De la reconnaissance faciale via l'analyse en composantes principales

Germain Van Bever*
germain.van.bever@ulb.ac.be

Résumé

Nous aborderons dans ce chapitre le problème de la reconnaissance faciale. Afin de résoudre celui-ci, nous présentons un outil statistique, l'analyse en composantes principales. Celui-ci utilise abondamment l'algèbre linéaire et permet des liens inattendus entre les quantités abstraites que sont vecteurs et valeurs propres et celles, physiques, du positionnement d'observations dans l'espace. Divers rappels de statistique sont effectués, avant un traitement rigoureux de l'analyse en composantes principales et de son application.

Sommaire

1 La statistique **66**
2 Analyse en composantes principales **68**
3 Reconnaissance faciale et compression d'images **75**
4 Bibliographie **79**

*Germain Van Bever est doctorant et Aspirant FNRS au Département de Mathématique de l'Université libre de Bruxelles et à ECARES. Il est titulaire d'un Master en Sciences Mathématiques, et travaille en statistiques non paramétriques.

Introduction

Trop souvent entend-on dire « *Les statistiques ne sont pas de vraies maths* », ou, plus radical encore, « *Les statisticiens ne sont pas de vrais mathématiciens* ». Ma présentation lors de la *Brussels Summer School of Mathematics* fut initialement prévue en réaction à ces adages.

Ces notes couvrent donc un exposé dont l'objectif était double :

1. présenter un problème statistique simple et rencontré dans la vie de tous les jours,
2. dont la résolution utilise des outils mathématiques simples (l'algèbre linéaire) et les élégantes propriétés que ceux-ci induisent lorsqu'utilisés en statistique.

Le but de ce chapitre est d'exposer les bases fondamentales de la « reconnaissance faciale ». Plus précisément, ce chapitre présente en grande partie l'article (considéré comme fondateur) de Turk et Pentland « *Eigenfaces for Recognition* » (1991). L'outil-clé de leur méthode est l'*analyse en composantes principales*.

L'ACP (abréviation classique pour analyse en composantes principales, qu'on utilisera par la suite) est un outil statistique permettant de déterminer les « directions principales » dans un jeu d'observations, c'est-à-dire les directions dans lesquelles la variabilité des données est la plus grande. Très exactement, nous verrons que ces directions sont déterminées par les vecteurs propres de l'opérateur linéaire associé à la matrice de variance-covariance des observations. La démonstration (et la compréhension) d'un tel résultat demande une connaissance de base de l'algèbre linéaire et de diverses notions statistiques.

Nous supposerons que le lecteur est familier avec les concepts généraux de l'algèbre linéaire (tels que vecteurs et valeurs propres d'une application linéaire, bases, matrices d'un opérateur, etc.) et nous ne nous bornerons qu'à de brefs rappels lorsque nécessaires. À l'inverse, nous partirons de l'hypothèse que le lecteur n'est que peu à l'aise avec certaines notions de statistique, plus particulièrement de statistique multivariée et nous introduirons certaines notions dans une section dédiée.

La reconnaissance faciale est un vaste sujet, aux multiples ramifications et écueils. Sous cette dénomination, on peut entendre deux types de problèmes. Le premier est celui du repérage du visage dans une photographie (le fameux « cadre » lors de la prise de photographies à l'aide d'un appareil). Le second cherche, sur base d'un ensemble de photographies prises au préalable, à associer une nouvelle image à un visage déjà existant (ou à déterminer s'il s'agit d'une nouvelle personne). Nous nous intéresserons à ce second problème.

Outre la présente introduction, ce chapitre est constitué de trois sections. Dans la première, une introduction rapide aux outils de la statistique multivariée est présentée. La seconde section est consacrée à l'exposition de l'ACP. La dernière partie présente le problème qu'est la reconnaissance faciale et une tentative de solution.

1 La statistique

Le but de cette section est d'introduire les diverses notions statistiques classiques utilisés par la suite. Dans un but d'exhaustivité, nous introduirons les concepts aussi bien dans le cadre empirique d'un jeu d'observations que dans le cadre population (c'est-à-dire celui tout à fait général de *variables aléatoires*). Supposons

donc disposer d'observations $\underline{X}_1,\dots,\underline{X}_n$ indépendantes et identiquement distribuées dans $\mathbb{R}^N$, de même loi de probabilité F [1]. Les mesures classiques univariées (N = 1) se passent de commentaires.

Définition 1. Les observations *centrées* sont $Z_i = X_i - \bar{X}$, où $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Définition 2. La *variance* des observations est donnée par $s_X^2 = \frac{1}{n}\sum_{i=1}^{n} z_i^2$.

Définition 3. L'*écart-type* est $s_X = \sqrt{s_X^2}$.

Lorsque l'on dispose de n couples d'observations $(X_1, Y_1),\dots,(X_n, Y_n)$, il est toujours possible d'utiliser tous les outils univariés sur chacune des composantes. En plus de cela, la définition suivante est classique.

Définition 4. La *covariance* entre les variables X (l'abscisse du couple) et Y (l'ordonnée) est donnée par $s_{XY} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$. La *corrélation* est donnée par $\mathrm{Corr}(X,Y) = \frac{s_{XY}}{s_X s_Y}$.

