

OZBAY Yasin

3T2

LHOEST Simon

3T1

VAN RENTERGHEM Tom

3T1

Rapport du Bureau d'Étude
N° 374

Partie 1:

1) On cherche ici à montrer que $\text{Var}(Xv) = v^T \text{Cov}(x) v$.

D'après les propriétés de la variance, on peut écrire :

$$\text{Var}(Xv) = \text{Var}(x) \cdot \text{Var}(v).$$

On, on sait que $\text{Var}(v) = v^T v$, donc :

$$\text{Var}(Xv) = \text{Var}(x) \cdot \text{Var}(v)$$

$$\text{Var}(Xv) = \text{Var}(x) \cdot v^T v$$

v étant un vecteur, le produit est commutatif, donc :

$$\text{Var}(x) \cdot v^T v = v^T \text{Var}(x) v$$

Il ne reste plus qu'à montrer que $\text{Var}(x) = \text{Cov}(x)$ et l'égalité sera démontrée.

Partons de la définition de la variance :

$$\text{Var}(x) = E(x^2) - E(x)^2 = E((x - E(x))^2) = E[(x - E(x)) \cdot (x - E(x))] = \text{Cov}(x, x) = \text{Cov}(x).$$

On a donc $\text{Var}(x) = \text{Cov}(x)$.

Finalement, on a bien démontré que :

$$\text{Var}(Xv) = v^T \text{Var}(x) v = v^T \text{Cov}(x) v.$$

2) La matrice de covariance de X ($\text{Cov}(X)$) est, par définition de la matrice de covariance carrée, symétrique et positive.

Grâce à ces trois critères, on peut dire que $\text{Cov}(X)$ se diagonalise dans une base orthonormée et :

$$\text{Cov}(X) = V D V^T$$

3) En partant de l'expression de la première composante principale :

$$Y_1 = X v_1 = \sum_{i=1}^n v_{1i} X_i,$$

On déduit l'expression de la k -ième composante :

$$Y_k = \sum_{i=1}^n v_{ki} X_i$$

On utilise la définition d'un produit vectoriel que nous développons pour k fixé :

$$\sum_{i=1}^n v_{ki} X_i = (X_1 \ X_2 \ \dots \ X_n) \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_k \\ v_{k+1} \\ \vdots \\ v_n \end{pmatrix}, \text{ on inverse } v \text{ et } X \text{ par commutativité du produit vectoriel.}$$

On reconnaît ici le vecteur ligne aléatoire X et le vecteur v_k un vecteur de \mathbb{R}^n , on peut donc dire que :

$$Y_k = \sum_{i=1}^n v_{ki} X_i = X \cdot v_k.$$

À partir de la question 1), on peut affirmer que :

$$\text{Var}(Y_k) = \text{Var}(X \cdot v_k)$$

On a donc :

$$\text{Var}(X \cdot v_k) = v_k^T \text{Cov}(X) v_k.$$

On sait que v_k est le k -ième vecteur de la matrice orthogonale V et que $\text{Cov}(X) = V D V^T$. Il vient :

$$v_k^T \text{Cov}(X) v_k = v_k^T V D V^T v_k.$$

On remarque ici que $v_k^T V = (V^T v_k)^T$ et que :

$$v_k^T \cdot v = \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \\ \hline 0 & \cdots & v_k^T \\ \vdots & \ddots & 0 \end{pmatrix} \begin{pmatrix} v_1 & v_2 & \dots & v_k & \dots & v_n \end{pmatrix} = \|v_k\|_2^2 = 1$$

car, d'après l'énoncé, $\|v\|_2 = 1$, $v \in \mathbb{Q}^n$; $v_k \in \mathbb{Q}^n$ donc $\|v_k\| = 1$.

On a donc :

$$\text{Var}(v_k) = v_k^T V D V^T v_k = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}^T \begin{pmatrix} \lambda_1 & & & & & \\ & \ddots & & & & \\ & & \lambda_n & & & \\ & & & \ddots & & \\ & & & & \lambda_n & \end{pmatrix} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \lambda_n$$

4) @ Pour montrer que $C = (E_{cn})^T \cdot (E_{cn})$, commençons par écrire le produit matriciel.

On a premièrement :

$$(E_{cn})_{j,n} = \frac{x_{jn} - \bar{x}_{nn}}{\sqrt{(m-1)} \sigma_{nn}} \quad \text{et} \quad (E_{cn})_{j,n}^T = (E_{cn})_{n,j} = \frac{x_{nj} - \bar{x}_{nn}}{\sqrt{(m-1)} \sigma_{nn}}$$

$$\text{On pose donc } C = (E_{cn})^T \cdot (E_{cn}) = \sum_{i=1}^j \left(\frac{x_{ki} - \bar{x}_{ni}}{\sqrt{(m-1)} \sigma_{ni}} \right) \left(\frac{x_{ik} - \bar{x}_{nn}}{\sqrt{(m-1)} \sigma_{nn}} \right)$$

