MAE 5776

ANÁLISE MULTIVARIADA

Júlia M Pavan Soler

pavan@ime.usp.br

1º Sem/2020 - IME

Análise Multivariada

Já vimos ☺

- Análise Descritiva Multivariada
- Elipsóides de Dispersão e de Confiança, MANOVA
- Metodologias Clássicas de redução de dimensionalidade: $Y_{n\times p}$; $\Re^p \rightarrow \Re^m$
  - ✓ ACP *(m≤min(n,p))*, ACoP *(m≤min(n,p))*, AC *(m≤min(I-1,J-1))*, AF *(m≤min(n,p))*
  - ✓ AG
  - ✓ AD *(m≤min(n,p,G-1))*, ACC *(m≤min(n,p,q))*
  - ✓ Soluções Duais ($\Re^{nxp}$, $\Re^{pxp}$, $\Re^{nxn}$), Representações Biplot
  - ✓ PCR (Regressão via CP), PLS (Regressão via MQ Parciais)
- Integração de Bancos de Dados - Diagrama de Caminhos (Grafos)

Var. quantitativas
n > p
Obs iid

Caso: **n << p (*big p*)**: Soluções Regularizadas e Penalizadas
⇒ Componentes Principais (ACP)
⇒ Análise Discriminante (AD)
⇒ Correlação Canônica (ACC)

Métodos via Fatoração de Matrizes
e via Modelos de Regressão!

Caso: observações "dependentes" ⇒ dados de famílias
(indivíduos e seus familiares)

⇒ Componentes Principais de Herdabilidade (PCH)

Matriz de Dados

	Variáveis					
Unidades Amostrais	1	2	…	j	…	p
1	Y_{11}	Y_{12}		Y_{1j}		Y_{1p}
2	Y_{21}	Y_{22}		Y_{2j}		Y_{2p}
…	…	…	…	…		…
i	Y_{i1}	Y_{i2}		Y_{ij}		Y_{ip}
…	…	…	…	…	…	…
n	Y_{n1}	Y_{n2}		Y_{nj}		Y_{np}

- Amostra Aleatória Simples de n-Vetores em $\Re^p$

$$Y_{n\times p} \in \Re^{n\times p}; \quad Y_i \in \Re^p \overset{iid}{\sim} (\mu;\Sigma)$$

Já vimos vários resultados sob a formalização de observações independentes (AASn)!

Casos mais gerais:

- Considere G Amostras Aleatórias de observações em $\Re^{n_g \times p}$

(AAS$_G$)

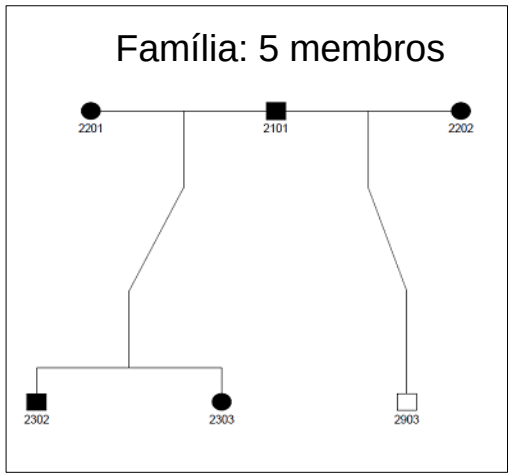
$$(g = 1,\ldots,G; \sum_g n_g = n)$$

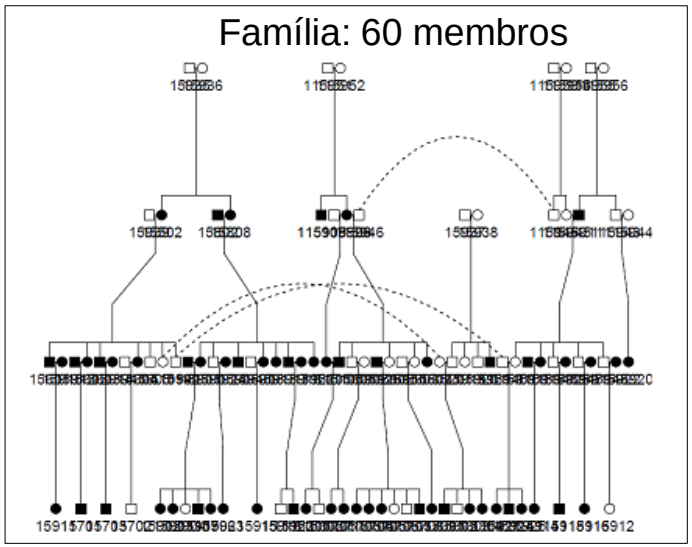
Motivação – Dados de Base Familiar
Projeto Corações de Baependi, MG

Mapear “Genes” associados a fatores de risco cardiovascular na População Brasileira

Baependi, MG

- Amostragem de domicílios de Baependi, MG (2006-)
- Amostra: 1.712 indivíduos de 119 famílias
- Tamanho das famílias: 21 ± 26
- Gerações: 2 - 4
- Número de filhos: 3 ± 2
- Núcleos familiares: 631
- Gênero: M(43,5%) F(56,5%)

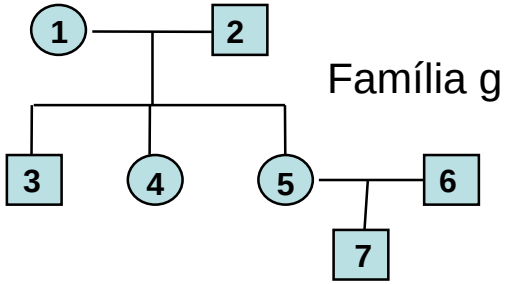




Nos membros de todas as famílias, muitas variáveis foram avaliadas (fenótipos, genótipos, dentre outras).

Motivação – Dados de Base Familiar

$Y_{n\times p}$ Observações em $\Re^p$ correlacionadas devido à estrutura familiar (grau de parentesco entre indivíduos)



Família	Unidade Amostral	Y_1	Y_2	...	Y_p
1	1	Y_{111}	Y_{112}		Y_{11p}
1	2	Y_{121}	Y_{122}		Y_{12p}
...	...				
1	n_1	Y_{1n11}	Y_{1n12}		Y_{1n1p}
Médias da Família 1		$\bar{Y}_{11}$	$\bar{Y}_{12}$		$\bar{Y}_{1p}$
...					
G	1	Y_{G11}	Y_{G12}		Y_{G1p}
G	2	Y_{G21}	Y_{G22}		Y_{G2p}
...	...				
G	n_G	$Y_{Gn_{G1}}$	$Y_{Gn_{G2}}$		$Y_{Gn_{Gp}}$
Médias da Família G		$\bar{Y}_{G1}$	$\bar{Y}_{G2}$		$\bar{Y}_{Gp}$
Vetor de Médias Geral		$\bar{Y}_{.1}$	$\bar{Y}_{.2}$		$\bar{Y}_{.p}$

