Data Mining 2
TP 2.6 - Recommandation

Thomas Robert

L'objectif du TP est d'étudier le problème Netflix en essayant de déterminer la note qu'une personne attribuerait à une film a partir des notes qu'elle a attribuée aux autres films.

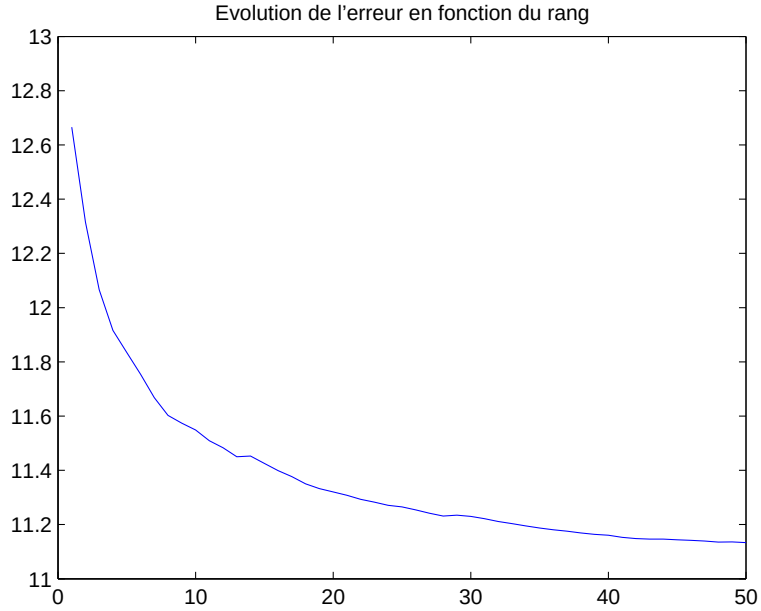
Pour cela, on applique une méthode factorielle qui consiste donc à factoriser la matrice de données sous la forme d'un produit de 2 matrices plus petites U et V.

Pour cela, on calcule les k premiers vecteurs singuliers de la matrice. On utilisera pour cela la fonciton `lansvd` de `PROPACK` plutot que `svds` de Matlab pour des raisons d'optimisation (3 fois plus rapide).

On estime ensuite les notes à "deviner" grâce à U et V et on les compare aux données de tests pour évaluer la qualité de la méthode.

On constate que plus on prends de vecteurs singuliers, plus les résultats sont bon, jusqu'à 30. Je n'ai pas testé plus loin pour des raisons de temps de calculs, mais il semblerai selon les résultats du challenge Netflix qu'il faille prendre beaucoup de vecteurs propres pour commencer à faire du sur-apprentissage.

