Apprentissage en grande dimension (Partie #5)

Thomas Verdebout

Université de Lille

Contenu du cours

1. Introduction.
2. Apprentissage supervisé: régression.
3. Apprentissage supervisé: classification.
4. Apprentissage non-supervisé

4: Apprentissage non-supervisé: analyse en composantes principales.

Supposons qu'on a un p-vecteur aléatoire $\mathbf{X}$. Nous aspirons à transformer $\mathbf{X}$ en un nouveau vecteur $\mathbf{Y}$ de dimension inférieure q (de préférence q est bien plus petit que p). Cette transformation devrait être opérée en minimisant la perte d'"information contenue" dans le vecteur original $\mathbf{X}$.

$\Longrightarrow$ *réduction de la dimension de données*

En fin de compte nous espérons que $\mathbf{Y}$ soit plus facile à interpréter que $\mathbf{X}$:

$\Longrightarrow$ *meilleure interprétation*

Exemple. Supposons qu'on ait sauvegardé pour une étude financière plusieurs centaines de prix d'actions sur base quotidienne (p.ex. S&P 500 contient 500 actions!) Il s'avère très difficile de tirer des conclusions statistiques à partir de vecteurs si larges. Comme bon nombre des prix d'actions sont fortement corrélés, on pourrait songer à en enlever une certaine quantité. Cependant, chaque prix peut contenir à lui seul des informations précieuses sur le marché, voilà pourquoi on ne peut pas les enlever de façon arbitraire!

Afin de “comprimer” le vecteur $\mathbf{X}$ nous allons utiliser une fonction H:

$$\mathbf{X} \mapsto H(\mathbf{X}) = \mathbf{Y}, \quad \mathbf{Y} \in \mathbb{R}^q,$$

où $q \leq p$. Souvent il est désirable que q soit beaucoup plus petit que p: $q \ll p$.

L'approche la plus simple consiste à choisir pour H une fonction linéaire, et alors $\mathbf{Y} = \mathbf{HX}$ avec $\mathbf{H} \in \mathbb{R}^{q \times p}$.

Question: quel pourrait être un choix intelligent pour $\mathbf{H}$?

L'ACP utilise la stratégie suivante: soit $\mathbf{H} = (\mathbf{h}_1, \ldots, \mathbf{h}_q)'$.

(1) Commençons par $q = 1$ et choisissons $\mathbf{h}_1$ tel que

$$\mathrm{Var}(Y_1) = \mathrm{Var}(\mathbf{h}_1'\mathbf{X})$$

soit maximal sous la constrainte $\|\mathbf{h}_1\| = 1$. (Sans contrainte ceci ne fait aucun sens!)

$$\mathbf{h}_1 = \mathrm{argmax} \left\{ \mathrm{Var}(\mathbf{h}'\mathbf{X}) : \mathbf{h} \in \mathbb{R}^p,\ \|\mathbf{h}\| = 1 \right\}.$$

Remarquons que

$$\mathrm{Var}(\mathbf{h}_1'\mathbf{X}) = \mathrm{Var}(\|\langle \mathbf{h}_1, \mathbf{X} \rangle \mathbf{h}_1\|).$$

En d'autres mots: projetons $\mathbf{X}$ sur le sous-espace 1-dimensionnel L_1 qui donne la plus grande (possible) variabilité pour la norme de la projection.

(2) Pour $q = 2$ choisissons $\mathbf{h}_1$ comme avant et définissons

$$\mathbf{h}_2 = \operatorname{argmax} \left\{ \operatorname{Var}(\mathbf{h}'\mathbf{X}) : \mathbf{h} \in \mathbb{R}^p,\ \|\mathbf{h}\| = 1, \mathbf{h} \perp \mathbf{h}_1 \right\}.$$

En d'autres mots: projetons $\mathbf{X}$ sur le sous-espace 1-dimensionnel L_2 qui donne la plus grande (possible) variabilité pour la norme de la projection, sous la contrainte que $L_2 \perp L_1$.

