Analiza componentelor principale in compresia imaginilor

Prelucratul direct cu date multi-dimensionale, cum ar fi imaginile, poate fi sursa unei serii de dificultati. Datele multi-dimensionale sunt greu de analizat si interpretat, vizualizarea lor in unele cazuri poate fi imposibila, iar pe langa asta, depozitarea lor poate fi foarte costisitoare. Reducerea dimensionalitatii datelor exploateaza structura si corelatia dintre acestea, permitandu-ne sa aducem datele intr-o forma mai compacta, ideal fara pierdere de informatie.

Analiza componentelor principale (PCA - (Karhunen-Loeve sau Hotelling transform ) este o procedura statistica care foloseste o transformare ortogonala pentru a transforma un set de observatii ale unor variabile posibil corelate (entitati, fiecare dintre acestea asumand diferite valori numerice) intr-un set de valori ale variabilelor necorelate liniar numite componente principale. Aceasta transformare este definita in asa fel incat prima componenta principala sa aiba cea mai mare varianta posibila (adica sa reprezinte cat mai mult variabilitatea datelor) si fiecare componenta succesiva are la randul ei cea mai mare varianta posibila sub constrangere ca este ortogonala la componentele precedente. Vectorii rezultati (fiecare fiind o combinatie liniara a variabilelor si care contine n observatii) sunt o baza ortogonala necorelata.

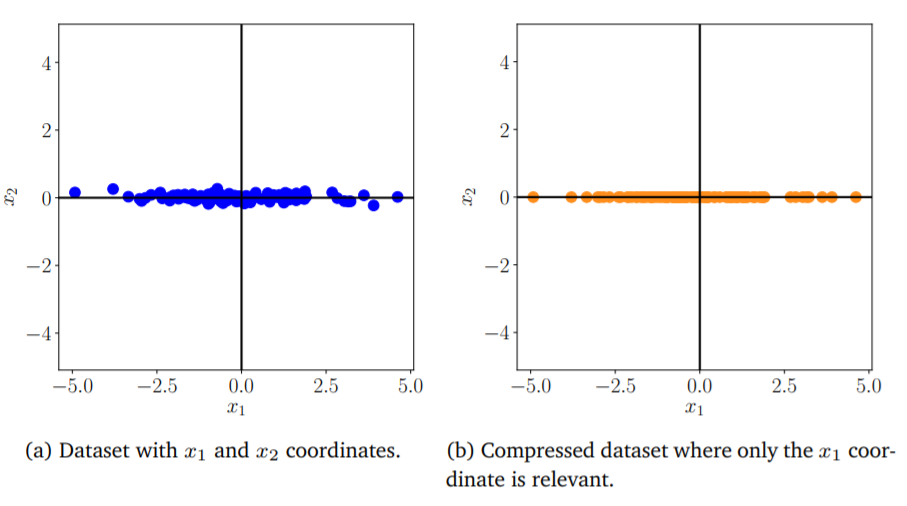


Fig 1 prezinta o ilustrare a reducerii dimensionalitatii. (a) Setul de date initial nu variaza foarte mult pe axa lui x2. (b) Datele din figura (a) sunt reprezentate folosind coordonata x1 cu o pierdere minimala de informatie.  
 PCA este folosit mai ales ca instrument în analiza datelor exploratorii și pentru realizarea de modele predictive. Este adesea folosit pentru a vizualiza distanța genetică și relația dintre populații.In PCA, suntem interesati sa gasim proiectii de puncte de date sunt la fel de similare cu punctele de date originale, dar care au o dimensionalitate intrinseca semnificativ mai mica.

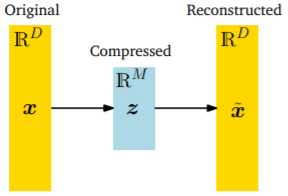


Fig 2 – ilustrarea grafica a PCA, unde z e versiune compresata a lui x, iar x este decompresia lui z care are aceeasi dimenisune ca originalul x

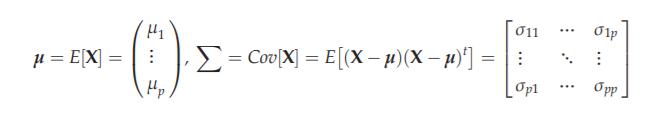
Aspecte teoretice

1. Vectorul de medii si matricea de covarianta

Fie X = o coloana random a unui vector de dimensiune p. Fiecare componenta, Xi, este o variabila random (v.r.) media:  E[Xi] = si varianta :  . Dandu-se doua variabile random Xi si Xj se poate define covarianta dintre ele ca fiind:



Valorile asteptate (mediile), varianta si covariantele dintre variabile aleatoare pot fi grupate in vectori si matrici numiti: vectorul de medii (mean vector µ), respectiv, matricea de covarianta (covariance matrice )



1. Eigenvalues si eigenvectors

Fie A o matrice matratica. Daca pentru oricare vector v, iar A este o matrice non negativa, atunci, daca Av = λv, cu v !=0, λ este valoarea eigen (eigenvalue) asociata vectorului eigen v (eigenvector).