Lorsque X et Y sont des variables aléatoires de distributions respectives F_X et F_Y et de distribution jointe $F_{(X,Y)}$ [2], les versions populations des définitions précédentes sont

$$\mu_X = E[X] = \int x \, dF_X(x),$$

$$\sigma_X^2 = E\left[(X-\mu)^2\right] = \int (x-\mu)^2 \, dF_X(x), \text{ et}$$

$$\sigma_{XY} = E[(X-\mu_X)(Y-\mu_Y)] = \int (x-\mu_X)(y-\mu_Y) \, dF_{(X,Y)}(x,y)$$

1.1 Variance-Covariance

Le cas d'un N général est bien plus intéressant. Supposons disposer d'observations

$$\underline{X}_1 = \begin{pmatrix} X_{11} \\ X_{21} \\ \vdots \\ X_{N1} \end{pmatrix}, \dots, \underline{X}_n = \begin{pmatrix} X_{1n} \\ X_{2n} \\ \vdots \\ X_{Nn} \end{pmatrix}; \qquad \underline{X}_i \in \mathbb{R}^N,$$

indépendantes et identiquement distribuées de loi de distribution F sur $\mathbb{R}^N$. L'ensemble des observations peut être collecté dans la matrice $X = (X_{ij})$, $1 \leqslant i \leqslant N$, $1 \leqslant j \leqslant n$:

$$X = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1n} \\ X_{21} & X_{22} & \dots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{N1} & X_{N2} & \dots & X_{Nn} \end{pmatrix}$$

La matrice de variance-covariance collecte les variances univariées et les covariances, pour chaque paire de variables.

1. le lecteur moins familier lira « On dispose d'observations $X_1,\dots,X_n$. »

2. Pour rappel, la fonction de répartition de X (respectivement de (X, Y)) est la fonction $F_X(x) = \mathbb{P}(X \leqslant x)$ (resp. $F_{(X,Y)}(x,y) = \mathbb{P}(X \leqslant x, Y \leqslant y)$)

Définition 5. La *matrice de variance-covariance* de X est la matrice

$$\mathrm{S} = \begin{pmatrix} s_1^2 & s_{12} & \cdots & s_{1\mathrm{N}} \\ s_{21} & s_2^2 & \cdots & s_{2\mathrm{N}} \\ \vdots & \vdots & \ddots & \vdots \\ s_{\mathrm{N}1} & s_{\mathrm{N}2} & \cdots & s_{\mathrm{N}}^2 \end{pmatrix},$$

où s_j^2 est la variance de la j^{e} composante des vecteurs $\underline{\mathrm{X}}_i$ et $s_{ij} = s_{ji}$ est la covariance entre la i^{e} et la j^{e} composante de ces vecteurs.

En notant $\bar{\underline{\mathrm{X}}} = \frac{1}{n}\sum_{i=1}^{n} \underline{\mathrm{X}}_i$, on peut voir que la définition 5 est équivalente à

$$\mathrm{S} = \frac{1}{n}\sum_{i=1}^{n} (\underline{\mathrm{X}}_i - \bar{\underline{\mathrm{X}}})(\underline{\mathrm{X}}_i - \bar{\underline{\mathrm{X}}})^{\mathrm{T}},$$

où, par convention, les vecteurs de $\mathbb{R}^{\mathrm{N}}$ sont des vecteurs-colonnes et T désigne la transposée matricielle. Le lemme suivant est utile pour un calcul rapide de la matrice de variance-covariance :

Lemme 6. *En construisant la matrice* W *telle que* $\mathrm{W}_{ij} = (\underline{\mathrm{X}}_j - \bar{\underline{\mathrm{X}}})_i$, *i.e.* $\mathrm{W} = \mathrm{X} - \underline{1}.\bar{\underline{\mathrm{X}}}^{\mathrm{T}}$, *où* $\underline{1} = (1,\ldots,1)^{\mathrm{T}} \in \mathbb{R}^{\mathrm{N}}$, *on a*

$$\mathrm{S} = \frac{1}{n}\mathrm{W}\mathrm{W}^{\mathrm{T}}.$$

La démonstration, facile et immédiate, est laissée au lecteur.

Si l'on dispose d'un N-vecteur aléatoire $\underline{\mathrm{X}}$, la matrice de variance-covariance de $\underline{\mathrm{X}}$ est définie comme

$$\Sigma = \mathrm{E}\left[(\underline{\mathrm{X}} - \mathrm{E}[\underline{\mathrm{X}}])^{\mathrm{T}}\,(\underline{\mathrm{X}} - \mathrm{E}[\underline{\mathrm{X}}])\right].$$

Nous allons voir dans la section suivante que la matrice de variance-covariance ne contient pas seulement les informations sur les variances et covariances des différentes variables, mais également sur la structure même du jeu d'observations.

2 Analyse en composantes principales

2.1 Introduction

Lorsqu'on cherche à analyser un ensemble de données, il arrive souvent que celles-ci se trouvent dans un espace de dimension « trop grande » pour permettre une analyse aisée. Plus particulièrement, dès la dimension trois, il devient difficile de déceler une quelconque structure dans les observations en se basant, par exemple, simplement sur les graphes cartésiens (par paire de composantes) de celles-ci.

Afin d'illustrer cela, nous donnons dans la figure 1 (à gauche) les graphes d'un ensemble d'observations. Rien dans ceux-ci ne permet d'affirmer que ces observations se trouvent dans un plan de l'espace, ce qui est pourtant le cas. En effet, les données ont été engendrées de la manière suivante : X et Y sont

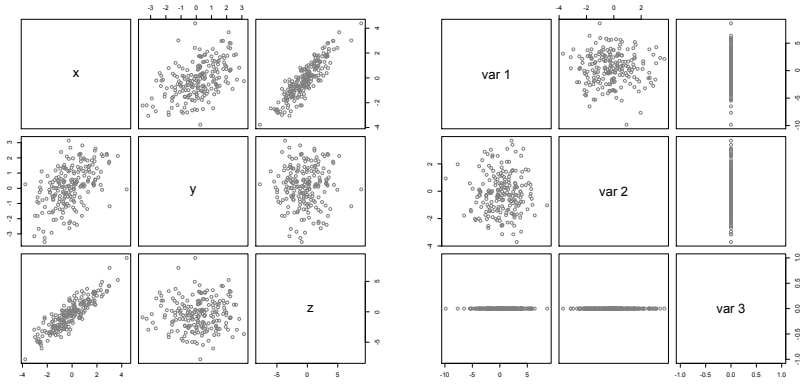

FIGURE 1 — Les graphes cartésiens d'observations générées aléatoirement dans le plan d'équation $z = 2x - y$ de $\mathbb{R}^3$. À droite, les observations après changement de base.

générées aléatoirement [3] selon une loi gaussienne dans le plan $z = 0$. La troisième composante Z est alors définie comme Z = 2X − Y.

