Imputation of Missing Values by Low Rank Matrix Approximation

MoonJung Cho*

Key Words: Auxiliary variables, Correlation, Rank estimation, Singular value decomposition.

1. Introduction

The analysis of sample survey data often requires adjustments for missing values in variables of interest. Standard adjustments based on item imputation or propensity weighting factors rely on the availability of auxiliary variables for both responding and non-responding units. However, the application of these can be challenging when the auxiliary variables are themselves subject to incomplete-data issues. This paper will demonstrate how low rank matrix approximation can be applied to impute missing auxiliary variables. The performance depends on the rank of the auxiliary variable matrix and the extent to missingness rates. We will evaluate the method in terms of bias and mean squared error.

2. Low Rank Matrix Approximation

In survey data, auxiliary variables are sometimes called predictor variables or explanatory variables. We consider auxiliary variables X in a matrix form and are interested in the imputation of auxiliary variables. Hence, observations are in rows, and variables are in columns. We noted that predictor variables are typically chosen because they are correlated to a dependent variable. The higher correlation ensures the better prediction or explanation that auxiliary variables can provide about the dependent variable. This results in high correlations among the auxiliary variables themselves. For example, when a dependent variable is a price of specific commodity, the 1-month and 3-month previous prices can be correlated against each other. This may lead X (auxiliary variable matrix) to be low rank.

The rank of matrix X is the number of independent columns (or rows). A matrix X is full rank if

$$\text{rank of } X = \min(n, p)$$

where n is the number of rows and p is the number of columns. X is low or deficient rank if it is not a full rank. In practice, statistical software such as MatLab estimate a rank of matrix by counting non-zero singular values after considerable numerical adjustment. We consider cases where the

*Office of Survey Methods Research, U.S. Bureau of Labor Statistics, 2 Massachusetts Avenue NE, Washington, DC 20212

rank of X is significantly low and show a way to impute missing observations of X. Let auxiliary variable matrix X be an $n \times p$ matrix with missing entries at $(i, j) \in \Omega$:

$$\Omega = \{(i, j) : \ X(i, j) \text{ is missing}\}.$$

Our goal is to find a low rank matrix which has the smallest sum of singular values (i.e., nuclear norm) while its non-missing entries are the same as non-missing entries of X.

3. Rank Estimation

Let a matrix X be any real $n \times p$ matrix with no missing entries. We can then decompose X uniquely as a product of orthogonal matrices U and V, and a diagonal matrix S:

$$X = USV'$$

where $U \in O(n)$, $V \in O(p)$ and S is $n \times p$ diagonal matrix. The diagonal elements of S are called the singular values and they satisfy $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq 0$.

X can be a full rank matrix even when variables are dependent among each other. In such cases, some of the singular values of X may be close to zero. When variables are highly dependent among each other, we may examine the magnitude of singular values instead of accepting its nominal rank. Suppose the rank $r_\epsilon(X)$ is the number of singular values greater than ϵ. If some of the singular values of X are close to zero, the rank $r_\epsilon(X)$ is much smaller than the nominal rank of X but a better choice for practical applications.

We now show how to estimate the rank of X when X has missing observations. Note that the singular values of $X'X$ are the square of the singular values of X. That means $X'X$ and X have the same number of non-zero singular values, and to estimate the rank of $X'X$ is to estimate the rank of X.

Let $Q = X'X$, i.e.,

$$Q_{ij} = \sum_{k=1}^{n} X_{ki} X_{kj}$$

for $1 \leq i, j \leq p$. Let Γ_{ij} be the set of indices of rows for which the i-th and j-th columns have no missing values:

$$\Gamma_{ij} = \{k : \ \text{neither of } X_{ki} \text{ and } X_{kj} \text{ is missing}\}.$$

Estimate Q_{ij} by

$$\hat{Q}_{ij} = \frac{n}{\|\Gamma_{ij}\|} \sum_{k \in \Gamma_{ij}} X_{ki} X_{kj}$$

where $\|\Gamma_{ij}\|$ is the size of Γ_{ij}. We then obtain singular values of $\hat{Q}_{ij}$. Estimates of singular values of X are positive square root of the singular values of $\hat{Q}_{ij}$.

We simulated a rank-3 auxiliary variable matrix where a number of observations is 500 and a number of variables is 10. Figure 1 shows the cumulative sum of singular values where there is no

missing observation: on the horizontal axis are singular values from the largest to the smallest; on the vertical axis is the cumulative sum of singular values. Since the first (largest) value of singular values is 56.25, the cumulative sum is 56.25; the second (largest) value of singular values is 6.21, hence the cumulative sum increases by 6.21. Since it is a rank-3 matrix, the rest of singular values are 0 after the third, and there is no increase in the cumulative sum.

