PRINCIPAL COMPONENT ANALYSIS

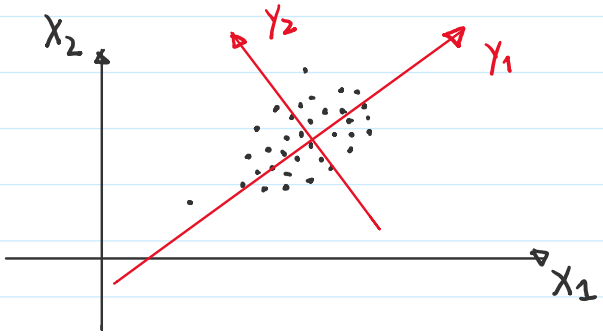
Sunday, April 7, 2019 11:34 AM

INTRODUCTION

THE PRINCIPAL COMPONENT ANALYSIS FOCUSES ON EXPLAINING THE COVARIANCE STRUCTURE OF A SET OF RANDOM VARIABLES (FROM WHICH OUR DATA ARE PRODUCED) THROUGH A FEW LINEAR COMBINATION OF THOSE RANDOM VARIABLES.

ITS GENERAL OBJECTIVES ARE: 1. DATA DIMENSIONAL REDUCTION. 2. DATA INTERPRETATION. ANYWAY, THE CORE OF PCA IS: FIND WHO CAUSES THE VARIANCE.

SUPPOSE THAT WE HAVE A RANOM VECTOR WITH 2 ELEMENTS $\underline{X}' = [X_1, X_2]$, WHERE X_1 AND X_2 ARE RANDOM VARIABLES. FROM THIS RANDOM VECTOR WE CAN GENERATE A COLECTION OF DATA, THAT IS, POINTS IN $\mathbb{R}^2$. LOOK THE FIGURE BELOW, IN WHICH THE BLACK POINTS REPRESENT OUR DATA, IT SEEMS THAT THEY ARE FROM A MULTI-VARIATE GAUSSIAN DISTRIBUTION.



NOW, OUR GOAL IS: FIND A DIRECTION (A VECTOR), ON WHICH THE VARIANCE OF PROJECTIONS OF DATA IS AS GREAT AS POSSIBLE; THEN REPEAT THE PROCEDURE ITERATIVELY, BUT THEY SHOULD BE MUTUALLY ORTHOGONAL. IN OUR EXAMPLE, OBVIOUSLY y_1 AND y_2, WHICH ARE DRAWN AS RED ARROWS IN THE FIGURE, ARE THE RIGHT ANSWER.

MATHEMATICAL FORMULATION

MATHEMATICAL FORMULATION

SUPPOSE THAT WE HAVE A RANDOM VECTOR

$$\underline{X}' = [X_1, X_2, \dots, X_p]$$

IN WHICH $X_1, \dots, X_p$ ARE REAL-VALUED RANDOM VARIABLES. MOREOVER, X' IS WITH COVARIANCE MATRIX Σ, THAT IS

$$\Sigma := \begin{bmatrix} \operatorname{cov}(X_1, X_1) & \cdots & \operatorname{cov}(X_1, X_p) \\ \vdots & \ddots & \vdots \\ \operatorname{cov}(X_p, X_1) & \cdots & \operatorname{cov}(X_p, X_p) \end{bmatrix}$$

THEN OUR PROBLEM CAN BE REWRITTEN AS AN OPTIMIZATION PROBLEM AS THE FOLLOWING:

P.C.A. PROBLEM

FIND $\underline{a}_1, \dots, \underline{a}_p \in \mathbb{R}^p$ WITH $\|\underline{a}_i\| = 1$, $\forall i \in \{1, \dots p\}$, SUCH THAT:

1) $$\underline{a}_1 = \underset{\underline{a} \in \mathbb{R}^p, \|\underline{a}\|=1}{\operatorname{argmax}} \operatorname{Var}(\underline{a}'\underline{X})$$

EXPLICITLY, $\underline{a}'\underline{X} = a_1X_1 + a_2X_2 + \dots + a_pX_p$ WHICH IS EXACTLY A LINEAR COMBINATION OF FORMER R.V.S

2) $$\underline{a}_2 = \underset{\underline{a} \in \mathbb{R}^p, \|\underline{a}\|=1}{\operatorname{argmax}} \operatorname{Var}(\underline{a}'\underline{X})$$
$$\operatorname{cov}(\underline{a}'\underline{X}, \underline{a}_1'\underline{X}) = 0$$

$\vdots$

3) $$\underline{a}_p = \underset{\underline{a} \in \mathbb{R}^p, \|\underline{a}\|=1}{\operatorname{argmax}} \operatorname{Var}(\underline{a}'\underline{X})$$
$$\operatorname{cov}(\underline{a}'\underline{X}, \underline{a}_k'\underline{X}) = 0, \forall k = 1, \dots, p-1$$

IN FACT THE LAST PROBLEM IS NOT NECESSARY, BECAUSE THE SPACE IS WITH DIMENSION p. WHEN YOU HAVE CALCULATED THE FIRST $p-1$ DIRECTIONS, THE LAST ONE NATURALLY APPEARS.