(3) Pour $q = 3$ choisissons $\mathbf{h}_1$ et $\mathbf{h}_2$ comme avant et définissons

$$\mathbf{h}_3 = \operatorname{argmax} \left\{ \operatorname{Var}(\mathbf{h}'\mathbf{X}) : \mathbf{h} \in \mathbb{R}^p,\ \|\mathbf{h}\| = 1, \mathbf{h} \perp \mathbf{h}_1, \mathbf{h}_2 \right\}.$$

(4) Continuons de la même manière pour $q = 4, \ldots, p$. Nous appelons $Y_i = \mathbf{h}_i'\mathbf{X}$ la *i-ème composante principale* de $\mathbf{X}$.

Remarquons que les Y_i ne dépendent pas du nombre de composantes que nous voulons inclure. D'où le fait que nous pouvons toujours supposer que $\mathbf{H} \in \mathbb{R}^{p \times p}$, $\mathbf{Y} \in \mathbb{R}^p$ et que par après nous rejetons les composantes Y_i, $i > q$.

En fin de compte nous avons

$$\mathbf{Y} = \mathbf{H}\mathbf{X},$$

où $\mathbf{H}$ est une matrice orthogonale, et donc $\mathbf{H}$ effectue une rotation ou bien une réflection de nos données.

Une première question est de savoir comment implémenter $\mathbf{H}$ en pratique.

Puisque $\mathrm{Var}(\mathbf{v}'\mathbf{X}) = \mathbf{v}'\mathbf{\Sigma}\mathbf{v}$, la détermination de $\mathbf{H}$ sera facile à l'aide des deux lemmes suivants.

Lemma

Soit $\mathbf{\Sigma}$ une matrice symétrique et définie positive, de valeurs propres $\lambda_1 > \lambda_2 > \ldots > \lambda_p (> 0)$ et soient $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p$ les vecteurs propres associés. Alors

$$\max \left\{ \mathbf{v}'\mathbf{\Sigma}\mathbf{v} : \ \mathbf{v} \in \mathbb{R}^p, \|\mathbf{v}\| = 1 \right\} = \lambda_1,$$

et

$$\mathrm{argmax} \left\{ \mathbf{v}'\mathbf{\Sigma}\mathbf{v} : \ \mathbf{v} \in \mathbb{R}^p, \|\mathbf{v}\| = 1 \right\} = \mathbf{e}_1.$$

Remarque. Ceci reste vrai si nous demandons seulement que $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$. Néanmoins le terme maximisant n'est alors plus unique.

Preuve.

Les vecteurs propres forment une base orthonormée de $\mathbb{R}^p$ et donc nous pouvons exprimer tout vecteur v comme $\mathbf{ev} = \sum_{i=1}^{p} \alpha_i \mathbf{e}_i$. La condition $\|\mathbf{v}\| = 1$ est équivalente par Pythagore à

$$1 = \|\mathbf{v}\|^2 = \sum_{i=1}^{p} \|\alpha_i \mathbf{e}_i\|^2 = \sum_{i=1}^{p} \alpha_i^2.$$

Remarquons que

$$\mathbf{\Sigma v} = \mathbf{\Sigma}\left(\sum_{i=1}^{p} \alpha_i \mathbf{e}_i\right) = \sum_{i=1}^{p} \alpha_i \mathbf{\Sigma e}_i = \sum_{i=1}^{p} \lambda_i \alpha_i \mathbf{e}_i$$

et donc par l'orthonormalité des vecteurs propres il suit que

$$\mathbf{v'\Sigma v} = \left(\sum_{i=1}^{p} \alpha_i \mathbf{e}_i\right)' \sum_{i=1}^{p} \lambda_i \alpha_i \mathbf{e}_i = \sum_{i=1}^{p} \lambda_i \alpha_i^2 \underbrace{\mathbf{e}_i' \mathbf{e}_i}_{=1}.$$

Clairement, par la contrainte $\sum_{i=1}^{p} \alpha_i^2 = 1$, cette somme est maximisée si $\alpha_1 = 1$ et $\alpha_i = 0$ pour $i = 2, \ldots, p$. Le maximum est λ_1 et le maximisant est e_1. (Discuter de l'unicité.) □