The red circles in Figure 2 display $X'X$ rank estimation after removing 10% of observations of X at random. We observed that the first three singular values were well in line with true values and had a slight increment for the rest of values. Figure 3 shows the performance of $X'X$ rank approximation as we increase a missing rate to 30% of observations. The black diamonds display $X'X$ rank estimation after having 30% of observations of X removed at random. We observed that values from $X'X$ rank estimation were farther away from true values as the missing rate increases. Table 1 shows the values of singular values with various missing rates.

Table 1: $X'X$ Rank Estimation With Various Missing Rates

TRUE	10%	30%	50%
56.25	56.07	56.35	56.21
6.21	6.30	6.81	7.15
4.12	4.24	4.38	6.14
0.00	3.24	4.29	5.31
0.00	2.06	3.35	3.84
0.00	1.40	3.00	3.38
0.00	1.38	2.34	2.68
0.00	1.15	1.60	2.59
0.00	0.67	1.20	1.41
0.00	0.33	1.18	0.63

4. Imputation Using Singular Value Decomposition

Recall the singular value decomposition of any real $n \times p$ matrix X without any missing entries. Then X can be decomposed as a product of orthogonal matrices and a diagonal matrix:

$$X = USV'$$

and the diagonal elements satisfy $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq 0$.

Now, let S_k be the $n \times p$ diagonal matrix with the first k singular values $\sigma_1, \ldots, \sigma_k$ and the rest of the diagonal set to zero. Then the rank-k approximation $X_k = US_kV'$ satisfies

$$\|X - X_k\| \leq \|X - A\|$$

for any rank-k matrix A. It means the difference between X_k and X is the smallest, and X_k is the best approximation of X among all rank-k matrices.

For a matrix X with missing entries, imputation of X can be done by an iterative procedure which shrinks the singular values of an initially imputed matrix. Let X be an $n \times p$ rank-r matrix with a number of random entries are missing. Let Z be X where missing values of X are replaced by initially imputed values. For example, one often imputes a missing value initially with the mean or median of non-missing entries of the column in which a missing entry falls: $Z(i,j) = mean(X(:,j))$ for $(i,j) \in \Omega$ and $Z(i,j) = X(i,j)$ for $(i,j) \notin \Omega$.

We modify Z to a matrix of low rank by shrinking its singular values (more specifically a nuclear norm). We then update missing values with values from modified Z. Each iteration step reduces the singular values of Z, while keeping the non-missing entries of X unchanged. The procedure is run iteratively until it meets the given criteria:

1. Find the singular value decomposition (SVD) of Z: $Z = USV'$
2. Shrink the singular values of Z by setting singular values

$$\sigma_i = \begin{cases} \sigma_i - \sigma_{r+1} & \text{for } i \leq r \\ 0 & \text{for } i > r. \end{cases}$$

3. Replace $Z_{i,j}$ with $(UTV')_{i,j}$ for $(i,j) \in \Omega$.
4. Repeat at 1.

The MatLab code is given below:

```
% I is an indicator matrix: 1 for missing; 0 otherwise
I =isnan(X);
Z =X;

% fill each missing value of Z initially
% with the mean of non-missing elements of its column
for j=1:p
Z(I(:,j),j)=nanmean(X(:,j));
end

% missing values of Z are updated iteratively
% by newly computed values of W using SVD
% repeat k times
for i=1:k
[U S V] = svd(Z);
S  = max(0, S-S(r+1,r+1));
W  = U*S*V';
Z(I) = W(I);
end
```

We compared the approximation with the nearest neighbor and column mean imputation methods. The nearest imputation method computes distances among observations using non-missing values across variables, and then choose the nearest observation to impute missing values. The column mean method fills a missing value with its column mean of non-missing entries

Figure 4 plots predicted y values against true y values when missing rate is 10% at random. If an imputation method were perfect, its predicted values would fall on $y = x$ line. We observed that the predicted values of the low rank approximation followed the line more closely compared to the other two imputation methods. Figure 5 shows boxplots of differences between predicted y and true y values when missing rate is 10% at random. All three methods centered at 0 but values of the low rank approximation were less variable and stayed closer to 0. Figure 6 shows that predicted values from all three methods became more variable when we increase missing rate from 10% to 30%.