Daca derivam: Av=λv => Av–λv=0 => (A-λI)v=0, unde I este matricea unitate, iar v este diferit de 0, atunci det(A-λI)=0 este numit polinomul characteristic al matricii A. Teorema fundamentala implica faptul ca polinomul characteristic poate fi factorizat:

det(A − λ I) = 0 = (λ1 − λ)(λ2 − λ). . .(λn − λ)

Valorile eigen λi nu sunt neaparat distincte.

3.Distante

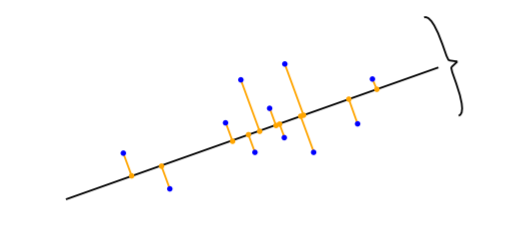
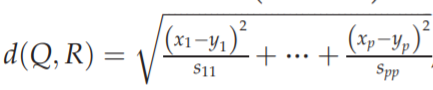


Figura 3 ilustreaza o proiectie a punctelor albastre pe dreapta astfel incat diferenta de distanta intre punctele proiectate (cele portocalii) si cele originale (albastre ) sa fie minima.

Multe tehnici de analiza statistica a datelor multidimensionale sunt bazate pe conceptul de distanta. Pentru doua puncte Q=(x1,x2,..,xp) si R=(y1,y2,…yp) distanta Euclideana dintre cele doua este data de formula:

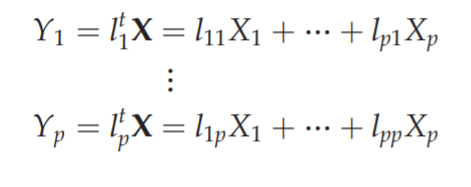
Uneori distanta Euclideana poate fi nesatisfacatoare deoarece fiecare coordinate contribuie in egala masura in calcularea distantei. Uneori este de preferat atribuirea unor ponderi asupra coordonatelor care sa stabileasca masura impactului acelei coordonate in calcularea mediei.

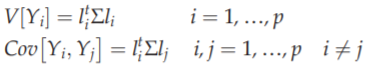


4. Componentele principale

Compontele principale sunt un caz particular de combinari lineare a variabilelor X1,..,Xp. Aceste combinatii lineare reprezinta un nou system de coordonate care este obtinut prin rotatia sistemului original de referinta care are X1,..,Xp ca axe de coordonate. Noile axe reprezinta directiile cu cea mai mare magnitudine si prezinta o descriere a covariantei.

Fie X = un vector aleatoriu de dimensiune p cu matricea de covarianta sigma si valorile eigen: λ1≥ λ2≥ λ3≥…≥ λp. Sa consideram urmatoare p combinatii lineare:

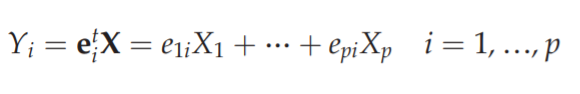




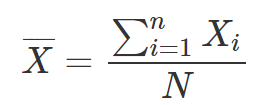
Care verifica ecuatiile:

Componentele principale sunt acele combinatii lineare care, nefiind corelate intre ele, au cea mai mare varianta posibila. Prima componenta principala reprezinta combinatia lineara cu cea mai mare varianta.