Sans changer de base, il est difficile de déceler le lien existant entre les variables. Cependant, en choisissant une base appropriée de $\mathbb{R}^3$, il devient alors clair que les observations se trouvent exactement dans ce plan. Nous verrons que cette base peut être choisie orthonormée. Dans notre cas, on a var1, var2 et var3 les composantes d'un vecteur de $\mathbb{R}^3$ dans une base $\{v_1, v_2, v_3\}$ telle que les vecteurs v_1 et v_2 se trouvent dans le plan d'équation $\pi \equiv 2x - y - z = 0$. Le vecteur $v_3 = (2/\sqrt{6}, -1/\sqrt{6}, -1/\sqrt{6})^{\mathrm{T}}$, la version normalisée du vecteur $(2, -1, -1)^{\mathrm{T}}$, est orthogonal au plan π. Il est alors évident que $var3 = 0$ pour toutes les observations (voir la figure 1 (à droite)).

Bien entendu, les observations ne se trouvent jamais exactement dans un plan de l'espace. Les données de la figure 1 sont reprises dans la figure 2 en y ajoutant un bruit. Une observation est également ajoutée, en dehors de ce plan (en noir). Cette dernière n'est absolument pas repérable lorsqu'on regarde les observations de manière classique (à gauche). Elle le sera dans une base adéquate (comme le montre la partie droite). Dans cet exemple, la même base a été utilisée.

L'analyse en composantes principales est la méthode cherchant à expliquer la variabilité des données au moyen d'un petit nombre de combinaisons linéaires des N composantes originales. Les possibilités offertes par une telle méthode sont nombreuses :

1. L'attrait principal reste dans la découverte des liens sous-tendant les données. En effet, lorsque les données se trouvent dans un sous-espace de $\mathbb{R}^{\mathrm{N}}$, la variabilité dans une direction orthogonale à ce sous-espace sera minimale et l'une des composantes principales indiquera ce fait. Remarquons que

3. De manière précise, le vecteur (X,Y) est aléatoire de loi normale de paramètres $\mu = (0,0)$ et $\Sigma = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$.

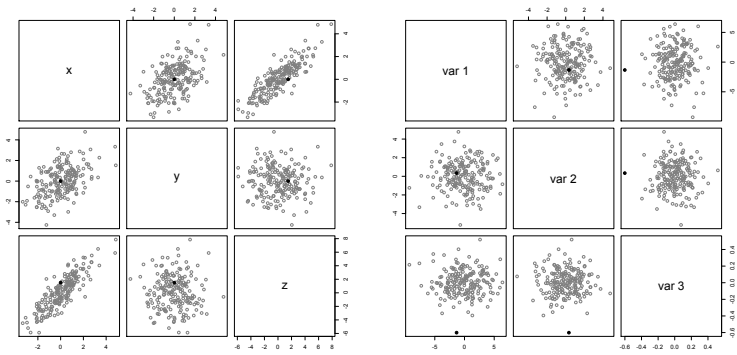

FIGURE 2 — Les graphes cartésiens d'observations générées aléatoirement. La cote z est donnée par $z = 2x - y + \epsilon$, où ϵ est une erreur gaussienne. Les bases utilisées sont les mêmes qu'en figure 1. Le point noir est clairement identifié comme valeur aberrante dans les graphes de droite.

les nouvelles variables utilisées dans les figures 1 et 2 sont exactement les composantes principales.

2. Conséquemment, la découverte de tels liens permettra une réduction de la dimension des observations. Souvent, $k << \mathrm{N}$ combinaisons linéaires expliqueront la plus grand partie de la variabilité. Il sera alors possible, au prix d'une perte d'information minimale, de réduire le jeu de données en ne considérant que les k premières composantes du jeu de données dans la base « principale » [4].
3. Last but not least, il sera également possible de détecter certaines observations aberrantes, c'est-à-dire ne suivant pas la tendance générale du jeu de données.

2.2 Définition

Soit $\underline{\mathrm{X}}$ un N-vecteur aléatoire [5] (comme toujours, le lecteur peu familier avec le concept de vecteur aléatoire pensera à la distribution empirique associée à n observations) dont la loi admet des moments finis d'ordre 2. L'ACP cherche à expliquer la variabilité des données au moyen d'un petit nombre de combinaisons linéaires des N composantes originales. Notons $\Sigma = \mathrm{Var}(\underline{\mathrm{X}})$ la matrice de variance-covariance telle que définie dans la section 1 (S dans le cas empirique).

Définition 7. Les *composantes principales* d'un vecteur aléatoire $\underline{\mathrm{X}}$ sont les combinaisons linéaires $\mathrm{Y}_i = \underline{v}_i^{\mathrm{T}}\underline{\mathrm{X}}$, $i = 1, \ldots, \mathrm{N}$, où

$$\underline{v}_1 = \underset{\underline{v} \in \mathbf{S}^{\mathrm{N}-1}}{\operatorname{argmax}} \operatorname{Var}\left(\underline{v}^{\mathrm{T}}\mathrm{X}\right) \tag{1}$$

4. Dans notre exemple, ceci revient à omettre la variable var3. Dans la figure 1, il n'y a aucune perte d'information (puisque la variable est uniformément nulle). La perte d'information sera minimale (voir plus loin) dans la figure 2.
5. Dans toute cette section, on supposera N « grand ».

et

$$\underline{v}_k = \underset{\underline{v}\in \mathbf{S}^{N-1}\cap T_{k-1}}{\operatorname{argmax}} \operatorname{Var}\left(\underline{v}^T \underline{X}\right), \quad k = 2,\ldots,N, \tag{2}$$

où

$$\mathbf{S}^{N-1} := \left\{\underline{v} \in \mathbb{R}^N \text{ t.q. } \left\|\underline{v}\right\| = 1\right\}$$

et

$$T_k = \left\{\underline{v} \in \mathbb{R}^N \text{ t.q. } \operatorname{Cov}\left(\underline{v}_i^T \underline{X}, \underline{v}^T \underline{X}\right) = 0, i = 1,\ldots,k\right\}. \tag{3}$$