We also compared relative errors of the methods. Relative Error is estimated:

$$\sqrt{\sum_{i}^{n}(y_i - \hat{y}_i)^2} \Big/ \sqrt{\sum_{i}^{n} {y_i}^2} \ .$$

Table 2 shows the relative errors of the methods. As the missing rate increases, the relative errors of all methods increase. The values of relative error of low rank approximation were smaller compare with the other two methods throughout varying missing rates.

Table 2: Relative Errors

	10%	30%	50%
Low Rank Approx	0.010	0.039	0.075
Nearest Neighbor	0.018	0.056	0.099
Column Mean	0.046	0.113	0.165

5. Discussion

We considered cases where the rank of X is significantly low and showed how the low rank matrix approximation could be applied to impute missing auxiliary variables. We also considered how to estimate the rank of X when X had missing values.

If X is assumed to be of low rank, the missing values can be imputed so that its nuclear norm (the sum of singular values) is minimized. Since the norm is a convex function, the imputation can be converted to a constrained convex optimization problem. Imputation can be done by solving a constrained convex optimization problem which finds a matrix with a minimum norm while having the same entries of non-missing values of X. The matrix can be found by any suitable optimization algorithm, for example, CVX package of MatLab.

REFERENCES

Loh, W.-Y., Eltinge, J., Cho, M. and Li, Y. (2019), "Classification and regression trees and forests for incomplete data from sample surveys," *Statistica Sinica*, vol. 29, 431-453.

Martinez, W.L. and Cho, M.J. (2015), *Statistics In Matlab A Primer*, CRC Press, New York.

Strang, G. (2019), *Linear Algebra and Learning from Data*, Wellesley-Cambridge Press.

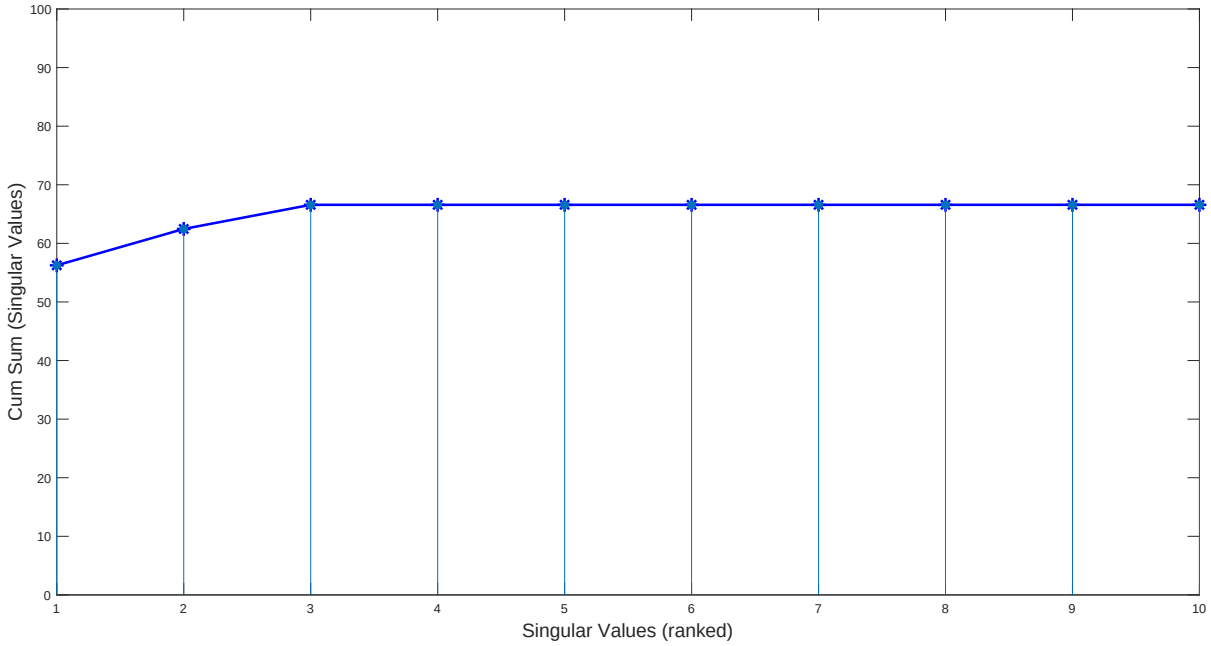

Figure 1: **Cumulative Sum of Singular Values (**$n = 500; p = 10; rank = 3$**)**

