Setup

We have $X = \{x_1, \dots, x_n\}$ where X is $p \times n$ with each col corresponding to an obs. We want to decompose this to M dimensions where $M < p$ by projecting X onto the subspace spanned by the col vectors of $B = \{b_1, b_2, \dots, b_M\}$. Assume B is orthonormal and all col vectors of X have mean 0 and standardised variance of 1.

We want to minimize info loss (variance loss) / squared errors by selecting appropriate basis vectors & coefficients

Deriving via calculus - Part 1

Consider x_i WLOG. $x_i = B_{i1} b_1 + B_{i2} b_2 + \dots + B_{iM} b_M + \dots + B_{ip} b_p$

$$x_i^* = B_{i1} b_1 + B_{i2} b_2 + \dots + B_{iM} b_M$$

Thus, we want to minimize $J = \frac{1}{n} \sum_{i=1}^{n} \|x_i - x_i^*\|^2$

We do this by orthogonally projecting each x_i onto our basis vectors. Thus $B_{ij} = x_i^T b_j = b_j^T x_i$. Order is arbitrary as B_{ij} is a scalar

Proof: $\frac{\partial J}{\partial B_{ij}} = \frac{\partial J}{\partial x_i^*} \frac{\partial x_i^*}{\partial B_{ij}} = \frac{-2}{n} (x_i - x_i^*)^T b_j = 0$

$$\implies x_i^T b_j = \sum_{j=1}^{m} B_{ij} b_j \cdot b_j = B_{ij} \quad //$$

So either via linear algebra ($r \cdot s = \|r\| \|s\| \cos\theta$) or calculus, we know $B = x^T b$

Now, $J = \frac{1}{n} \sum_{i=1}^{n} \| \sum_{j=M+1}^{p} B_{ij} b_j \cdot b_j \|^2 = \frac{1}{n} \sum \| \sum b_j^T B_{ij} b_j \|^2 = \frac{1}{n} \sum \| \sum B_{ij} \|^2$

$$= \frac{1}{n} \sum_{i=1}^{n} \| \sum_{j=M+1}^{p} x_i^T b_j \|^2 \overset{ONB}{=} \frac{1}{n} \sum \sum \| x_i^T b_j \|^2 = \frac{1}{n} \sum \sum b_j^T x_i x_i^T b_j$$

$$= \sum_{j=M+1}^{p} b_j^T \underbrace{\frac{1}{n} \sum_{i=1}^{n} x_i x_i^T}_{\text{The covariance matrix of } X! \text{ Found by taking Var}(X)!} b_j = \sum_{j=M+1}^{p} b_j^T \Sigma b_j$$

Deriving via calculus - Part 2

$J = \sum_{j=M+1}^{p} b_j^T \Sigma b_j$, subject to $b_j \cdot b_j = 1$ & $b_i \cdot b_j = 0$

$L = \sum_{j=M+1}^{p} b_j^T \Sigma b_j - \lambda_{M+1}(b_{M+1} \cdot b_{M+1} - 1) - \dots$

$\frac{dL}{db_j} = 2b_j^T \Sigma - 2\lambda_j b_j^T - \lambda_{j+1} b_{j+1}^T - \dots = 0$

(j is arbitrary & $j+1, j+2, \dots$ represent diff basis vectors)

Multiply by all basis vectors where $b_i \neq b_j$ individually to show under minimal conditions, their corresponding factor $\lambda = 0$. Since $b_i \cdot b_j = 0 \Leftrightarrow b_j^T b_i = 0$, $i \neq j$

left with $b_j^T \Sigma - \lambda_j b_j^T = 0 \Leftrightarrow \Sigma b_j = \lambda_j b_j$ (since Σ is symmetric)
Repeat this for $j = M+1$ till p

So $J = \sum_{j=M+1}^{p} b_j^T \lambda_j b_j = \sum_{j=M+1}^{p} \lambda_j$

Thus, the λ's should the the smallest e-values of Σ, and b_j the e-vectors and $B = \{b_1, \dots, b_M\}$ should be the e-vectors corresponding to the largest e-values! We have found B!

Deriving via Variance

You could also start by attempting to find B by maximising the variance of the projection of X onto the subspace spanned by B. Let us define C to be $\{c_1, \dots, c_p\}$

maximise $\text{Var}(B^T X)$ subject to $B^T B = BB^T = I_p$
Let $L = \frac{1}{n} B^T X X^T B - \Lambda(B^T B - I_p)$

$\frac{\partial L}{\partial B} = B^T \Sigma - \Lambda B^T = 0 \Leftrightarrow \Sigma B = B\Lambda^T \Leftrightarrow \Lambda^T = B^T \Sigma B$

We know the covariance matrix is always diagonalizable ie. $\Sigma = C V C^T$ where C is the square matrix of the e.vectors of Σ and V the diagonal matrix whose elements are the corresponding e.values. Since $V = C^T \Sigma C$ and we have above $\Lambda^T = B^T \Sigma B$, $\Lambda = \Lambda^T$ must be the diagonal matrix consisting of the M largest e.values of Σ and B their corresponding e.vectors. However, we cannot write $\Sigma = B \Lambda B^T$ as B is not a square matrix and $BB^T \neq I$

Result

$B = [b_1\ b_2 \cdots b_M]$ is the matrix whose col vectors are the eigenvectors of the covariance matrix of X corresponding to the M-largest e.values. Given X is $p \times n$, define X^* as $X^* = BB^TX$, where the ith col corresponds to $x_i^* = \sum_{j=1}^{M} B_{ij} b_j = \sum_{j=1}^{M} b_j b_j^T x_i$

$B^TX = \begin{bmatrix} b_1^T x_1 & \cdots & b_1^T x_n \\ \vdots & & \\ b_M^T x_1 & \cdots & b_M^T x_n \end{bmatrix}$, a $m \times n$ matrix where each element represents the coordinates of the projection of x onto the space (1D) spanned by the basis vector, wrt the basis vector.

Our solution $X^* = BB^TX$ is a $p \times n$ matrix which is a p-dim representation of our projected M-dim data. We have successfully orthogonally projected X onto the subspace spanned by B in a way that maximises the remaining info (variance) and minimizes the average squared reconstruction error.