Final exam: Tuesday April 19 7–9 pm
Coverage: Ch 1–6 of text + classification (mainly LDA, QDA)
Allowed 2 8.5 × 11 sheet (both sides) & a calculator

Office hour: Next week Wed 10am – noon
Thurs 1–3 pm
Mon April 18 1–3 pm

Structured Multivariate Models

Last time: Repeated Measures

$$\underset{\sim}{X}_i = \begin{pmatrix} X_{i1} \\ \vdots \\ X_{ik} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix} + A_i \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + \begin{pmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{ik} \end{pmatrix}$$

$\mathrm{Var}(A_i) = \sigma_A^2$; $\mathrm{Var}(\varepsilon_{ij}) = \sigma_\varepsilon^2$

Longitudinal data

n subjects measured at time $t_1, \cdots, t_k$

$$X_{ij} = X_i(t_j) \Rightarrow \underset{\sim}{X}_i = \begin{pmatrix} X_i(t_1) \\ \vdots \\ X_i(t_k) \end{pmatrix}$$

i: subject

Examples

① Growth curves: normal v.s. abnormal growth
② Biological measurements: cholesterol level over time

Question: How do longitudinal data differ from, say, repeated measure data?

For each subject, we would expect that

$$\mathrm{Var}(X_i(t_j) - X_i(t_k)) = \mathrm{Var}(X_i(t_j)) + \mathrm{Var}(X_i(t_k)) - 2\mathrm{Cov}(X_i(t_j), X_i(t_k))$$

increases as $|t_j - t_k|$ increases

Equivalently, $\mathrm{Cov}(X_i(t_j), X_i(t_k))$ decreases as $|t_j - t_k|$ increases

– need to incorporate this into modeling.

Modeling: Starting pt.

$$X_i(t_j) = \underset{\text{random}}{A_i} + \underset{\text{fixed}}{\mu(t_j)} + \underset{\text{noise}}{W_i(t_j)}$$

or $X_i(t_j) = A_i + \mu(t_j) + W_i(t_j) + \varepsilon_{ij}$ ← measurement error

Assumptions:

① $A_1, \cdots, A_n$ are independent $N(0, \sigma_A^2)$
② $W_1, \cdots, W_n$ are independent stochastic process with $E[W_i(t)] = 0$
and $\mathrm{Cov}(W_i(t_j), W_i(t_k)) = \gamma(|t_j - t_k|)$

$\gamma(s) \to 0$ as $s \to 0$

γ: autocovariance function

③ $\{\varepsilon_{ij}\}$ are independent $N(0, \sigma_\varepsilon^2)$
④ All random components are independent

$$\underset{\sim}{X}_i = \begin{pmatrix} X_i(t_1) \\ \vdots \\ X_i(t_k) \end{pmatrix} \sim N_k(\underset{\sim}{\mu}, C)$$ where elements of $\underset{\sim}{\mu}, C$ are given by:

$$\underset{\sim}{\mu} = \begin{pmatrix} \mu(t_1) \\ \vdots \\ \mu(t_k) \end{pmatrix}$$

$$\mu(t_j) = E[X_i(t_j)] = E[(X_i(t_j) - \mu(t_j))^2]$$

$$\begin{aligned} Var(X_i(t_j)) &= Var(A_i + W_i(t_j) + \varepsilon_{ij}) \\ &= Var(A_i) + Var(W_i(t_j)) + Var(\varepsilon_{ij}) \\ &= \sigma_A^2 + \gamma(0) + \sigma_\varepsilon^2 \end{aligned}$$

$$\boxed{X_i(t_j) = A_i + \mu(t_j) + W_i(t_j) + \varepsilon_{ij}}$$

$$\begin{aligned} Cov(X_i(t_j), X_i(t_k)) &= E[(A_i + W_i(t_j) + \varepsilon_{ij})(A_i + W_i(t_k) + \varepsilon_{ik})] \\ &= \sigma_A^2 + \gamma(|t_j - t_k|) \end{aligned}$$

Need to specify $\underset{\sim}{\mu}$, $\gamma(s)$ more precisely

Simple estimate for $\underset{\sim}{\mu}$

$$\hat{\underset{\sim}{\mu}} = \begin{pmatrix} \hat{\mu}(t_1) \\ \vdots \\ \hat{\mu}(t_k) \end{pmatrix} = \frac{1}{n}\sum_{i=1}^{n} \underset{\sim}{X}_i \rightarrow \quad \hat{\mu}(t_j) = \frac{1}{n}\sum_{i=1}^{n} X_i(t_j)$$

We can also specify $\mu(t)$ parametrically

$$\mu(t) = \beta_0 + \sum_{i=1}^{p} \beta_i \phi_i(t)$$

β_0, β_i: unknown; ϕ_i: known

- $\phi_1, \cdots, \phi_p$ are some specified functions e.g. polynomials, spines, sinusoids.