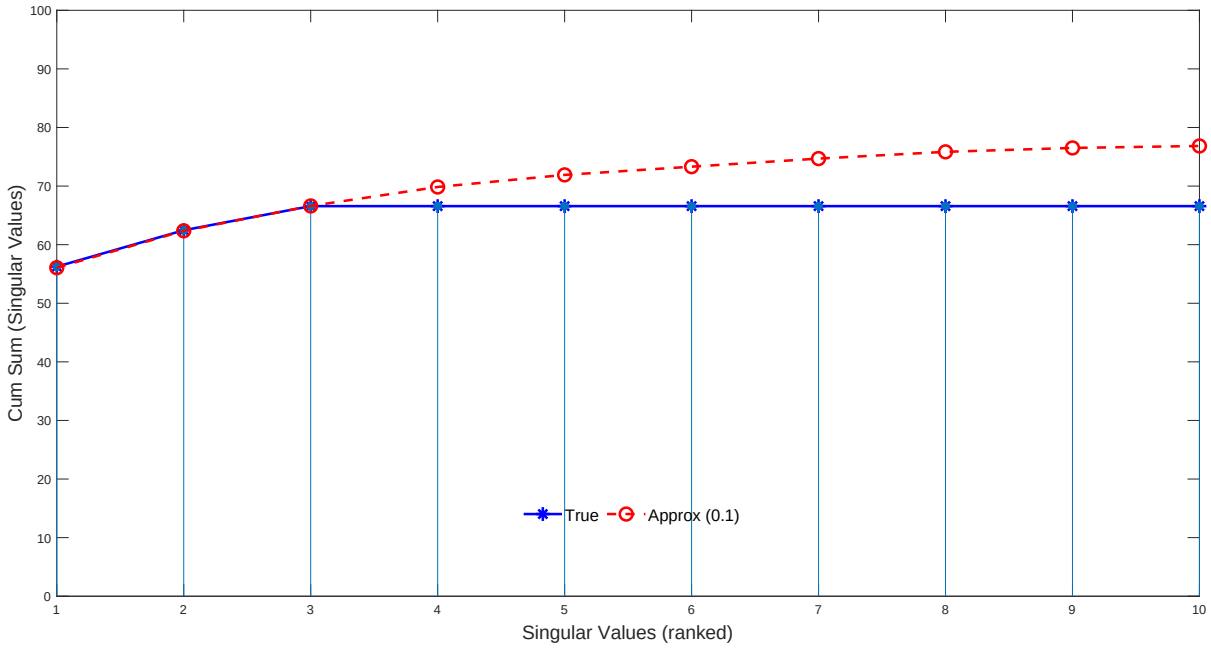

Figure 2: **Cumulative Sum of Singular Values of Simulated X with Missing 10% ($n = 500; p = 10; rank = 3$)**

