X has variance larger than λ_1, which is the maximum eigenvalue, i.e., the variance of the first PC. If $\alpha_{k+1} = a'(X-\mu)$ such that $a'a = 1$ and $\text{cov}(\alpha_{k+1}, Y_i) = 0$ for $i = 1, \ldots, k < p$ and Y_i being the i-th PC of X, then the variance of α_{k+1} is maximized when $\alpha_{k+1} = Y_{k+1}$.

  Proof: We have $Y_i = \gamma_i'(X-\mu)$, where γ_i' is the i-th eigenvector of Σ s.t. $\gamma_i'\gamma_i = 1$ for $i = 1, \ldots, k$. Let λ_i is the i-th eigenvalues corresponding to γ_i. Then,

$$0 = \text{cov}(\alpha_{k+1}, Y_i) = \text{cov}[a'(X-\mu), \gamma_i'(X-\mu)] = a'\Sigma\gamma_i = \lambda_i a'\gamma_i \;\Rightarrow\; a'\gamma_i = 0$$

  so that we maximize

$$f(a, \eta_1, \ldots \eta_k) = a'\Sigma a - \eta_{k+1}(a'a - 1) - 2\sum_{i=1}^{k} \eta_i a'\gamma_i,$$

  where η_i's are the Lagrange multipliers. Solving $f'(a) = 2\Sigma a - 2\eta_{k+1}a - 2\sum_i \eta_i\gamma_i = 0$, we have

$$(\Sigma - \eta_{k+1}I_p)a = \sum_i \eta_i\gamma_i \quad \Rightarrow \quad a'(\Sigma - \eta_{k+1}I_p)a = \sum_i \eta_i a'\gamma_i = 0.$$

  Thus, η_{k+1} is $(k+1)$-th eigenvalue and hence $a = \gamma_{k+1}$, meaning that $\alpha_{k+1} = \gamma_{k+1}'(X-\mu) = Y_{k+1}$.

- Relationship between X and Y. WLOG. consider $E(X) = \mu = 0 \in \mathbb{R}^p$. Then

$$\text{Cov}(X, Y) = E(XY') = E(XX'\Gamma) = \Sigma\Gamma = \Gamma\Lambda\Gamma'\Gamma = \Gamma\Lambda$$

  and hence $\text{cov}(X_i, Y_j) = \gamma_{ij}\lambda_j$, where γ_{ij} is j-th element in $\gamma_i \in \mathbb{R}^p$. Hence, the correlation is given by

$$\text{corr}(X_i, Y_j) = \frac{\text{cov}(X_i, Y_j)}{\text{var}(X_i)^{1/2}\,\text{var}(Y_j)^{1/2}} = \frac{\gamma_{ij}\lambda_j}{(\sigma_{ii}\lambda_j)^{1/2}} = \gamma_{ij}\sqrt{\frac{\lambda_j}{\sigma_{ii}}} = \rho_{ij}.$$

  The squared correlation $\rho_{ij}^2 = \gamma_{ij}^2\lambda_j/\sigma_{ii}$ can be interpreted as the proportion of variation in X_i explained by j-th PC, Y_j. This notion extends to a set of D PC as $\rho_{iD}^2 = \sum_{j\in D}\rho_{ij}^2$.

- By the above, the squared sample correlations are simply defined as $r_{ij} = g_{ij}^2\ell_j/S_{ii}$.

- *Multivariate* version: Suppose $X : n \times p$ with iid rows, instead of $p \times 1$, then PCs are given by

$$Y = (X - \mu)\Gamma$$

  with $\text{Cov}(Y) = I_n \otimes \Gamma'\Sigma\Gamma = I_n \otimes \Lambda$.

