Bare minimum on matrix algebra

Psychology 588: Covariance structure and factor models

Matrix multiplication

- Consider three notations for *linear combinations*

$$\begin{bmatrix} y_{11} & \cdots & y_{1m} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nm} \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} b_{11} & \cdots & b_{1m} \\ \vdots & \ddots & \vdots \\ b_{p1} & \cdots & b_{pm} \end{bmatrix}$$

$$y_{ij} = x_{i1}b_{j1} + x_{i2}b_{j2} + \cdots + x_{ip}b_{jp}, \quad i = 1,\ldots,n, \quad j = 1,\ldots,m$$

$$\mathbf{y}_j = \mathbf{X}\mathbf{b}_j, \quad j = 1,\ldots,m$$

$$\mathbf{Y} = \mathbf{X}\mathbf{B}$$

- Matrix multiplication is a very efficient way of writing simultaneous equation systems (i.e., linear combinations)

Inner (scalar) product

- Suppose $\mathbf{Y}$ contains p DVs as columns, $\mathbf{X}$ contains q IVs, and $\mathbf{B}$ contains regression weights as:

$$\begin{bmatrix} \boxed{y_{11}} & \cdots & y_{1p} \\ \vdots & \ddots & \vdots \\ y_{N1} & \cdots & y_{Np} \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1q} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{Nq} \end{bmatrix} \begin{bmatrix} b_{11} & \cdots & b_{1p} \\ \vdots & \ddots & \vdots \\ b_{q1} & \cdots & b_{qp} \end{bmatrix}$$

- Then an arbitrary y-entry for subject i and variable j is a linear combination of subject i's X-scores weighted for the j-th variable:

$$\mathbf{Y} \equiv \left\{ y_{ij} \right\}_{N \times p}, \quad y_{ij} = x_{i1}b_{1j} + x_{i2}b_{2j} + \cdots + x_{iq}b_{qj}$$

Outer products

- If all entries in $\mathbf{Y}$ are considered simultaneously, $\mathbf{Y}$ can be shown as a sum of q outer products:

$$\mathbf{Y} = \sum_{k=1}^{q} \mathbf{Y}_k = \sum_{k=1}^{q} \mathbf{x}_k \mathbf{b}'_k$$

$$\begin{bmatrix} y_{11} & \cdots & y_{1p} \\ \vdots & \ddots & \vdots \\ y_{N1} & \cdots & y_{Np} \end{bmatrix} = \begin{bmatrix} \boxed{\begin{matrix} x_{11} \\ \vdots \\ x_{N1} \end{matrix}} & \begin{matrix} \cdots \\ \ddots \\ \cdots \end{matrix} & \begin{matrix} x_{1q} \\ \vdots \\ x_{Nq} \end{matrix} \end{bmatrix} \begin{bmatrix} \boxed{b_{11} \quad \cdots \quad b_{1p}} \\ \begin{matrix} \vdots & \ddots & \vdots \end{matrix} \\ b_{q1} \quad \cdots \quad b_{qp} \end{bmatrix}$$

- $\mathbf{Y}_k$ accounts for a fraction of the DVs' variances, explained by the k-th IV $\mathbf{x}_k$ with its weights $\mathbf{b}_k$ for the p DVs

Algebraic properties of matrix multiplication

- Suppose all following multiplications are defined:

$$\mathbf{AB} \neq \mathbf{BA} \quad \text{in general}$$

$$(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$$

$$\mathbf{A}(\mathbf{B}+\mathbf{C}) = \mathbf{AB} + \mathbf{AC}$$

$$c(\mathbf{A}+\mathbf{B}) = c\mathbf{A} + c\mathbf{B}$$

Trace

- Trace is defined for square matrices as sum of diagonal elements:

$$\mathrm{tr}\left(\mathbf{A}_{n\times n}\right) \equiv \sum_{i=1}^{n} a_{ii}$$

- Useful for operations of sum of squares (typically of discrepancy of a model from the data) or weighted SS, along with the following properties:

$$\mathrm{tr}\left(\mathbf{A}\right) = \mathrm{tr}\left(\mathbf{A}'\right)$$

$$\mathrm{tr}\left(\mathbf{AB}\right) = \mathrm{tr}\left(\mathbf{BA}\right), \quad \mathrm{tr}\left(\mathbf{ABC}\right) = \mathrm{tr}\left(\mathbf{CAB}\right) = \mathrm{tr}\left(\mathbf{BCA}\right)$$

$$\mathrm{tr}\left(\mathbf{A}+\mathbf{B}\right) = \mathrm{tr}\left(\mathbf{A}\right) + \mathrm{tr}\left(\mathbf{B}\right)$$