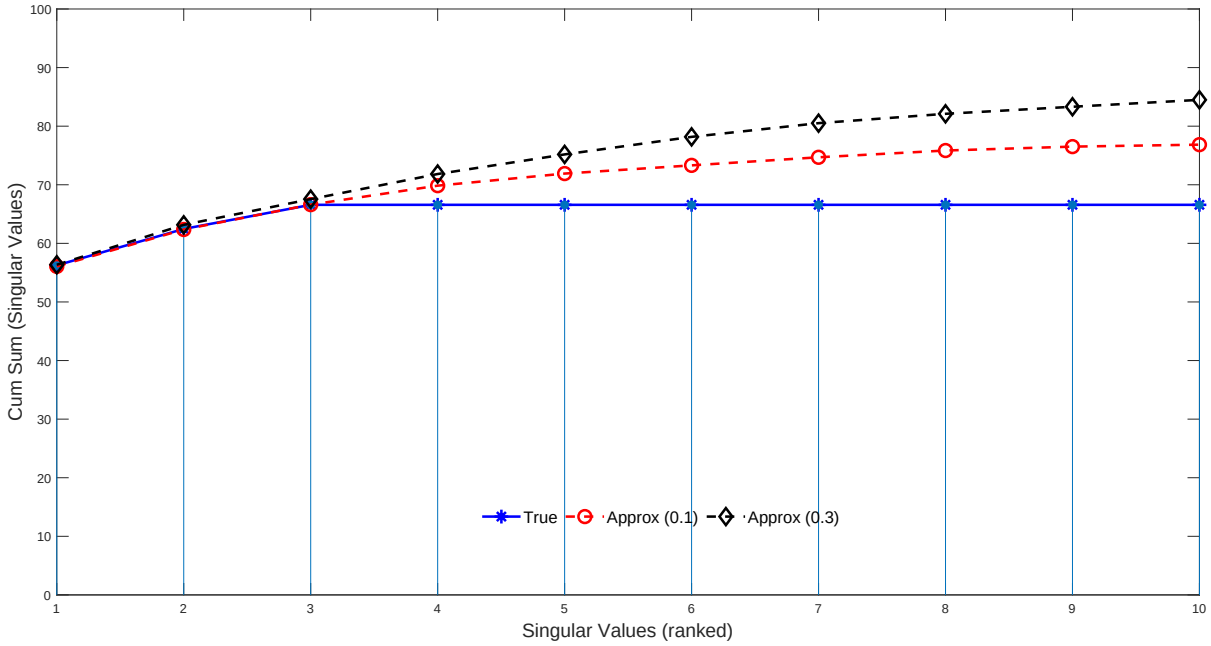

Figure 3: **Cum Sum of Singular Values with Missing** 10%**,** 30%

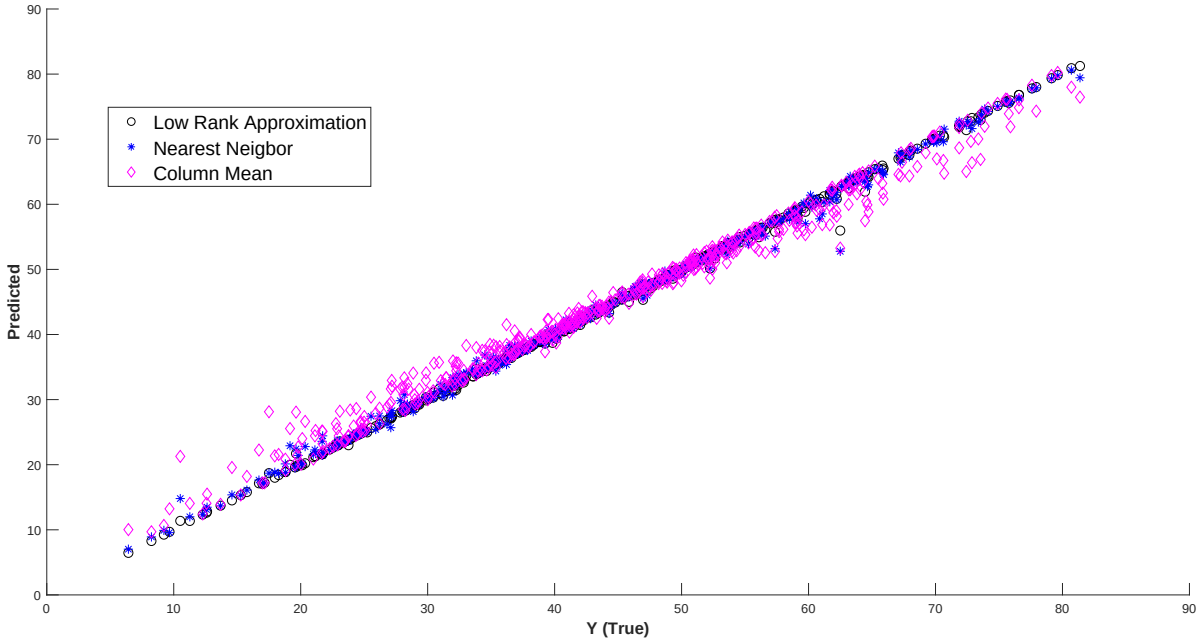

Figure 4: **Predicted (Y) n = 500; p = 10; rank = 3, missing rate = 10%**