- **Sample Principal Components**: Let $X : n \times p$. The sample variance is defined as

$$S = \frac{1}{n}X'HX = \frac{1}{n}X'\left(I_n - \frac{1}{n}1_n 1_n'\right)X.$$

  After spectral decomposition of $S = GLG'$, the sample PC of X may be defined as

$$\underbrace{Y}_{n\times p} = \underbrace{(X - 1_n\overline{X}')}_{n\times p}\underbrace{G}_{p\times p}.$$

  Then the column covariance of Y is given by

$$S_Y = G'\,\text{Cov}(X - 1_n\overline{X}')G = \frac{1}{n}G'X'HXG = G'SG = L \;\Rightarrow\; \text{Cov}(Y) = I_n \otimes L.$$

- PCs are **not scale invariant**, which motivates standardizing all the p variables before applying PCA.
- How many PCs? The simplest rule is to retain the first k PCs explaining an $\alpha\%$ of the total variance:

$$100\alpha = \frac{\lambda_1 + \cdots + \lambda_k}{\lambda_1 + \cdots + \lambda_p}$$

  or Kaiser rule retains all PCs with $\lambda_j > 1$.
- **Singular Value Decomposition** of the centered data X is more numerically stable than spectral decomposition and is usually used in practice:
  - SVD: For *any* matrix $X : n \times p$, there exist a diagonal matrix D and two diagonal matrices U and V such that $X = UDV'$. Then $X' = VDU'$, $XX' = UD^2U'$, and $X'X = VDV'$.
  - PCs are simply defined as $Y = XV'$.
  - For highly collinear X, this procedure is sometimes used to perform the PCV regression.

22 Canonical Correlation Analysis

- Consider two sets of variables $X \in \mathbb{R}^r$ and $Y \in \mathbb{R}^s$ such that

$$E\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \quad \text{Cov}\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix}$$

- Canonical Correlation analysis aims to replace the two sets of correlated variables (X and Y) with $t < \min(r, s)$ pairs of linear projections

$$\xi_j = g_j'X, \quad \omega_j = h_j'Y, \quad j = 1, \ldots, t.$$

- We want to seek linear projections, ξ and ω such that $\text{cov}(\xi_j, \xi_k) = 0$, $\text{cov}(\omega_j, \omega_k) = 0$, and

$$\text{corr}(\xi_j, \omega_j) = \frac{g_j'\Sigma_{XY}h_j}{(g_j'\Sigma_{XX}g_j)^{1/2}(h_j'\Sigma_{YY}h_j)^{1/2}} = \rho_j$$

  with $\rho_1 \geq \cdots \geq \rho_t$.

23 Factor analysis

- Factor analysis is based on a model in which the observed vector is partitioned into an unobserved systematic part and an unobserved error part.
- The components of the error vector are considered as uncorrelated or independent, while the systematic part is taken as a linear combination of a relatively small number of unobserved factor variables.
- Let X be a random observable with $E(X) = \mu$ and $\text{Cov}(X) = \Sigma$.
- The k-factor models are written as

$$\underbrace{X}_{k\times 1} = \underbrace{\mu}_{k\times 1} + \underbrace{A}_{p\times k}\underbrace{f}_{k\times 1} + \underbrace{\epsilon}_{k\times 1}, \quad f \perp \epsilon,$$

  - μ would be the average of X, $\overline{X}$.
  - The components of f (latent factors) are linear combinations of the components of X.
  - The elements of A (**factor loadings**) are the coefficients of f.

- Assumptions: $E(\epsilon) = 0$ and $\text{Cov}(\epsilon) = E(\epsilon\epsilon') = \Psi = (\phi_1, \ldots, \phi_p)$, $E(f) = 0$ and $\text{Cov}(f) = I_k$, and $\text{Cov}(f, \epsilon) = 0$. Then $\text{Cov}(X) = AA' + \Psi = \Sigma$.
- The model is **invariant** under changes in the units of measurement of X: Let C be orthogonal with i-th diagonal element c_i $(i = 1, \ldots, p)$ and $Y = CX$, then

$$Y = C\mu + CAf + C\epsilon = \mu^* + A^* f + \epsilon^*$$

with $\text{Cov}(Y) = CAA'C + C\Psi C = A^*(A^*)' + \Psi^*$.

- There is the identifiability problem. Let G be a $k \times k$ orthogonal matrix, then

$$X = \mu + (AG)(G'f) + \epsilon = \mu + A^* f^* + \epsilon$$

with $E(f^*) = 0$ and $\text{Cov}(f^*) = I_k$, and $\text{Cov}(f^*, \epsilon) = 0$. That is, the model expression is *not* unique.