SOLUTION OF PROBLEM

FOR SOLVING THIS OPTIMAZION PROBLEM, WE NEED SOME LEMMAS:

LEMMA 1 THE COVARIANCE MATRIX Σ OF A RANDOM VECTOR $\underline{X}' = [X_1, X_2, \dots, X_p]$ IS POSITIVE SEMI-DEFINITE.

PROOF LET US USE $\underline{\mu}$ TO DENOTE THE EXPECTED VALUES OF THE RANDOM VECTOR, THAT IS

$$\underline{\mu}' = (\mu_1, \mu_2, \dots, \mu_p)$$

$$\mu' = (\mu_1, \mu_2, \dots, \mu_p) := (\mathbb{E}[X_1], \mathbb{E}[X_2], \dots, \mathbb{E}[X_p])$$

FOR $y \in \mathbb{R}^p$, WE HAVE

$$\begin{aligned} y'\Sigma y &= y'\,\mathbb{E}[(X-\mu)(X-\mu)']\,y \\ &= \mathbb{E}[y'(X-\mu)(X-\mu)'y] \\ &= \mathbb{E}[((X-\mu)'y)'\cdot(X-\mu)'y] \\ &= \mathbb{E}[((X-\mu)'y)^2] \geq 0 \end{aligned}$$

NOTICE $(X-\mu)'y \in \mathbb{R}$

BECAUSE y IS ARBITRARY, AND Σ IS SYMMETRIC, WE FINISH THE PROOF. ■

LEMMA 2 LET B BE A $p \times p$ POSITIVE SEMI-DEFINITIVE MATRIX WITH EIGENVALUES $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$ AND ASSOCIATED NORMALIZED EIGENVECTORS $e_1, e_2, \dots, e_p$. THEN

(1) $\max_{x \neq 0} \frac{x'Bx}{x'x} = \lambda_1$ (OBTAINED WHEN $x = e_1$)

(2) $\min_{x \neq 0} \frac{x'Bx}{x'x} = \lambda_p$ (OBTAINED WHEN $x = e_p$)

MOREOVER,

(3) $\max_{x \perp e_1, \dots, e_k} \frac{x'Bx}{x'x} = \lambda_{k+1}$ (OBTAINED WHEN $x = e_{k+1}$, WHERE $k = 1, 2, \dots, p-1$)

PROOF THANKS TO REAL SPECTRAL THEOREM, WE CAN WRITE Σ AS:

$$B = P \Lambda P'$$

WHERE $\Lambda = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \dots & \\ & & & \lambda_p \end{bmatrix}$ AND $P = [e_1, e_2, \dots, e_p]$,

IN PARTICULAR $PP' = P'P = I$.

$B^{1/2}B^{1/2} = P\Lambda^{1/2}P'P\Lambda^{1/2}P' = P\Lambda P' = B$

DENOTE $B^{1/2} := P\Lambda^{1/2}P'$, WHERE $\Lambda^{1/2} := \begin{bmatrix} \sqrt{\lambda_1} & & & \\ & \sqrt{\lambda_2} & & \\ & & \dots & \\ & & & \sqrt{\lambda_p} \end{bmatrix}$,

ALSO $y := P'x$; IF $x \neq 0$, OBVIOUSLY $y \neq 0$. THUS

$$\frac{x'Bx}{x'x} = \frac{x'B^{1/2}B^{1/2}x}{x'x} = \frac{x'P\Lambda^{1/2}P'P\Lambda^{1/2}P'x}{x'x} = \frac{y'\Lambda y}{y'y} = \dots$$

$$\frac{x'Bx}{x'x} = \frac{y'\Lambda y}{x'x} = \frac{y'\Lambda y}{y'y} = \frac{\sum_{i=1}^{p}\lambda_i y_i^2}{\sum_{i=1}^{p} y_i^2} \leq \lambda_i \frac{\sum_{i=1}^{p} y_i^2}{y_i^2} = \lambda_i$$

SETTING $x = e_1$ GIVES

$$y = P'e_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

THANKS A BUNCH TO REAL SPECTRAL THEOREM

UNDER THIS CONDITION, WE GET $\frac{y'\Lambda y}{y'y} = \lambda_1$, WHICH COMPLETES THE PROOF OF (1).