En plus de quelques amélioration mémoire dans le calcul de l'erreur, j'ai essayé d'implémenter une méthode de soft-shrinkage. Malheureusement, cette méthode n'a apporté aucune modification de l'erreur supérieure à 10^{-14}, donc rien de signifiant.




```
% nombre de vecteurs singuliers
k = 50;

% Calcul des vecteurs singuliers
tic
[U,D,V] = svds(netflix_data_app, k);
disp(['Time to compute SVD with svds : ' num2str(toc) 's']);
U=0;V=0;D=0; % clear RAM
tic
[U, D, V] = lansvd(netflix_data_app, k, 'L');
diagD = diag(D);
```

```
disp(['Time to compute SVD with propack : ' num2str(toc) 's']);

% recherche des éléments non nuls dans probe (éléments à estimer)
[i,j,s] = find(netflix_data_probe);
nt = length(s);

% Reconstructions
for nbVS = 1:k;

    % hard shrinkage
    tic
    Err(nbVS) = 0;
    d = D(1:nbVS,1:nbVS);
    for ii=1:nt
        rec = U(i(ii),1:nbVS)*d*V(j(ii),1:nbVS)';
        err = (rec - s(ii))^2;
        Err(nbVS) = Err(nbVS) + err;
    end
    Err(nbVS) = Err(nbVS) / nt;
    disp(['Time to reconstruct for ' num2str(nbVS) ' rank without soft-shrinkage : ' num2str(toc) 's
        - Err : ' num2str(Err(nbVS))]);

    % soft shrinkage
    if (nbVS < k)
        ErrSS(nbVS) = 0;
        tic
        d = diagD(1:nbVS);
        d = diag(soft_shrinckage(d, D(nbVS+1), d(ceil(nbVS*2/3))));
        for ii=1:nt
            rec = U(i(ii),1:nbVS)*d*V(j(ii),1:nbVS)';
            err = (rec - s(ii))^2;
            ErrSS(nbVS) = ErrSS(nbVS) + err;
        end
        ErrSS(nbVS) = ErrSS(nbVS) / nt;
        disp(['Time to reconstruct for ' num2str(nbVS) ' rank with    soft-shrinkage : ' num2str(toc)
            's - Err : ' num2str(ErrSS(nbVS))]);

    end

end

% Evolution de l'erreur
figure;
plot(Err);
title('Evolution de l''erreur en fonction du rang');
```

```
Time to compute SVD with svds : 516.3756s
Time to compute SVD with propack : 170.6699s
Time to reconstruct for 1 rank without soft-shrinkage : 29.1694s - Err : 12.6657
Time to reconstruct for 1 rank with    soft-shrinkage : 29.1986s - Err : 12.6657
Time to reconstruct for 2 rank without soft-shrinkage : 18.4652s - Err : 12.3152
Time to reconstruct for 2 rank with    soft-shrinkage : 17.9975s - Err : 12.3152
Time to reconstruct for 3 rank without soft-shrinkage : 22.9109s - Err : 12.0659
Time to reconstruct for 3 rank with    soft-shrinkage : 19.2011s - Err : 12.0659
Time to reconstruct for 4 rank without soft-shrinkage : 20.393s - Err : 11.9157
Time to reconstruct for 4 rank with    soft-shrinkage : 20.3951s - Err : 11.9157
Time to reconstruct for 5 rank without soft-shrinkage : 19.905s - Err : 11.8342
Time to reconstruct for 5 rank with    soft-shrinkage : 21.4893s - Err : 11.8342
Time to reconstruct for 6 rank without soft-shrinkage : 21.296s - Err : 11.7546
Time to reconstruct for 6 rank with    soft-shrinkage : 23.4792s - Err : 11.7546
Time to reconstruct for 7 rank without soft-shrinkage : 19.4804s - Err : 11.6681
Time to reconstruct for 7 rank with    soft-shrinkage : 21.3661s - Err : 11.6681
Time to reconstruct for 8 rank without soft-shrinkage : 21.4469s - Err : 11.6028
Time to reconstruct for 8 rank with    soft-shrinkage : 19.7388s - Err : 11.6028
Time to reconstruct for 9 rank without soft-shrinkage : 19.6264s - Err : 11.5743
Time to reconstruct for 9 rank with    soft-shrinkage : 19.9285s - Err : 11.5743
Time to reconstruct for 10 rank without soft-shrinkage : 19.7391s - Err : 11.5485
Time to reconstruct for 10 rank with    soft-shrinkage : 20.9487s - Err : 11.5485
```

```
Time to reconstruct for 11 rank without soft-shrinkage : 21.005s - Err : 11.5088
Time to reconstruct for 11 rank with    soft-shrinkage : 20.8367s - Err : 11.5088
Time to reconstruct for 12 rank without soft-shrinkage : 20.6141s - Err : 11.4828
Time to reconstruct for 12 rank with    soft-shrinkage : 19.4083s - Err : 11.4828
Time to reconstruct for 13 rank without soft-shrinkage : 20.0996s - Err : 11.4499
Time to reconstruct for 13 rank with    soft-shrinkage : 22.2349s - Err : 11.4499
Time to reconstruct for 14 rank without soft-shrinkage : 21.075s - Err : 11.4527
Time to reconstruct for 14 rank with    soft-shrinkage : 20.7026s - Err : 11.4527
Time to reconstruct for 15 rank without soft-shrinkage : 20.7379s - Err : 11.4251
Time to reconstruct for 15 rank with    soft-shrinkage : 24.5853s - Err : 11.4251
Time to reconstruct for 16 rank without soft-shrinkage : 23.7491s - Err : 11.3988
Time to reconstruct for 16 rank with    soft-shrinkage : 21.1658s - Err : 11.3988
Time to reconstruct for 17 rank without soft-shrinkage : 19.6393s - Err : 11.3769
Time to reconstruct for 17 rank with    soft-shrinkage : 19.3621s - Err : 11.3769
Time to reconstruct for 18 rank without soft-shrinkage : 19.9756s - Err : 11.3502
Time to reconstruct for 18 rank with    soft-shrinkage : 21.5878s - Err : 11.3502
Time to reconstruct for 19 rank without soft-shrinkage : 20.8397s - Err : 11.3324
Time to reconstruct for 19 rank with    soft-shrinkage : 20.5016s - Err : 11.3324