- **Identification** and **Dimension reduction**:
  - Assume $X \sim MN(1_n\mu', I_n, \Sigma = AA' + \Psi)$ with $A'\Psi^{-1}A = \Gamma = (\gamma_1, \ldots, \gamma_k)$ (diagonal), which is an identifiability constraint.
  - The number of free elements in Σ is $p(p+1)/2$, the number of parameters in A and Ψ is $pk + p$, and the constraint $A'\Psi^{-1}A = \Gamma = (\gamma_1, \ldots, \gamma_k)$ reduces by $k(k+1)/2 - k = k(k-1)/2$ parameters.
  - Hence, the difference in parameters between the saturated model and the k-factor model is

$$s_k = p(p+1)/2 - (pk + p - k(k-1)/2) = \frac{(p-k)^2 - (p+k)}{2},$$

    which can be positive and negative, but we expect $s_k > 0$ as $p >> k$.

- **LRT test statistic** and null distribution under H : reduced (k-factor) model:

$$\Lambda = \frac{L(\widehat{A}, \widehat{\Psi})}{L(\widehat{\Sigma})} \quad , -2\log\Lambda \xrightarrow{D} \chi^2_r, \quad r = s_k.$$

  Note that $\widehat{\mu} = \overline{X}$ is the same so that this estimator is canceled out.
  - Rather than maximizing $\ell(AA' + \Psi)$, minimizing

$$\mathcal{F}(A, \Psi) = \text{tr}(AA' + \Psi) = \log|(AA' + \Psi)| - \log|S| - p$$

    is convenient, where $S = \widehat{\Sigma} = X'QX \sim IW_p(n - p, \Sigma)$.
  - For fixed Ψ, obtain $\widehat{A}(\Psi)$ analytically, but $\widehat{\Psi}$ must be obtained numerically.

- **Principal components Factor analysis**
  - Since $\Sigma - \Psi = AA' \succeq O$ with $\text{rank}(AA') = k$, we have $M'(\Sigma - \Psi)M = \text{diag}(d_1, \ldots, d_k, 0, \ldots, 0) = D$ by spectral decomposition, so that we can write

$$\Sigma - \Psi = \underbrace{M}_{p\times p}\underbrace{D}_{p\times p}M' = \underbrace{M_1}_{p\times k}\underbrace{D_1}_{k\times k}M_1' = (M_1 D_1^{1/2})(M_1 D_1^{1/2})'$$

  - Given a suitable estimator $\widehat{\Psi}$, we can take the first k standardized PCs of $S - \widehat{\Psi}$, say $\widehat{M_1}$, and define $\widehat{A} = \widehat{M_1}\widehat{D}_1^{1/2}$. The well-known $\widehat{\Psi}$ is $\widehat{\psi}_j = 1/[S^{-1}]_{jj}$.

- **Estimation of Factor Scores** F: Joint MLEs of (A, Ψ, F) do not exist, so one often operates under the assumption that A and Ψ are known.
  - GLS estimate: $\widehat{F} = (A'\Psi^{-1}A)^{-1}A'\Psi^{-1}X$, which is unbiased but less efficient than the next.
  - Ridge estimate: $\widetilde{F} = (I + A'\Psi^{-1}A)^{-1}A'\Psi^{-1}X$, which is biased but has lower variance.

24 Cluster analysis

- Cluster analysis aims to estimate partitions of the sample space $\mathcal{X}$ identifying homogeneous classes of obs, which is often used for dimension reduction.
- Consider N p-dimensional observations $(X_1, \ldots, X_N)$. We assume that each observation is sampled from one of K subgroups $C_1, \ldots, C_k$, i.e., if X_i is sampled from C_j, then we write $x_i \sim f_j(x_i)$.
- In contrast with discriminant analysis, clusters or class levels $(\gamma_1, \ldots, \gamma_N) \in \{1, \ldots, K\}$ are *not observed*.
- Clustering tasks.
  - Clustering Observations. Cluster samples into K homogeneous classes.
  - Cluster Variables. We may partition p variables into K distinct groups. It is often considered as a special case of dimension reduction.
- 3 ideas: Model-based Clustering, Partition Methods (K-means clustering), Hierarchical clustering.
- We would use two sums of squares as performance/accuracy metrics