$$\text{Il vient: } C = \sum_{i=1}^j \left(\frac{(x_{ki} - \bar{x}_{ni})(x_{ik} - \bar{x}_{nn})}{\sqrt{(m-1)} \sigma_{ni} \sqrt{(m-1)} \sigma_{nn}} \right) = \sum_{i=1}^j \frac{1}{m-1} \left(\frac{x_{ki} - \bar{x}_{ni}}{\sqrt{\sigma_{ni}}} \right) \left(\frac{x_{ik} - \bar{x}_{nn}}{\sqrt{\sigma_{nn}}} \right)$$

$$\text{Finalement, } C = \frac{1}{m-1} \cdot \sum_{i=1}^j \left(\frac{x_{ki} - \bar{x}_{ni}}{\sqrt{\sigma_{ni}}} \right) \cdot \left(\frac{x_{ik} - \bar{x}_{nn}}{\sqrt{\sigma_{nn}}} \right)$$

On peut ici reconnaître les variables centrées, notées x_i , composantes du vecteur aléatoire X de l'énoncé.

Les deux termes de la somme sont très ressemblants à ceux que l'on trouve dans l'espérance pour définir la covariance d'une composante x_i de X . Leur somme donne donc la covariance de X tout entier.

Prenons l'espérance de ce terme :

$$E(C) = E \left(\frac{1}{m-1} \sum_{i=1}^j \left(\frac{x_{ki} - \bar{x}_{ni}}{\sqrt{\sigma_{ni}}} \right) \left(\frac{x_{ik} - \bar{x}_{nn}}{\sqrt{\sigma_{nn}}} \right) \right)$$

$$= \frac{1}{m-1} \cdot E \left(\sum_{i=1}^j \left(\frac{x_{ki} - \bar{x}_{mi}}{\sqrt{s_{mi}}} \right) \left(\frac{x_{ik} - \bar{x}_{mk}}{\sqrt{s_{mk}}} \right) \right)$$

On peut ici faire rentrer l'espérance dans la somme et avec les propriétés de l'espérance, on a :

$$= \frac{1}{m-1} \cdot \sum_{i=1}^j \frac{E(x_{ki} - \bar{x}_{mi})(x_{ik} - \bar{x}_{mk})}{E(\sqrt{s_{mi}}) \cdot E(\sqrt{s_{mk}})}$$

On reconnaît au dénominateur la covariance d'un vecteur x_i centré.

Pour que l'on ait la covariance d'un x_i de X , il faut que x_i soit centrée-réduite.

On utilise donc la formule de l'énoncé :

$$= \frac{1}{m-1} \cdot \sum_{i=1}^j E \left(\left(\frac{x_{ki} - \bar{x}_{mi}}{\sqrt{s_{mi}}} \right) \cdot \left(\frac{x_{ik} - \bar{x}_{mk}}{\sqrt{s_{mk}}} \right) \right)$$

On a alors la covariance d'un x_i . En faisant rentrer la somme dans l'espérance, on aura la covariance de X , car on aura tous les termes le composant et plus la covariance d'un unique x_i .

Finalement, on obtient :

$$E(C) = \frac{1}{m-1} \cdot E \left(\sum_{i=1}^j \left(\frac{x_{ki} - \bar{x}_{mi}}{\sqrt{s_{mi}}} \right) \left(\frac{x_{ik} - \bar{x}_{mk}}{\sqrt{s_{mk}}} \right) \right) = \frac{1}{m-1} \cdot \text{Cov}(X)$$

On peut donc dire que $C = (E_C)_T \cdot E_C$ est bien une approximation d'un estimateur avec blocs de la matrice de covariance de X . Il est obtenu en centrant et en réduisant les termes de E pour approximer les x_i de X puis en prenant les sommants pour constituer X tout entier pour obtenir la covariance en prenant l'espérance de ces termes.

(b) R_m étant une réalisation de E_m , qui contient des x_i pouvant estimer ceux de X , en trouvant une décomposition en valeurs singulières de $R_m = U \Sigma V^T$, on retrouve le vecteur V^T de la question 2.

Ce dernier contient les v_k (k -ième vecteur de V^T) qui nous permettent d'obtenir les K -directions principales.

Comme démontré à la question 3, $y_k = x v_k$ nous donne la k -composante principale de X .

Dans la SVD de R_m , on obtient V^T qui nous donne les v_k avec lesquels on peut trouver la k -composante principale.

On sait, d'après la question 7, que $\text{Var}(y_k) = \text{Var}(xv_k) = v^T \text{Cov}(x)v$.

D'après la question 9), $E_{\text{en}}^T E_{\text{en}}$ est une approximation d'un estimateur de $\text{Cov}(x)$.

Étant donné que R_{en} forme une réalisation de E_{en} , on peut approximer $\text{Cov}(x) \approx R_{\text{en}}^T R_{\text{en}}$.