- For example, consider a residual matrix under the principal component model

$$\mathbf{D} = \mathbf{X} - \mathbf{F}\mathbf{A}'$$

- Then, sum of squares of all residuals is:

$$\begin{aligned} F &= \operatorname{tr}\left(\mathbf{D}'\mathbf{D}\right) \\ &= \operatorname{tr}\left(\mathbf{X}' - \mathbf{A}\mathbf{F}'\right)\left(\mathbf{X} - \mathbf{F}\mathbf{A}'\right) \\ &= \operatorname{tr}\left(\mathbf{X}'\mathbf{X}\right) - 2\operatorname{tr}\left(\mathbf{X}'\mathbf{F}\mathbf{A}'\right) + \operatorname{tr}\left(\mathbf{A}\mathbf{F}'\mathbf{F}\mathbf{A}'\right) \end{aligned}$$

- The least-squares estimator of $\mathbf{A}$ is the one that minimizes F

Determinant

- Determinant of an $n \times n$ matrix $\mathbf{A} = \{a_{ij}\}$, denoted by $|\mathbf{A}|$ is defined as:

$$|\mathbf{A}| = a_{11}, \quad \text{if} \quad n = 1$$

$$|\mathbf{A}| = \sum_{j=1}^{n} a_{ij} |\mathbf{A}_{ij}| (-1)^{i+j} \quad \text{for any } i, \quad \text{if} \quad n > 1$$

where $|\mathbf{A}_{ij}|$, called “minor”, is the determinant of a submatrix of $\mathbf{A}$ without row i and column j, and $|\mathbf{A}_{ij}|(-1)^{i+j}$ is called “cofactor” for element (i, j)

Inverse of square matrices

- If all rows of $\mathbf{A}$ are linearly independent, there exists a <u>unique</u> $n \times n$ matrix $\mathbf{B}$ such that:

$$\mathbf{AB} = \mathbf{BA} = \mathbf{I}$$

- $\mathbf{B}$ is denoted by $\mathbf{A}^{-1}$ and called "the inverse of $\mathbf{A}$"

- The (j,i)-th entry of $\mathbf{A}^{-1}$ is $\dfrac{|\mathbf{A}_{ij}|}{|\mathbf{A}|}(-1)^{i+j}$ --- note the reversed subscripts

- Thus, it's obvious that $|\mathbf{A}| \neq 0$ for $\mathbf{A}^{-1}$ to be defined

Rank

- Rank of an $m \times n \; (m \geq n)$ matrix $\mathbf{A}$ is defined as the number of linearly independent columns of $\mathbf{A}$, with the following crucial properties

$$\operatorname{rank}(\mathbf{A}) \leq \min(m, n)$$

$$\operatorname{rank}(\mathbf{AB}) \leq \min\left(\operatorname{rank}(\mathbf{A}), \operatorname{rank}(\mathbf{B})\right)$$

- A square matrix must be “full-rank” for its inverse to exist --- necessary for estimation of parameters in SEM since it involves the inverse of the data covariance matrix and its determinant

- Following are all equivalent: “full-rank”, “nonsingular”, “positive definite”, $|\mathbf{A}| \neq 0$

Spectral decomposition

- For an $n \times n$ matrix $\mathbf{A}$, an eigenvalue e is defined as

$$\mathbf{A}\mathbf{v} = e\mathbf{v}, \quad \mathbf{v} \neq \mathbf{0}, \quad \mathbf{v}'\mathbf{v} = 1 \quad \leftrightarrow \quad \mathbf{v}'\mathbf{A}\mathbf{v} = e$$

- For any symmetric matrix $\mathbf{A}$, all n eigenvalues are nonnegative (i.e., positive semi-definite or nonnegative definite); if collectively written,

$$\mathbf{AV} = \mathbf{VE}, \quad \mathbf{V}'\mathbf{V} = \mathbf{I}, \quad \mathbf{E} = \mathrm{diag}\left(e_1, \ldots, e_n\right)$$

$$\mathbf{A} = \mathbf{VEV}' \quad \leftrightarrow \quad \mathbf{V}'\mathbf{AV} = \mathbf{E}$$ --- spectral decomposition

where eigenvalues are successively maximum, with

$$\mathrm{tr}\left(\mathbf{E}\right) = \mathrm{tr}\left(\mathbf{A}\right) \qquad cf.\ \prod_{j=1}^{n} e_j = \left|\mathbf{A}\right|$$

Singular value decomposition

- While eigenvalue decomposition is defined for square matrices (spectral decomposition as a special case), SVD is defined more generally for any rectangular matrix as:

$$\mathbf{X} = \mathbf{UTV}', \quad \mathbf{U}'\mathbf{U} = \mathbf{V}'\mathbf{V} = \mathbf{I}$$

  where $\mathbf{T}$ is a diagonal matrix with <u>nonnegative</u> “singular values” on the diagonal and columns of $\mathbf{U}$ and $\mathbf{V}$ are orthonormal singular vectors