⇒ Modelar a dependência entre observações do mesma família (*cluster*): Matriz de parentesco

$$\Psi_g = \begin{array}{c|ccccccc} & \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} & \mathbf{5} & \mathbf{6} & \mathbf{7} \\ \hline \mathbf{1} & 1 & 0 & ½ & ½ & ½ & 0 & ¼ \\ \mathbf{2} & 0 & 1 & ½ & ½ & ½ & 0 & ¼ \\ \mathbf{3} & ½ & ½ & 1 & ½ & ½ & 0 & ¼ \\ \mathbf{4} & ½ & ½ & ½ & 1 & ½ & 0 & ¼ \\ \mathbf{5} & ½ & ½ & ½ & ½ & 1 & 0 & ½ \\ \mathbf{6} & 0 & 0 & 0 & 0 & 0 & 1 & ½ \\ \mathbf{7} & ¼ & ¼ & ¼ & ¼ & ½ & ½ & 1 \end{array}$$

$$\Psi_{n\times n} = I_G \otimes_{g=1}^{G} \Psi_g$$

⇒ Modelar a dependência entre observações do mesmo indivíduo (vetor em $\Re^p$): Matriz de correlação entre variáveis (Σ_{pxp})

$$\Omega_{np\times np} = \Psi_{n\times n} \otimes \Sigma_{p\times p}$$

Variáveis Aleatórias Multidimensionais

- Matriz aleatória (Gupta and Nagar, 2000):

Formulações alternativas

Formulação matricial

$$Y_{n\times p} = \left(Y_{ij}\right) \in \Re^{n\times p}; \quad Y_{n\times p} \sim N_{n,p}\left(M; \Psi \otimes \Sigma\right);$$

Formulação vetorial

$$vec\left(Y\right)_{np\times 1} \sim N_{np}\left(vec(M); \Psi \otimes \Sigma\right)$$

$M_{n\times p} = 1_n \mu'_{p\times 1}$: matriz de médias

$vec\left(M\right)_{np\times 1} = 1_n \otimes \mu_{p\times 1}$: vetor de médias de n observações em p variáveis

$\left(\Psi_{n\times n} \otimes \Sigma_{p\times p}\right)_{np\times np}$: matriz de covariâncias

Formulação flexível para diferentes modelagens

Matrizes de covariância Estruturadas: entre indivíduos (Ψ) e entre variáveis (Σ)

$\Psi = I_n ; \quad \Sigma = I_p$

Observações e variáveis independentes

$\Psi = I_n ; \quad \Sigma = \left(1-\rho\right) I_p + \rho 1_p 1_p{}'$

Observações independentes e correlação uniforme entre as variáveis

$\Psi = \oplus_{g=1}^{G} \left[\left(1-\rho\right) I_{n_g} + \rho 1_{n_g} 1_{n_g}{}'\right]; \; \Sigma = \left(\sigma_{jl}\right)$

Correlação uniforme entre observações agrupadas em G grupos

Correlação não estruturada entre variáveis

Matriz Aleatória

- Amostra Aleatória de n-Vetores em $\Re^p$

Já vimos!

$$Y_{n\times p} \in \Re^{n\times p}; \quad Y_i \in \Re^p \overset{iid}{\sim} \left(\mu_{p\times 1}; \Sigma_{p\times p}\right)$$

- Amostra Aleatória de *n*-Vetores em $\Re^p$ tal que $n = \sum_{g=1}^{G} n_g$

n-Vetores estratificados em G grupos e com
Correlação entre Observações dentro dos Estratos (grupos)

Grupos	Unidades amostrais	Variáveis					
		1	2	…	j	…	p
1	1	Y_{11}	Y_{12}		Y_{1j}		Y_{1p}
	2	Y_{21}	Y_{22}		Y_{2j}		Y_{2p}
	…	…	…	…	…		…
…	i	Y_{i1}	Y_{i2}		Y_{ij}		Y_{ip}
	…	…	…	…	…	…	…
G	n	Y_{n1}	Y_{n2}		Y_{nj}		Y_{np}

$$Y_{n\times p} = \begin{pmatrix} Y_1 \\ \dots \\ Y_G \end{pmatrix}$$

$$Y_{g\,(n_g\times p)}$$

$$n = \sum_{g=1}^{G} n_g$$

$$Y_{n\times p} \in \Re^{n\times p}; \quad Y_g \in \Re^{n_g\times p} \overset{iid}{\sim} \left(\mu_g; \Omega_g\right)$$

$$\begin{cases} \mu_{g\,(n_g\times p)} = 1_{n_g} \otimes \mu'_{p\times 1} \\ \Omega_{g\,(n_g p\times n_g p)} = \Psi_g \otimes \Sigma_{p\times p} \end{cases}$$

Matriz Aleatória – Componentes de Covariâncias em $\Re^{pxp}$

- Amostra Aleatória de G-Grupos em $\Re^{n_g \times p}$ $(g = 1,...,G; \sum_g n_g = n)$

$$Y_{n\times p} \in \Re^{n\times p}; \quad Y_g \in \Re^{n_g \times p} \overset{iid}{\sim} \left(\mu_g ; \Omega_g\right) \begin{cases} \mu_{g\,(n_g \times p)} = 1_{n_g} \otimes \mu'_{p\times 1} \\ \Omega_{g\,(n_g p \times n_g p)} = \Psi_g \otimes \Sigma_{p\times p} \end{cases}$$

$$Y_{n\times p} \sim \left(\mu_{n\times p}; \Omega_{np\times np}\right) \begin{cases} \mu_{n\times p} = \left(\oplus_{g=1}^{G} 1_{n_g}\right) \otimes \mu'_{p\times 1} \\ \Omega_{np\times np} = \oplus_{g=1}^{G}\left(\Psi_g \otimes \Sigma_{p\times p}\right) = \left(\oplus_{g=1}^{G} \Psi_g\right) \otimes \Sigma_{p\times p} \end{cases}$$

Decompor $\Re^{pxp}$ em Componentes de Covariâncias

$$\Sigma_{p\times p} = \Sigma_B + \Sigma_W$$

$$\Omega_{np\times np} = \left(\oplus_{g=1}^{G} \Psi_g\right) \otimes \Sigma \quad \Rightarrow \quad \Omega = \left(\oplus_{g=1}^{G} \Psi_g\right) \otimes \Sigma_B + I_n \otimes \Sigma_W$$

Σ_B: componente da covariância associado à dependência entre as *n*-observações
Σ_W: componente da covariância associado à independência entre as *n*-observações

Componentes de Covariâncias em $\Re^{pxp}$

$$Y_{n\times p} \sim \left(\mu_{n\times p}; \Omega_{np\times np}\right) \begin{cases} \mu_{n\times p} = \oplus_{g=1}^{G}\left(1_{n_g} \otimes \mu'_{g\ p\times 1}\right) \\ \Omega_{np\times np} = \left(\oplus_{g=1}^{G}\Psi_g\right)\otimes\Sigma_B + I_n \otimes \Sigma_W \end{cases}$$

$$\Omega_{np\times np} = \begin{bmatrix} \Psi_1 & & 0 \\ & \Psi_2 & \\ 0 & \ldots & \Psi_G \end{bmatrix} \otimes \Sigma_B + \begin{bmatrix} I_{n_1} & & 0 \\ & I_{n_2} & \\ 0 & \ldots & I_{n_G} \end{bmatrix} \otimes \Sigma_W$$

