

# Folosirea PCA pentru recunoasterea imaginilor cu ajutorul librăriei Accord

Motrescu Radu

# Contents

<b>1</b>	<b>Libraria Accord.NET</b>	<b>3</b>
<b>2</b>	<b>Ce este PCA?</b>	<b>4</b>
<b>3</b>	<b>Jurnalul dezvoltarii aplicatiei</b>	<b>7</b>
3.1	Aplicatia AccordPCA . . . . .	7
3.2	Aplicatia PlotPCA . . . . .	9
3.2.1	Setul 3-cloud-points . . . . .	9
3.2.2	Setul two-moon si abordarea Kernel PCA . . . . .	9
3.2.3	Cros-validare 10-fold . . . . .	10
3.3	Aplicatia AccordFace . . . . .	10

## 1 Libraria Accord.NET

Libraria Accord.Net contine clase pentru:

- Calcul stiintific: matematica, statistica si machine learning
- Procesare de imagini si semnale: imagini, semnale audio si recunoastere si urmarire faciala in timp real
- Librarii suport pentru controale specifice: histograme, scatter-plots, controale pentru fiecare clasa de procesare imagini si semnale

In contextul cerintelor pe care vrem noi sa le indeplinim, vom folosi clase dedicate matematicii, statisticii, si ale unor controale de afisare a datelor.

## 2 Ce este PCA?

PCA, sau Principal Component Analysis, sau pe romaneste, Analiza Componentelor Principale este o unealta matematica aplicata din algebra liniara.

Este o metoda non-parametrica (care nu depinde de statistici) de a extrage informatiile relevante dintr-un set de date complex sau confuz.

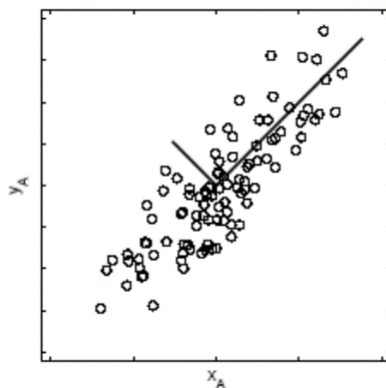
**2.0.0.1 Un scurt exemplu:** Sa presupunem ca am extras informatii pentru 100 de parametri pentru un student: inaltime, varsta, greutate, nota obtinuta la un test, culoarea parului etc. Vrem sa gasim cele mai importante caracteristici care definesc studentul. Cum facem asta? Folosim PCA pentru a selecta numai cele mai importante caracteristici.

- Ce parametri dorim sa indepartam:
  - Constanti: numarul de capete, care este 1 pentru toti studentii
  - Aproape constanti: grosimea firului de par: 0.003, 0.002, 0.0005 etc.
  - Care depind de alti parametri
- Ce parametri dorim sa pastram:
  - Care nu depind de alti parametri: culoarea ochilor
  - Care se schimba mult, au variatie mare: notele

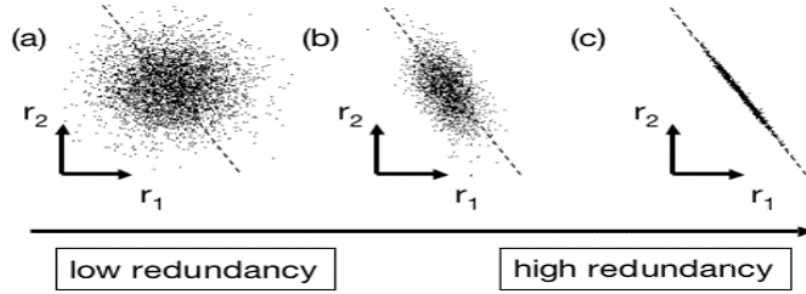
A putea elimina parametrii irelevanti si a-i pastra pe cei relevanti este usor pentru un om, el putand vedea clar care parametri nu exprima informatii relevante despre subiectul respectiv, dar cum putem face calculatorul sa isi dea seama de acesti parametri? Folosind matematica, desigur!

Dorim sa minimizam "sunetul de fundal" si redundanta datelor si sa maximizam variatia dintre parametri.

Se poate vedea in imaginea de mai jos maximizarea variatiei dintre axele norului de puncte respectiv.



In imaginile de mai jos se pot vedea inregistrările unor informații. In imaginea (a) și (b) se poate vedea cum informațiile nu sunt corelate, având redundanța mică spre medie (ex: înălțimea unui student și media lui), iar in imaginea (c) se poate vedea o redundanță mare, însemnând că ambii parametrii pot fi exprimați unul în funcție de celălalt.