Cette définition brute demande quelques explications. Premièrement, remarquons que chaque composante principale est une combinaison linéaire de variables de $\underline{X}$. Chaque composante principale est donc une *variable univariée*. Il y a N composantes principales. La première composante principale, Y_1, est la combinaison linéaire de composantes de $\underline{X}$, i.e. $\underline{v}_1^T \underline{X}$, qui maximise la variabilité d'une telle combinaison. La seconde composante principale est non corrélée avec la première (ceci est imposé par la condition dans l'équation (2) et la définition de T_k (3)). Elle est alors définie comme la combinaison linéaire de variabilité maximale parmi celles restantes. Par récurrence, la k^e composante principale est non corrélée avec les $(k-1)$ précédentes et est de variabilité maximale sous la contrainte.

Commençons par signaler que cette définition peut être mal posée. En effet, rien n'assure que ces différents maxima sont uniques [6]. Cela ne pose pas de problème supplémentaire : chaque utilisateur peut définir une règle lui permettant de trancher entre les différents maxima. De manière générale, la définition (7) reste valable dans le cas empirique. Dans ce cas, l'unicité des maximas arrive avec probabilité 1 lorsque la distribution engendrant les observations est continue. On aura par ailleurs convergence des composantes principales empiriques vers leur version population.

On appellera parfois *composantes principales* le vecteur $\underline{v}_i$ telle que $\underline{v}_i^T \underline{X}$ possède une variance maximale. Nous utiliserons indifféremment ces deux définitions par la suite, le contexte clarifiant la situation.

Il est également important de signaler que la matrice Σ induit un produit scalaire sur $\mathbb{R}^N$ défini par :

$$\underline{v} \odot_\Sigma \underline{w} = \underline{v}^T \Sigma \underline{w}. \tag{4}$$

Muni de ce produit scalaire, il est alors aisé de voir que $\underline{w} \in T_k$ si et seulement si $\underline{w} \odot_\Sigma \underline{v}_i = 0$ pour tout $i = 1,\ldots,k$. De plus, la recherche des composantes principales revient à chercher des extrema sous contraintes de la fonction

$$f_{\text{Var}} : \mathbb{R}^N \to \mathbb{R} : \underline{v} \mapsto \underline{v} \odot_\Sigma \underline{v},$$

où $\underline{v}$ est un vecteur de la sphère unité (pour la distance euclidienne). Les composantes principales sont donc les points « les plus éloignés de l'origine » pour la distance induite par le produit scalaire.

La maximisation de la fonction f_{Var} peut être pénible. Le résultat suivant est fondamental et permet un calcul immédiat des composantes principales.

6. L'existence est immédiate… puisque cherchant un extremum d'une fonction continue (en v) sur un compact.

Théorème 8. *Supposons que* Σ *est symétrique et définie positive*[7]. *Notons* $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_N (> 0)$ *ses valeurs propres et prenons* $\underline{e}_1, \ldots, \underline{e}_N$ *une base orthonormée de vecteurs propres correspondante*[8]. *Alors*

1. $$\max_{\underline{v} \in \mathbf{S}^{N-1}} f_{Var}(\underline{v}) = \lambda_1 \text{ et } \operatorname*{argmax}_{\underline{v} \in \mathbf{S}^{N-1}} f_{Var}(\underline{v}) = \underline{e}_1$$

2. *pour tout* $k = 2, \ldots, N$,

$$\max_{\underline{v} \in \mathbf{S}^{N-1} \cap \{\underline{e}_1, \ldots, \underline{e}_{k-1}\}^{\perp}} f_{Var}(\underline{v}) = \lambda_k, \text{ et } \operatorname*{argmax}_{\underline{v} \in \mathbf{S}^{N-1} \cap \{\underline{e}_1, \ldots, \underline{e}_{k-1}\}^{\perp}} f_{Var}(\underline{v}) = \underline{e}_k.$$

Démonstration. 1. Comme Σ est symétrique et définie positive, elle peut être diagonalisée dans une base orthonormée. Écrivons dès lors $\Sigma = P \Lambda P^T$, où P est une matrice orthogonale contenant les vecteurs propres $\underline{e}_1, \ldots, \underline{e}_N$ associés aux valeurs propres (classées) $\lambda_1, \ldots, \lambda_N$. Λ est la matrice diagonale contenant les valeurs propres ordonnées sur sa diagonale. La forme à maximiser est

$$f_{\mathrm{Var}}(\underline{v}) = \underline{v}^T \Sigma \underline{v} = (P^T \underline{v})^T \Lambda (P^T \underline{v}) = \underline{u}^T \Lambda \underline{u} = \sum_{i=1}^{N} \lambda_i u_i^2 = \frac{1}{\sum_{i=1}^{N} u_i^2} \sum_{i=1}^{N} \lambda_i u_i^2 \leqslant \lambda_1.$$

Ce maximum est atteint lorsque $\underline{u} = (1, 0, \ldots, 0)^T$, i.e. en $\underline{v} = P\underline{u} = \underline{e}_1$.

2. Sous la contrainte $\underline{v}^T \underline{e}_i = 0$ pour tout $i = 1, \ldots, (k-1)$, on trouve

$$\underline{v}^T \underline{e}_i = 0 \iff (P\underline{u})^T \underline{e}_i = 0 \iff \underline{u}^T P^T \underline{e}_i = 0 \iff u_i = 0.$$

Ainsi, sous la contrainte,

$$f_{\mathrm{Var}}(\underline{v}) = \underline{v}' \Sigma \underline{v} = \sum_{i=k}^{N} \lambda_i u_i^2 \leqslant \lambda_k,$$

valeur atteinte en $\underline{v} = \underline{e}_k$.