A matriz de covariância Σ é decomposta em dois componentes, um associado à suposição de correlação entre observações do mesmo grupo (Σ_B: matriz de covariância ENTRE grupos) e outro associado à independência condicional entre observações dado o grupo (Σ_W: matriz de covariância DENTRO de grupos)

Componentes Principais em Modelos de Componentes de Covariâncias

$$Cov\left(Y_g\right) = \Omega_g = \Psi_g \otimes \Sigma_B + I_{n_g} \otimes \Sigma_W; \qquad \Sigma_B + \Sigma_W = \Sigma$$

$PC_B \Rightarrow \max_{\|a\|=1} \dfrac{a'\Sigma_B a}{a'a}$: direção com máxima variação entre grupos

$PC_W \Rightarrow \max_{\|a\|=1} \dfrac{a'\Sigma_W a}{a'a}$: direção com máxima variação dentro dos grupos

$PCH \Rightarrow \max_{\left\|\Sigma_W^{1/2} a\right\|=1} \dfrac{a'\Sigma_B a}{a'\Sigma\ a} = \max_{\left\|\Sigma_W^{1/2} a\right\|=1} \dfrac{a'\Sigma_B a}{a'\hat{\Sigma}_W a}$: direção com máxima variação entre e mínima variação dentro de grupos

Obter estimativas dos Componentes de Covariância!

$\hat{\Sigma}_B, \quad \hat{\Sigma}_W$

Como são essas estimativas sob o modelo MANOVA clássico?

ANOVA – Efeito Fixo e Aleatório

Como obter estimativas de componentes de "variância"?
Recordando o caso p=1

$$Y_{gi} = \mu_g + e_{gi}$$
$$= \mu + \tau_g + e_{gi}; \quad i = 1,...,n_g,\ g = 1,...,G$$

Modelo ANOVA de Efeitos Fixos

$\sum_{g=1}^{G} \tau_g = 0$ Restrições de identificabilidade.

$$e_{gi} \overset{iid}{\sim} \left(0\,;\sigma^2\right)$$

$$\Rightarrow Y_{gi} \overset{iid}{\sim} \left(\mu_g\,;\sigma^2\right)$$

$$Cov\left(Y_{n\times 1}\right) = \begin{bmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & \sigma^2 \end{bmatrix} = \sigma^2 I_n$$

Modelo ANOVA de Efeitos Aleatórios

$$\tau_g \sim \left(0\,;\sigma_B^2\right) \qquad \mu_g = \mu + \tau_g \sim \left(\mu\,;\sigma_B^2\right)$$
$$e_{gi} \sim \left(0\,;\sigma_W^2\right) \qquad \tau_g \perp e_{gi}$$

σ_B^2: é componente de variância e de covariância

$$\Rightarrow Y_{gi} \sim \left(\mu\ ;\sigma^2 = \sigma_B^2 + \sigma_W^2\right)$$

$$Cov\left(Y_{gi};Y_{g'i'}\right) = \begin{cases} \sigma_B^2 + \sigma_W^2; & i = i', g = g' \\ \sigma_B^2; & i \neq i', g = g' \\ 0; & i \neq i', g \neq g' \end{cases}$$

$$Cov\left(Y_{n\times 1}\right) = \begin{bmatrix} \sigma^2 & \sigma_B^2 & 0 & 0 \\ \sigma_B^2 & \sigma^2 & 0 & 0 \\ 0 & 0 & \ldots & \sigma_B^2 \\ 0 & 0 & \sigma_B^2 & \sigma^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & \rho & 0 & 0 \\ \rho & 1 & 0 & 0 \\ 0 & 0 & \ldots & \rho \\ 0 & 0 & \rho & 1 \end{bmatrix}$$

$$\rho = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2}$$ Coeficiente de correlação intraclasse

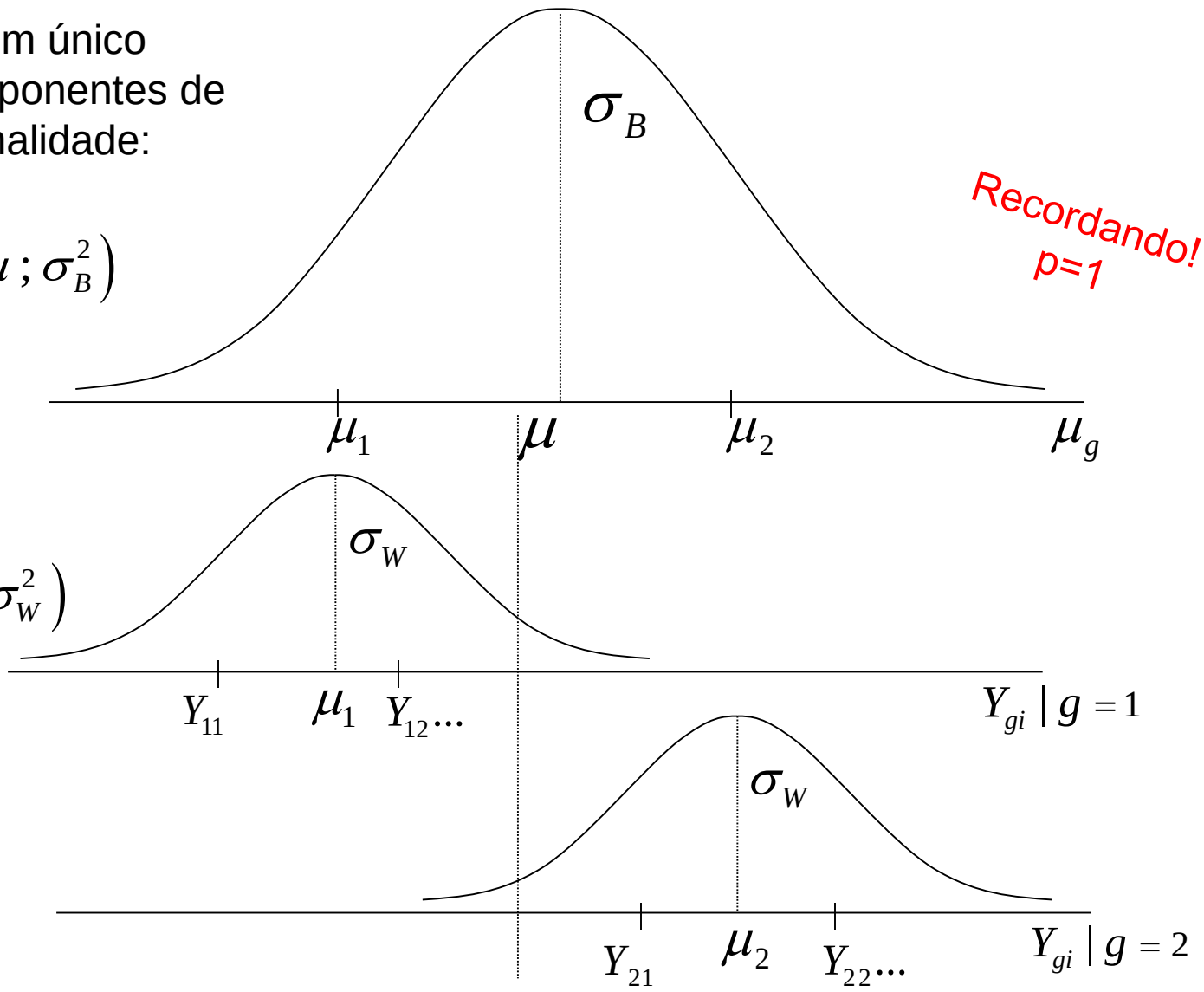
Modelo ANOVA de um único Fator Aleatório (componentes de variância), sob Normalidade:

$$\mu_g = \mu + \tau_g \sim N\left(\mu ; \sigma_B^2\right)$$

Recordando! p=1

Distribuição condicional de Y:

$$\Rightarrow y_{gi} \mid g \sim N\left(\mu_g ; \sigma_W^2\right)$$



$$H_0 : \sigma_B^2 = 0 \quad \Leftrightarrow \quad H_0 : \mu_g = \mu$$

Sob H_0, os modelos de ANOVA com um fator aleatório ou um fator fixo são equivalentes (mesma tabela ANOVA)

Tabela de ANOVA

Recordando o caso p=1. Tabela ANOVA é equivalente para os modelos de um fator fixo ou um fator aleatório!

$$H_0 : \sigma_B^2 = 0 \quad \Leftrightarrow \quad H_0 : \mu_g = \mu$$

F.V.	g.l.	SQ	QM	F
Grupo	G-1	$S_b = \sum r\,(\bar{Y}_g - \bar{Y}\,)^2$	$QMGr = S_b / (G-1)$	$\dfrac{QMGr}{QM\ \mathrm{Res}}$
Resíduo	n-G	$S_w = \sum_{gi} (Y_{gi} - \bar{Y}_g)^2$	$QM\ \mathrm{Res} = S_w / (n-G)$	
TOTAL	n-1	$S_T = \sum_{gi} (Y_{gi} - \bar{Y})^2$		

$n_g = r$: Dados balanceados

$$E\left(QM\ \mathrm{Res}\right) = \sigma_W^2 \quad \Rightarrow \quad \hat{\sigma}_W^2 = QM\ \mathrm{Res}$$

$$E\left(QMTr\right) = \sigma_W^2 + r\,\sigma_B^2 \quad \Rightarrow \quad \hat{\sigma}_B^2 = \frac{QMTr - QM\ \mathrm{Res}}{r}$$

Estimadores ANOVA dos componentes de variância

$$F \sim F_{(G-1),(n-G)}$$

Componentes de Covariâncias em $\Re^{pxp}$ Estruturas para Ψ em $\Re^{nxn}$

- Correlação Uniforme entre Observações (Simetria Composta ou Equicorrelação)

(Konish and Rao, 1992)

$$Y_{n\times p} \in \Re^{n\times p};\quad Y_g = Y_{n_g\times p} \in \Re^{n_g\times p} \overset{iid}{\sim} \left(\mu_g;\Omega_g\right)\begin{cases}\mu_{g\,(n_g\times p)} = 1_{n_g}\otimes \mu'_{g\ p\times 1}\\ \Omega_{g\,(n_g p\times n_g p)} = \Psi_g \otimes \Sigma_{p\times p}\end{cases}$$

$$Cov\left(Y_{gi}, Y_{g'i'}\right)_{(p\times p)} = \begin{cases}\Sigma = \Sigma_B + \Sigma_W & \text{se } g = g', i = i'\\ \Sigma_B & \text{se } g = g', i \neq i'\\ 0 & \text{se } g \neq g', i \neq i'\end{cases} \longleftarrow \begin{cases}\Psi_g = 1_{n_g}1_{n_g}'\\ \Sigma = \Sigma_B + \Sigma_W\end{cases}$$

$$\Rightarrow Cov\left(Y_g\right)_{n_g p\times n_g p} = \Omega_g = \left(1_{n_g}1_{n_g}'\right)\otimes\Sigma_B + I_{n_g}\otimes\Sigma_W$$

$$\Rightarrow Y_g \overset{iid}{\sim} \left(1_{n_g}\otimes\mu'_g;\Omega_g = (1_{n_g}1_{n_g}')\Sigma_B + I_{n_g}\otimes\Sigma_W\right)$$

$$\Rightarrow Y_{n\times p} \sim \left(\oplus_{g=1}^{G}(1_{n_g}\otimes\mu'_g);\Omega = \oplus_{g=1}^{G}\Omega_g\right)$$

Estimadores MANOVA

Equivalência analítica entre MANOVA com um Fator Fixo e MANOVA com um Fator Aleatório (sob equicorrelação)

$$\Rightarrow Y_g \overset{iid}{\sim} \left(1_{n_g} \otimes \mu_g' ; \Omega_g = (1_{n_g} 1_{n_g}')\Sigma_B + I_{n_g} \otimes \Sigma_W\right)$$

Tabela de MANOVA:

F.V.	g.l.	Matriz de SQPC
Trat	G-1	$H_{p\times p} = \sum_{g=1}^{G} n_g \left(\bar{Y}_g - \bar{Y}\right)\left(\bar{Y}_g - \bar{Y}\right)' = S_b$
Resíduo	n-G	$E_{p\times p} = \sum_{g=1}^{G}\sum_{i=1}^{n_g}\left(Y_{gi} - \bar{Y}_g\right)\left(Y_{gi} - \bar{Y}_g\right)' = S_w$
TOTAL	n-1	$H + E = \sum_{g=1}^{G}\sum_{i=1}^{n_g}\left(Y_{gi} - \bar{Y}\right)\left(Y_{gi} - \bar{Y}\right)'$

Sob $H_0 : \mu_g = \mu, \quad g = 1,\ldots,G$

Estimadores MANOVA dos componentes de covariância

$$E\left(\frac{S_w}{n-G}\right) = \Sigma_W ; \quad E\left(\frac{S_b}{G-1}\right) = \Sigma_W + n_0 \Sigma_B$$

$$n_0 = \frac{n - \left(\sum_g n_g^2 / n\right)}{G-1}$$

Componentes de Covariâncias em $\Re^{pxp}$ Sob Correlação Uniforme entre Observações

Estimadores MANOVA dos componentes de covariância

$$\Rightarrow \hat{\Sigma}_W = \frac{S_w}{n-G}; \qquad \hat{\Sigma}_B = n_0^{-1}\left\{\frac{S_b}{G-1} - \frac{S_w}{n-G}\right\}$$

$$\Rightarrow \hat{\Sigma} = \hat{\Sigma}_W + \hat{\Sigma}_B = n_0^{-1}\left\{\frac{S_b}{G-1} + \frac{(n_0-1)S_w}{n-G}\right\}$$

$$n_0 = \frac{n - \left(\sum_g n_g^2 / n\right)}{G-1}$$

Componentes de Covariâncias em $\Re^{pxp}$ Sob Correlação Uniforme entre Observações