$$\text{Within cluster variation: } W(C_K) = \sum_{j=1}^{K} \sum_{i \in C_j} (x_i - \overline{x}_j)(x_i - \overline{x}_j)'$$

$$\text{Between cluster variation: } B(C_K) = \sum_{j=1}^{K} n_j (\overline{x}_j - \overline{x})(\overline{x}_j - \overline{x})'$$

  and then Calinski and Harabasz's (CH) variance (a statistic) is defined as

$$CH(C_K) = \frac{\text{tr}(B(C_k))/(K-1)}{\text{tr}(W(C_k))/(n-K)}.$$

- Cluster validation generally refers to exploring the quality of a clustering. This task is quite difficult when compared with regression or classification. In model-based settings, significant testing is possible.

25 Model-based Clustering

- Assume $x_i \sim f_j(x_i \mid \theta_j)$ with probability π_j $(j = 1, \ldots, K)$ s.t. $\sum_j \pi_j = 1$.
- A common sampling model assumes the **finite mixture**

$$p(x_i \mid \pi_{1:K}, \theta_{1:K}) = \sum_{j=1}^{K} \pi_j f_j(x_i \mid \theta_j).$$

- The conditional probability of a sample i belonging to cluster j given the observed feature vector x_i is

$$p(x_i \in C_j \mid x_i) = \frac{p(C_j) p(x_i \mid x_i \in C_j)}{p(x_i)} = \frac{\pi_j f_j(x_i \mid \theta_j)}{\sum_{g=1}^{K} pi_g f_g(x_i \mid \theta_g)}$$

  Inference for the proportions $(\pi_1, \ldots, \pi_K)$ and model parameters $(\theta_1, \ldots, \theta_K)$ is often based on MLE.

$$L(\pi_{1:K}, \theta_{1:K} \mid x_{1:N}) = \prod_{i=1}^{N} \sum_{j=1}^{K} \pi_j f_j(x_i \mid \theta_j),$$

$$\ell(\pi_{1:K}, \theta_{1:K} \mid x_{1:N}) = \sum_{i=1}^{N} \log \left[\sum_{j=1}^{K} \pi_j f_j(x_i \mid \theta_j) \right],$$

  which is not easy to solve. Hence, a data augmentation strategy and EM estimation are often used.

26 K-means clustering

- Consider clustering as a partition $C = \{C_j\}_{j=1}^K$ of $\mathbb{R}^p$.
- Define dissimilarity function between data x_i and **centroids** m_j, e.g., $d(x_i, m_j) = \|x_i - m_j\|^2$.
- The K-centroids square criterion is defined as summary distance:

$$W(C, m) = \sum_{j=1}^{K} \sum_{i \in C_j} \|x_i - m_j\|^2.$$

- We wish to find optimal partitions C and centers m that minimize $W(C, m)$, but the direct optimization of $W(C, m)$ is computationally challenging. Instead, iterative computation is often used.

  0) Set an initial partition C

  1) Given a C, K-means clustering identifies the cluster centroid as the cluster means

$$m_j = \frac{1}{n_j} \sum_{i \in C_j} x_i, \quad j = 1, \ldots, K.$$

  2) Given centroids m, the partition C is updated in relation to the minimum distance allocation rule, i.e., assign x_i to the cluster with nearest center m_j s.t. $C_j = \{i : \ d(x_i, m_j) = \min_\ell d(x_i, m_\ell)\}$.

  3) Re-compute centroids and relocate samples till no re-allocation is performed.