**2.0.0.2 Variatia** Este un mijloc de realizare a variabilității datelor dintr-un set de date cu media  $\bar{X}$ :

$$\sigma^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1} \quad (1)$$

**2.0.0.3 Covariația** Reprezintă variabilitatea fiecărei dimensiuni în relație cu celelalte, și este măsurată între 2 dimensiuni pentru a se putea observa relația dintre cele 2, spre exemplu numărul de ore studiate și nota obținută la examen.

$$\text{var}(X) = \frac{\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})}{n - 1} \quad (2)$$

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1} \quad (3)$$

**2.0.0.4 Matricea de covariație** Considerăm setul de date din care extragem valoarea medie (zero-mean data) și având setul de vectori  $\{x_1, x_2, \dots, x_m\}$  care reprezintă liniile unei matrici  $X_{m,n}$ .

Fiecare linie a matricii reprezintă toate măsurătorile unui anumit parametru, iar fiecare coloană reprezintă măsurătorile care s-au întâmplat la un moment dat.

De aici ajungem la definiția matricii de covariație:

$$S_x \equiv \frac{1}{n - 1} X X^T \text{ where } X = \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix} \quad (4)$$

$S_x$  este o matrice simetrica  $m \times m$ , termenii de pe diagonala reprezentand variatia din acelasi parametru, iar termenii care nu sunt de pe diagonala reprezinta covariatia dintre parametri diferiti. Calculand  $S_x$ , cuantificam corelatia dintre toate posibilele perechi de masuratori. Observand elementele din matrice, o covariatie mare reprezinta un caz de redundanta mare, iar o covariatie egala cu 0 reprezinta date complet necorelate.

$$C = \begin{bmatrix} cov(X, X) & cov(X, Y) & cov(X, Z) \\ cov(Y, X) & cov(Y, Y) & cov(Y, Z) \\ cov(Z, X) & cov(Z, Y) & cov(Z, Z) \end{bmatrix} \quad (5)$$

**2.0.0.5 Valorile si vectorii proprii** Urmatorul pas in calcularea PCA este aflarea valorilor si a vectorilor proprii ale matricii de covariatie. Extragand aceste informatii, ele ne vor arata componentele principale ale setului de date: vectorul propriu cu cea mai mare valoare proprie este componenta principala a setului de date. Se obisnuieste sortarea vectorilor proprii in functie de valoarea proprie pentru a determina ordinea de semnificativitate.

Vectorii si valorile proprii reies din probleme de urmatoarea forma:

$$A.v = \lambda.v \quad (6)$$

**A**: matrice  $m \times m$

**v**: vector  $m \times 1$  nenul

$\lambda$ : constanta

Pentru orice valoare a lui  $\lambda$  pentru care ecuatia are solutie se numeste valoarea proprie a lui A, si vectorul **v** care corespunde acestei valori se numeste vectorul propriu a lui A.

**2.0.0.6 Pasul final PCA** este sa aflam valorile finale ale setului de date. Aici avem optiunea sa ignoram o parte din dimensiunile pe care le avem, deoarece ele pot fi nesemnificative, avand valori proprii mici, sau dorim afisarea lor in 1, 2 sau 3 dimensiuni. Dupa stabilirea vectorilor proprii doriti, aflarea datelor finale este simpla, proiectam punctele in spatiul lor:

$FinalData = RowFeatureVector \times RowZeroMeanData$

**RowFeatureVectore** este matricea cu vectorii proprii transpusi, cu cel mai semnificati vector propriu pe prima linie.

**RowZeroMeanData** este matricea care contine setul de date initial din care s-a scazut valoarea medie.

Matricea rezultata **FinalData** va contine setul de date dupa aplicarea algoritmului PCA.