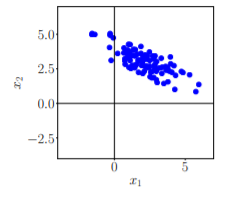
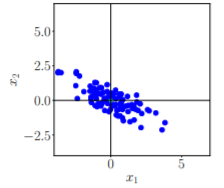
Fie Σ matricea de covarianta a unui vector aleatoriu X = . Presupunem ca Σ are p perechi de valori eigen si vectori eigen: (λ1,e1), …, (λ1,ep), cu : λ1≥ λ2≥ λ3≥…≥ λp. Atunci a i-a componenta principala este data de formula:



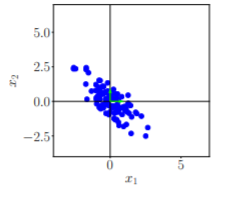
Implementarea algoritmului PCA

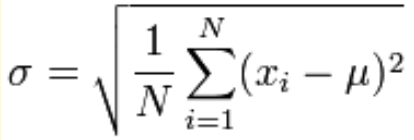
1. **Scaderea mediei.** Incepem prin centrarea datelor prin calcularea medie µ din setul de date si scazand-o din fiecare punct de date. Acest lucru asigura ca setul de date are media 0. Scaderea medie nu este strict necesara, dar reduce riscul problemelor numerice.

Formula pentru medie:

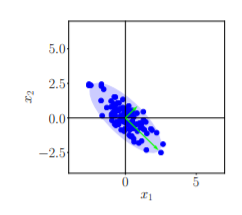


1. Setul initial de date b) Datele centrate dupa scaderea mediei
2. **Standardizarea datelor**. Punctele sunt impartite la abaterea standard σ a setului de date pentru fiecare dimensiune d= 1,..,D. Dupa acest pas datele au varianta egala cu 1.



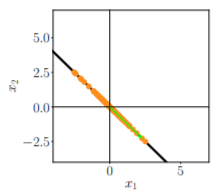


Formula pentru calcularea variantei c) Divizarea datelor la varianta calculata

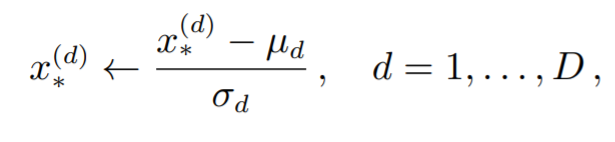


1. **Decompozita Eigen a matricei de covarianta.** Se calculeaza matricea de covarianta si valorile vectoriilor eigen corespunzatori. In figura matricea de covarianta a datelor sunt reprezentate printr-o eclipsa.

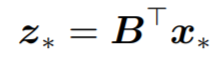
d) Calcularea vectorilor eigen



1. **Proiectia.** Se poate proiecta orice punct pe subspatiul principal. Mai intai, trebuie standardizat, utilzand media si abaterea standard calculate anterior.



e) proiectia datelor pe spatial principal

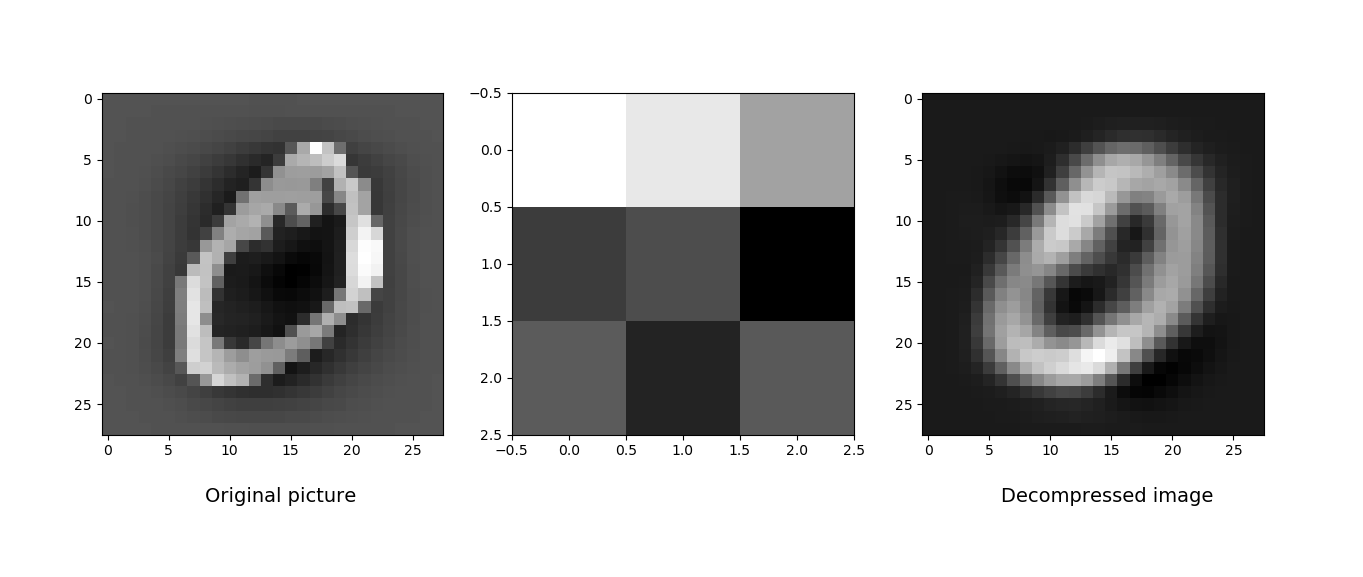
 unde este a d-a componenta a lui .