□

Pour rappel,

$$f_{\mathrm{Var}}(\underline{v}) = \underline{v}^T \Sigma \underline{v} = Var\left(\underline{v}^T X\right).$$

Ainsi, il est immédiat que la première composante principale $\underline{v}_1 = \underline{e}_1$. De plus, comme

$$\mathrm{Cov}\left(\underline{v}^T X, \underline{v}_1^T X\right) = \underline{v} \odot_{\Sigma} \underline{v}_1 = \underline{v} \Sigma \underline{v}_1 = \lambda_1 (\underline{v}^T \underline{e}_1),$$

on a

$$\mathrm{Cov}\left(\underline{v}' X, \underline{v}_1 X\right) = 0 \iff \underline{v} \perp \underline{e}_1.$$

En conclusion, on obtient le résultat suivant

7. Ce qui sera le cas pour toute matrice de variance-covariance, sauf dans de rares cas (qui n'existeront jamais dans le cas continu).

8. Pour rappel, une matrice symétrique possède toujours une base orthonormée de vecteurs propres, associés à des valeurs propres positives lorsque la matrice est définie positive.

Corollaire 9. *Les composantes principales du* N*-vecteur aléatoire* $\underline{X}$ *admettant des moments finis d'ordre deux sont les variables*

$$Y_i = \underline{e}_i^T X,$$

où les $\underline{e}_i$ *sont les vecteurs propres (ordonnés par valeurs décroissantes de valeurs propres) de la matrice de variance-covariance* Σ *de* $\underline{X}$.

2.3 Variabilité expliquée

Les vecteurs propres de la matrice de variance-covariance nous donnent donc les directions de variabilité maximale. Nous allons voir que les valeurs propres ont également une signification d'importance.

Remarquons que la *variabilité totale* de $\underline{X}$, $\sum_{i=1}^{N} \mathrm{Var}(\underline{X}_i)$, est égale à la trace de la matrice de variance-covariance. Par propriété de la trace [9], nous avons donc, puisque $\mathrm{Var}(Y_i) = \lambda_i$,

$$\mathrm{Tr}(\Sigma) = \mathrm{Tr}(P\Lambda P^T) = \sum_{i=1}^{N} \lambda_i.$$

Ainsi, la part de la variabilité totale expliquée par les k première composantes principales est donnée par

$$p_k = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{N} \lambda_i}.$$

Si pour un petit nombre k_0 de composantes principales, une grande proportion de la variabilité totale est expliquée, celles-ci peuvent remplacer avantageusement les variables X_i d'origine *sans grande perte d'information*. Nous tenons ici l'une des principales utilités de l'analyse en composantes principales : la *réduction de dimension*. Remarquons que $p_k = 1$ si et seulement si $k = N$ ou la matrice de covariance est dégénérée (signifiant que toutes les observations sont contenues dans un même sous-espace de $\mathbb{R}^N$). Pratiquement, l'utilisateur regardera un graphe des p_k en fonction de k et décidera du nombre de composantes principales à utiliser.

Les vecteurs propres disposent également d'une interprétation supplémentaire. Notons $\underline{e}_i = (e_{i1}, \ldots, e_{iN})^T$, le i^e vecteur propre. Le résultat suivant ne sera pas prouvé ici.

Lemme 10. *Pour tout* i *et* $k = 1, \ldots, N$,

$$\mathrm{Corr}\,(Y_i, X_k) = \sqrt{\frac{\lambda_i}{\Sigma_{kk}}}\, e_{ik}.$$

La taille de e_{ik} mesure l'importance de X_k dans la composante principale $Y_i = \underline{e}_i^T \underline{X}$!

9. Rappelons que, pour des matrices A, B et C de dimensions compatibles, on a Tr(ABC) = Tr(BCA).

2.4 Approximation par des sous-espaces

Soit $\mathrm{W} = (\underline{w}_1, \ldots, \underline{w}_k)$ une matrice $\mathrm{N} \times k$ ($k < \mathrm{N}$) telle que $\mathrm{W}^{\mathrm{T}}\mathrm{W} = \mathrm{I}_k$. Notons $\pi_{\underline{a},\mathrm{W}}$ le sous-espace affin passant par $\underline{a}$ et de vecteurs générateurs $\underline{w}_1, \ldots, \underline{w}_k$, i.e.

$$\pi_{a,\mathrm{W}} = \left\{ a + \sum_{i=1}^{k} \alpha_i \underline{w}_i \text{ t.q. } \alpha_i \in \mathbb{R}, i = 1, \ldots, k \right\}.$$

Soit $\mathrm{P}_{a,\mathrm{W}}$ la projection orthogonale sur $\pi_{a,\mathrm{W}}$. Cette application linéaire est donnée par

$$\mathrm{P}_{\underline{a},\mathrm{W}}(\underline{x}) = \underline{a} + \mathrm{W}\left[\mathrm{W}^{\mathrm{T}}\mathrm{W}\right]^{-1} \mathrm{W}^{\mathrm{T}}(\underline{x} - \underline{a}) = \underline{a} + \mathrm{W}\mathrm{W}^{\mathrm{T}}(\underline{x} - \underline{a}).$$

Les composantes principales permettent de donner une réponse rapide à la question suivante :

Question. Soit k une dimension fixée. Quel est le sous-espace affin qui approxime le mieux l'échantillon $\underline{\mathrm{X}}_1, \ldots, \underline{\mathrm{X}}_n$ au sens qu'il minimise

$$\sum_{i=1}^{n} \left\| \underline{\mathrm{X}}_i - \mathrm{P}_{\underline{a},\mathrm{W}}(\underline{\mathrm{X}}_i) \right\|^2 ?$$

Le théorème suivant montre que le meilleur sous-espace est le translaté en $\bar{\underline{\mathrm{X}}}$ du sous-espace de dimension k engendré par les k premières composantes principales.