On a donc :

$$\text{Var}(y_k) = v^T (U\Sigma v^T)^T U\Sigma v^T \cdot v = v^T V \Sigma U^T U\Sigma v^T \cdot v = v^T \cdot v \cdot \Sigma^2 \cdot v^T v. \Sigma = \Sigma^T, \text{ matrice diagonale}$$

D'après la question 3, $v^T D v^T v = \lambda_k$, ici $D = \Sigma^2$, donc $\lambda_k = \sigma_k^2$.

On a démontré à la question 3 que $\text{Var}(y_k) = \lambda_k$ et nous venons de démontrer que $\lambda_k = \sigma_k^2$.

Nous avions bien montré que $\text{Var}(y_k) = \sigma_k^2$

③ Nous cherchons ici à projeter R_{en} dans l'espace engendré par les vecteurs de directions principales $\text{Vec}(v_1, \dots, v_n)$, les vecteurs de V .

On cherche donc à calculer $R_{\text{en}} \cdot V$.

Pour ce faire, partons de la SVD de R_{en} :

$$R_{\text{en}} = U\Sigma v^T \Rightarrow R_{\text{en}} \cdot V = U \underbrace{\Sigma}_{\Sigma} (v^T \cdot V) \Rightarrow R_{\text{en}} \cdot V = U \cdot \Sigma$$

Notons $\text{proj}_V(R_{\text{en}})$ le projection de R_{en} sur V soit $R_{\text{en}} \cdot V$, on obtient:

$$\text{proj}_V(R_{\text{en}}) = U \cdot \Sigma = (u_1 \dots u_n) \begin{pmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & & \\ \vdots & & \ddots & \\ 0 & & & \sigma_k \end{pmatrix}$$

On obtient bien ! $\text{proj}_V(R_{\text{en}}) = (\sigma_1 u_1, \sigma_2 u_2, \dots, \sigma_k u_k)$.

Les σ_k étant des réels, on peut décrire :

$$\text{proj}_V(R_{\text{en}}) = (\sigma_1 u_1, \sigma_2 u_2, \dots, \sigma_k u_k).$$

Ce résultat rappelle la décomposition en valeurs singulières d'une matrice.

Pour composer la matrice U de la SVD d'une matrice A , on ait :

$$u_k = \frac{1}{\sigma_k} A \cdot v_k \Rightarrow A \cdot v_k = \sigma_k u_k, \text{ où } A = R_{\text{en}} \text{ dans notre cas.}$$

5) On travaille ici avec des variables y_i centrées réduites.

On peut donc dire que $\text{Var}(y_i) = 1$, ce qui implique:

$$\sum_{i=1}^n \text{Var}(y_i) = \sum_{i=1}^n 1 = n, \text{ d'où : } \frac{1}{n} \cdot \sum_{i=1}^n \text{Var}(y_i) = \frac{1}{n} \sum_{i=1}^n 1 = \frac{1}{n} \cdot n = 1.$$

6) Pour étudier le vecteur Con_i , nous allons étudier ses composantes.

Prenons x_i une composante de x que nous voulons corrélée à y_k , une composante de y pour voir si une dépendance existe entre les VAR que nous avons étudiées. On a alors:

$$\text{Cor}(y_k, x_i) = \frac{\text{Cov}(y_k, x_i)}{\sqrt{\text{Var}(y_k)} \sqrt{\text{Var}(x_i)}} = \frac{\text{Cov}(y_k, x_i)}{\sqrt{y_k} \cdot \sqrt{x_i}}.$$

On peut reconnaitre la covariance sous forme de produit scalaire, et les écart-type sous forme de normes, on a donc:

$$\text{Cor}(y_k, x_i) = \frac{\|y_k\| \cdot \|x_i\| \cdot \cos(\gamma_k, x_i)}{\|y_k\| \cdot \|x_i\|} = \cos(\gamma_k, x_i)$$

Chaque composante de Con_i est donc un cosinus. Étant donné que $\text{Con}_i \in \mathbb{B}_K$ et que l'on étudie la boule unité \mathbb{B}_K , Cor_i contient les coordonnées d'un point de \mathbb{B}_K .

Les coordonnées étant toutes des cosinus, on peut dire que le point appartient à la boule unité \mathbb{B}_K .

Pour ce qui est de l'interprétation du positionnement de ce point, on distingue deux grands critères:

-> la norme du vecteur, plus elle se rapproche de 1, plus le point est proche du cercle de corrélation. Pour \mathbb{B}_K , cela signifie que le point est proche de la bordure de la sphère.

Plus le point se rapproche de la bordure de la sphère, meilleure est sa représentation de la caractéristique étudiée. On peut voir la norme comme une valeur de confiance à mettre en pourcentage.

-> l'angle entre les vecteurs, plus l'angle de corrélation est faible, plus la corrélation entre les VAR est forte.

Un point à l'intérieur du cercle et avec un angle faible indique que le paramètre représenté

par les VAR sera important pour différencier les données.