

STAT 580 Final Report

Yifan Zhu, Gang Han, Lijin Zhang, Lingnan Yuan

Department of Statistics, Iowa State University

1 Summary of Soft-Impute

[1, 2]

2 Comparison of SVD implementations

3 Application

3.1 Lena

Based the comparison between these 4 implementations, we choose the `original svd` in R to do the image imputing. The reason is that from **Figure** we can see the `original svd` and `rcpparmadillo` seems to be more steady than `irlba` (the gray region that represents confidence interval is smaller). Also these two methods seems to have smaller test error around the true matrix rank. Because the data for lena is not large, thus we do not care about the speed of implementations. And we can expect `original svd` and `rcpparmadillo` would give almost the same results from the comparison part, hence we choose to use `original svd`.

First we randomly removed the 40% of the lena and got an image with only 60% being observed. Now we use the 60% as our data to restore the image with soft-impute. In order to obtain an optimal restoration, we need to tune the parameter λ in the soft-impute algorithm. The strategy we took is to further divide the data into two parts: 70% as training data and 30% as validation data. We did the imputation with a series of λ 's (we used $\lambda = 500, 490, \dots, 10$). Then we compared the corresponding validation part of these imputed matrix with the true validation set. The error term we used is

$$\frac{\|P_{\Omega_{\text{validation}}}(Z_{\text{true}} - \hat{Z}_{\lambda})\|_F^2}{\|P_{\Omega_{\text{validation}}}Z_{\text{true}}\|_F^2} \quad (1)$$

where $\Omega_{\text{validation}}$ is the set of matrix entries that validation set has.

We chose the λ that gave us the smallest test error and applied soft-impute again on the whole data set with the optimal λ we got. In this case, with the test error we defined, the optimal λ we chose is $\lambda = 190$.

After we imputed the matrix, we wanted to see how close it is to the true one. We looked at the rank of recovered matrix and the true matrix, and also the error of the imputed one compared with

the true one. The error we used is

$$\frac{\|P_{\Omega_{\text{test}}}(Z_{\text{true}} - \hat{Z}_{\lambda})\|_F^2}{\|P_{\Omega_{\text{test}}}(Z_{\text{true}})\|_F^2} \quad (2)$$

where Ω_{test} is the set of matrix entries that test set has, which is the rest 40% of the true lena image that is not taken as data.

With the optimal λ , the rank of recovered matrix is

$$\text{rank}(\hat{Z}_{\lambda_{\text{opt}}}) = 77$$

While the true rank is

$$\text{rank}(Z_{\text{true}}) = 254$$

The test error of the imputed matrix when compared with the true matrix in the rest 40% of the image is

$$\text{test error} = 0.031$$

Finally, we compare the true image, incomplete image and the restored image visually, which is shown in Figure 1.



(a) true image

(b) incomplete image

(c) restored image

Figure 1: Visually comparing the imputed image with the true one

We can see that the restored image is not so good. It is not like a natural image with good smoothness. Also the restored image has a much smaller rank than the true one. However, soft-impute is intended to impute low-rank matrix, and true image actually has a quite big rank. That might also be a reason why the final result is not good.

3.2 MovieLens

In this section, we use MovieLen 100K Dataset to apply our soft-impute algorithm. As the data set is not complete when we convert the original data into a matrix form, we need to divide it into three parts: 70% as training set, 15% as validation set and 15% as test set. What we did is similar as we did in lena part. We first use the training set as observed entries to impute the matrix, then we compare the validation set with the corresponding entries of the imputed matrix. Again we use (1) as the error term and we pick the λ with the smallest validation error. After we obtained the optimal λ , we use the 70% training set and 15% validation set combined as our data, or observed

entries to impute the matrix. Then we compare the test set with the corresponding entries of the imputed matrix and calculate the test error using (2). Besides, we calculated the RMSE of the imputed matrix with respect to the test set in corresponding parts. The RMSE formula is

$$\text{RMSE} = \sqrt{\frac{\|P_{\Omega_{\text{test}}}(Z_{\text{true}} - \hat{Z}_{\lambda_{\text{opt}}})\|_F^2}{|\Omega_{\text{test}}|}} \quad (3)$$

As the dataset is pretty big, we chose `irlba` and `rcpparmadillo` as two implementations to be applied to this dataset. There two implementations are relatively faster than others ,thus with the big dataset we could save a lot of time.

In choosing the optimal λ , as the big matrix takes time to impute, we used a strategy to find the optimal one using 3 grids of λ 's by assuming the test error as a function of λ is unimodal. The first grid is $\lambda = 100, 90, \dots, 10$ with step size 10. From this grid we can find the interval where the optimal λ located. We found $\lambda = 10$ was the smallest in this grid. Then we knew the optimal λ is between 20 and 0. Then we used the second grid $\lambda = 19, 16, \dots, 4, 1$ with step size 3. We found $\lambda = 13$ is the smallest in this grid. Then we tried the third grid $\lambda = 15, 14, 12, 11$. In this grid, $\lambda = 12$ gave us the smallest test error. But comparing it with the test error when $\lambda = 13$, we found the test error for $\lambda = 13$ was smaller. Hence, the optimal λ we chose is $\lambda = 13$.

References

- [1] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity*. CRC press, 2015.
- [2] Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of machine learning research*, 11(Aug):2287–2322, 2010.