STAT 580 Final Report

Yifan Zhu, Gang Han, Lijin Zhang, Lingnan Yuan

Department of Statistics, Iowa State University

1 Summary of Soft-Impute

First we want to fit a matrix $\widehat{\mathbf{Z}}$ that approximates $\mathbf{Z}$ in a suitable sense. By doing this, we can find an approximation that has simple structure or fill in any missing entries in $\mathbf{Z}$ which is our purpose.[1]

Our general approach is to consider estimators based on optimization problem of the form:

$$\widehat{\mathbf{Z}} = \arg\min_{\mathbf{M}\in\mathbb{R}^{m\times n}} \|\mathbf{Z}-\mathbf{M}\|_F^2 \qquad \text{subject to} \qquad \Phi(\mathbf{M}) \leq c, \tag{1}$$

$\Phi(\cdot)$ is a constraint function that encourages $\hat{\mathbf{Z}}$ to be sparse in some general sense. There are some kinds of constraint methods, one is the singular value decomposition that we'll discuss later.

The SVD provides a solution to the rank-q matrix matrix approximation problem. Then our optimization problem (1) has the form:

$$\arg\min_{rank(\mathbf{M})=r} \|\mathbf{Z}-\mathbf{M}\|_F. \tag{2}$$

It has a closed form solution $\widehat{\mathbf{Z}}_r = UD_rV^T$. It's sparse in the sense that all but r singular values are 0 and it provides an effective method for matrix completion.

Suppose our observed entries of the matrix $\mathbf{Z}$ indexed by the subset $\Omega \subset 1,\cdots,m \times 1,\cdots,n$. A natural approach is to seek the lowest rank approximating matrix $\widehat{\mathbf{Z}}$. But forcing the estimate $\mathbf{M}$ to interpolate each of the observed entries z_{ij} will be too harsh and can lead to overfitting. So it would be better to allow some errors. Consider the optimization problem:

$$\arg\min_{\text{rank}(\mathbf{M})\leq r} \sum_{(i,j)\in\Omega} (z_{ij}-m_{ij})^2. \tag{3}$$

However, the optimization problem (3) is nonconvex, the exact solutions are in general not available. The nuclear norm is a convex relaxation of the rank of a matrix, so we can give a convex relaxation with noise:

$$\arg\min_{\mathbf{M}} \left\{ \frac{1}{2} \sum_{(ij)\in\Omega} (z_{ij}-m_{ij})^2 + \lambda\|\mathbf{M}\|_* \right\}. \tag{4}$$

Algorithm 1 SOFT-IMPUTE FOR MATRIX COMPLETION[2]

```
Initialize $Z^{\text{old}} = 0$ and create a decreasing grid $\lambda_1 > \cdots > \lambda_n$
for each $k = 1, \cdots, K$, set $\lambda = \lambda_k$ and iterate until convergence: do
    Compute $\widehat{\mathbf{Z}}_\lambda \leftarrow S_\lambda(P_\Omega(\mathbf{Z} + P_\Omega^\perp(Z^{\text{old}})))$.
    Update $Z^{\text{old}} \leftarrow \widehat{\mathbf{Z}}_\lambda$
end for
Output the sequence of solutions $\widehat{\mathbf{Z}}_{\lambda_1}, \cdots, \widehat{\mathbf{Z}}_{\lambda_K}$.
```

Then we can use Soft-Impute Algorithm to solve this problem.

Define the projection operator $P_\Omega : \mathbb{R}^{m\times n} \mapsto \mathbb{R}^{m\times n}$ as follows:

$$[P_\Omega(\mathbf{Z})]_{ij} = \begin{cases} z_{ij} & \text{if} \quad (i,j) \in \Omega \\ 0 & \text{if} \quad (i,j) \notin \Omega \end{cases} \tag{5}$$

Define the soft-thresholded version of SVD as:

$$S_\lambda(\mathbf{W}) \equiv \mathbf{U}\mathbf{D}_\lambda\mathbf{V}^\mathbf{T} \quad \text{where} \quad \mathbf{D}_\lambda = \text{diag}[(d_1 - \lambda)_+, \cdots, (d_r - \lambda)_+]. \tag{6}$$

2 Comparison of SVD implementations

We've got different functions of doing soft-impute implementations using 4 svd methods. Now we want to compare which of these methods is the fastest and which of these methods is the most accurate.