Konishi and Rao, 1992; Oualkacha et al., 2012)

$$\Rightarrow Y_g \overset{iid}{\sim} \left(1_{n_g} \otimes \mu_g'; \Omega_g = (1_{n_g} 1_{n_g}')\Sigma_B + I_{n_g} \otimes \Sigma_W\right); \quad \hat{\Sigma}_{p\times p} = \hat{\Sigma}_{B_{p\times p}} + \hat{\Sigma}_{W_{p\times p}}$$

$$PC_g \Rightarrow \max_a \frac{a'\hat{\Sigma}_B a}{a'a}, \qquad a'a = 1$$

Direção com máxima variação Entre grupos

$$PC_e \Rightarrow \max_a \frac{a'\hat{\Sigma}_W a}{a'a}, \qquad a'a = 1$$

Direção com máxima variação Dentro de grupos

$$PC_T \Rightarrow \max_a \frac{a'\hat{\Sigma}\, a}{a'a} = \max_a \frac{a'\left[\hat{\Sigma}_B + \hat{\Sigma}_W\right]a}{a'a}, \qquad a'a = 1$$

Direção com máxima variação Total

$$PCH \Rightarrow \max_a \frac{a'\hat{\Sigma}_B a}{a'\left[\hat{\Sigma}_B + \hat{\Sigma}_W\right]a} = \max_a \frac{a'\hat{\Sigma}_B a}{a'\hat{\Sigma}_W a}, \qquad a'\hat{\Sigma}_W a = 1$$

Direção com máxima variação Entre grupos e mínima variação Dentro

Componente Principal de Herdabilidade

Componentes Principais em Modelos de Componentes de Covariâncias

Correlação uniforme (equicorrelação)

$$\Rightarrow Y_g \overset{iid}{\sim} \left(1_{n_g} \otimes \mu_g' ; \Omega_g = (1_{n_g} 1_{n_g}')\Sigma_B + I_{n_g} \otimes \Sigma_W\right) \quad (**)$$

$$PCH \Rightarrow \max_a \frac{a'\hat{\Sigma}_B a}{a'\hat{\Sigma}_W a}$$

: Componentes Principais de Herdabilidade (PCH: direção com máxima variação Entre e mínima variação Dentro de grupos)

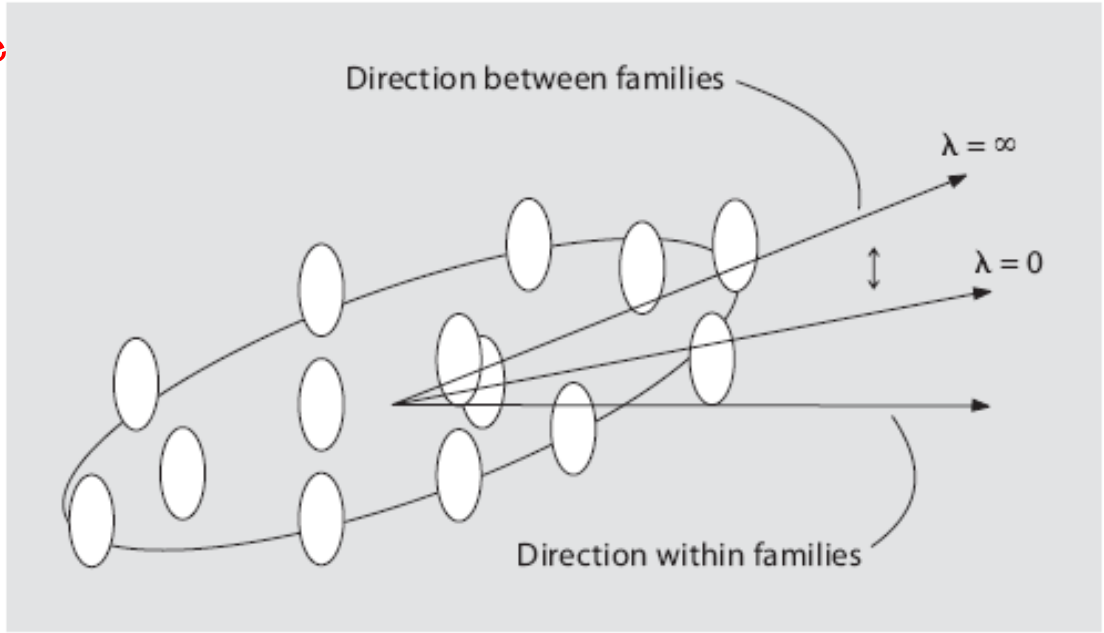
$$\overset{**}{\Rightarrow} \hat{\Sigma}_W = \frac{S_w}{n-G}; \qquad \hat{\Sigma}_B = n_0^{-1}\left\{\frac{S_b}{G-1} - \frac{S_w}{n-G}\right\} \qquad n_0 = \frac{n - \left(\sum_g n_g^2 / n\right)}{G-1}$$

Os PCH de dados correlacionados (sob **) correspondem às direções da Análise Discriminante sob obs. independentes!

Obter os autovalores e autovetores (*V*) de $\hat{\Sigma}_W^{-1}\hat{\Sigma}_B \Rightarrow PCH = YV$

Componentes Principais-Componentes de Covariância Soluções Regularizadas

Ilustração do padrão de dispersão entre e dentro de grupos



(Wang, 2007)

Elipse vertical: corresponde à variabilidade dentro dos grupos (famílias)
Elipse maior: corresponde à variação entre grupos (famílias)

$$PCH_{\lambda} \Rightarrow \max_{a,\lambda>0} \frac{a'\hat{\Sigma}_B a}{a'\hat{\Sigma}_W a + \lambda \|a\|^2} = \max_{a,\lambda>0} \frac{a'\hat{\Sigma}_B a}{a'\left[\hat{\Sigma}_W + \lambda I_p\right] a}$$

Solução Regularizada (*n*<<*p*). Como estimar o parâmetro de regularização λ?

λ=0: solução não regularizada do PCH
λ=∞: solução do PCH próxima à solução para Σ_B.(maximização entre grupos)

Componentes Principais-Componentes de Covariância Soluções Regularizadas

Algoritmo para estimação do parâmetro de regularização λ

Passo 1: Partição dos grupos em dois sub-grupos: Grupo 1 e Grupo 2.
(Repetir L=50 vezes)

Passo 2: Grupo1: para λ=0.01, obter o j-ésimo autovetor $V_{j\lambda}^{(1)l}$ de $\left(\hat{\Sigma}_W + \lambda I_p\right)^{-1/2}$.

Será necessário substituir os autovalores negativos por "0" (Amemiya, 1985).

Grupo2: obter as estimativas $\hat{\Sigma}_B^{(2)l}$, $\hat{\Sigma}_W^{(2)l}$;

Passo 3. Repetir para λ = 0.01, 2, 4, ..., 1000. O parâmetro de regularização é estimado como:

$$\lambda_{CV} = \arg\max_{\lambda} \frac{1}{L}\sum_{l=1}^{L} \frac{V_{j\lambda}^{(1)l\prime}\hat{\Sigma}_B^{(2)l}V_{j\lambda}^{(1)l}}{V_{j\lambda}^{(1)l\prime}\hat{\Sigma}_W^{(2)l}V_{j\lambda}^{(1)l}};$$

$$PCH_{\lambda_{CV}} = Ya'; \max_a \frac{a'\hat{\Sigma}_B a}{a'\left[\hat{\Sigma}_W + \lambda_{CV} I_p\right]a}$$

Componentes Principais-Componentes de Covariâncias
Correlação Familiar

Matriz de covariância mais geral entre obs.!

