1. Introduction

This report is a preliminary test and application of an interpolation method based on principal component analysis (PCA). Data interpolation, as always, is broadly applied in various studies including raster data process, remote sensing, signal process, etc. LU et al. (2017) has applied a PCA-coKriging method interpolated station based PM2.5 concentration map of Beijing. And Yang et al (20020) used PCA together with inverse distance weighted (IDW) as interpolation method dealing with water quality in Xin Jiang province. Above these studies, PCA was cooperated with another classical interpolation method, while this report focus on an application of pure PCA data reconstruction. In order to solve the interpolation problem, especially for raster data with regional bulk missing or massive missing, we one on head applied this method for data interpolation; on the other head, we test its statistical feasibility and adaptability for some special conditions. This method was proposed by Oliveira & Gomes in 2009. And a similar method was applied in picture sharpness by Jiji & Chaudhuri in 2004, although they had nearly the same PCA process for interpolation. In this report, the second part is statistical deduction of PCA based interpolation process. And the following part is two application of this method: one is a remote sensing data application and the other is a grayscale image interpolation test. The forth part is conclusion about specific conditions which adapted to this interpolation method.

This project aimed to specify there problems:

1) Accomplish PCA based interpolation method within R.
2) Usage of Taiyi Super Perform Computer for parallel computing.
3) Test two missing conditions, single point missing and region missing, for this method.

2. Methods

2.1 PCA interpolation process

PCA based reconstruction is an unbiased estimation method when deal with interpolation (Jolliffe, 1986). This method compress a set of high dimensional data into a set of lower dimension. Its ensemble covariance contains the most information even with missing data. And given the model parameters, projection into and from the bases, represent computationally inexpensive operations (Gomes & Oliveira, 2008). The following part will discuss the process of how PCA works for interpolation.

Suppose we have an observed dataset matrix X. Each column represent a observed result series:

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1M} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{NM} \end{pmatrix}_{N\times M}$$

$$= (\boldsymbol{x}_1 \quad \boldsymbol{x}_2 \quad \cdots \quad \boldsymbol{x}_M)$$

And an indicator index matrix A which indicate any element in X is available or not with 1 and 0 means valid and missing:

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1M} \\ \vdots & \ddots & \vdots \\ a_{N1} & \cdots & a_{NM} \end{pmatrix}_{N\times M}$$

$$= (\boldsymbol{a}_1 \quad \boldsymbol{a}_2 \quad \cdots \quad \boldsymbol{a}_M)$$

This two counter indicate the number of available data. Vector $\boldsymbol{c}$ indicates the valid counter, or saying number of non-missing data, of each data columns and C indicate the valid data count for ensemble dataset X:

$$\boldsymbol{c} = \sum_{i=1}^{M} \boldsymbol{a}_i \quad C = \sum_{i=1}^{M} \boldsymbol{a}_i \cdot \boldsymbol{a}_i^T$$

And we can calculate the valid mean of each column:

$$\boldsymbol{m}_x(j) = \frac{1}{\boldsymbol{c}(j)} \sum_{i=1}^{M} \boldsymbol{a}_i(j) \times \boldsymbol{x}_i(j) \quad (j = 1,\ 2,\ \ldots,\ N)$$

According to PCA requirement, centralization of the x[th] observed data column $\boldsymbol{x}_i$ was applied by subtracting the mean:

$$\boldsymbol{y}_i = \boldsymbol{x}_i - (\boldsymbol{m}_x(i) \quad \boldsymbol{m}_x(i) \quad \cdots \quad \boldsymbol{m}_x(i))^T$$

Due to the missing value inside X, we need to deduce the covariance matrix R for the observed data X without using missing ones. The centralization also help to take missing data as 0, which is the mean of data after centralization. Mean inside covariance element does not change the correlation:

$$R(j,k) = \frac{1}{C-1}\sum_{i=1}^{M}[\boldsymbol{a}_i(j)\cdot\boldsymbol{x}_i(j)]\times[\boldsymbol{a}_i(k)\cdot\boldsymbol{x}_i(k)]$$

Calculate the eigenvectors U and eigenvalues λ_i. Assume $N > M$, so we have N eigenvectors in U:

$$U = (\boldsymbol{u}_1 \quad \boldsymbol{u}_2 \quad \cdots \quad \boldsymbol{u}_N)$$

Based on PCA theory, we need to pick up n numbers of eigenvalues (together with eigenvectors) to represent the original data, which is called principle component. This process stem from how many percentage of principle component we want to persist, which usually larger than 80% (Yang et al., 2020):

$$ratio = \frac{\sum_{i=1}^{n}\lambda_i}{\sum_{i=1}^{N}\lambda_i}$$

$$\widetilde{U} = (\boldsymbol{u}_1 \quad \cdots \quad \boldsymbol{u}_n)$$

So far we've finished all the process of PCA. The $ratio$ percentage information was stored inside the eigenvector matrix $\widetilde{U}$. Assume one column of realistic measurement data is $\tilde{\boldsymbol{r}}_i$, corresponding to $\boldsymbol{x}_i$. And the data after interpolated is $\boldsymbol{r}_i$. We want the difference between $\tilde{\boldsymbol{r}}_i$ and $\boldsymbol{r}_i$ as small as possible, which is the second norm of vectors reduction:

$$\min\|L_i\cdot(\tilde{\boldsymbol{r}}_i-\boldsymbol{r}_i)\|_2^2 = (L_i(\tilde{\boldsymbol{r}}_i-\boldsymbol{r}_i))^T\cdot(L_i(\tilde{\boldsymbol{r}}_i-\boldsymbol{r}_i))$$

Within: $\tilde{\boldsymbol{r}}_i = \widetilde{U}\cdot(\tilde{\boldsymbol{v}}_i+\boldsymbol{m}_x(i))$, $L_i = digonal(\boldsymbol{a}_i)$.