First of all, we simulated several matrices using the underlying model:

$$\mathbf{Z}_{m\times n} = \mathbf{U_{m\times r}}\mathbf{V^T_{r\times n}} + \epsilon$$

Where $\mathbf{U}$ and $\mathbf{V}$ are random matrices with standard normal Gaussian entries, and ϵ is i.i.d. Guassian. Ω is uniformly random over the indices of the matrix with p % of missing entries. Define the SNR, test error and training error as:

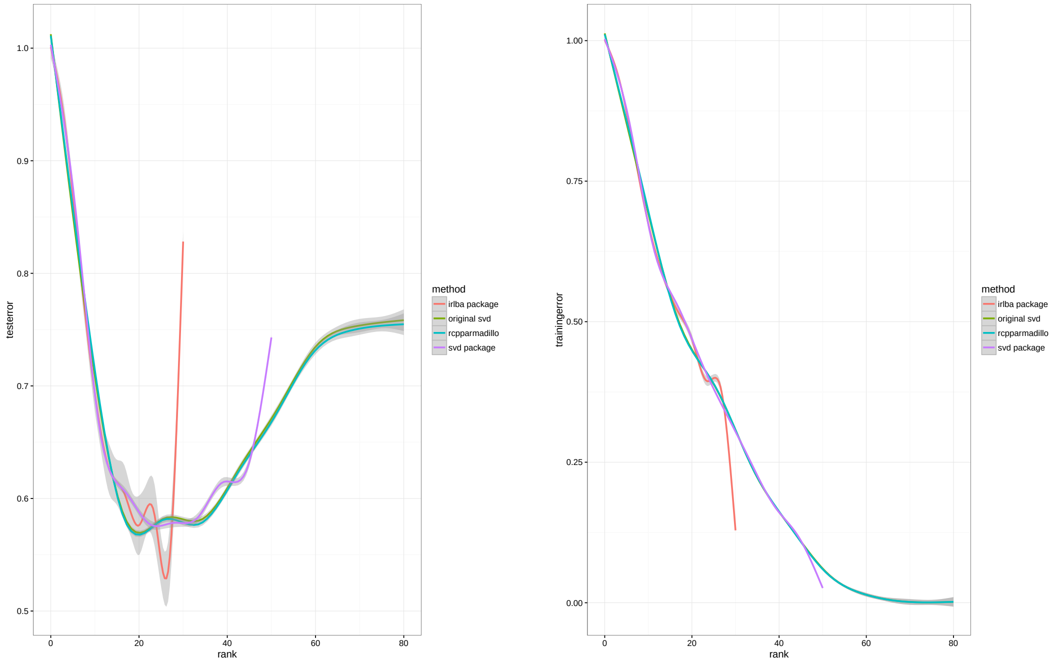
$$\text{SNR} = \sqrt{\frac{\text{var}(UV^T)}{\text{var}(\epsilon)}}$$

$$\text{Test error} = \frac{\|P_\Omega^\perp(UV^T - \widehat{Z})\|_F^2}{\|P_\Omega^\perp(UV^T)\|_F^2}$$

$$\text{Training error} = \frac{\|P_\Omega(Z - \widehat{Z})\|_F^2}{\|P_\Omega(Z)\|_F^2}$$

In the following figures, we show the training and test error for the 4 methods as a function of rank.

As we can see in the figures, the soft-impute using svd methods `original svd` and `rcpparmadillo` perform very similarly. While the method of `irlba package` seems to have lowest Test error at some rank, but it seems fluctuate a lot. the methods `original svd` and `rcpparmadillo`. According to the training error, the method of `irlba package` recovers a matrix with a relatively low rank which is closer to our true rank. So it has less evidence of overfitting.



(a) Test error

(b) Training error

Figure 1: Compare the test error against the rank among the 4 methods with $p = 50\%, \mathrm{SNR} = 1, \text{true rank} = 10, m = n = 100$.

Then we tried different properties of the generated matrix. Let $p = 0.3, 0.5, 0.8$; $\mathrm{SNR} = 1, 10$; true rank $= 5, 10, 20.$,then we get the table that compares the running time and test error among the 4 methods.