$Y_{n\times p}$ Observações correlacionadas devido à estrutura familiar (grau de parentesco entre indivíduos)

Família	Unidade Amostral	Y_1	Y_2	...	Y_p
1	1	Y_{111}	Y_{112}		Y_{11p}
1	2	Y_{121}	Y_{122}		Y_{12p}
...	...				
1	n_1	Y_{1n11}	Y_{1n12}		Y_{1n1p}
Médias da Família 1		$\overline{Y}_{11}$	$\overline{Y}_{12}$		$\overline{Y}_{1p}$
...					$\overline{Y}_{1p}$
G	1	Y_{G11}	Y_{G12}		Y_{G1p}
G	2	Y_{G21}	Y_{G22}		Y_{G2p}
...	...				
G	n_G	$Y_{G}n_{G1}$	$Y_{G}n_{G2}$		$Y_{G}n_{Gp}$
Médias da Família G		$\overline{Y}_{G1}$	$\overline{Y}_{G2}$		$\overline{Y}_{Gp}$
Vetor de Médias Geral		$\overline{Y}_{.1}$	$\overline{Y}_{.2}$		$\overline{Y}_{.p}$

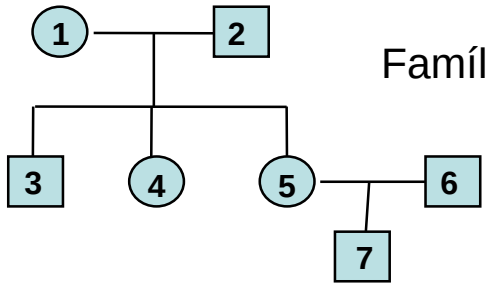
Modelo de Componentes de Covariância

$$\Rightarrow Cov\left(Y_g\right) = \Omega_g = \Psi_g \otimes \Sigma_B + I_{n_g} \otimes \Sigma_W$$

Matriz de parentesco (não é uniforme)

$$\Psi_g =$$

	1	2	3	4	5	6	7
1	1	0	½	½	½	0	¼
2	0	1	½	½	½	0	¼
3	½	½	1	½	½	0	¼
4	½	½	½	1	½	0	¼
5	½	½	½	½	1	0	½
6	0	0	0	0	0	1	½
7	¼	¼	¼	¼	½	½	1



Família g

Componentes Principais-Componentes de Covariâncias Correlação Familiar

Estimadores MANOVA do Modelo de Componentes de Covariância: são funções lineares de S_b e S_w (Oualkacha et al., 2012)

$$\hat{\Sigma}_B = \frac{S_b/(G-1) - S_w/(n-G)}{(\tau_c - \tau_b/n)/(G-1) - (\tau_a - \tau_c)/(n-G)}$$

$$\hat{\Sigma}_W = \frac{1}{(n-G)} S_w - \frac{(\tau_a - \tau_c)}{(n-G)} \hat{\Sigma}_B$$

$$n = \sum_{g=1}^{G} n_g, \quad \tau_a = \sum_{g=1}^{G} \tau_{a_g}, \quad \tau_b = \sum_{g=1}^{G} \tau_{b_g}, \quad \tau_c = \sum_{g=1}^{G} \frac{1}{n_g} \tau_{b_g}$$

$$\tau_{a_g} = 2Trace\left[\Phi_g\right], \quad \tau_{b_g} = 2\sum_{\substack{i'=1 \\ i'>i}}^{n_g} \sum_{i=1}^{n_g} \left(\Phi_g\right)_{ii'}$$

Componentes Principais-Componentes de Covariância Correlação Familiar (de Andrade et al., 2015)

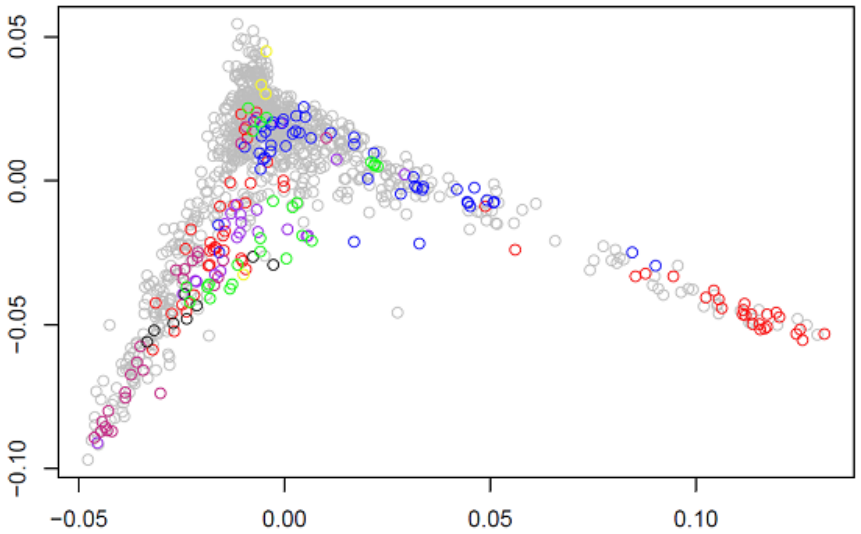
Aplicação: Projeto Corações de Baependi (MG)

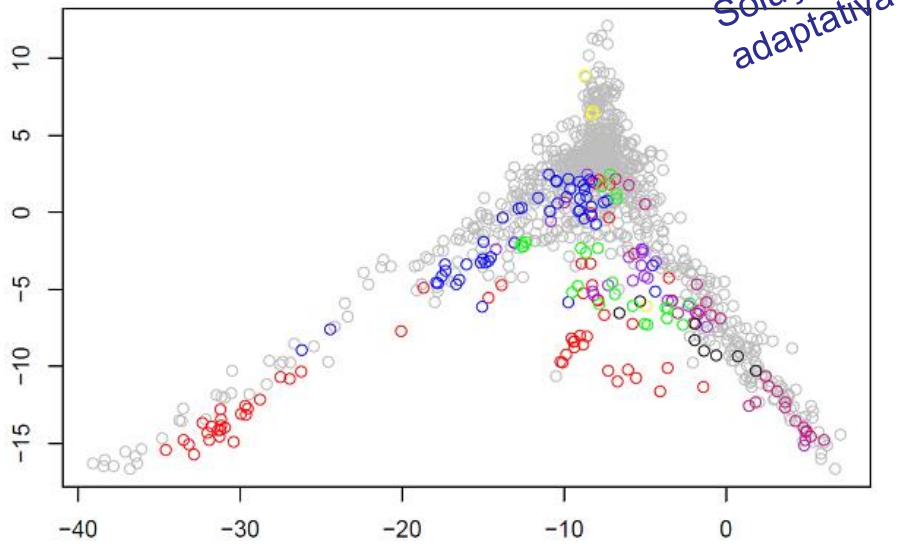
n=1.109 indivíduos de G=80 famílias e p=8.764 variáveis genéticas (SNPs)