Lemma

Soit $\mathbf{\Sigma}$ *une matrice symétrique et définie positive, de valeurs propres* $\lambda_1 > \lambda_2 > \ldots > \lambda_p (> 0)$ *et soient* $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p$ *les vecteurs propres associés. Alors pour* $k = 2, \ldots, p$

$$\max \left\{ \mathbf{v}' \mathbf{\Sigma} \mathbf{v} : \ \mathbf{v} \in \mathbb{R}^p, \|\mathbf{v}\| = 1, \mathbf{v} \perp \mathbf{e}_1, \ldots, \mathbf{e}_{k-1} \right\} = \lambda_k,$$

et

$$\operatorname{argmax} \left\{ \mathbf{v}' \mathbf{\Sigma} \mathbf{v} : \ \mathbf{v} \in \mathbb{R}^p, \|\mathbf{v}\| = 1, \mathbf{v} \perp \mathbf{e}_1, \ldots, \mathbf{e}_{k-1} \right\} = \mathbf{e}_k.$$

Preuve. La preuve est presque la même. Il nous suffit juste de remarquer que les conditions

$$\mathbf{v} \perp \mathbf{e}_1, \ldots, \mathbf{e}_{k-1} \quad \text{et} \quad \|\mathbf{v}\| = 1$$

impliquent que $\mathbf{v} = \sum_{i=k}^{p} \alpha_i \mathbf{e}_i$ avec $\sum_{i=k}^{p} \alpha_i^2 = 1$. □

Il est commun de centrer d'abord les données avant que les composantes principales ne soient implémentées. Cela mène à la définition suivante:

Definition

Soit $\mathbf{X}$ *un p-vecteur aléatoire de moyenne* $\boldsymbol{\mu}$ *et de matrice de variance-covariance définie positive* $\boldsymbol{\Sigma}$*. Alors la transformation*

$$\mathbf{X} \mapsto \mathbf{Y} = \mathbf{H}(\mathbf{X} - \boldsymbol{\mu})$$

est appelée transformation des composantes principales*.*

Proposition

Soit Σ une matrice de variance-covariance non-singulière d'un certain p-vecteur aléatoire X et supposons que $\lambda_1 > \lambda_2 > \ldots > \lambda_p$ sont ses valeurs propres et que $\beta = (\beta_1, \ldots, \beta_p)$ est la matrice des vecteurs propres correspondants. Soit Y le vecteur des composantes principales de X. Alors

$$(i) \quad \mathbf{H} = \beta',$$

$$(ii) \quad \mathrm{Var}(\mathbf{Y}) = \mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_p),$$

$$(iii) \quad \mathrm{tr}(\mathbf{\Lambda}) = \mathrm{tr}(\mathbf{\Sigma}) = E\|\mathbf{X} - \boldsymbol{\mu}\|^2,$$

$$(iv) \quad |\mathbf{\Lambda}| = |\mathbf{\Sigma}|.$$

Preuve. Nous avons déjà démontré (i). Le théorème spectral nous donne

$$\mathrm{Var}(\mathbf{Y}) = \mathrm{Var}(\mathbf{HX}) = \mathbf{H\Sigma H}' = \mathbf{H}\boldsymbol{\beta}\boldsymbol{\Lambda}\boldsymbol{\beta}'\mathbf{H}' = \boldsymbol{\Lambda},$$

ce qui prouve (ii). De même,

$$\mathrm{tr}(\boldsymbol{\Sigma}) = \mathrm{tr}(\boldsymbol{\beta}\boldsymbol{\Lambda}\boldsymbol{\beta}') = \mathrm{tr}(\boldsymbol{\beta}'\boldsymbol{\beta}\boldsymbol{\Lambda}) = \mathrm{tr}(\boldsymbol{\Lambda}),$$

et

$$|\boldsymbol{\Sigma}| = |\boldsymbol{\beta}\boldsymbol{\Lambda}\boldsymbol{\beta}'| = |\boldsymbol{\beta}||\boldsymbol{\Lambda}||\boldsymbol{\beta}'| = |\boldsymbol{\Lambda}|,$$

ce qui établit (iii) et (iv). □

Les composantes principales sont le plus facilement intelligibles dans le contexte d'un p-vecteur aléatoire $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Souvenez-vous que les contours de la densité de X sont donnés par

$$\{\mathbf{x} |\, (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = c^2\}.$$

