Tensor factorization and its application to multidimensional seismic data recovery

Mauricio D Sacchi, University of Alberta, Jianjun Gao, China University of Geosciences (Beijing), Aaron Stanton and Jinkun Cheng, University of Alberta*

SUMMARY

Research in the area of data analytics and recommendation systems have lead to important efforts toward solving the problem of matrix completion. The latter entails estimating the missing elements of a matrix by assuming a low-rank matrix representation. The aforementioned problem can be extended to the recovery of the missing elements of a multilinear array or tensor. Prestack seismic data in midpoint-offset domain can be represented by a 5th order tensor. Therefore, tensor completion methods can be applied to the recovery of unrecorded traces. Furthermore, tensor completion methodologies can also be applied for multidimensional signal-to-noise-ratio enhancement. We discuss the implementation of the Parallel Matrix Factorization (PMF) algorithm, an SVD-free tensor completion method that we applied to 5D seismic data reconstruction. The Parallel Matrix Factorization (PMF) algorithm expands our first generation of 5D tensor completion codes based on High Order SVD and Nuclear norm minimization. We review the PMF method and explore its applicability to processing industrial data sets via tests with synthetic and field data.

INTRODUCTION

In recent years, the development of recommendation systems has become an important area of research for data scientists (Koren et al., 2009). A recommendation system (or recommender system) is an algorithm that attempts to predict the rating that a user or costumer will give to an item. Recommendation systems have become quite popular in e-commerce for predicting ratings of movies, books, news, research articles etc. In Figure 1, we provide a simplified example of a data matrix with ratings of a series of movies. It is clear that recommendation systems use thousands of users to rate thousands of items and that our figure is merely for illustrative purposes. A rating of 5 means that the user liked the movie, a rating of 1 means that he/she did not like the movie. Question marks are used to indicate that the movie has not been rated by the user. This is a table (matrix) where one can immediately infer that the data could be predicted by simple examination of patterns or relationships between users and movies. For instance, users who liked romantic movies appear not to like action movies. The main task for the recommendation algorithm is to extract patterns that might exist in the data and use them to predict the rating a user would have given to an item he/she did not rate. The unknown ratings can be found by solving the so called Matrix Completion problem (Recht, 2011). A similar problem is also present in seismic data processing (Kreimer and Sacchi, 2011). Figure 2 presents a simple example of data recovery via matrix completion. For this particular example we adopted a reduced rank matrix completion algorithm that operates in the $t-x$ domain. However, it is clear that seismic data are much more complicated than the example portrayed in Figure 2. Reconstruction methods based on rank-reduction techniques for prestack seismic data must operate on full 5D patches of seismic data. Techniques that can cope with multidimensional seismic data reconstruction can mainly be divided into two categories. One category of methods applies rank reduction to block Hankel matrices formed by the entries of observed seismic data in the frequency-space domain. Methods in this category are often named Cadzow (Trickett et al., 2010) or Multichannel Singular Spectrum Analysis reconstruction (Oropeza and Sacchi, 2011; Gao et al., 2013). A second category of methods are based on dimensionality reduction of multilinear arrays or tensors. Examples of the latter are High Order SVD (HOSVD) reconstruction (Kreimer and Sacchi, 2011, 2012), Tucker decomposition (Herrmann and Silva, 2013), the nuclear norm minimization method (Kreimer et al., 2013) and the tensor SVD method (Ely et al., 2013). The common feature of these method is that they all utilize the SVD algorithm to reduce the rank of the data tensor. For large-scale seismic data reconstruction problems, the cost of the SVD algorithm prevents the use of low rank tensor completion methods for industrial applications.

In this paper we analyze the Parallel Matrix Factorization (PMF) algorithm proposed by Xu et al. (2013). The PMF method does not utilize SVDs. We show that PMF is an effective algorithm to recover missing traces from large 5D volumes.

Movie →

User ↓

	Taxi Driver	Sense and Sensibility	Battleship Potemkin	Raging Bull	Titanic	Alexander Nevsky
John	5	1	5	4	1	3
Mary	1	4	?	1	4	?
Pepe	4	2	2	3	4	?
Adrian	3	1	?	3	3	?
Tony	?	?	?	?	4	?
Kevin	3	3	?	3	2	?
Jianjung	2	1	?	2	4	?
Natasha	?	?	3	?	5	3

Figure 1: Example of data used by a recommendation system. Question marks are ratings that the recommendation system should predict based on available data (ratings).