- Let M: $K \times p$ be a matrix of K cluster centroids and $Z : n \times K$ be a cluster membership matrix with $z_{ij} = I(x_i \in C_j)$. Consider the model: $X = ZM + E$, then the K-means square criterion minimizing $W(C, m)$ is equivalent to minimizing SSE $= \|X - ZM\|^2 = \|E\|^2$.
- We can decompose $\|X\|^2 = B(C, m) + \|E\|^2$. This decomposition allows us to define the criterion as

$$R^2 = \frac{B(C, m)}{\|X\|^2} = 1 - \frac{W(C, m)}{\|X\|^2}$$

- Assume $x_i \sim N_p(\mu_j, \Sigma_j)$. Maximizing the log-likelihood function for $C_{1:K}, \mu_{1:K}, \Sigma_{1:K}$:

$$\ell(C_{1:K}, \mu_{1:K}, \Sigma_{1:K} \mid x_{1:N}) = \text{const.} - \sum_{j=1}^{K} \frac{n_j}{2} \log |\Sigma_j| - \sum_{j=1}^{K} \sum_{i \in C_j} \frac{1}{2} (x_i - \mu_j)' \Sigma_j^{-1} (x_i - \mu_j)$$

  is equivalent to minimizing $W(C, m)$.
- The number of clusters K can be determined by detecting elbow points of percent variance.
- Since K-means works with Euclidean distances, results are highly dependent on measurement scales.
- Let w_K be the minimum values of $W(C, m)$ then $W_{K+1} \leq w_K$, i.e., the more cluster, the less distance.

27 Hierarchical Clustering

- One potential drawback of K-means is that it requires us to specify the number of clusters in advance.
- Hierarchical clustering operates on the premise of **bottom-up** grouping.
- Given a distance metric between observations $d(x_j, x_k)$
  - Start clumping together the closest observations
  - Keep clumping together groups of observations till everything is one group.

- How do we merge groups of observations?
  - Complete Linkage: maximal inter-cluster distance.
  - Single Linkage: minimal inter-cluster distance.
  - Average Linkage: mean inter-cluster distance.

28 Graphical Model

- Graphical model is a graph to express conditional independence among random variables:
  - Undirected Graph (UG) = Markov Network (MN)
  - Directed Acyclic Graph (DAG) = Bayesian Network (BN)
- Consider three nodes X, Y, Z.
  - (a) $P(Y)P(X|Y)P(Z|Y)$; (b) $P(X)P(Y|X)P(Z|Y)$; (c) $P(Z)P(Y|Z)P(X|Y)$ are equivalent and can be expressed in both DAG and UG. In contrast,
  - (d) $P(X)P(Y|X)P(Y|Z)$ is not equivalent to the above, as it has a collider at Y in DAG, which cannot be represented in UG (MN).

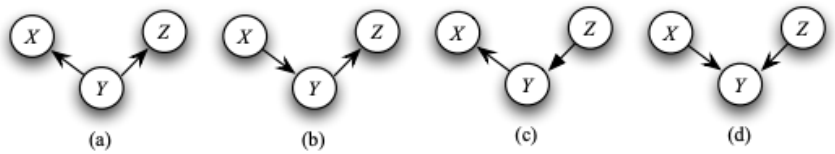


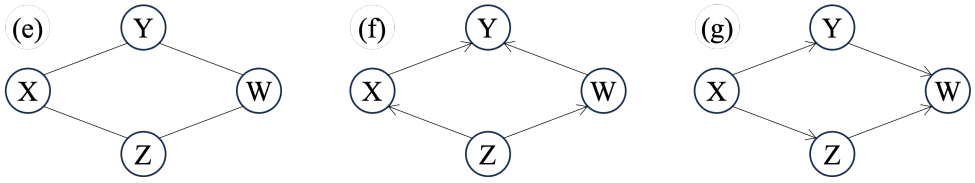
Figure 18.12. *The first three DAGs have no colliders. The fourth DAG has a collider at Y.*

The Rules of d-Separation

Consider the DAGs in Figures 18.12 and 18.4.

1. When Y is not a collider, X and Z are *d-connected*, but they are *d-separated* given Y.
2. If X and Z collide at Y, then X and Z are *d-separated*, but they are *d-connected* given Y.
3. Conditioning on the descendant of a collider has the same effect as conditioning on the collider. Thus in Figure 18.4, X and Z are *d-separated* but they are *d-connected* given W.