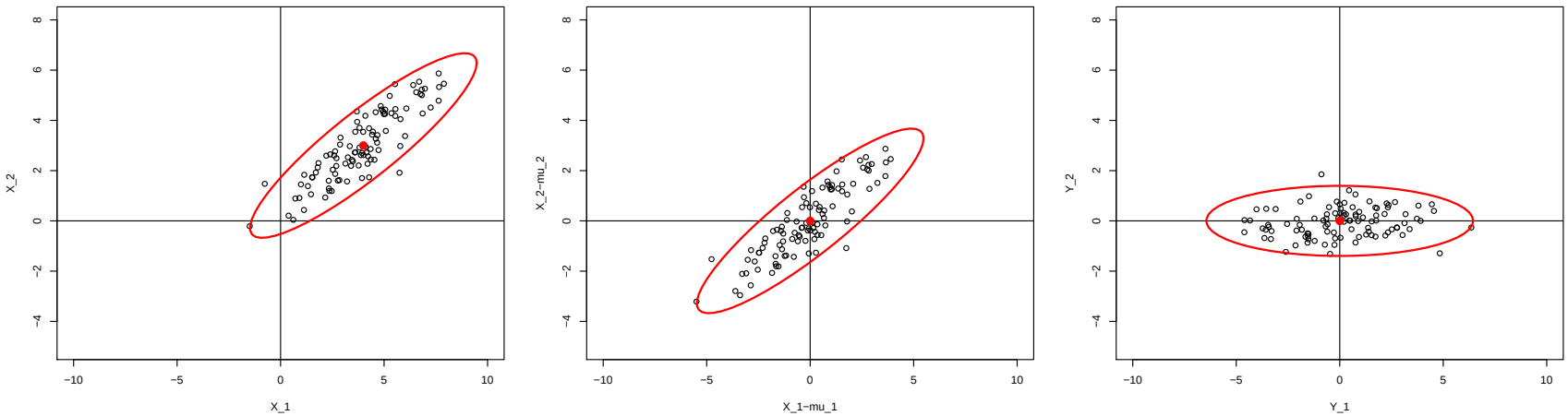
Vous avez vu/verrez aux exercices que ces contours forment des ellipsoïdes de centre $\boldsymbol{\mu}$ et d'axes principaux $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$, où les $\boldsymbol{\beta}_i$ sont des vecteurs propres de $\boldsymbol{\Sigma}$. Les rayons de ces axes sont $c\sqrt{\lambda_i}$.

Les composantes principales $\mathbf{Y} = \mathbf{H}(\mathbf{X} - \boldsymbol{\mu}) \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Lambda})$, et de ce fait les nouveaux contours correspondent à

$$\{\mathbf{y} |\, \mathbf{y}' \boldsymbol{\Lambda}^{-1} \mathbf{y} = c^2\}.$$

Nous avons donc des ellipsoïdes de centre $\mathbf{0}$, des axes principaux canoniques et les rayons sont les mêmes que précédemment.

La figure ci-dessous montre la transformation des composantes principales appliquée à 100 données/observations issues d'une loi $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ avec $\boldsymbol{\mu} = \begin{pmatrix} 3 \\ 4 \end{pmatrix}$ et $\boldsymbol{\Sigma} = \begin{pmatrix} 5 & 3 \\ 3 & 2.25 \end{pmatrix}$.



Jusqu'à présent nous n'avons pas encore retiré d'observations de notre jeu de données. Nous pouvons toujours réobtenir $\mathbf{X}$ tout entier à partir de $\mathbf{Y}$:

$$\mathbf{X} = \mathbf{H}'\mathbf{Y} + \boldsymbol{\mu}. \tag{1}$$

Nous souhaitons remplacer le vecteur $\mathbf{Y} = (Y_1, \ldots, Y_p)$ par un vecteur de plus petite dimension $\mathbf{Y}^{(q)} = (Y_1, \ldots, Y_q)$, $q \leq p$.