time1	time2	time3	time4	error1	error2	error3	error4	true rank	SNR	p
2.1650	2.5915	1.3096	1.7030	0.2592	0.2592	0.2592	0.2592	5	1	0.3
2.0902	2.6631	1.4791	1.6333	0.4338	0.4338	0.4339	0.4338	10	1	0.3
2.0032	2.7237	1.6506	1.5678	0.6502	0.6502	0.6506	0.6502	20	1	0.3
1.9709	1.9810	0.6514	1.5437	0.0046	0.0046	0.0046	0.0046	5	10	0.3
1.9875	1.9486	0.6748	1.5096	0.0113	0.0113	0.0111	0.0113	10	10	0.3
2.0435	1.9584	0.6629	1.5443	0.0443	0.0432	0.0312	0.0443	20	10	0.3
2.8228	2.8239	1.3798	2.2199	0.3492	0.3492	0.3492	0.3492	5	1	0.5
2.6000	2.8008	1.6049	2.0534	0.5798	0.5798	0.5798	0.5798	10	1	0.5
2.4068	2.5195	1.7681	1.8531	0.7970	0.7970	0.7970	0.7970	20	1	0.5
2.8192	2.8098	0.9243	2.2188	0.0097	0.0097	0.0097	0.0097	5	10	0.5
2.7298	2.7674	0.9625	2.1510	0.0319	0.0319	0.0306	0.0319	10	10	0.5
2.6173	2.6013	0.9362	2.0141	0.1913	0.1896	0.1488	0.1913	20	10	0.5
4.1483	4.1360	2.3177	3.2476	0.7382	0.7382	0.7382	0.7382	5	1	0.8
4.4981	4.4856	2.5795	3.5294	0.9040	0.9040	0.9040	0.9040	10	1	0.8
3.8848	3.8947	2.4525	3.0592	0.9829	0.9829	0.9829	0.9829	20	1	0.8
3.9987	3.9948	1.5107	3.1125	0.1739	0.1739	0.1746	0.1739	5	10	0.8
4.1852	4.1783	1.9447	3.2685	0.5261	0.5261	0.5259	0.5261	10	10	0.8
4.6070	4.5803	2.4109	3.6193	0.8458	0.8458	0.8458	0.8458	20	10	0.8

Table 1: Performance of different methods for different problem instances, in terms of test error and running time.

This table shows that the minimum test errors among the 4 methods are quite similar, but the running time of method `irlba package` is the shortest followed by `rcpparmadillo`.

According to the comparison above, we'll use the method `irlba package` in our application part, because it's the fastest one and in the figure it turns out to have minimum test error. We'll also use `rcpparmadillo`, because it's also very fast and is more stable in the figure.

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 `orginal 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 $\lambda = 500, 490, \ldots, 10$). Then we compared the corresponding validation part of these imputed matrix with the true validation set. The error term we used is

$$\frac{\|P_{\Omega_{\text{validation}}}(Z_{\text{true}} - \hat{Z}_{\lambda})\|_F^2}{\|P_{\Omega_{\text{validation}}} Z_{\text{true}}\|_F^2} \tag{7}$$

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

$$\frac{\|P_{\Omega_{\text{test}}}(Z_{\text{true}} - \hat{Z}_{\lambda})\|_F^2}{\|P_{\Omega_{\text{test}}}(Z_{\text{true}})\|_F^2} \tag{8}$$

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

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

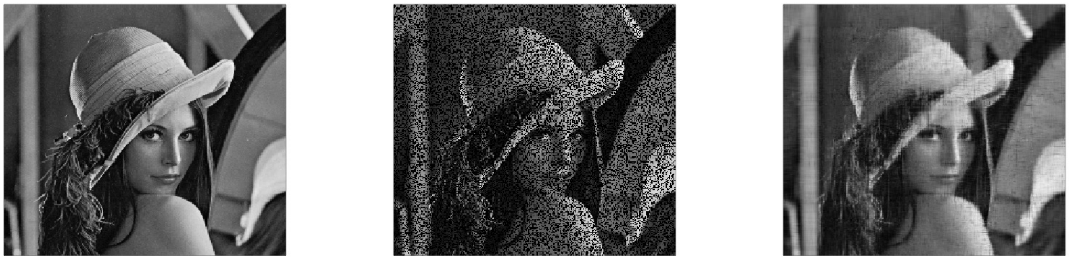
While the true rank is

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

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 2.



(a) true image

(b) incomplete image

(c) restored image

Figure 2: 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 (7) 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 (8). Besides, we calculated the RMSE of the imputed matrix with respect to the test set in corresponding parts. The RMSE formula is

$$\text{RMSE} = \sqrt{\frac{\|P_{\Omega_{\text{test}}}(Z_{\text{true}} - \hat{Z}_{\lambda_{\text{opt}}})\|_F^2}{|\Omega_{\text{test}}|}} \tag{9}$$

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 $\lambda = 100, 90, \ldots, 10$ 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 $\lambda = 19, 16, \ldots, 4, 1$ 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.