$\tilde{\boldsymbol{v}}_i$ is the re-constructed data (one column) based on PCA. By import the result of linearly centralization:

$$\min\|L_i\cdot\left(\widetilde{U}\tilde{\boldsymbol{v}}_i+\boldsymbol{m}_x(i)\right)-\boldsymbol{x}_i\|_2^2 = \|L_i\widetilde{U}\tilde{\boldsymbol{v}}_i-\boldsymbol{x}_i\|_2^2$$

Fortunately, it has an analytic solution:

$$\tilde{\boldsymbol{v}}_i = (\widetilde{U}^T L_i^T\cdot L_i\widetilde{U})^{-1}\cdot U^T L_i^T\left[\boldsymbol{x}_i - L_i\cdot vectorize\left(\boldsymbol{m}_x(i)\right)\right]$$

And the re-construct of $\boldsymbol{x}_i$ is:

$$\tilde{\boldsymbol{r}}_i = \widetilde{U}\tilde{\boldsymbol{v}}_i + \boldsymbol{m}_x(i)$$

Replacing the missing value by the reconstruction result, a PCA based interpolation was accomplished.

2.2 Interpolation experiments methodology

We conduct two application for this interpolation method. One is a 241×241 observed sea surface temperature (SST) raster created by NOAA NMFS SWFSC ERD. This data was observed in 2020-11-28 10:30 from remote sensing MODIS with multiple satellites. Missing rate of this raster is 2.27% distributed irregularly. The application is to treat the ensemble data as observed matrix X. Column was chosen along latitude due to lower SST mean in the same latitude compared with longitude. Result will be discussed in the next section.

The other experiment is a 726×726 picture of SUSTech campus. This picture was converted to grayscale data and manually delete 20 squares with 11×11 pixels in order to represent the missing condition. Regions of missing were randomly selected but following these rules that each missing square must has its outside at least 7 width of valid data boundary. Missing squares were interpolated from four directions in order to reduce the iteration length as short as possible. In this region missing interpolation, one PCA reconstruction was based on previous reconstruction result. Thus the accuracy will decrease as the interpolated data approaching to the geometric center of missing region. Four directions, the four corners of missing region, of interpolating iteration is somehow a better solution. PCA based interpolation matrix was set to 7×7 because of an appropriate balance between the variance of a column, and the effective weight from the nearby data. If the PCA region is too large, inefficient centralization mean would be induced. While region that is too small would lead to less regional correlation towards the nearby data. After one missing data was found, a 7×7 region was sent to interpolation and refill the missing from the PCA reconstruction result.

3. Result and discussion

3.1 Single missing experiment

Figure 1 shows the result of SST raster interpolation, with (a) the original data, (b) the PCA reconstructed data, and (c) the interpolated data. Result shows a good interpolation when missing is neighbored by close temperature, but failed to reveal a water mixing at region where isotherm is complex like the southwest part in (b). A good correlation was obtained from the interpolation that strong cold water input at southeast part near latitude 3S. And result near 115E to 117E, 3S shows a well interpolated water distribution. This reveal the PCA based interpolation will have better result in the region where data has a regular variance. If water mixing is too complex, in this case, this method lose its accuracy. It is a remarkable fact that reconstruction result has no cold water input at northwest part. Same as warm water input from south at southeast part in Figure 1 (a), reconstruction will lose the longitude trend due to the decentralization by adding mean value.

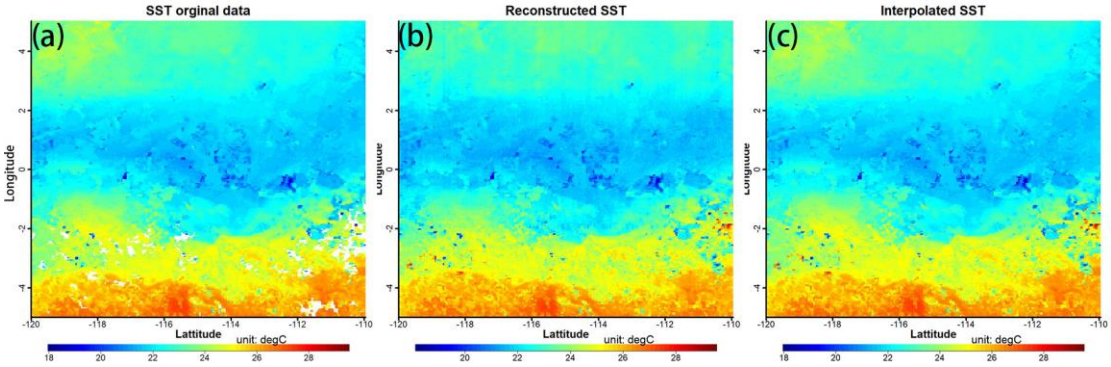

Figure 1: SST interpolation experiment. Subfigure (a) is the original SST raster, (b) is the reconstruction result, and (c) is the interpolated result.