Comme $\mathbf{H}' = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p)$ nous obtenons de (1) que

$$\mathbf{X} - \boldsymbol{\mu} = Y_1\boldsymbol{\beta}_1 + Y_2\boldsymbol{\beta}_2 + \cdots + Y_p\boldsymbol{\beta}_p.$$

Si nous gardons seulement $Y_1, \ldots, Y_q$, nous pourrions utiliser l'approximation

$$\mathbf{X} - \boldsymbol{\mu} \approx Y_1\boldsymbol{\beta}_1 + Y_2\boldsymbol{\beta}_2 + \cdots + Y_q\boldsymbol{\beta}_q.$$

La proposition suivante montre que l'approximation

$$\mathbf{X} - \boldsymbol{\mu} \approx Y_1\beta_1 + Y_2\beta_2 + \cdots + Y_q\beta_q.$$

est “la meilleure" en termes de q vecteurs de base

Proposition

Soit $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_p)$ *une quelconque base orthonormale de* $\mathbb{R}^p$. *Alors*

$$\mathbf{X} - \boldsymbol{\mu} = W_1\mathbf{b}_1 + W_2\mathbf{b}_2 + \cdots + W_p\mathbf{b}_p,$$

et pour tout $1 \leq q \leq p$

$$E\left\|(\mathbf{X} - \boldsymbol{\mu}) - \left(\sum_{i=1}^{q} W_i\mathbf{b}_i\right)\right\|^2 \geq E\left\|(\mathbf{X} - \boldsymbol{\mu}) - \left(\sum_{i=1}^{q} Y_i\beta_i\right)\right\|^2.$$

Cette proposition montre que les composantes principales retiennent “le maximum" de l'information contenue dans le vecteur original $\mathbf{X}$. Aucune autre base orthonormale jouit d'une meilleure puissance d'approximation que la fonction propre de $\mathbf{\Sigma}$.

Notons que

$$E \left\| (\mathbf{X} - \boldsymbol{\mu}) - \left(\sum_{i=1}^{q} Y_i \boldsymbol{\beta}_i \right) \right\|^2 = \lambda_{q+1} + \ldots + \lambda_p.$$

Le ratio

$$\frac{\lambda_{q+1} + \ldots + \lambda_p}{\lambda_1 + \ldots + \lambda_p} = \frac{\lambda_{q+1} + \ldots + \lambda_p}{\mathrm{tr}(\mathbf{\Sigma})} \tag{2}$$

peut donc être vu comme une mesure de la quantité d'information expliquée par les q premières composantes principales. On dit que (2) est *la part de la variabilité de* $\mathbf{X}$ *expliquée par les q premières CP.*

Preuve. Par simplicité, prenons $\boldsymbol{\mu} = \mathbf{0}$ (sans perte de généralité).Tout d'abord, il est clair que $\mathbf{X}$ peut être réécrit sous la forme $\sum_{i=1}^{p} W_i \mathbf{b}_i$. Cela suit du fait que $\mathbf{B}$ forme une base orthonormale. Remarquons aussi que $W_i = \mathbf{b}_i' \mathbf{X}$ (puisque la transformation des composantes principales nous dit que $\mathbf{W} = \mathbf{B}'\mathbf{X}$).

On sait que

Maintenant, par le théorème de Pythagore, nous avons

$$E \left\| (\mathbf{X} - \left(\sum_{i=1}^{q} W_i \mathbf{b}_i \right) \right\|^2 = E \left\| \sum_{i=q+1}^{p} W_i \mathbf{b}_i \right\|^2$$
$$= \sum_{i=q+1}^{p} E W_i^2 \underbrace{\|\mathbf{b}_i\|^2}_{=1} = \sum_{i=q+1}^{p} \mathbf{b}_i' \boldsymbol{\Sigma} \mathbf{b}_i = \mathrm{tr}(\tilde{\mathbf{B}}' \boldsymbol{\Sigma} \tilde{\mathbf{B}}),$$

où $\tilde{\mathbf{B}} = (\mathbf{b}_{q+1}, \ldots, \mathbf{b}_p)$ est une matrice $p \times (p - q)$ orthonormale.