Théorème 11. *Posons* $\mathrm{W}_{\mathrm{CP}}^{(k)} = (\underline{e}_1, \ldots, \underline{e}_k)$. *Alors*

$$\operatorname*{argmin}_{\pi_{\underline{a},\mathrm{W}}} \left\| \underline{\mathrm{X}}_i - \mathrm{P}_{\underline{a},\mathrm{W}}(\underline{\mathrm{X}}_i) \right\|^2 = \pi_{\bar{\underline{\mathrm{X}}},\mathrm{W}_{\mathrm{CP}}^{(k)}}.$$

Démonstration. Puisque $\mathrm{I}_{\mathrm{N}} - \mathrm{W}\mathrm{W}^{\mathrm{T}}$ est symétrique et idempotente [10], on trouve

$$\begin{aligned}
\sum_{i=1}^{n} \left\| \underline{\mathrm{X}}_i - \mathrm{P}_{\underline{a},\mathrm{W}}(\underline{\mathrm{X}}_i) \right\|^2 &= \sum_{i=1}^{n} \left\| \underline{\mathrm{X}}_i - (\underline{a} + \mathrm{W}\mathrm{W}^{\mathrm{T}}(\underline{\mathrm{X}}_i - \underline{a})) \right\|^2 \\
&= \sum_{i=1}^{n} \left\| (\underline{\mathrm{X}}_i - \underline{a}) - \mathrm{W}\mathrm{W}^{\mathrm{T}}(\underline{\mathrm{X}}_i - \underline{a}) \right\|^2 \\
&= \sum_{i=1}^{n} \left\| \left(\mathrm{I}_{\mathrm{N}} - \mathrm{W}\mathrm{W}^{\mathrm{T}}\right)(\underline{\mathrm{X}}_i - \underline{a}) \right\|^2 \\
&= \sum_{i=1}^{n} (\underline{\mathrm{X}}_i - \underline{a})^{\mathrm{T}} (\mathrm{I}_{\mathrm{N}} - \mathrm{W}\mathrm{W}^{\mathrm{T}})(\underline{\mathrm{X}}_i - \underline{a}).
\end{aligned}$$

En écrivant [11] $(\underline{\mathrm{X}}_i - \underline{a}) = (\underline{\mathrm{X}}_i - \bar{\underline{\mathrm{X}}}) + (\bar{\underline{\mathrm{X}}} - \underline{a})$, on trouve que cette dernière quantité vaut

$$\sum_{i=1}^{n} (\underline{\mathrm{X}}_i - \bar{\underline{\mathrm{X}}})^{\mathrm{T}} (\mathrm{I}_{\mathrm{N}} - \mathrm{W}\mathrm{W}^{\mathrm{T}})(\underline{\mathrm{X}}_i - \bar{\underline{\mathrm{X}}}) + n \sum_{i=1}^{n} (\bar{\underline{\mathrm{X}}} - \underline{a})^{\mathrm{T}} (\mathrm{I}_{\mathrm{N}} - \mathrm{W}\mathrm{W}^{\mathrm{T}})(\bar{\underline{\mathrm{X}}} - \underline{a}).$$

10. Une matrice A est idempotente si $\mathrm{A}^2 = \mathrm{A}$.
11. On se rappelera que $\sum_{i=1}^{n} (\underline{\mathrm{X}}_i - \bar{\underline{\mathrm{X}}}) = 0$.

En utilisant les différentes propriétés de la trace et le fait que $\mathrm{Tr}(c) = c$ quand c est un scalaire, on voit que l'expression ci-dessus n'est autre que

$$n\left(\mathrm{Tr}\left[\Sigma(\mathrm{I_N} - \mathrm{WW^T})\right]\right) + n\left\|\left(\mathrm{I_N} - \mathrm{WW^T}\right)\left(\underline{\bar{X}} - \underline{a}\right)\right\|^2,$$

ou encore

$$n\left(\mathrm{Tr}\left[\Sigma\right]\right) - n\left(\mathrm{Tr}\left[\Sigma\mathrm{WW^T}\right]\right) + n\left\|\left(\mathrm{I_N} - \mathrm{WW^T}\right)\left(\underline{\bar{X}} - \underline{a}\right)\right\|^2,$$

quantité à minimiser. Le premier terme ne dépend ni de $\underline{a}$ ni de W. Le dernier impose $\underline{a} = \underline{\bar{X}}$. Le deuxième, quant à lui, est

$$\mathrm{Tr}\left[\Sigma\mathrm{WW^T}\right] = \mathrm{Tr}\left[\Sigma\sum_{i=1}^{n}\underline{w}_i\underline{w}_i^{\mathrm{T}}\right] = \sum_{i=1}^{n}\underline{w}_i^{\mathrm{T}}\Sigma\underline{w}_i$$

et est maximisé lorsque $\underline{w}_i = \underline{e}_i$. □

2.5 Quelques mots sur l'inférence

De manière générale, on dispose d'observations et il faut dès lors faire une différence entre les versions empiriques et populations. On donne ici, par souci d'exhaustivité, deux résultats permettant de préciser la convergence des estimateurs des variances des composantes principales (i.e. les valeurs propres de la matrice de variance-covariance empirique) vers leur version population. On peut montrer que

Théorème 12. *Soient* $\hat{\lambda} = \left(\hat{\lambda}_1, \ldots, \hat{\lambda}_{\mathrm{N}}\right)$ *et* $\lambda = (\lambda_1, \ldots, \lambda_{\mathrm{N}})$. *Alors, si* $n \to \infty$, *en loi,*

$$\sqrt{n}\left(\hat{\lambda} - \lambda\right) \to \mathcal{N}\left(\underline{0}, 2\Lambda^2\right).$$

Corollaire 13. *Soit* $f : \mathbb{R}^{\mathrm{N}} \to \mathrm{R}$ *une fonction de classe* C^1 *dans un voisinage de* λ. *Alors, si* $n \to \infty$,

$$\sqrt{n}\left(f(\hat{\lambda}) - f(\lambda)\right) \to \mathcal{N}_1\left(\underline{0}, 2\tau^2\right),$$

où $\tau^2 \coloneqq \sum_{i=1}^{\mathrm{N}} \lambda_i^2 \left(\frac{\partial f}{\partial \lambda_i}(\lambda)\right)^2$.