- Consider the following four-node graph: (e) UG and (f)(g) DAG with a collider.
  - (e) expresses both $Y \perp\!\!\!\perp Z \mid (X, W)$ and $X \perp\!\!\!\perp W \mid (Y, Z)$. In contrast,
  - (f) or (g) expresses only $Y \perp\!\!\!\perp Z \mid (X, W)$ or $X \perp\!\!\!\perp W \mid (Y, Z)$ due to a collider.



29 Copula

- A copula is a multivariate cumulative distribution function for which the marginal probability distribution of each variable is uniform on the interval $[0,1]$, i.e.,

$$C : [0,1]^p \longrightarrow [0,1]$$

- Copulas are used to describe or model the dependence (inter-correlation) between random variables.
- Consider $X \in \mathbb{R}^p$ with joint cdf $F(x) = Pr(X \leq x)$, which is right continuous and $U = F(X) \sim U(0,1)$.
- We aim to decompose F into univariate margins $F_1, \ldots, F_p$ and a Copula C.
- Define the generalizes inverse CDF or **Generalized Quantile Function**:

$$F^-(u) = \inf\{x \in R \mid F(x) \geq u\}, \ 0 < u < 1.$$

with (1) $F[F^-(u)] \geq u$, (2) $F(x) \geq u \ \Leftrightarrow \ x \geq F^-(u)$. (3) If $U \sim U(0,1)$, then $X = F^-(U)$ has cdf F.

- **Definition**. A p-variate copula $C : [0,1]^p \to [0,1]$ is the cdf of a random vector $(U_1, \ldots, U_p) = (F(X_1), \ldots, F_p(X_p))$ with $U(0,1)$ margins

$$C(u_1, \ldots, u_p) = Pr(U_1 \leq u_1, \ldots, U_p \leq u_p),$$

where $Pr(U_j \leq u_j) = F(u_j) = u_j$ as $U_j \sim U(0,1)$ for $j = 1, \ldots, p$.

- **Sklar's Theorem**. Let C be a p-variate copula and let $F_1, \ldots, F_p$ be univariate CDFs. The function

$$F(x_1, \ldots, x_p) = C(F_1(x_1), \ldots, F_p(x_p)) = Pr(U_1 \leq F_1(x_1), \ldots, U_p \leq F_1(x_p))$$

is a p-variate CDF with margins $F_1, \ldots, F_p$. Conversely, if F is a p-variate CDF, then there exists a copula C such that the above equation holds.

If the margins are continuous, then C is unique and is equal to

$$C(u_1, \ldots, u_p) = C(F_1^-(u_1), \ldots, F_p^-(u_p))$$

- Product (independence) copula: If X and Y have continuous CDFs, $X \perp\!\!\!\perp Y$ iff $C(u,v) = uv$.
- **Gaussian Copula**: Let Σ be a correlation matrix. $C = C(u_1, \ldots, u_p)$ is a Gaussian Copula iff

$$C(u_1, \ldots u_p \mid \Sigma) = \Phi_p(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_p)),$$

where Φ_p is the CDF of a $N(0, \Sigma)$ random variable, and Φ^{-1} are univariate normal quantile functions.

By Sklar's theorem, given arbitrary margins $F_1 \ldots, F_p$, the joint distribution

$$F(x_1, \ldots, x_p) = C(F_1(x_1), \ldots, F_p(x_p)) = \Phi_p(\Phi^{-1}(F_1(x_1)), \ldots, \Phi^{-1}(F_p(x_p)))$$

is a multivariate distribution with margins F_j $(j = 1, \ldots, p)$ connected by the Gaussian copula $C(\cdot \mid \Sigma)$.