Nous cherchons donc une matrice $\mathbf{B}^*$ qui minimise $\mathrm{tr}(\tilde{\mathbf{B}}'\boldsymbol{\Sigma}\tilde{\mathbf{B}})$ sur l'ensemble des matrices $(p-q)\times q$ orthonormales $\tilde{\mathbf{B}}$. Comme les e_i forment une base de $\mathbb{R}^p$, nous avons que $\tilde{\mathbf{B}} = \mathbf{HC}$ où $\mathbf{C} = (c_{jk})$ est $p \times (p-q)$. Nous avons que

$$\mathrm{tr}(\tilde{\mathbf{B}}'\boldsymbol{\Sigma}\tilde{\mathbf{B}}) = \mathrm{tr}(\mathbf{C}'\boldsymbol{\Lambda}\mathbf{C}) = \sum_{j=1}^{p} \lambda_j \left(\sum_{k=1}^{p-q} c_{jk}^2 \right) \tag{3}$$

et que $\mathbf{C}' = \tilde{\mathbf{B}}'\mathbf{H}$ est telle que $\mathbf{C}'\mathbf{C} = \tilde{\mathbf{B}}'\mathbf{H}\mathbf{H}'\tilde{\mathbf{B}} = \mathbf{I}_{p-q}$ (orthonormale). Puisque les valeurs propres sont supposées être bien ordonnées, la matrice orthonormale C qui minimise (3) est obtenue en prenant les vecteurs propres associés dans $\mathbf{H}$ associés aux plus petites valeurs propres.

□

Exemple. Supposons que $\mathbf{X} = (X_1, X_2, X_3)$ est de moyenne nulle et de matrice de variance-covariance

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix}.$$

Alors les valeurs propres sont

$$(\lambda_1, \lambda_2, \lambda_3) = (5.83, 2, 0.17)$$

avec comme vecteurs propres associés

$$\boldsymbol{\beta}_1 = \begin{pmatrix} 0.383 \\ -0.924 \\ 0 \end{pmatrix} \quad \boldsymbol{\beta}_2 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \quad \boldsymbol{\beta}_3 = \begin{pmatrix} 0.923 \\ -0.383 \\ 0 \end{pmatrix}.$$

Cela nous donne

$$Y_1 = 0.383X_1 - 0.923X_2$$
$$Y_2 = X_3$$
$$Y_3 = 0.924X_1 + 0.383X_2$$

Les coordonnées des vecteurs propres nous révèlent combien de poids est attribué à chaque X_i!

Dans cet exemple la proportion de la variance totale expliquée par les premières CP équivaut à $\lambda_1/\mathrm{tr}(\mathbf{\Sigma}) = 5.83/8 = 0.73$, celle expliquée par les 2 premières CP correspond à $(\lambda_1 + \lambda_2)/\mathrm{tr}(\mathbf{\Sigma})$
$= (5.83 + 2)/8 = 0.98$.

Une propriété défavorable de l'ACP est que cette méthode *n'est pas échelle-invariante*. Cela veut dire que si $\mathbf{A}$ est une matrice d'échelle et si P_{tr} représente la transformation des composantes principales, alors

$$P_{tr}(\mathbf{AX}) \neq \mathbf{A}P_{tr}(\mathbf{X}).$$

Exemple. Soient $E\mathbf{X} = \mathbf{0}$ et $\mathrm{Var}(\mathbf{X}) = \begin{pmatrix} 1 & 4 \\ 4 & 100 \end{pmatrix}$. Les valeurs propres sont $\lambda_1 = 100.16$ and $\lambda_2 = 0.84$ avec comme vecteurs propres associés

$$\beta_1 = \begin{pmatrix} 0.04 \\ 0.999 \end{pmatrix} \quad \text{et} \quad \beta_2 = \begin{pmatrix} 0.999 \\ -0.04 \end{pmatrix}.$$

Nous avons $\lambda_1/\mathrm{tr}(\mathbf{\Sigma}) = 0.992$.