3.2 Region missing experiment

Figure 2 is the result of picture interpolation result. Within this figure, (a) is the original RGB picture, (b) is the grayscale image, (c) is the interpolated image, and circles in (c) shows the position of missing squares. Several missing squares were marked with number in order to have a better discussion. As showed in the figure, regions that have close grayscale are easy to interpolate, for example circle 1, 2 and 3, despite it is inside water or sky. And for circles 4 to 7, good interpolation result was obtained. These regions have a same property, which is the vertical or horizontal

regularity. Circles 4 and 5 are water reflection, circle 6 is building texture, and circle 7 is grass. This condition will receive a not good interpolation when using method like IDW. However, interpolation shows bad result when dealing with region that has a slanting regularity like circles 8 and 9, which are slanting sun shadows. This can be modify by rotating the PCA reconstruction region makes the import column parallel to its trend. Circles 11 totally failed to be interpolated while the similar circle 6 has a good result. Probably reason is the interpolation scripts rotate the entire image at once instead of rotate the missing region and its neighbor data, which may lead to unequal iteration from different directions. This part may be interpolated from the roof for too much times during iteration.

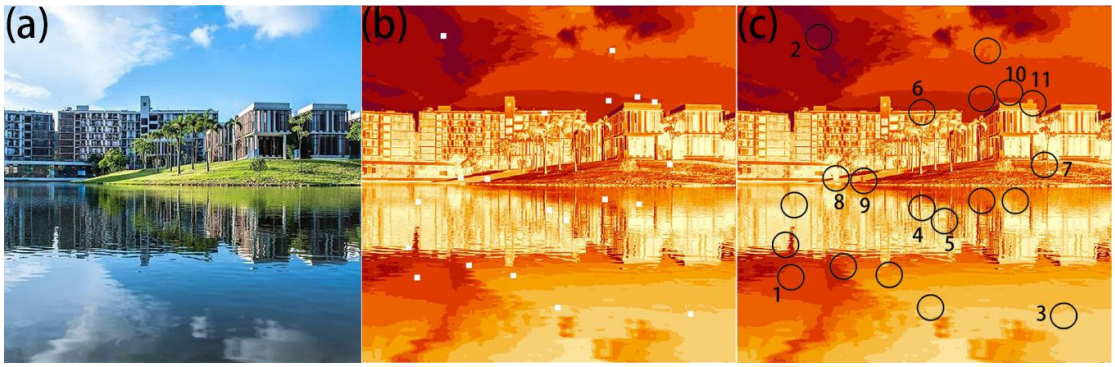

Figure 2: Picture interpolation experiment. Subfigure (a) is the original RGB picture, (b) is the grayscale image, and (c) is the interpolated image.

4. Conclusion

A PCA based interpolation method was applied in this report. This method is efficient when dealing with data has linear regularity. If the regularity is not arranged vertically or horizontally, rotating the data set parallel to its trend might be a better modification. In addition, this method is better than a lot of data process because it has the analytic solution, better than most numerical solution. Unfortunately, this method failed to interpolate data with massive trend, or missing data that has low correlation with nearby region. There is another disadvantage that this method requires a lot of linear algebra, matrix calculation. As a result, programing platforms with better optimization algorithm, especially for matrix, would have a better performance.

References

Gomes, L., & Oliveira, P. (2008). Bathymetric data fusion: PCA based Interpolation and regularization, sea tests, and implementation. OCEANS 2008. doi: 10.1109/oceans.2008.5151973.

Jiji, C., & Chaudhuri, S. (2004). PCA based Generalized Interpolation for Image Super-Resolution. ICVGIP. doi: 10.1.1.533.9859.

Jolliffe, I. (1986). Principal component analysis. New York: Springer-Verlag.

LU, Y., WANG, L., QIU, A., ZHANG, Y., & ZHAO, Y. (2017). A CoKriging Interpolation Method Based on Principal Component Analysis. Bulletin Of Surveying And Mapping, (11). doi: 10.13474/j.cnki.11-2246.2017.0347.

Oliveira, P., & Gomes, L. (2009). Interpolation of signals with missing data using Principal Component Analysis. Multidimensional Systems And Signal Processing, 21(1), 25-43. doi: 10.1007/s11045-009-0086-3.

Yang, W., Zhao, Y., Wang, D., Wu, H., Lin, A., & He, L. (2020). Using Principal Components Analysis and IDW Interpolation to Determine Spatial and Temporal Changes of Surface Water Quality of Xin'anjiang River in Huangshan, China. International Journal Of Environmental Research And Public Health, 17(8), 2942. doi: 10.3390/ijerph17082942