Se obtine proiectia:

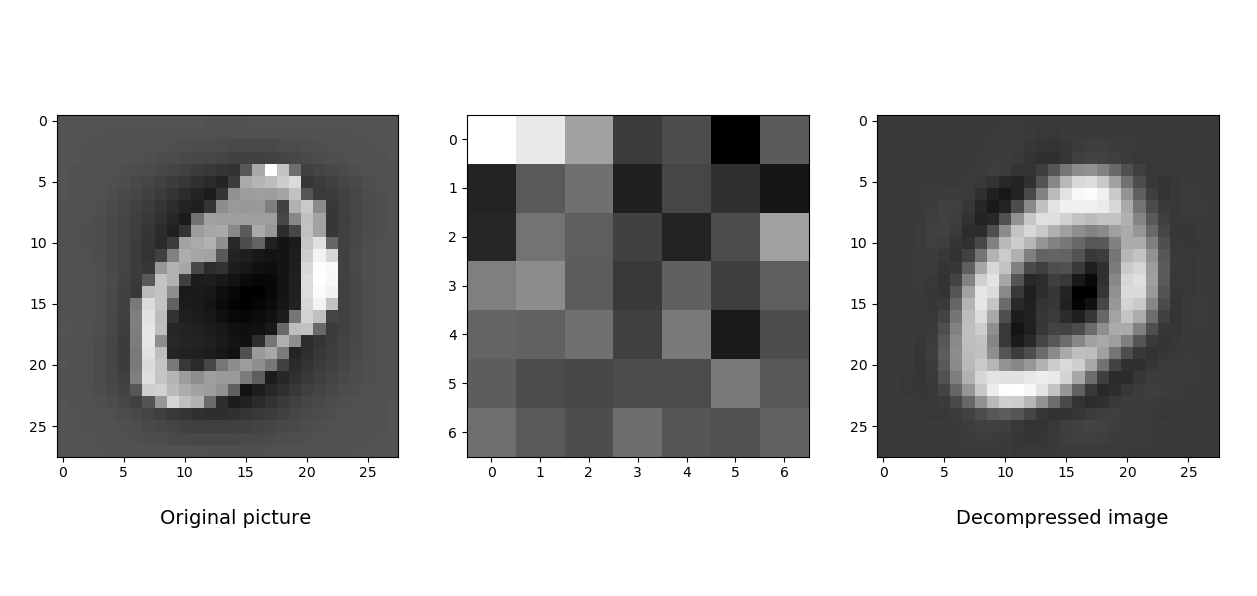
Rezultate

Ca si imagini de test s-a folosit setul de date MNIST care contine 60.000 de imagini cu cifre scrise de mana de la 0 la 9. Fiecare imagine fiind o imagine greyscale de dimensiune 28 x 28. Fiecare imagine prelucrata ca fiind un vector de 784 pixeli.