- To generate a random vector with any continuous margins and dependence by a Gaussian Copula C:

$$\begin{array}{rcl} (z_1, \ldots, z_p) & \sim & N_p(0, \Sigma) \\ (u_1, \ldots, u_p) & = & [\Phi(z_1), \ldots, \Phi(z_p)] \\ (x_1, \ldots, x_p) & = & [F_1^-(u_1), \ldots, F_p^-(u_p)] \end{array}$$

- Gaussian Copula estimation (Parameterized Margins). If margins $F_1, \ldots, F_p$ are parameterized, s.t. $F_j = F_j(x_j \mid \theta_j)$, we estimate Σ as follows:

  1) Obtain $\widehat{F}(x_{ij}) = F_j(x_j \mid \widehat{\theta}_j)$

2) Define $z_{ij} = \Phi^{-1}[\widehat{F}_j(x_{ij})]$ for $j = 1, \ldots, p$, $i = 1, \ldots, n$.

3) $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p} z_{ij}^2 = \frac{1}{n}\sum_{i=1}^{n} Z_i Z_i'$, where $Z_i = (z_{i1}, \ldots, z_{ip})'$.

If $F_1, \ldots, F_p$ are fully known, step (1) is not required and change $\widehat{F}(x_{ij})$ to $F(x_{ij})$.

- There are a limited number of parametric multivariate *discrete* distributions with specific margins and dependence structures. We can get a copula C even if some univariate CDFs $(F_j(x_j))$ are discrete, but that copula C is not unique.

30 Exercise

- (Midterm) Consider $Y \sim MN(XB, I_n, \Sigma)$ with $\text{rank}(X) = q$ (full) and $\Sigma \succ O \in \mathbb{R}^{p\times p}$. We consider the hypothesis of mutual independence in a b-partition of Σ, s.t.

$$H_0 : \Sigma = \begin{bmatrix} \Sigma_{11} & O & \cdots & O \\ O & \Sigma_{22} & & \\ \vdots & & \ddots & \\ O & & & \Sigma_{bb} \end{bmatrix}; \quad H_1 : \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \cdots & \Sigma_{1b} \\ \Sigma_{21} & \Sigma_{22} & & \\ \vdots & & \ddots & \\ \Sigma_{b1} & & & \Sigma_{bb} \end{bmatrix},$$

where we may assume $\Sigma_{jj} : k \times k$ with $bk = p$. Derive a suitable union/intersection test procedure and describe the form of the rejection region for your test. Explain how an empirical null distribution for your test may be obtained using Monte Carlo methods.

Solution: Let $H_{0\gamma_{ij}} : \Sigma_{ij} = O$ $(1 \le i < j \le b)$, more specifically let

$$H_0 : \Sigma_{(i,j)} = \begin{bmatrix} \Sigma_{ii} & O \\ O & \Sigma_{jj} \end{bmatrix}, \quad H_1 : \Sigma_{(i,j)} = \begin{bmatrix} \Sigma_{ii} & \Sigma_{ij} \\ \Sigma_{ji} & \Sigma_{jj} \end{bmatrix}$$

then we can rewrite the hypotheses as the *intersection* of a single null hypothesis $H_{0\gamma_{ij}}$:

$$H_0 : \bigcap_{1\le i<j\le b} H_{0\gamma_{ij}}, \quad H_1 : H_0^c.$$

Each null hypothesis $H_{0\gamma_{ij}}$, we have a LRT statistic

$$\lambda(\gamma_{ij}) = \left(\frac{|\Sigma_{ii}||\Sigma_{jj}|}{|\Sigma_{(i,j)}|}\right)^{-n/2}.$$

We reject $H_{0\gamma_{ij}}$ if $-2\log\lambda(\gamma_{ij}) > c_{\gamma_{ij}}$ or $\lambda(\gamma_{ij}) < d_{\gamma_{ij}}$ for a suitable $c_{\gamma_{ij}}$ so that the rejection region for H_0 takes the form

$$R = \bigcup_{1\le i<j\le b} \left\{Y : -2\log\lambda(\gamma_{ij}) > c_{\gamma_{ij}}\right\} = \{Y : \max(-2\log\lambda(\gamma_{ij})) > c_0\} = \{Y : \min(\lambda(\gamma_{ij})) < d_0\}$$

For the empirical Null Distribution: You can use Monte Carlo methods:

1. Simulate a large number (say, B) of datasets under $H_{0\gamma_{ij}}$.
2. For each simulated dataset, calculate $\lambda(\gamma_{ij})$.
3. You use these B test statistics to estimate the null distribution of the LR test statistic.
4. Compare your observed LRT statistic to the empirical null distribution to determine the p-value.
5. Iterate (1) - (4) for all $1 \le i < j \le b$.