THEORY

The PMF tensor reconstruction method is implemented in midpoint-offset frequency domain. We denote the data by $D(\omega, x, y, h_x, h_y)$, where x, y, h_x and h_y indicate the spatial coordinates and in the

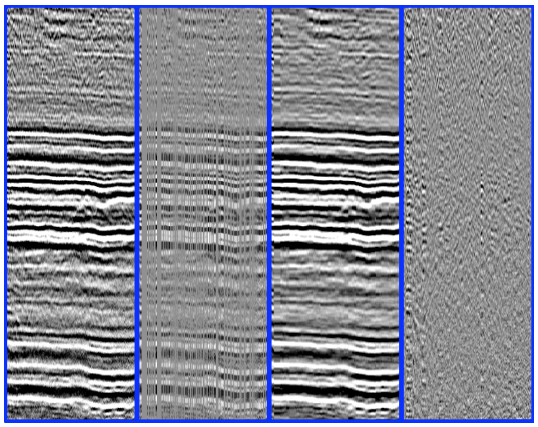

Figure 2: Seismic data reconstruction can be posed as matrix completion problem. From left to right: Complete data, decimated data, recovered data and recovery error. In this example we used a projected gradient algorithm where a rank reduction and a lateral smoothing constraint were simultaneously applied to recover the data.

inline midpoint, cross-line midpoint, in-line offset and cross-line offset. After binning the data into a midpoint-offset grid, a frequency slide can be denoted as $D(\omega,x,y,h_x,h_y)$. The latter can be represented by a 4th-order tensor $\mathscr{D}$ with elements D_{i_1,i_2,i_3,i_4}, where i_1,i_2,i_3,i_4 are bins indices for the spatial coordinates x,y,h_x and h_y, respectively. We remove the dependency on ω to simplify the notation. The reconstructed data are obtained by minimizing the following cost function

$$\Phi \quad = \quad \Phi_C + \mu\,\Phi_M\,, \tag{1}$$

where, Φ_M is the data misfit term, $\Phi_M = \frac{1}{2}\|\mathscr{P}\circ\mathscr{Z}-\mathscr{D}\|_F^2$, $\mathscr{P}$ is the Nth-order sampling operator tensor with elements 1 for the observed samples and 0 for the missing samples. The tensor $\mathscr{Z}$ is the Nth-order low rank tensor representing the reconstructed data (the unknown of our problem). The functional Φ_C is the low-rank constraint term

$$\Phi_C \quad = \quad \frac{1}{2}\sum_{k=1}^{N}\|\,\mathbf{X}_{(k)}\mathbf{Y}_{(k)}-\mathbf{Z}_{(k)}\|_F^2\,, \tag{2}$$

where, $\mathbf{Z}_{(k)}$ is the mode-k unfolding matrix of the tensor $\mathscr{Z}$. Figure 3 portrays the process of unfolding and folding an arbitrary tensor $\mathscr{X}$. A low-rank matrix factorization is applied to each mode unfolding of $\mathscr{Z}$ by seeking matrices $\mathbf{X}_{(k)}\in\mathbb{C}^{I_k\times r_k}$ and $\mathbf{Y}_{(k)}\in\mathbb{C}^{r_k\times I_1\ldots I_{k-1}I_{k+1}\ldots I_N}$ such that $\mathbf{Z}_{(k)}\approx\mathbf{X}_{(k)}\mathbf{Y}_{(k)}$ for $k=1,\ldots,N$, where r_k is the rank of the unfolding matrix $\mathbf{Z}_{(k)}$. In order to solve $\mathbf{X}_{(k)}$, $\mathbf{Y}_{(k)}$ and $\mathscr{Z}$, we minimize the cost function Φ via an alternating least-squares algorithm:

$$\mathbf{X}_{(k)}^{i+1} = \mathbf{Z}_{(k)}^{i}(\mathbf{Y}_{(k)}^{i})^H,\ k=1,\ldots,N, \tag{3a}$$

$$\mathbf{Y}_{(k)}^{i+1} = ((\mathbf{X}_{(k)}^{i+1})^H\mathbf{X}_{(k)}^{i+1})^{\dagger}(\mathbf{X}_{(k)}^{i+1})^H\mathbf{Z}_{(k)}^{i},\ k=1,\ldots,N, \tag{3b}$$

$$\mathscr{Z}^{i+1} = (\mathscr{I}-\alpha\mathscr{P})\circ\mathscr{C}+\alpha\mathscr{D}\,, \tag{3c}$$