Time to reconstruct for 20 rank without soft-shrinkage : 19.7294s - Err : 11.3203
Time to reconstruct for 20 rank with    soft-shrinkage : 22.058s - Err : 11.3203
Time to reconstruct for 21 rank without soft-shrinkage : 19.6424s - Err : 11.308
Time to reconstruct for 21 rank with    soft-shrinkage : 19.8191s - Err : 11.308
Time to reconstruct for 22 rank without soft-shrinkage : 20.8974s - Err : 11.2931
Time to reconstruct for 22 rank with    soft-shrinkage : 20.1793s - Err : 11.2931
Time to reconstruct for 23 rank without soft-shrinkage : 22.6677s - Err : 11.2824
Time to reconstruct for 23 rank with    soft-shrinkage : 19.7633s - Err : 11.2824
Time to reconstruct for 24 rank without soft-shrinkage : 19.5498s - Err : 11.2707
Time to reconstruct for 24 rank with    soft-shrinkage : 19.4062s - Err : 11.2707
Time to reconstruct for 25 rank without soft-shrinkage : 19.787s - Err : 11.2652
Time to reconstruct for 25 rank with    soft-shrinkage : 19.7354s - Err : 11.2652
Time to reconstruct for 26 rank without soft-shrinkage : 19.6752s - Err : 11.2542
Time to reconstruct for 26 rank with    soft-shrinkage : 19.654s - Err : 11.2542
Time to reconstruct for 27 rank without soft-shrinkage : 20.1012s - Err : 11.2419
Time to reconstruct for 27 rank with    soft-shrinkage : 20.0472s - Err : 11.2419
Time to reconstruct for 28 rank without soft-shrinkage : 20.0119s - Err : 11.2314
Time to reconstruct for 28 rank with    soft-shrinkage : 19.6804s - Err : 11.2314
Time to reconstruct for 29 rank without soft-shrinkage : 20.1769s - Err : 11.2345
Time to reconstruct for 29 rank with    soft-shrinkage : 19.9856s - Err : 11.2345
Time to reconstruct for 30 rank without soft-shrinkage : 20.0354s - Err : 11.2301
Time to reconstruct for 30 rank with    soft-shrinkage : 20.0675s - Err : 11.2301
Time to reconstruct for 31 rank without soft-shrinkage : 20.3278s - Err : 11.2213
Time to reconstruct for 31 rank with    soft-shrinkage : 20.2999s - Err : 11.2213
Time to reconstruct for 32 rank without soft-shrinkage : 20.2381s - Err : 11.2111
Time to reconstruct for 32 rank with    soft-shrinkage : 20.0575s - Err : 11.2111
Time to reconstruct for 33 rank without soft-shrinkage : 20.3078s - Err : 11.2033
Time to reconstruct for 33 rank with    soft-shrinkage : 20.3414s - Err : 11.2033
Time to reconstruct for 34 rank without soft-shrinkage : 20.2426s - Err : 11.1951
Time to reconstruct for 34 rank with    soft-shrinkage : 20.3614s - Err : 11.1951
Time to reconstruct for 35 rank without soft-shrinkage : 20.6654s - Err : 11.1873
Time to reconstruct for 35 rank with    soft-shrinkage : 20.7896s - Err : 11.1873
Time to reconstruct for 36 rank without soft-shrinkage : 20.5084s - Err : 11.1806
Time to reconstruct for 36 rank with    soft-shrinkage : 20.7254s - Err : 11.1806
Time to reconstruct for 37 rank without soft-shrinkage : 21.2362s - Err : 11.1755
Time to reconstruct for 37 rank with    soft-shrinkage : 20.6975s - Err : 11.1755
Time to reconstruct for 38 rank without soft-shrinkage : 20.6672s - Err : 11.1691
Time to reconstruct for 38 rank with    soft-shrinkage : 20.6363s - Err : 11.1691
Time to reconstruct for 39 rank without soft-shrinkage : 21.0527s - Err : 11.1636
Time to reconstruct for 39 rank with    soft-shrinkage : 21.0487s - Err : 11.1636
Time to reconstruct for 40 rank without soft-shrinkage : 20.6513s - Err : 11.1608
Time to reconstruct for 40 rank with    soft-shrinkage : 20.6754s - Err : 11.1608
```

```
Time to reconstruct for 41 rank without soft-shrinkage : 21.0006s - Err : 11.153
Time to reconstruct for 41 rank with    soft-shrinkage : 21.2379s - Err : 11.153
Time to reconstruct for 42 rank without soft-shrinkage : 20.8004s - Err : 11.1485
Time to reconstruct for 42 rank with    soft-shrinkage : 21.0708s - Err : 11.1485
Time to reconstruct for 43 rank without soft-shrinkage : 21.2824s - Err : 11.1464
Time to reconstruct for 43 rank with    soft-shrinkage : 21.3163s - Err : 11.1464
Time to reconstruct for 44 rank without soft-shrinkage : 20.6856s - Err : 11.1462
Time to reconstruct for 44 rank with    soft-shrinkage : 20.6959s - Err : 11.1462
Time to reconstruct for 45 rank without soft-shrinkage : 21.2721s - Err : 11.1434
Time to reconstruct for 45 rank with    soft-shrinkage : 21.1986s - Err : 11.1434
Time to reconstruct for 46 rank without soft-shrinkage : 21.0003s - Err : 11.1419
Time to reconstruct for 46 rank with    soft-shrinkage : 21.0633s - Err : 11.1419
Time to reconstruct for 47 rank without soft-shrinkage : 21.4714s - Err : 11.1391
Time to reconstruct for 47 rank with    soft-shrinkage : 21.5156s - Err : 11.1391
Time to reconstruct for 48 rank without soft-shrinkage : 21.0165s - Err : 11.1352
Time to reconstruct for 48 rank with    soft-shrinkage : 21.1159s - Err : 11.1352
Time to reconstruct for 49 rank without soft-shrinkage : 21.4003s - Err : 11.1358
Time to reconstruct for 49 rank with    soft-shrinkage : 21.4995s - Err : 11.1358
Time to reconstruct for 50 rank without soft-shrinkage : 21.3881s - Err : 11.1333
```