Autocovariance function

- typically $\gamma(s) \rightarrow 0$ as $s \rightarrow 0$
- parametric family: $\gamma(s) = \gamma(0)\exp(-\alpha s^\lambda) \quad (\alpha > 0, \lambda > 0)$

$\gamma(0), \alpha, \lambda$: unknown

- special case: $\lambda = 1 \Rightarrow$ autoregressive autocovariance function

One approach: Throw everything into likelihood function and maximize $\Rightarrow$ black box

$$L(\sigma_A^2, \sigma_\varepsilon^2, \gamma(0), \alpha, \lambda, \beta_0, \beta_1, \cdots, \beta_p)$$

Sample variogram: Estimating $\gamma(s)$

What is it? Look at $\frac{1}{2}Var(X_i(t_j) - X_i(t_l))$

$$= \underbrace{\frac{1}{2}Var(X_i(t_j)) + \frac{1}{2}Var(X_i(t_l))}_{\sigma_A^2 + \gamma(0) + \sigma_\varepsilon^2} - \underbrace{Cov(X_i(t_j), X_i(t_l))}_{\sigma_A^2 + \gamma(|t_j - t_l|)}$$

$$= \sigma_\varepsilon^2 + \gamma(0) - \gamma(|t_j - t_l|)$$

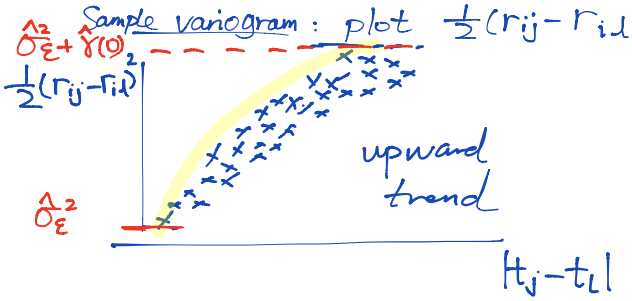
$$\begin{cases} = \sigma_\varepsilon^2 & \text{if } t_j = t_l \\ \doteq \sigma_\varepsilon^2 + \gamma(0) & \text{if } |t_j - t_l| \text{ is sufficiently large} \end{cases}$$

前面的 t_j, t_k → 採成 t_l

Given some estimate $\hat{\underline{\mu}} = \begin{pmatrix} \hat{\mu}(t_1) \\ \vdots \\ \hat{\mu}(t_k) \end{pmatrix}$ define

$$r_{ij} = r_i(t_j) = X_i(t_j) - \hat{\mu}(t_j)$$

Sample variogram: plot $\frac{1}{2}(r_{ij} - r_{il})^2$ vs. $|t_j - t_l|$



How to determine $\frac{\gamma(s)}{\gamma(0)} = \rho(s)$?

Spps $\gamma(s) = \gamma(0)\exp(-\alpha s^\lambda)$

$\therefore \rho(s) = \exp(-\alpha s^\lambda)$

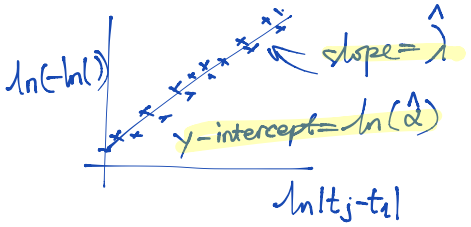
Estimate $\rho(s)$ (i.e. α, λ) from variogram plot

$$\frac{\hat{\sigma}_\varepsilon^2 + \hat{\gamma}(0) - \frac{1}{2}(r_{ij} - r_{il})^2}{\hat{\gamma}(0)} \doteq \exp(-\alpha |t_{ij} - t_{il}|^\lambda)$$

$$\approx \hat{\gamma}(|t_{ij} - t_{il}|)$$

$$\ln\left(\frac{\hat{\sigma}_\varepsilon + \hat{\gamma}(0) - \frac{1}{2}(r_{ij} - r_{il})^2}{\hat{\gamma}(0)}\right) \doteq -\alpha |t_{ij} - t_{il}|^\lambda$$

$$\ln(-\ln(\cdots)) \doteq \ln(\alpha) + \lambda \ln|t_{ij} - t_{il}|$$



- use estimates to get std error for estimates of $\hat{\mu}(t_j)$
- use estimates from variogram as initial estimates in maximum likelihood estimation.