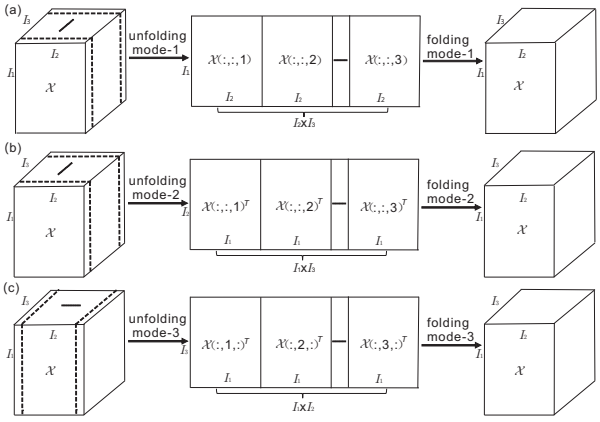

Figure 3: Process of unfolding a tensor into a matrix and folding back a matrix into a tensor. This figure exemplifies the process for a 3rd-order tensor.

where, the parameter $\alpha=\frac{\mu}{N+\mu}$, $\mathscr{I}$ is the Nth order tensor with all entries equal to 1 and $\mathscr{C}$ is given by

$$\mathscr{C} = \frac{1}{N}\sum_{k=1}^{N}\text{fold}_k[\mathbf{X}_{(k)}^{i+1}\mathbf{Y}_{(k)}^{i+1}]\,. \tag{4}$$

The preceding analysis corresponds to the case where data are contaminated with noise. The noise-free data reconstruction case is tackled by finding the minimum of the following cost function

$$\Phi = <\mathscr{W},\mathscr{P}\circ\mathscr{Z}-\mathscr{D}> + \Phi_M \tag{5}$$

where $<\mathscr{A},\mathscr{B}>=\sum_{i_1=1}^{I_1}\ldots\sum_{i_N=1}^{I_N}\bar{A}_{i_1\ldots i_N}B_{i_1\ldots i_N}$. Using the method of Lagrange multipliers, the solution of equation 5 is given by

$$\mathscr{Z} = (\mathscr{I}-\mathscr{P})\circ\mathscr{C}+\mathscr{D}\,. \tag{6}$$

Expression 6 is equal to expression 3c for the particular case when α=1. Interestingly, 3c resembles the typical imputation algorithm used for reconstruction via POCS (Abma and Kabir, 2005) and Cazdow (Trickett et al., 2010; Gao et al., 2013) methods.

SYNTHETIC EXAMPLES

The first example is a 5D seismic data that consist of $I_1\times I_2\times I_3\times I_4$ spatial traces with $I_k=6,8,10,12,14$, $k=1,2,3,4$ and 301 time samples per trace. The data include three linear events and S/N $=\infty$. We randomly remove 50% of the traces and perform the reconstruction using the proposed PMF algorithm, the HOSVD algorithm and the nuclear norm minimization method. For the PMF and HOSVD methods, we adopt a rank r_k=3 for all modes ($k=1,2,3,4$), the maximum number of iterations is set to $N_{iter}=100$, and an iteration stopping error $tol=10^{-4}$ is adopted for each frequency, respectively. For the nuclear norm method, we set N_{iter}=100, $tol=10^{-4}$ and the parameters $\lambda=2.5$, $\beta=15$ (see, Kreimer et al. (2013)) Table 1 shows the comparison of the computational cost of the

I_k	Cost (secs)		
	PMF	HOSVD	Nuclear norm
8	49.8	814.1	74.1
10	74.7	919.3	159.3
12	117.9	1077.1	307.4
14	195.1	1259.4	569.6

Table 1: Computational time comparison of the proposed PMF reconstruction method, the HOSVD method and nuclear norm method for different 5D volumes with size of $301 \times I_1 \times I_2 \times I_3 \times I_4$, $I_k = 8, 10, 12, 14$, $k = 1, 2, 3, 4$.