- For example, if $\mathbf{Z}$ is a matrix of deviation scores, the principal component model of its sample covariance matrix has simple relationship to its SVD as:

$$\mathbf{Z} = \mathbf{UTV}' = \mathbf{FV}'$$

- In terms of the data covariance matrix,

$$\mathbf{S} = (N-1)^{-1}\mathbf{Z}'\mathbf{Z} = (N-1)^{-1}\mathbf{VTU}'\mathbf{UTV}'$$

$$= \mathbf{V}\left((N-1)^{-1}\mathbf{T}^2\right)\mathbf{V}' = \mathbf{VEV}'$$

- If the factor-analysis convention of scaling is desired (i.e., components/factors scaled to have variance of one),

$$\mathbf{Z} = \mathbf{UTV}' = \left(\sqrt{N-1}\,\mathbf{U}\right)\left(\sqrt{N-1}^{\,-1}\mathbf{VT}\right)' = \tilde{\mathbf{F}}\tilde{\mathbf{V}}'$$

- Alternatively, SVD of $\tilde{\mathbf{Z}} = \sqrt{N-1}^{\,-1}\mathbf{Z} = \mathbf{U}\tilde{\mathbf{T}}\mathbf{V}'$ removes the constant scaling factor so that $\tilde{\mathbf{T}}^2 = \mathbf{E}$

Lower-rank approximation (Eckart-Young theorem)

- Suppose we want to extract R principal components from an $N \times p$, rank-p data matrix $\mathbf{X}$ $(N \geq p > R)$, then they are given by the largest R singular values and their singular vectors

$$\mathbf{X} = \mathbf{U}\mathbf{T}\mathbf{V}' = \mathbf{U}_1\mathbf{T}_1\mathbf{V}_1' + \mathbf{U}_2\mathbf{T}_2\mathbf{V}_2',$$

$$\mathbf{U} = [\mathbf{U}_1, \mathbf{U}_2], \quad \mathbf{V} = [\mathbf{V}_1, \mathbf{V}_2], \quad \mathbf{T} = \begin{bmatrix} \mathbf{T}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_2 \end{bmatrix}$$

  where $\mathbf{U}_1\mathbf{T}_1\mathbf{V}_1'$ represents the R-dimensional subspace where the data variance is maximally captured and $\mathbf{U}_2\mathbf{T}_2\mathbf{V}_2'$ indicates $(p-R)$-dimensional subspace orthogonal to $\mathbf{U}_1\mathbf{T}_1\mathbf{V}_1'$

- The R columns in $\mathbf{V}_1$ are orthogonal reference axes in the R-dimensional space and the rows of $\mathbf{U}_1\mathbf{T}_1$ are coordinates of the N observations projected onto this space

- Likewise, the rank-R approximation can be shown for covariance matrix $\mathbf{S} = (N-1)^{-1}\mathbf{Z}'\mathbf{Z}$ by the spectral decomposition:

$$\mathbf{S} = \mathbf{VEV}' = \mathbf{V}_1\mathbf{E}_1\mathbf{V}_1' + \mathbf{V}_2\mathbf{E}_2\mathbf{V}_2',$$

$$\mathbf{V} = [\mathbf{V}_1, \mathbf{V}_2], \quad \mathbf{E} = \begin{bmatrix} \mathbf{E}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{E}_2 \end{bmatrix}$$

  where $\mathbf{V}_1\mathbf{E}_1\mathbf{V}_1'$ is the R-dimensional approximation to $\mathbf{S}$ that minimizes SS of the residuals $\mathbf{S} - \mathbf{V}_1\mathbf{E}_1\mathbf{V}_1' = \mathbf{V}_2\mathbf{E}_2\mathbf{V}_2'$

- In addition, the left singular vectors of $\mathbf{Z}$ can be found by the spectral decomposition of $(N-1)^{-1}\mathbf{ZZ}'$ as

$$(N-1)^{-1}\mathbf{ZZ}' = \mathbf{UEU}' = \mathbf{U}_1\mathbf{E}_1\mathbf{U}_1' + \mathbf{U}_2\mathbf{E}_2\mathbf{U}_2',$$

$$\mathbf{U} = [\mathbf{U}_1, \mathbf{U}_2]$$

Further operations

- “vec” operation vectorizes a matrix stacking columns one below another

$$\operatorname{vec}\mathbf{A}_{m\times n} \equiv \begin{bmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_n \end{bmatrix}_{mn\times 1}$$

- Kronecker product of $\mathbf{A}$ and $\mathbf{B}$ of any order is defined as

$$\mathbf{A}_{m\times n} \otimes \mathbf{B}_{p\times q} \equiv \begin{bmatrix} a_{11}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{bmatrix}_{mp\times nq}$$