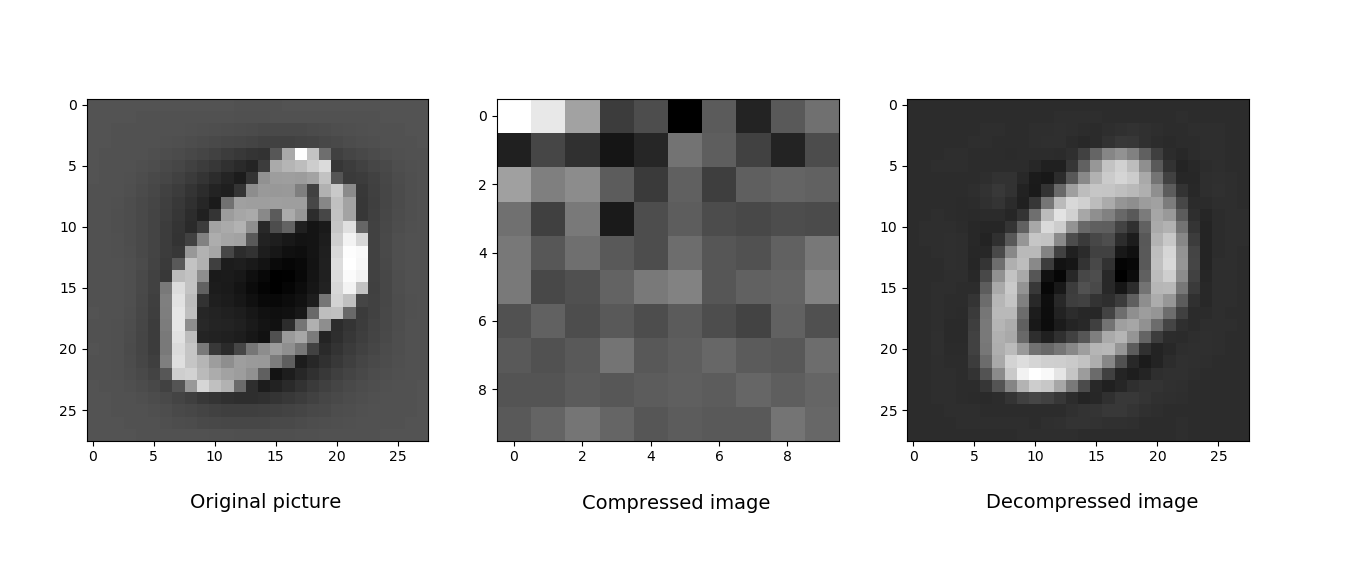
Compactare din 784 in 9 dimensiuni:



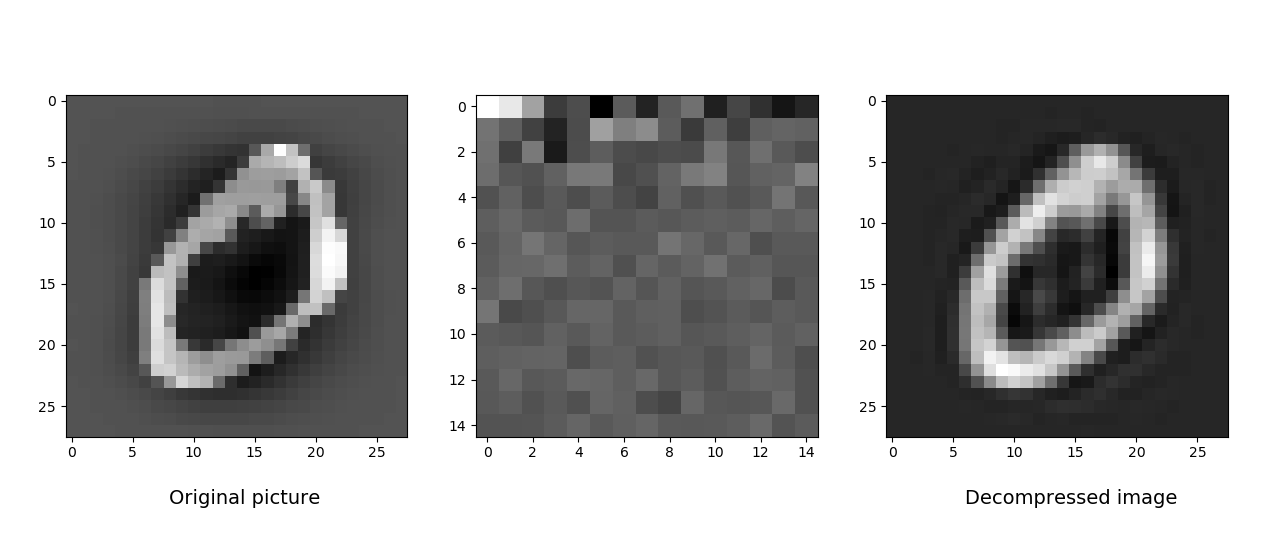
Compactare din 784 in 49 de dimensiuni:



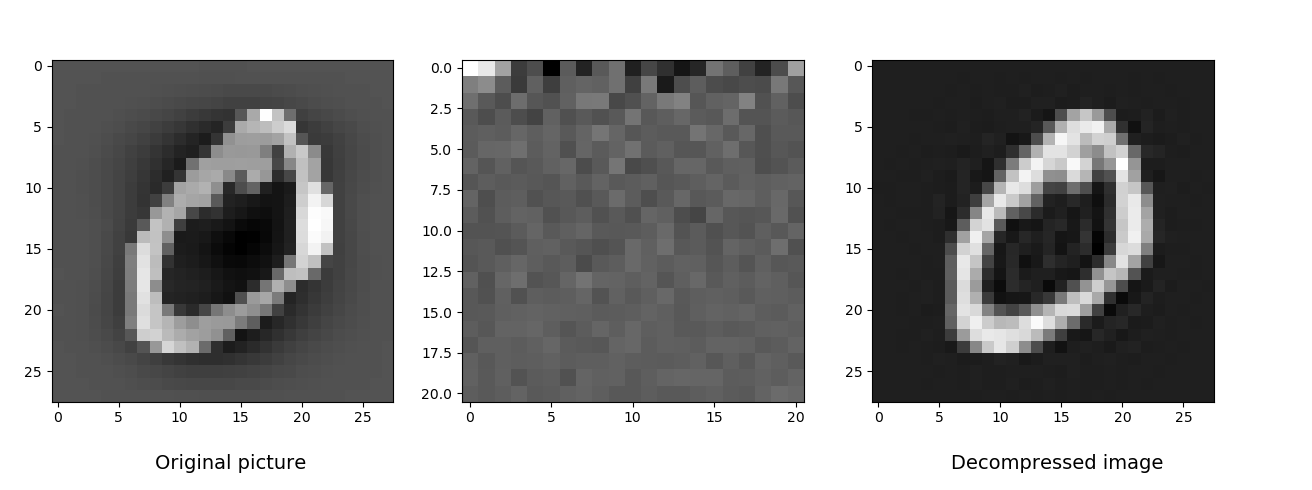
Compactare din 784 in 100 de dimensiuni:



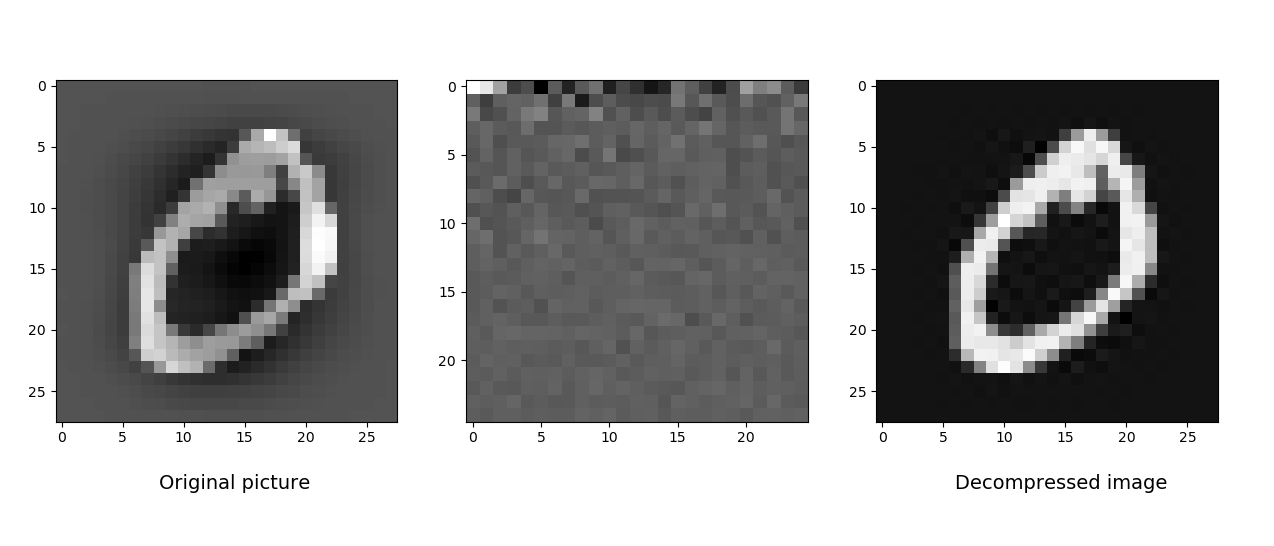
Compactare din 784 in 225 dimensiuni:



Compactare din 784 in 441 dimensiuni:



Compactare din 784 in 576 de dimensiuni:



Bibliografie

1. [Mathematics for Machine Learning](https://mml-book.github.io/book/mml-book.pdf)
2. [Application of Principal Component Analysis to Image Compression](https://www.researchgate.net/publication/328798241_Application_of_Principal_Component_Analysis_to_Image_Compression)
3. [PCA Wikipedia](http://people.ciirc.cvut.cz/~hlavac/TeachPresEn/11ImageProc/15PCA.pdf)