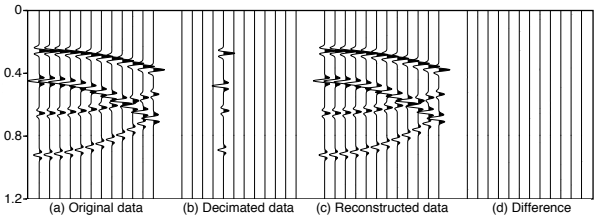

Figure 4: Recovery of a 5D volume via the PFM method. One slice of the 5D volume is shown ($SNR = \infty$)

three methods. For each iteration, the computational cost of the PMF method is $O(N(3mnr + mr^2 + \frac{2nr^3}{3}))$. For the nuclear norm method, the cost is $O(N(2m^2n + 2m^3))$ per iteration and for the case of the HOSVD, the cost is $O(N(4m^2n + 13m^3 + mr^{3(N-1)}))$. Where $m = I_k = max\{I_1, I_2, \ldots, I_N\}$, $r = r_k = \max\{r_1, r_2, \ldots, r_N\}$, $n = I_1 I_2 \ldots I_{k-1} I_{k+1} \ldots I_N$ and N represents the order of seismic data tensor. From table 1, we observe that the PMF algorithm is faster than the nuclear norm minimization algorithm and the HOSVD algorithm. We also choose a synthetic data model containing $12 \times 12 \times 12 \times 12$ traces in the spatial directions and 301 time samples per trace which is also used in Table 2 to examine the reconstruction quality of the proposed PMF algorithm, HOSVD reconstruction and the nuclear norm minimization reconstruction method. We define the reconstruction quality via the expression $Q = 10 \log_{10}(\frac{\|\mathscr{D}^{true}\|^2}{\|\mathscr{D}^{true} - \mathscr{D}^{recon}\|^2})$ where $\mathscr{D}^{true}$ and $\mathscr{D}^{recon}$ represent the true noise-free complete data and reconstructed data in the time-space domain. Table 2 shows the comparison of the reconstruction quality versus the percentage of missing traces. From Table 2, we find that the reconstruction quality obtained by the proposed PMF method and HOSVD algorithm are very similar. They both perform better than the nuclear norm minimization method. For the third example, we synthesize a noise-free data with four events with strong curvature. The spatial size of the data is $12 \times 12 \times 12 \times 12$ with 301 time samples per trace and S/N $= \infty$. We randomly decimated 90% of the traces and set the rank $r_1 = r_2 = r_3 = 5$ for modes $1, 2, 3$ and $r_4 = 4$ for mode 4. We also set $N_{iter} = 300$, $tol < 10^{-4}$ and $\alpha = 1$. Figure 4 shows the reconstruction result. From error section in Figure 4, one can observe that missing traces were accurately recovered. We also add random noise to the noise-free data in Figure 5 to analyze the reconstruction capability our the algorithm in the presence of noise. In this example, we set S/N $= 1$, $N_{iter} = 300$, $tol < 10^{-4}$ and $\alpha = 0.51$.

Decimation [%]	Reconstruction quality Q		
	PMF	HOSVD	Nuclear norm
60	62.6	62.7	16.8
70	61.1	61.3	9.7
80	60.6	60.7	6.5
90	39.9	31.3	3.0

Table 2: Reconstruction quality Q versus percentage of missing traces for the PMF, HOSVD and Nuclear norm reconstruction methods with data size $I_k = 12$.

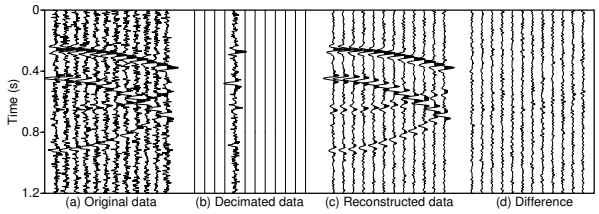

Figure 5: Recovery of a 5D volume via the PFM method. One slice of the 5D volume is shown ($SNR = 1$)

FIELD DATA EXAMPLE

Based on the above synthetic data analysis, we tested the performance of the PMF reconstruction method on a land data set obtained from a heavy oil field in the WCB (Figure 6). The data are first binned on a $5m \times 5m$ CMP grid and a $100m \times 100m$ offset-x-y grid prior to interpolation. The fold map of the survey is shown in Figure 7. The reconstruction area includes 300 CMP$_x$ bins and 220 CMP$_y$ bins. We divide the whole survey data into 2640 overlapping blocks. Each block has about 85% missing traces. We set $r_k = 4$, $k = 1, 2, 3, 4$, $N_{iter} = 100$ and $\alpha = 0.40$ for the PMF reconstruction. Figure