Ce dernier résultat signifie en particulier que les estimateurs des proportions de variance expliquée sont également convergent et donnent un sens à la prise de décision basée sur ceux-ci.

3 Reconnaissance faciale et compression d'images

On présentera dans cette section deux applications de l'analyse en composantes principales : la reconnaissance faciale et la compression d'images. Ces deux applications se basent sur la même idée.

3.1 Reconnaissance faciale

Avant de débuter, il est nécessaire de préciser ce qui est entendu par *reconnaissance faciale*. En effet, plusieurs méthodes tombent sous cette dénomination. Nous ne parlerons pas ici du problème de détection consistant à localiser le visage lors de la prise d'une photographie (ou par les logiciels de détection sur les réseaux sociaux par exemple). Le problème d'intérêt ici est le suivant. On dispose au préalable d'un ensemble de n visages de différentes personnes (parfois plusieurs pour une même personne). On cherche à déterminer si une nouvelle image que l'on reçoit appartient à une des personnes connues et, si c'est le cas, à laquelle. Ce problème est célèbre et est, depuis 20 ans, une aire de recherche extrêmement active. En effet, sur les deux dernière années, plus d'une centaine d'articles scientifiques sont parus sur ce sujet. Le lecteur intéressé par ce problème trouvera une masse d'informations importante en ligne, par exemple sur l'excellent site `http://www.face-rec.org/`.

De nombreuses tentatives de solution existent à ce jour. Celles-ci se classent en deux catégories bien distinctes. La première demande à l'utilisateur de signaler sur chaque photographies des points-clés : le bout du nez, les pupilles, les coins de la bouche et autres caractéristiques morphologiques. Ces méthodes possèdent de nombreux avantages et inconvénients. Elles sont en général rapide et relativement facile à implémenter. Elles seront les plus efficaces pour des échantillons de très grandes tailles (plusieurs milliers, voire millions de visages). En contrepartie, ne disposant que des caractéristiques et non de l'image, elle ne permettra à aucun moment de revenir en arrière et de reconstruire une approximation du visage par exemple. Elles seront en général moins efficace que les algorithmes de seconde catégorie. Ceux-ci cherchent en effet à reconnaître les visages, sans apport extérieur aucun, en se basant uniquement sur les images. Bien que plus efficaces, ces méthodes sont en général plus gourmandes en temps de calcul et donc souvent limitées à de petites tailles d'échantillon (allant jusqu'à quelques milliers). Elles sont également considérées comme plus sûres.

Notons $\underline{\mathrm{V}}_i$, $i = 1,\ldots,n$ les visages. Chacun de ceux-ci est une image, décrite dans son fichier, pixel par pixel. De manière exacte, chaque pixel est un mélange des trois couleurs primaires et est décrit par conséquent par trois nombres. Afin de simplifier la chose, nous supposerons disposer d'une image en noir et blanc, composée seulement, pour chaque pixel, d'un pourcentage indiquant l'intensité lumineuse de ce pixel [12]. Les différentes images sont supposées être de la même taille. Sous cette hypothèse, chaque image est alors transformée en vecteur. Chacun de ceux-ci est de même dimension N. On notera également $\underline{\mathrm{V}}_i$ ces vecteurs.

L'idée décrite ci-après est tirée de Turk et Pentland (1991). Ceux-ci emploient l'analyse en composantes principales afin de déterminer des *visages propres*, directions de $\mathbb{R}^{\mathrm{N}}$ sur lesquelles projeter. Une fois ces projections faites, une simple analyse de distance est réalisée. Clarifions ceci.

La moyenne de visage est

$$\underline{\bar{\mathrm{V}}} = \frac{1}{n}\sum_{i=1}^{n}\underline{\mathrm{V}}_i.$$

Pour chaque visage, on définit

$$\underline{\mathrm{W}}_i = \underline{\mathrm{V}}_i - \underline{\bar{\mathrm{V}}}.$$

12. Une autre solution consiste à se limiter simplement à l'une des trois couleurs ; alternative parfois choisie dans la littérature.

Nous avons vu dans le lemme 6 que la matrice de variance-covariance des visages Σ est proportionnelle à AA^T, où A est la matrice $N \times n$ définie par

$$A = \left(\underline{W}_1, \ldots, \underline{W}_n\right).$$

Les quantités d'intérêts étant les vecteurs propres, il est équivalent de chercher ceux de la matrice de AA^T. Dans le cas qui nous intéresse, la matrice est de taille gigantesque. Pour une « petite » taille d'image (par exemple, un rectangle de 400×300 pixels), la matrice sera de taille $400 * 300 \times 400 * 300 = 120000 \times 120000$. Ceci demande un temps de calcul extrêmement long (et une patience infinie) ! Cependant, le rang de cette matrice est « petit » comparé à sa taille (n est en général bien inférieur à 120000, et seules $(n-1)$ valeurs propres seront non nulles). Il est alors possible de contourner le problème de la manière suivante.

Considérons la matrice $L = A^T A$ de taille $n \times n$. Soit $\underline{f}_i$ un vecteur propre de L (associé à la valeur propre λ_i).

Lemme 14. *Si $\underline{f}$ est un vecteur propre de $A^T A$, alors $A\underline{f} = \underline{e}$ est un vecteur propre de AA^T, associé à la même valeur propre* [13].

Démonstration. La preuve est immédiate au vu de l'implication suivante :

$$L\underline{f}_i = \lambda_i \underline{f}_i \Rightarrow AA^T A\underline{f}_i = \lambda_i A\underline{f}_i.$$

□

Ainsi, de chaque vecteurs propres $\underline{f}_i$ de L, on tire un vecteur propre $\underline{e}_i = A\underline{f}_i$ de Σ. Puisqu'on trouvera n vecteurs propres de la matrice L et que les vecteurs propres associés aux valeurs propres non nulles de Σ sont au nombre maximum de n, on trouvera l'ensemble des vecteurs propres de valeurs propres non nulles de Σ.