La grande variance de X_2 domine complètement la première composante principale obtenue à partir de $\mathbf{\Sigma}$!

Nous utilisons maintenant le même scénario, mais nous standardisons d'abord X en divisant ses composantes par leur dispersion. Cela veut dire que nous utilisons la matrice de corrélation

$$\mathbf{P} = \begin{pmatrix} 1 & 0.4 \\ 0.4 & 1 \end{pmatrix}$$

afin de déterminer les composantes principales. Les valeurs propres sont maintenant $\lambda_1 = 1.4$ and $\lambda_2 = 0.6$ avec comme vecteurs propres associés

$$\boldsymbol{\beta}_1 = \begin{pmatrix} 0.707 \\ 0.707 \end{pmatrix} \quad \text{et} \quad \boldsymbol{\beta}_2 = \begin{pmatrix} 0.707 \\ -0.707 \end{pmatrix}.$$

Nous avons $\lambda_1/\mathrm{tr}(\mathbf{P}) = 0.7$.

Nous attribuons donc la même importance à X_1 et à X_2 lors du calcul de Y_1.

Cet exemple montre que l'ACP doit être utilisée avec prudence.

L'ACP peut être utilisée au mieux si toutes les variables se trouvent à une échelle comparable.

Les variables devraient donc être proprement standardisées si elles sont mesurées sur des échelles fortement différentes.

Bien sûr, en pratique, nous ne connaissons ni $\boldsymbol{\mu}$ ni $\boldsymbol{\Sigma}$ et par conséquent nous devons baser nos analyses sur leurs estimateurs

$$\bar{\mathbf{X}} \quad \text{et} \quad \mathbf{S},$$

basés eux-mêmes sur l'échantillon $\mathbf{X}_1, \ldots, \mathbf{X}_n$.

Soient $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)$ les vecteurs propres standardisés de $\mathbf{S}$ correspondant aux valeurs propres $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_p$. La transformation des composantes principales empiriques est donnée par

$$\mathbf{X}_k \mapsto \hat{P}_{tr}(\mathbf{X}_k) = \hat{\mathbf{X}}_k = \hat{\beta}(\mathbf{X}_k - \bar{\mathbf{X}}), \quad 1 \leq k \leq n.$$

Posant $\hat{\mathbf{Y}}_k = (\hat{Y}_{k,1}, \ldots, \hat{Y}_{k,p})'$, les variables $\hat{Y}_{k,i}$ sont appelées les i-èmes scores des composantes principales.

Proposition

Soit $\mathbf{S}$ *une matrice de variance-covariance non-singulière correspondant à un certain échantillon* $\mathbf{X}_1, \ldots, \mathbf{X}_n$ *de p-vecteurs. Supposons que* $\hat{\lambda}_1 > \hat{\lambda}_2 > \ldots > \hat{\lambda}_p$ *soient ses valeurs propres et que* $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_p)$ *soit la matrice des vecteurs propres correspondants. Soient* $Y_1, \ldots, Y_n$ *les composantes principales. Alors*

$$(i) \quad \mathbf{S}_{\mathbf{Y}} = \hat{\boldsymbol{\Lambda}} = \operatorname{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_p),$$

$$(ii) \quad \operatorname{tr}(\hat{\boldsymbol{\Lambda}}) = \operatorname{tr}(\mathbf{S}) = \frac{1}{n-1} \sum_{k=1}^{n} \|\mathbf{X}_k - \bar{\mathbf{X}}\|^2,$$

$$(iii) \quad |\hat{\boldsymbol{\Lambda}}| = |\mathbf{S}|.$$

Preuve. *Exercice.* □

Comme dans le cas de la population il est possible d'établir une propriété d'optimalité des fonctions propres.