BY THE SIMILAR METHOD, WE CAN PROVE (2).

FOR (3), OBSERVE THAT $y = P'x$, THEN

$$x = (PP')x = Py = y_1e_1 + y_2e_2 + \cdots + y_pe_p$$

SO $x \perp e_1, \ldots, e_k$ IMPLIES

$$0 = e_i'x = y_1e_i'e_1 + y_2e_i'e_2 + \cdots + y_pe_i'e_p = y_i$$

THEN WE KNOW $y_i = 0, \forall i \leq k$. THEREFORE WE HAVE

$$\frac{x'Bx}{x'x} = \frac{\sum_{i=1}^{p}\lambda_i y_i^2}{\sum_{i=1}^{p} y_i^2} = \frac{\sum_{i=k+1}^{p}\lambda_i y_i^2}{\sum_{i=k+1}^{p} y_i^2} \leq \lambda_{k+1}$$

TAKING $y_{k+1} = 1, y_{k+2} = \cdots = y_p = 0$ GIVES THE MAXIMUM. ■

NOW, WE CAN FORMLY STATE OUR FINAL RESULT AS

RESULT 3 THE SOLUTION OF P.C.A. PROBLEM IS

$$(a_1, a_2, \ldots, a_p) = (e_1, e_2, \ldots, e_p)$$

WHICH IS EQUIVALENT TO SAY:

THE i-TH PRINCIPAL COMPONENT OF X IS $y_i := e_i'X$.

PROOF FROM LEMMA 1 WE KNOW Σ IS POSITIVE SEMI-DEFINITIVE, AND $\max_{a \in \mathbb{R}^p} \mathrm{Var}(a'X) = \max_{a \in \mathbb{R}^p} a'\Sigma a = \max_{a \in \mathbb{R}^p} \frac{a'\Sigma a}{a'a}$

AND $\max_{\underline{a}\in\mathbb{R}^p,\ \|\underline{a}\|=1} \mathrm{Var}(\underline{a}'\underline{X}) = \max_{\underline{a}\in\mathbb{R}^p,\ \|\underline{a}\|=1} \underline{a}'\Sigma\underline{a} = \max_{\underline{a}\in\mathbb{R}^p} \frac{\underline{a}'\Sigma\underline{a}}{\underline{a}'\underline{a}}$

(COVARIANCE UNDER LINEAR TRANSFORM; NORMALIZATION)

THEN, IF WE CAN SHOW THE EQUIVALENT RELATIONSHIP:

$$\mathrm{Cov}(Y_i, Y_k) = 0 \Leftrightarrow \underline{e}_i \perp \underline{e}_k$$

THEN WE CAN DERICTLY APPLY LEMMA 2 TO GET THE RESULT; LUCKILY,

$$\mathrm{Cov}(Y_i, Y_k) = \underline{e}_i'\Sigma\underline{e}_k = \underline{e}_i'\lambda_k\underline{e}_k = \lambda_k\,\underline{e}_i'\underline{e}_k = 0.$$

WHICH COMPLETES THE PROOF. ∎

CHARACTERIZATION OF TOTAL VARIANCE

HERE WE GIVE ALSO AN EASY BUT IMPORTANT RESULT ABOUT THE CHARACTERIZATION OF TOTAL VARIANCE, WHICH WILL BE VERY USEFUL IN SELECTING NUMBER OF PRINCIPAL COMPONENTS.

RESULT 4 THE TOTAL VARIANCE EQUALS TO THE SUM OF EIGENVALUES OF COVARIANCE MATRIX. THAT IS:

$$\sum_{i=1}^{p} \mathrm{Var}(X_i) = \sum_{i=1}^{p} \lambda_i$$

PROOF THE PROOF IS ALMOST TRIVIAL:

$$\sum_{i=1}^{p} \mathrm{Var}(X_i) = \mathrm{tr}(\Sigma) = \mathrm{tr}(P\Lambda P') = \mathrm{tr}(\Lambda P'P) = \mathrm{tr}(\Lambda) = \sum_{i=1}^{p}\lambda_i$$

(CYCLIC PROPERTY OF TRACE.)

∎