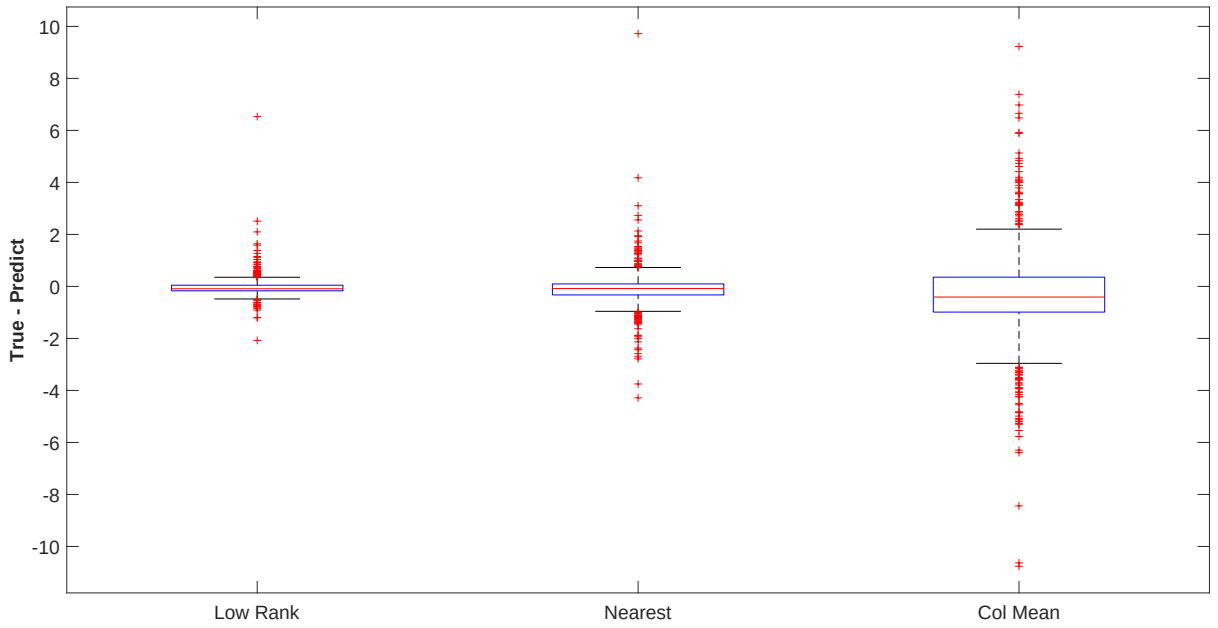

Figure 5: **Boxplot of Difference between True (Y) and Predicted (Y) when missing rate is 10%**

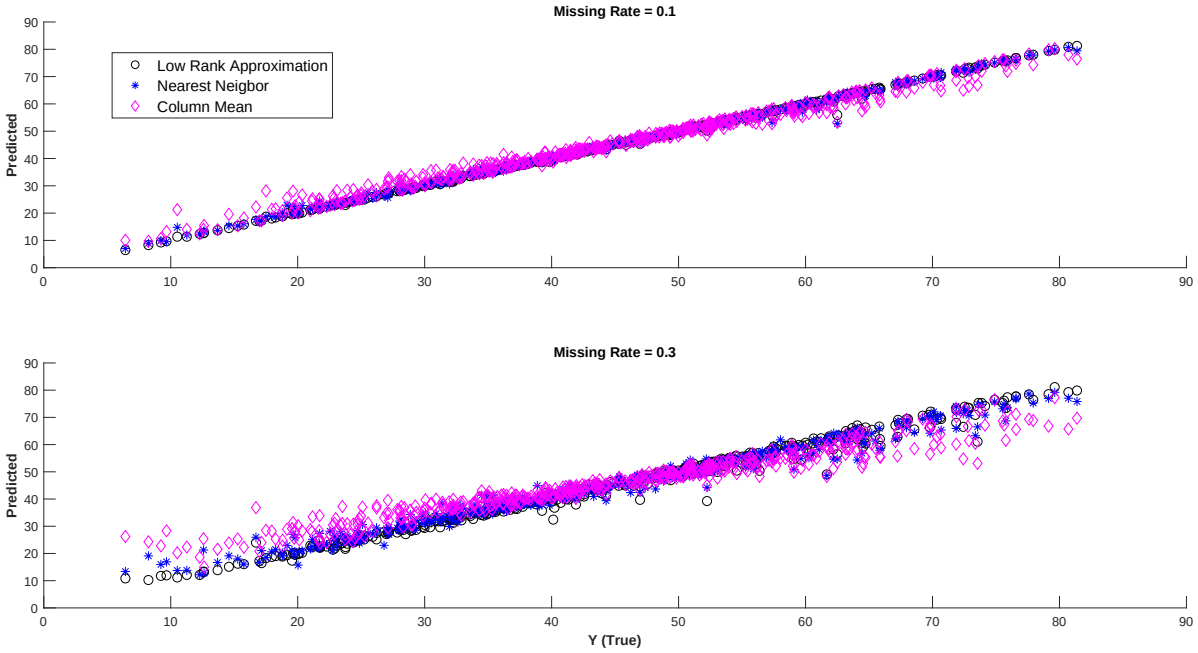

Figure 6: **Predicted Y with missing rate = 10% and 30%**