Proposition

Soit $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_p)$ *une base orthonormale de* $\mathbb{R}^p$. *Alors*

$$\mathbf{X}_k - \bar{\mathbf{X}} = W_{k,1}\mathbf{b}_1 + W_{k,2}\mathbf{b}_2 + \cdots + W_{k,p}\mathbf{b}_p,$$

et pour tout $1 \leq q \leq p$

$$\sum_{k=1}^{n} \left\| (\mathbf{X}_k - \bar{\mathbf{X}}) - \left(\sum_{i=1}^{q} W_{k,i}\mathbf{b}_i \right) \right\|^2$$
$$\geq \sum_{k=1}^{n} \left\| (\mathbf{X}_k - \bar{\mathbf{X}}) - \left(\sum_{i=1}^{q} \hat{Y}_{k,i}\hat{\beta}_i \right) \right\|^2 .$$

Considérons le modèle :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon, \text{ avec } E(\varepsilon|\mathbf{X}) = 0$$

à estimer via des données observées $S_n = (x_1, y_1), ..., (x_n, y_n)$ dans $\mathbb{R}^p \times \mathcal{Y}$, $p \geq 1$, issues d'une suite de variables aléatoires i.i.d.

- But : remplacer les variables $X_1, ..., X_p$ par des variables $Z_1, ..., Z_r$ $(r < p)$
- $Z_1, ..., Z_r$ sont les r premières composantes principales de l'ACP associées aux variables centrées (pour éviter des problèmes d'hétérogéneité des variances) de $X_1, ..., X_p$
- les $Z_1, ..., Z_r$ sont orthogonales entre elles et de normes $\lambda_1 > \lambda_2 > ... > \lambda_r$ respectivement

- Estimateur de β: $\widehat{\beta} = \left(\mathbf{Z}^{\mathsf{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathsf{T}}\mathbf{Y}$.
- $Cov(\widehat{\beta}_i, \widehat{\beta}_j) = 1/\lambda_j$ si $i = j$, 0 sinon.
- $r = p$ donne l'estimateur des moindres carrés
- $r < p$ élimine les composantes de variances faibles ou nulles (et élimine le problème de co-linéarité mais également quand $n > p$)
- Le choix de r peut se faire par validation croisée
- Incovénient : les premières composantes principales ne sont pas forcément corrélées avec la variable réponse Y

Nous avons vu que les composantes principales sont obtenues en calculant les vecteurs propres $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)$ de

$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})'.$$

Pour rappel, la première composante empirique est obtenue comme

$$\hat{\beta}_1 := \mathrm{argmax}_{\mathbf{h}} \mathbf{h}'\mathbf{S}\mathbf{h},$$

avec comme contrainte $\|\mathbf{h}\|^2 = 1$

Gros inconvénient de l'ACP, lorsque $p = p_n$ est tel que $p_n/n \to c > 0$, $\hat{\beta}_1$ n'est pas un estimateur convergent de β_1.

Dans la PCA sparse, on considère

$$\hat{\beta}_{1,\mathrm{sp}}(\rho) := \mathrm{argmax}_{\mathbf{h}} \mathbf{h}'\mathbf{S}\mathbf{h},$$

avec comme contrainte $\|\mathbf{h}\|^2 = 1$ et $\mathrm{Card}(\mathbf{h}) \leq k(\rho)$, où $\mathrm{Card}(\mathbf{h})$ est le nombre de coefficients non-nuls dans $\mathbf{h}$.

```
###
states <- row.names(USArrests)
###
pr.out <- prcomp(USArrests, scale = TRUE)
###
names(pr.out)
###
pr.out$center
pr.out$scale
###
pr.out$rotation
###
dim(pr.out$x)
###
biplot(pr.out, scale = 0)
###
pr.out$rotation = -pr.out$rotation
pr.out$x = -pr.out$x
```

```
biplot(pr.out, scale = 0)
###
pr.out$rotation = -pr.out$rotation
pr.out$x = -pr.out$x
biplot(pr.out, scale = 0)
###
pr.out$sdev
###
pr.var <- pr.out$sdev^2
pr.var
###
pve <- pr.var / sum(pr.var)
pve
###
```

Sparse PCA with the "sparsepca" package