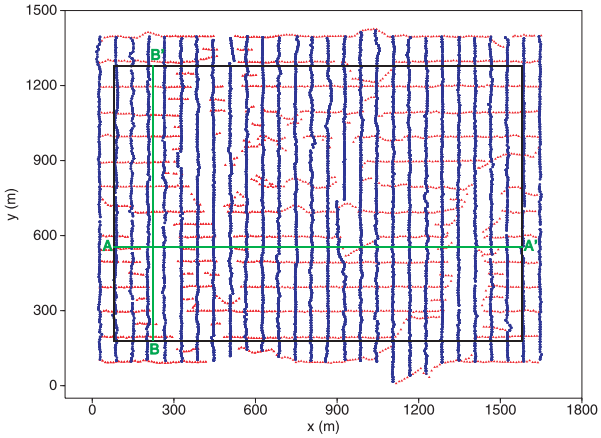

Figure 6: Field data example (WCB). Distribution of sources and receivers.

8 shows the input and reconstructed data for a slice of the 5D volume where we fixed CMPy, offset x and offset y (near offset) versus CMPx. Similarly, in Figure 9 we fixed CMPx and selected mid range offset bins x and y and display the data versus CMPy. Finally, Figure 10 displays a fixed CMPx bin

versus CMPy for fixed far offset bins x and y.

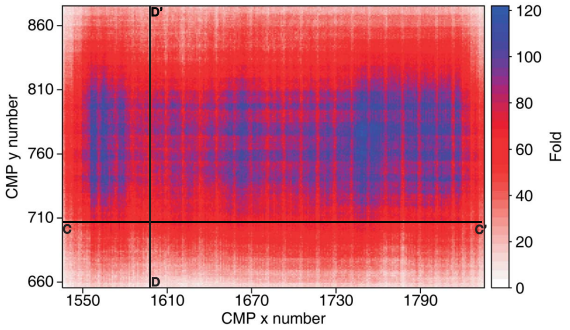

Figure 7: Field data example (WCB). Fold map.

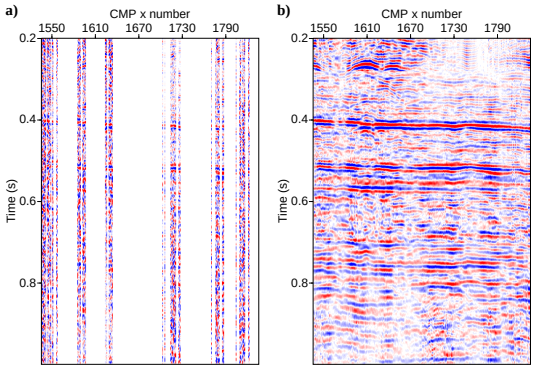

Figure 8: Near offset gather for a constant CMPy versus CMPx. (a) Observed data. (b) Reconstructed data.

CONCLUSIONS

We have presented a SVD-free method for multidimensional seismic data reconstruction. The proposed PMF method applies low rank matrix factorization to mode unfoldings of the seismic data tensor and applies an alternating minimization algorithm to estimate the complete data tensor. Contrary to other low rank reconstruction methods, PMF does not require the SVD algorithm. The latter makes the PMF algorithm attractive for industrial implementations. We compared the proposed method to two methods developed by our group (HOSVD and minimum Nuclear Norm reconstruction). We conclude that the proposed 5D data completion PMF method is faster than our previously reported algorithms for tensor completion. We stress that one of the main obstacles that might prevent industrial applications of tensor reconstruction methods is the computational cost of classical factorization methods based on the SVD. Our current research focuses on the randomized QR decomposition methods to gain further efficacy in our tensor-based 5D reconstruction techniques.