PC sob Independência: $PC \Rightarrow \max_{\|a\|=1} \frac{a'Sa}{a'a}$

$PCH \Rightarrow \max_{\|S_W^{1/2}a\|=1} \frac{a'S_Ba}{a'S_Wa}$

Solução mais adaptativa



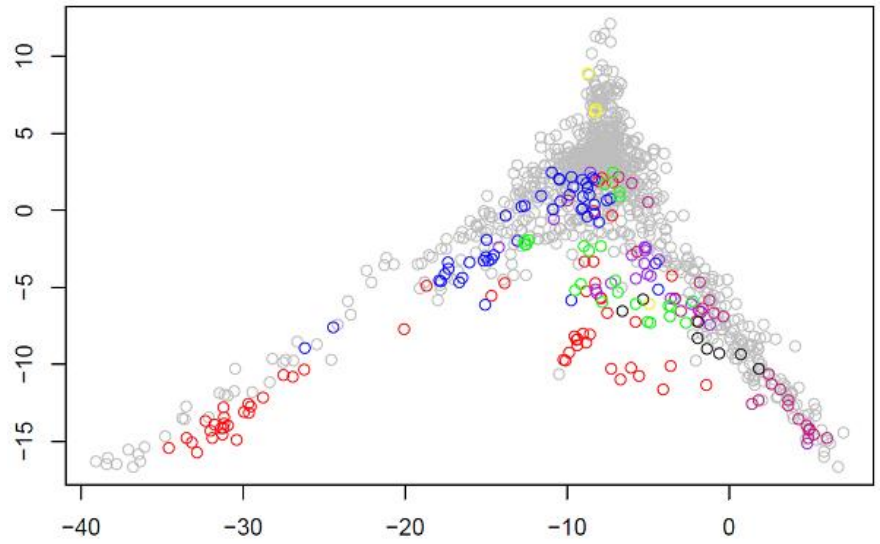


Proporção da variância explicada pelos PC1 e PCH1

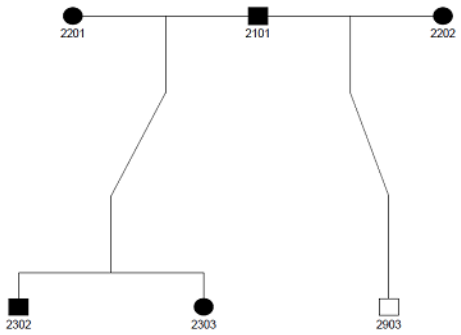
CP	1	2	3	4	5	6	7	8	9	10
S	0.022	0.014	0.008	0.0069	0.0068	0.0061	0.0059	0.0055	0.0053	0.005
$S_W^{-1}S_B$	0.086	0.070	0.035	0.031	0.028	0.026	0.0255	0.0249	0.0246	0.0239

Componentes Principais-Componentes de Covariância Correlação Familiar

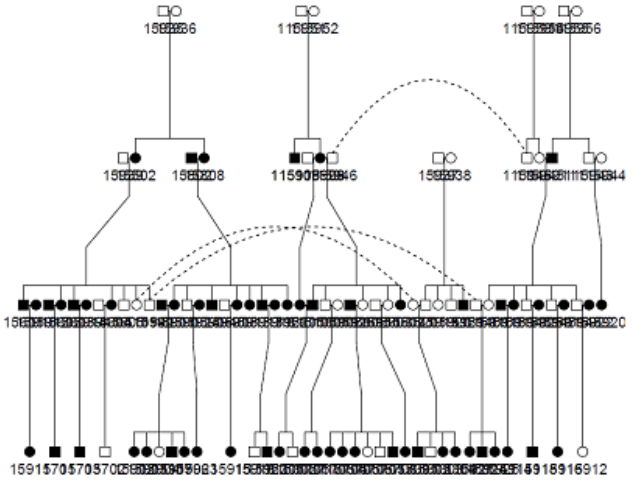
PCH1 x PCH2



- Família: 5 membros (homogênea)

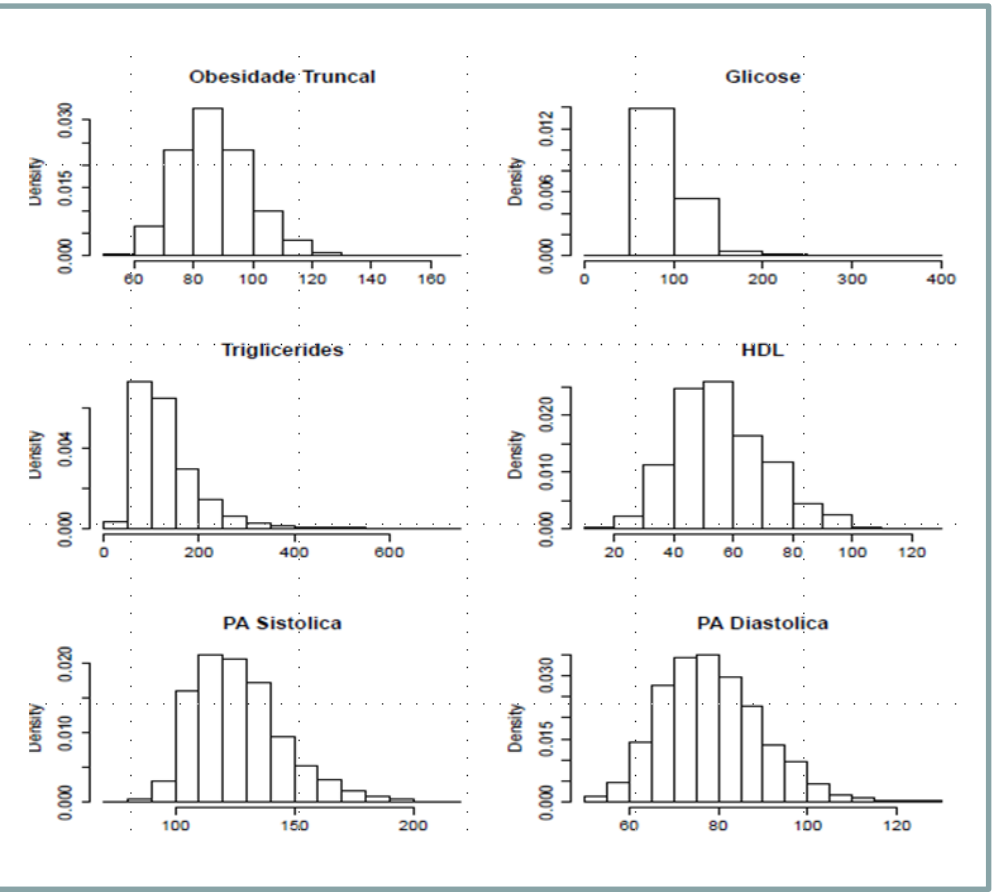


- Família: 60 membros (heterogênea)



Componentes Principais-Componentes de Covariâncias
Correlação Familiar

Aplicação: Projeto Corações de Baependi (MG)
Componentes Principais da Síndrome Metabólica (doença multifatorial: p=6)



