Proof of Exercise 7 Demo

Let V be the matrix defined as in lecture slides 7. Show that

$$\text{Trace}(V) = \frac{K}{P} - 1,$$

where K is the total number of modalities and P is the number of qualitative variables.

$n =$ sample size, $K =$ total number of modalities

$K =$ number of modalities, $K_p =$ num. of modalities of pth variable,

$$\sum_{p=1}^{P}\sum_{l=1}^{K_p} n_{pl} = nP, \qquad \sum_{l=1}^{K_p} n_{pl} = n,$$

$$T \in \mathbb{R}^{n\times K}, \qquad V = T^T T \in \mathbb{R}^{K\times K},$$

$$T = \begin{pmatrix} t_{11} & \dots & t_{1K} \\ \vdots & \ddots & \vdots \\ t_{n1} & \dots & t_{nK} \end{pmatrix}, \qquad T^T = \begin{pmatrix} t_{11} & \dots & t_{n1} \\ \vdots & \ddots & \vdots \\ t_{1K} & \dots & t_{nK} \end{pmatrix},$$

$$T_{i,pl} = t_{ipl} = \frac{x_{ipl} - n_{pl}/n}{\sqrt{Pn_{pl}}}, \qquad x_{ipl} \in \{0,1\},$$

$$\text{diag}(V) = \begin{pmatrix} t_{11}^2 + t_{21}^2 + \ldots + t_{n1}^2 \\ t_{12}^2 + t_{22}^2 + \ldots + t_{n2}^2 \\ \vdots \\ t_{1K}^2 + t_{2K}^2 + \ldots + t_{nK}^2 \end{pmatrix}.$$

Then,

$$\text{Trace}(V) = \sum_{m=1}^{K}\sum_{i=1}^{n} t_{im}^2 = \sum_{i=1}^{n}\sum_{p=1}^{P}\sum_{l=1}^{K_p}\left(\frac{x_{ipl} - \frac{n_{pl}}{n}}{\sqrt{Pn_{pl}}}\right)^2 = \sum_{i=1}^{n}\sum_{p=1}^{P}\sum_{l=1}^{K_p}\left(\frac{x_{ipl}^2 - 2x_{ipl}\frac{n_{pl}}{n} + \frac{n_{pl}^2}{n^2}}{Pn_{pl}}\right)$$

$$= \frac{1}{P}\sum_{i=1}^{n}\sum_{p=1}^{P}\sum_{l=1}^{K_p}\left(\frac{x_{ipl}^2}{n_{pl}} - 2\frac{x_{ipl}}{n} + \frac{n_{pl}}{n^2}\right)$$

Then consider the terms of the sum separately. For the second term, see the complete disjunctive table:

$$\frac{1}{P}\sum_{i=1}^{n}\sum_{p=1}^{P}\sum_{l=1}^{K_p}\left(-2\frac{x_{ipl}}{n}\right) = \frac{-2}{Pn}\sum_{i=1}^{n}\sum_{p=1}^{P}\sum_{l=1}^{K_p}(x_{ipl}) = \frac{-2}{Pn}nP = -2.$$

Likewise for the third term:

$$\frac{1}{P}\sum_{i=1}^{n}\sum_{p=1}^{P}\sum_{l=1}^{K_p}\frac{n_{pl}}{n^2} = \frac{1}{Pn^2}\sum_{i=1}^{n}\sum_{p=1}^{P}\sum_{l=1}^{K_p} n_{pl} = \frac{1}{Pn^2}\sum_{i=1}^{n} nP = 1.$$

The first term is the most difficult one here. Note that $x_{ipl} = x_{ipl}^2$, since $x_{ipl} \in \{0, 1\}$. By opening the sums, we get:

$$\begin{aligned}
\frac{1}{P}\sum_{i=1}^{n}\sum_{p=1}^{P}\sum_{l=1}^{K_p}\frac{x_{ipl}}{n_{pl}} &= \frac{1}{P}\sum_{i=1}^{n}\sum_{p=1}^{P}\left(\frac{x_{ip1}}{n_{p1}} + \frac{x_{ip2}}{n_{p2}} + \ldots + \frac{x_{ipK_p}}{n_{pK_p}}\right) \\
&= \frac{1}{P}\sum_{i=1}^{n}\left(\frac{x_{i11}}{n_{11}} + \frac{x_{i12}}{n_{12}} + \ldots + \frac{x_{i1K_1}}{n_{1K_1}} + \frac{x_{i21}}{n_{21}} + \ldots + \frac{x_{iPK_p}}{n_{PK_P}}\right) \\
&= \frac{1}{P}\left(\frac{1}{n_{11}}\sum_{i=1}^{n}x_{i11} + \frac{1}{n_{12}}\sum_{i=1}^{n}x_{i12} + \ldots + \frac{1}{n_{PK_P}}\sum_{i=1}^{n}x_{iPK_P}\right) \\
&= \frac{1}{P}\left(\frac{n_{11}}{n_{11}} + \frac{n_{12}}{n_{12}} + \ldots + \frac{n_{PK_P}}{n_{PK_P}}\right) = \frac{K}{P},
\end{aligned}$$

since K is the total number of modalities. Combine the terms and we get

$$\text{Trace}(V) = \frac{K}{P} - 2 + 1 = \frac{K}{P} - 1.$$

Table 1: Complete disjunctive table.

	X_1				X_P			
	X_{11}	$\ldots$	X_{1K_1}	$\ldots$	X_{P1}	$\ldots$	X_{PK_P}	$\sum_{p=1}^{P}\sum_{l=1}^{K_p} x_{ipl}$
1	x_{111}	$\ldots$	x_{11K_1}	$\ldots$	x_{1P1}	$\ldots$	x_{1PK_P}	P
2	x_{211}	$\ldots$	x_{21K_1}	$\ldots$	x_{2P1}	$\ldots$	x_{2PK_P}	P
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	x_{i11}	$\ldots$	x_{i1K_1}	$\ldots$	x_{iP1}	$\ldots$	x_{iPK_P}	P
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_{n11}	$\ldots$	x_{n1K_1}	$\ldots$	x_{nP1}	$\ldots$	x_{nPK_P}	P
$\sum_{i=1}^{n} x_{ipl}$	n_{11}	$\ldots$	n_{1K_1}	$\ldots$	n_{P1}	$\ldots$	n_{PK_P}	nP