## 3 Jurnalul dezvoltarii aplicatiei

### 3.1 Aplicatia AccordPCA

Primii pasi pe care i-am facut in cadrul acestui proiect a fost implementarea algoritmului PCA. Acest lucru a fost realizat usor, folosind libraria Accord, mai specific, pentru a realiza diferitele operatii cu matrici.

```
public void Compute()
{
    mean = initialData.Mean(0);

    dataAdjusted = initialData.Subtract(mean, 0);
    covarianceMatrix = dataAdjusted.Covariance();

    evd = new EigenvalueDecomposition(covarianceMatrix);

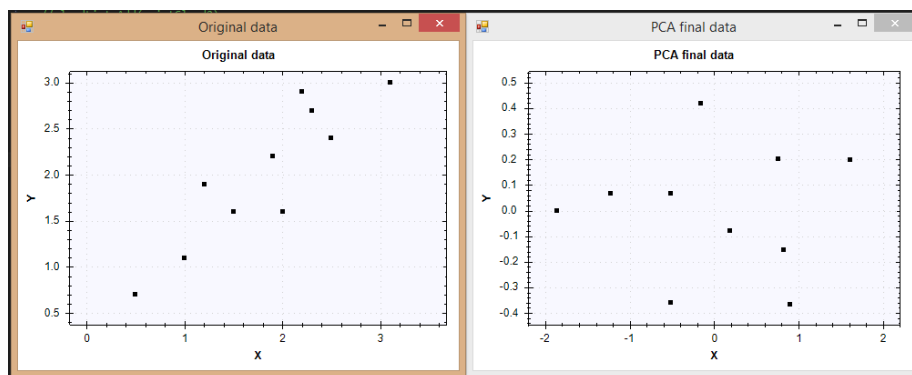
    eigenvalues = evd.RealEigenvalues;
    eigenvectors = evd.EigenVectors;

    eigenvectors = Matrix.Sort(eigenvalues, eigenvectors, new GeneralComparer(ComparerDirection.Descending, true));

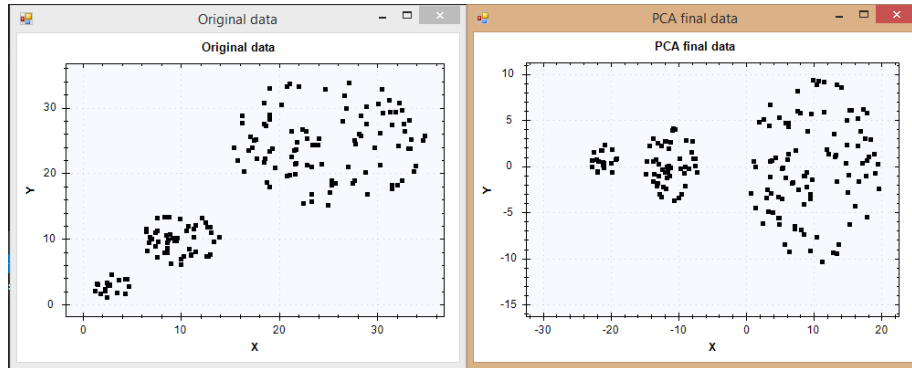
    finalData = dataAdjusted.Dot(eigenvectors);

    // data2D = finalData.GetColumns(0, 1);
}
```

Pentru a testa acest algoritm, am introdus prima data un set de date deja testat de catre altcineva, unde se poate observa clar rezultatul algoritmului. (referinta catre <http://dai.fmph.uniba.sk/courses/ml/sl/PCA.pdf>)



Urmatorul lucru pe care l-am dorit sa il facem a fost sa vedem cum mai multe noruri de puncte, de diferite dimensiuni, se modifica dupa aplicarea algoritmului. Pentru acest lucru, am creat o clasa **PointCloud** care descrie un nor de puncte. In interiorul clasei am generat norul de puncte in jurul unor coordonate  $(x, y)$ , sub forma unui disc, in care punctele au distributie normala. Rezultatele au fost urmatoarele:



La fel ca in primul test, se poate observa aranjarea dupa prima componenta principala a norurilor.



## 3.2 Aplicatia PlotPCA

In acest moment, am decis sa realizam o aplicatie care are un GUI usor de folosit si unde fiecare nor de puncte este reprezentat in alta culoare, pentru a se vedea mai clar ce se intampla, acest lucru fiind mai important pentru seturile de date in care masuratorile pentru fiecare entitate sunt intrepatruse.

### 3.2.1 Setul 3-cloud-points

Primele teste pe care le-am facut, au fost pe 3 nori de puncte generati aleator, la fel ca in exemplul de mai sus, si se poate vedea mai clar separarea norilor atat pe in spatiu 2D, cat si pe axa primei componente, cea cu cea mai mare variatie. De asemenea, atunci cand norii se suprapun putin de-a lungul axei Y, algoritmul PCA reuseste sa separe norii complet de-a lungul axei primei componente. Pe de alta parte, daca norii se suprapun mai mult, atunci se poate observa ca algoritmul PCA nu mai poate separa norii la fel de bine ca si pana acum.