Calcular o PCH1 (variável latente da SM)

$$Y_g \sim \left(1_{n_g}\mu_g';\Psi_g \otimes \Sigma_B + I_n \otimes \Sigma_W\right)$$

$$\max_a \frac{a'\hat{\Sigma}_B a}{a'\hat{\Sigma}_W a} \qquad PCH = a'Y_{6\times 1}$$

Var.	ObTr	Glic	Triglic	HDL	SBP	DBP	**PCH1**
ρ	0,16	0,12	0,35	0,30	0,18	0,13	**0,36**
ρ(Y,PCH1)	-0,23	-0,34	-0,93	-0,28	-0,35	-0,39	

$$\rho = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2}$$

Coef. de correlação intraclasse

Componentes Principais em Dados de Famílias
Minicurso: Oficina - R

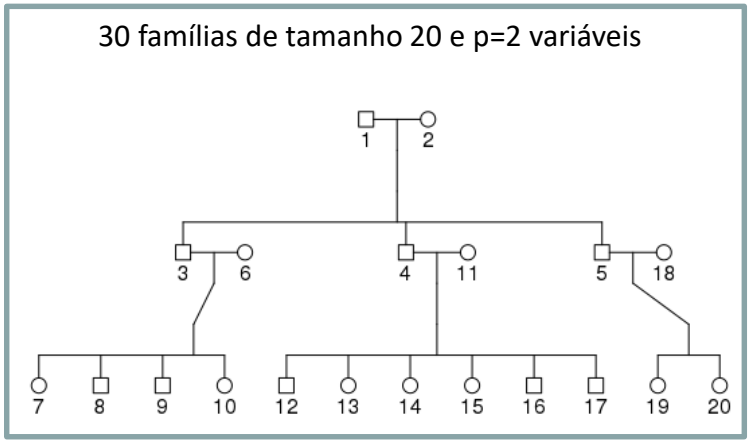
Visualização do PCH!

- Gerar dados com **estrutura familiar** aleatória **com vetor de médias** e **matrizes de covariância** conhecidas (p=2)

$$Y_g \in \Re^2;\ Y_g \sim N\left(1_{n_g}\,\mu_g{}';\Psi_g \otimes \Sigma_B + I_n \otimes \Sigma_W\right)$$

- Obter as estimativas das matrizes de covariância e os seguintes Componentes Principais (a'Y):

$$\max_a \frac{a'\hat{\Sigma}_B a}{a'\hat{\Sigma}_W a}$$



Exemplo1: Σ_B e Σ_W com correlações positivas (0,30 e 0,25, respectivamente) e variáveis (fenótipos) com correlação intraclasse moderada (0,40 a 0,70).

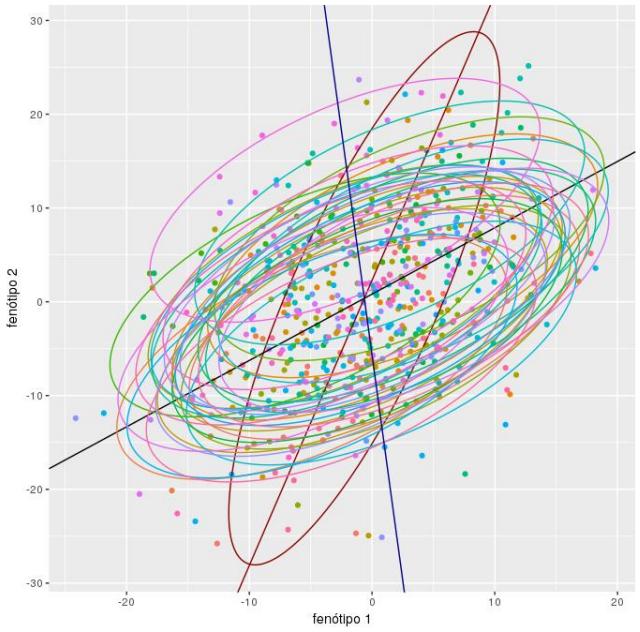
Exemplo 2: Σ_B e Σ_W com correlações de sinais opostos (0,30 e -0,25, respectivamente) e variáveis (fenótipos) com correlação intraclasse moderada (0,40 a 0,70).

Exemplo 3: Σ_B e Σ_W com correlações negativas (-0,90 e -0,80, respectivamente) e variáveis (fenótipos) com correlação intraclasse moderada (0,40 a 0,70).

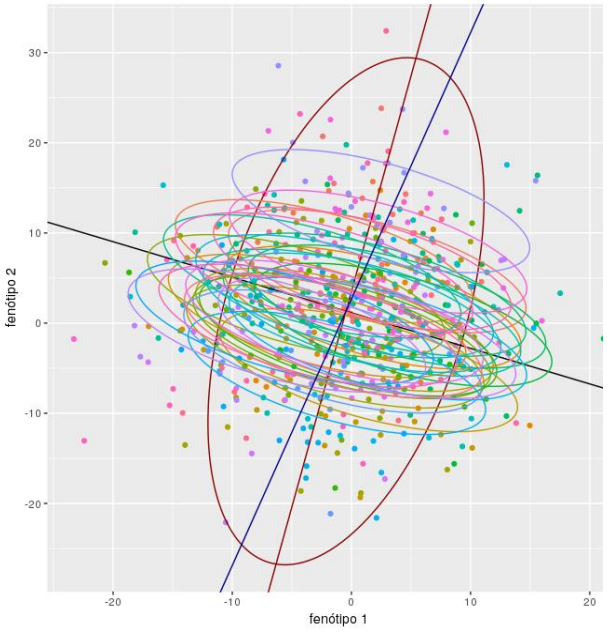
Componentes Principais em Dados de Famílias
Oficina - R

Visualização do PCH!

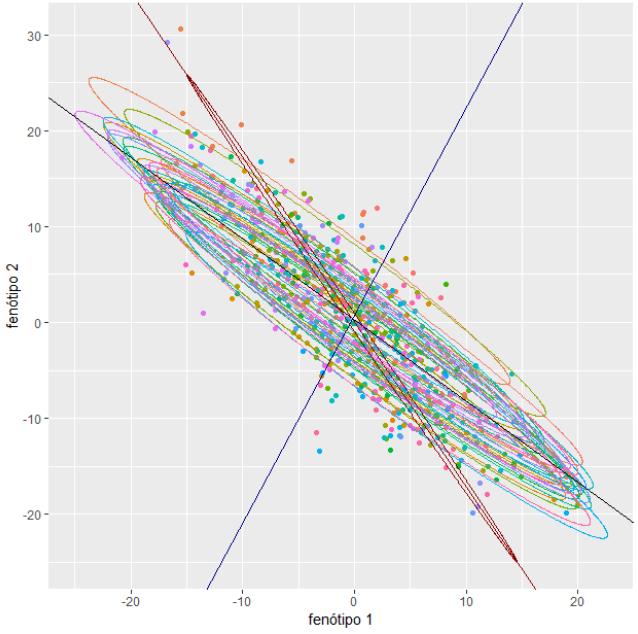
Exemplo 1



Exemplo 2



Exemplo 3



Preto: reta de MQO Vermelho: CP maximizando a variabilidade ENTRE famílias (S_B)
Azul: Componente Principal de Herdabilidade (PCH: maximiza $S_W^{-1}S_B$)