- A Gauss–Markov Theorem for Dependent Data: Suppose

$$E(Y) = X\beta, \quad \text{Var}(Y) = V$$

with $Y = [Y_1', \ldots, Y_m']' : N \times 1$, where $N = \sum_{i=1}^m n_i$, and where $Y_i = [Y_{i1}, ..., Y_{1n_i}]' : n_i \times 1$ and $x = [x_1, ..., x_m]'$ is $N \times p$ with $X_i = [X_{i1}, \ldots, X_{in_i}]$ and $X_{ij} = [1, x_{ij1}, \ldots, x_{ij(p-1)}]$, and β is the $k \times 1$ vector of regression coefficients. Consider linear estimators of the form

$$\widehat{\beta}_W = (X'W^{-1}X)^{-1}X'W^{-1}y, \quad W = W' \succ O.$$

with $E(\widehat{\beta}_W) = \beta$ (unbiased for all W, including $W = V$). Show that $\text{Var}(\widehat{\beta}_W) \succ \text{Var}(\widehat{\beta}_V)$.

Solution: Suppose $\widehat{\beta}_V = Ay$ and $\widehat{\beta}_W = By$, where

$$A = (X'V^{-1}X)^{-1}X'V^{-1}, \quad B = (X'W^{-1}X)^{-1}X'W^{-1}.$$

Then $E(Ay) = E(By) = \beta$ leads to $AX = BX = I$, and $\text{Var}(\widehat{\beta}_V) = AVA'$ and $\text{Var}(\widehat{\beta}_W) = BVB'$.

Show $BVB' - AVA' \succ O$:

$$\begin{aligned} BVB' &= (B - A + A)V(B - A + A)' \\ &= (B - A)V(B - A)' + AV(B - A)' + (B - A)VA' + AVA' \\ &= (B - A)V(B - A)' + AVA' \succ AVA' \end{aligned}$$

since $(B - A)VA' = (B - A)X(X'V^{-1}X)^{-1} = O$ and $AV(B - A)' = O$.

- (2020 Q3 Qual) Suppose we have a posterior for (B, Σ) from a likelihood $p(Y|B, \Sigma)$ and a prior $p(B, \Sigma) = MN \times IW(C, V, \nu, S)$. Provide an algorithm for sampling (B, Σ) from $p(B, \Sigma|Y)$ using *only* random number generators that draw from Inverse-Wishart and Multivariate Normal.

Solution 1) Set fixed parameters (C, V, ν, S); 2) Sample a initial $(B^{(0)}, \Sigma^{(0)})$; For $m = 0, \ldots, M$, 3) Generate Y from a $p(Y|B^{(m)}, \Sigma^{(m)})$; 4) Sample $(B^{(m+1)}, \Sigma^{(m+1)})$ from $p(B, \Sigma|Y)$. Repeat (3) and (4) until converge.

- Gamma-Poisson mixture. Suppose

$$\begin{aligned} y \mid \lambda &\sim \text{Poiss}(\lambda), \\ \lambda \mid a, b &\sim \text{Ga}(a, b). \end{aligned}$$

Then y given a and b has a (generalized) negative binomial distribution:

$$\begin{aligned} p(y \mid a, b) &= \int_\lambda p(y \mid \lambda)\ p(\lambda \mid a, b) d\lambda \\ &= \frac{b^a}{y!\Gamma(a)} \frac{\Gamma(y + a)}{(b + 1)^{y+a}} \int \text{Ga}(\lambda \mid y + a, b + 1) d\lambda \\ &= \frac{\Gamma(y + a)}{y!\Gamma(a)} \left(\frac{1}{b + 1}\right)^y \left(1 - \frac{1}{b + 1}\right)^a. \end{aligned}$$

c.f) if a count Y has a negative binomial with the number of success r with a probability p, then

$$\Pr(Y = k \mid r, p) = \binom{r + k - 1}{k} p^k (1 - p)^r$$