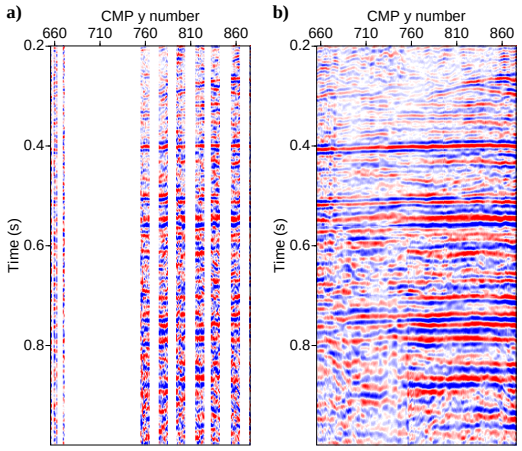

Figure 9: Mid offset range gather for a constant CMPx versus CMPy (a) Observed data. (b) Reconstructed data.

ACKNOWLEDGEMENTS

We thank the sponsors of the Signal Analysis and Imaging Group (SAIG) at the University of Alberta and NSERC (Discovery Grant, Fundamental and Applied Studies in Seismic Data Preconditioning and Inversion). We also thank Nadia Kreimer for developing 5D HOSVD and Minimum Nuclear Norm reconstruction methods. Jianjun Gao also thanks the financial support from the National Natural Science Foundation of China, Fundamental Research Fund for Central Universities and Fundamental Research Fund for the Key Laboratory of Geo-detection, China University of Geosciences, Beijing.

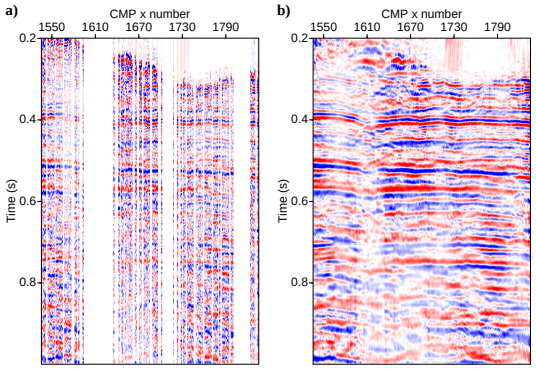

Figure 10: Far offset gathers for a constant CMPy versus CMPx (a) Observed data. (b) Reconstructed data.

REFERENCES

Abma, R. and N. Kabir, 2005, Comparison of interpolation algorithms: The Leading Edge, **24**, no. 10, 984–989.

Ely, G., S. Aeron, N. Hao, and M. Kilmer, 2013, 5D and 4D pre-stack seismic data completion using tensor nuclear norm (TNN): SEG Technical Program Expanded Abstracts, 3639–3644.

Gao, J., M. Sacchi, and X. Chen, 2013, A fast reduced-rank interpolation method for prestack seismic volumes that depend on four spatial dimensions: Geophysics, **78**, no. 1, V21–V30.

Herrmann, F. J. and C. D. Silva, 2013, Structured tensor missing-trace interpolation in the hierarchical Tucker format: SEG Technical Program Expanded Abstracts, **702**, 3623–3627.

Koren, Y., R. Bell, and C. Volinsky, 2009, Matrix factorization techniques for recommender systems: Computer, 30–37.

Kreimer, N. and M. D. Sacchi, 2011, Evaluation of a new 5d seismic volume reconstruction method - Tensor Completion versus Fourier reconstruction: CSPG CSEG CWLS Convention, 5 pages.

——— 2012, A tensor higher-order singular value (HO-SVD) decomposition for prestack seismic data noise reduction and interpolation: Geophysics, **77**, V113–V122.

Kreimer, N., A. Stanton, and M. D. Sacchi, 2013, Tensor completion based on nuclear norm minimization for 5D seismic data reconstruction: Geophysics, **78**, V273–V284.

Oropeza, V. and M. D. Sacchi, 2011, Simultaneous seismic data denoising and reconstruction via multichannel singular spectrum analysis: Geophysics, **76**, no. 3, V25–V32.

Recht, B., 2011, A simpler approach to matrix completion: The Journal of Machine Learning Research, **12**, 3413–3430.

Trickett, S., L. Burroughs, A. Milton, L. Walton, and R. Dack, 2010, Rank-reduction-based trace interpolation: SEG, Expanded Abstracts, **29**, no. 1, 3829–3833.

Xu, Y. Y., R. R. Hao, W. T. Yin, and Z. X. Su, 2013, Parallel matrix factorization for low rank tensor completion: UCLA Computational and Applied Mathematics, 13–77.