\*\* IMAGINI AICI \*\*

### 3.2.2 Setul two-moon si abordarea Kernel PCA

Pentru a aborda problema seturilor de date liniar inseparabile, nu putem folosi abordarea PCA simpla, rezultatul nu va fi cel dorit, dupa cum se poate vedea aplicand algoritmul PCA pentru setul two-moon: cele doua seturi de puncte se amesteca de-a lungul primei componente principale, astfel o clasificare a unui nou punct proiectat in spatiul respectiv va fi dificila, deoarece nu putem seta un punct de delimitare intre cei doi nori, acestia fiind intrepatrusi.

Abordarea Kernel PCA rezolva acest lucru prin proiectarea datelor intr-un spatiu dimensional mai mare, unde ele vor deveni liniar separabile, iar acest lucru se face cu o "functie kernel".

Pentru a implementa Kernel PCA, vom avea in considerare urmatoarele 2 lucruri:

#### 1. Calcularea matricii kernel

Pentru fiecare pereche de puncte se calculeaza :

$$k(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|_2^2) \quad (7)$$

Daca avem un set de date cu 100 de mostre, matricea kernel va fi o matrice simetrica  $100 \times 100$ .

Alegerea valorii lui  $\gamma$  este foarte importanta, in functie de ea rezultatul va fi cel dorit sau nu.

#### 2. Calcularea vectorilor si a valorilor proprii

Deoarece nu putem garanta ca matricea e centrata, vom aplica urmatoarea formula pentru a o centra:

$$K' = K - 1_N K - K 1_N + 1_N K 1_N \quad (8)$$

unde  $1_N$  este o matrice  $N \times N$  cu toate valorile egale cu  $\frac{1}{N}$ . Acum, putem afla vectorii si valorile proprii pentru matricea centrata, iar vectorii proprii vor reprezenta setul de date proiectat pe componentele principale respective.

In imaginile urmatoare se poate observa cum algoritmul PCA si algoritmul KPCA afecteaza setul de date two-moon, rezultatul obtin prin folosirea KPCA este mult mai bun.

**\*\* IMAGINI AICI \*\***

### 3.2.3 Cros-validare 10-fold

Pentru a ne asigura ca algoritmul se comporta corect si in alte situatii, am decis sa facem o cros-validare 10-fold (10-fold crossvalidation). Acest lucru implica urmatoarele:

- avem 100 de puncte, distribuite in setul de date two-moon
- din acestea, vom alege pe rand, cate 10 puncte, diferite de fiecare data, si vom retine norul din care face parte fiecare
- cu celelalte 90 de puncte vom "antrena" algoritmul pentru a putea proiecta celelalte 10 puncte pe componentele principale respective, in cazul nostru, dorim sa facem proiectarea pe prima componenta principala
- avand cele 90 de puncte proiectate in spatiul nou, putem determina din nou granitele celor 2 nori de puncte, iar mijlocul acestor granite va fi punctul nostru de separare a norilor
- cu punctul de separare gasit, cele 10 puncte proiectate pe prima componenta principala pot fi categorizate: daca se afla in stanga punctului de separare el va fi in norul albastru de puncte, iar daca este la dreapta punctul de separare, va apartine norului rosu
- stiind din ce nori au provenit punctele si in ce nori au fost proiectati, putem verifica eroarea algoritmului pentru cele 10 puncte: 
$$\frac{\text{nr. puncte proiectate gresit}}{\text{nr. puncte totale}}$$
- repetam acest proces pana cand toate punctele au fost proiectate, iar eroarea de final este cea cautata
- In cazul testelor noastre, eroarea a variat intre 1% si 4%, fiind un rezultat foarte bun.

### 3.3 Aplicatia AccordFace

Dupa ce am realizat testele KPCA, urmatorul pas a fost sa crestem dimensiunile cu care lucram. Mai exact, am inceput sa folosim imagini pentru a vedea cum actioneaza algoritmul PCA pe ele.

Imaginile folosite au fost cele din baza de date Yale, care a fost folosita in multe alte cercetari de genul acesta, continand destui subiecti si destule imagini pentru a putea testa orice algoritm.

Pentru a putea procesa aceste imagini, ele trebuiau sa aiba toate aceeasi dimensiune, in cazul nostru  $168 \times 192$ , iar pentru a le procesa algoritmul,