Ces vecteurs propres sont des vecteurs de taille N et correspondent à une image. Ces images (ainsi que ces vecteurs) sont appelés les *visages propres*. En particulier, tout visage peut être résumé par la données de ses coordonnées dans l'espace engendré par les visages propres (qu'on appellera *l'espace des visages*).

Tout nouveau visage $\underline{\tilde{V}}$ peut être projeté dans l'espace des visages. On retiendra alors ses coordonnées dans celui-ci :

$$\omega_i = \underline{e_i}^T \left(\underline{\tilde{V}} - \underline{\bar{V}}\right).$$

On choisira de manière générale de ne garder que $k < n$ visages propres. Ainsi, tout visage (nouveau ou ancien) peut être décrit (sans trop de perte) par

$$\Omega(\underline{V}) = (\omega_1(\underline{V}), \ldots, \omega_k(\underline{V})).$$

Maintenant, comment associer les nouvelles images aux visages existants ? Si l'on dispose d'un nouveau visage $\underline{\tilde{V}}$, on calculera sa distance à l'image i ($i = 1, \ldots, n$) par

$$\epsilon_i^2 = \left\|\Omega(\underline{\tilde{V}}) - \Omega(\underline{V}_i)\right\|^2$$

13. $\underline{e}$ est donc également un vecteur propre de Σ, pour la même valeur propre.

On décidera d'associer le nouveau visage à l'individu j tel que $j = \text{argmin}_i \epsilon_i^2$.

Jusque maintenant, toute nouvelle image est associée à un des visages existant. Deux détails restent à régler en pratique. Cette règle de décision suppose évidemment que le nouveau visage appartient à la famille de visages déjà enregistrés. Que se passe-t-il s'il s'agit du visage d'une personne n'appartenant pas à la base de données ? Pire encore, que se passe-t-il s'il ne s'agit même pas d'un visage ?

Si l'on dispose d'un nouveau visage n'appartenant pas à l'ensemble des visages existants, il sera, dans l'espace de visages, loin de tous les autres $\Omega(\underline{V}_i)$. Il convient dès lors d'imposer une règle supplémentaire. Lorsque les ${\epsilon_i}^2$ sont tous grand, plus grand qu'une bonne fixée disons K, on déclarera la présence d'un nouveau visage. La valeur K à utiliser peut être déterminée de manière empirique par diverses méthodes.

De même, en présence d'une image ne représentant pas un visage, il deviendra intéressant de mesurer la distance existante entre l'image et sa projection dans l'espace des visages. Celle-ci sera grande dans le cas d'une image n'étant pas un visage… permettant sa détection. Un nouveau visage sera donc éloigné des visages existant (les ϵ_i^2 seront grands), mais sera proche de l'espace des visages (peu de différence entre $\underline{\tilde{V}}$ et sa projection). Ce ne sera pas le cas pour l'image, par exemple, d'une fleur.

Signalons également une grande force d'une telle méthode. Très rapidement, même pour de petites valeurs de k, les images que l'on obtient « ressemblent » à l'original. Ceci signifie que, même si l'on ne dispose plus de l'image originale, il est encore possible de se faire une idée de ce à quoi elle ressemblait. Ceci est totalement à l'opposé des méthodes dites « du premier type » (voir plus haut).

Remarquons que beaucoup de problèmes n'ont pas été abordés ici. Pour commencer, nous n'avons à aucun moment parlé de l'arrière-plan se trouvant sur les images. Celui-ci peut parfois jouer un rôle important et mener à de mauvaises définitions de l'espace des visages. Un arrière-plan neutre est bien entendu souhaité.

Déjà abordé, le problème de détection de la tête reste d'importance. Il est impératif que les visages soient de même taille. Ceci est d'autant plus difficile si l'on reste sous la contrainte d'absence d'actions à effectuer de la part de l'utilisateur.

Il peut par ailleurs être intéressant de s'intéresser aux différentes positions que peut prendre un visage. En particulier, il peut être utile d'inclure dans la base de données plusieurs images d'une même personne prise dans des positions différentes. Une nouvelle photographie sera alors à une distance faible d'une image dans une position « similaire » (voir, par exemple, les photos prises lors d'une arrestation).

La luminosité peut également demander sa part de correction. Une luminosité trop grande peut absorber à elle seule une ou plusieurs composantes. Il convient d'y prendre garde.

3.2 Compression d'images

Les mêmes idées sont utilisées lors de la compression d'images. Plusieurs méthodes existent, les plus performantes utilisant l'analyse en composantes principales. Le principe est exactement le même que dans la section précédente. On dispose d'images $\underline{V}_i$, qui, au départ, demandent un espace de stockage de taille $n * \mathrm{N}$. Après réduction, les seules quantités à stocker sont les différents visages propres (rappelons qu'on en garde k), ainsi que les coordonnées des images $\underline{V}_i$ dans l'espace des visages. Cette information est de taille $k * (\mathrm{N} + n)$. Lorsque n

est négligeable par rapport à N (un petit calcul montrera que c'est effectivement le cas avec les résolutions de photographies actuelles), on disposera alors d'un facteur de compression k/n.

Remarquons également que, de manière générale, plus les images sont « proches » (i.e. images représentant le même type d'objet), plus il sera intéressant d'utiliser l'analyse en composantes principales. Si l'on dispose d'images extrêmement différentes, l'étude de la variabilité de celles-ci aboutira à des valeurs propres globalement toutes identiques, ce qui demandera un grand nombre de composantes afin de permettre une résolution acceptable.

Cette méthode possède les mêmes avantages et inconvénients que décrits plus haut et fait encore aujourd'hui l'objet de nombreuses recherches.

4 Bibliographie

[1] M. Turk et A. Pentland, « Eigenfaces for recognition », *Journal of Cognitive Neuroscience*, vol. 3, no. 1, p. 71–86, 1991.