STATISTICAL RANK SELECTION FOR INCOMPLETE LOW-RANK MATRICES

Rui Zhang, Alexander Shapiro, Yao Xie

H. Milton Stewart School of Industrial and Systems Engineering,
Georgia Institute of Technology, Atlanta, GA

ABSTRACT

We consider the problem of determining the rank in the low-rank matrix completion. We propose a statistical model for noisy observation. It is important for many existing algorithms and sometimes has practical meanings. We construct a test statistics for the low rank approximation problem. Under this model, we derive the distribution of the test statistics. By applying the test statistics, we propose a sequential rank test procedure to determine the rank with statistical inference. In the numerical section, we illustrate our theoretical results and give examples of our proposed rank test procedure.

***Index Terms*—** Rank selection, matrix completion, statistical inference

1. INTRODUCTION

Matrix completion is of great interest in machine learning, data mining, mathematics and signal processing. It studies a problem to recover the entire matrix from limiting samples. A well-known example in the field of recommendation system is the Netflix Problem [1], which uses the sparse ratings to predict user's like of the entire set of movies and make proper recommendations. In signal processing, one application of matrix completion is Sensor Network Localization [2, 3], which aims to infer the positions of all sensors from a few sensors, of which position is known.

Rank selection is an important part in matrix completion. From the angle of optimization, in the problem which there is noise in the observation, matrix completion need to be done with the tradeoff between the fitness and the rank of the matrix. A proper rank needs to guarantee the goodness of fit and avoid overfitting [3, 4]. From the angle of algorithm, in many existing algorithms, they need to specified the rank first [5, 6], such as alternating minimization which factorized the matrix into two low-rank matrix and minimize them alternatively. From the angle of practice, rank represent different practical meanings. For the recommendation systems, the rank of low-rank matrix is the number of the latent factors that affect the rating patterns [7]. For the source detection problem, the rank is close related to the number of sources [8, 9]. Therefore, knowing the rank is significant in low-rank matrix completion.

Though the rank selection is important, the study of it is limited compared with the massive study of matrix completion. A common technique is to apply SVD and see whether there is outstanding structures [10]. However, when the singular values are close, there is no clear way of thresholding.

Our main contribution in this paper is to give a statistical way to do the inference of rank selection. With our assumed model for the noisy observation and the well-posedness condition [11], we are able to construct a test statistics for low-rank matrix approximation problem. We prove that the test statistics follows a noncentral χ^2 distribution. Then, with this statistics, we propose a sequential testing procedure to test the true ranks for the low-rank matrix completion. To the best of our knowledge, this is the first paper to present a rigorous statistical procedure to test the true rank of a matrix, given noisy observations for a subset of its entries.

2. PROBLEM SET-UP

Consider the problem of recovering an $n_1 \times n_2$ data matrix of low rank when observing a small number m of its entries, which are denoted as M_{ij}, $(i,j) \in \Omega$. We assume that $n_1 \geq 2$ and $n_2 \geq 2$. Here $\Omega \subset \{1, ..., n_1\} \times \{1, ..., n_2\}$ is the observation index set of cardinality m.

Consider the following low rank matrix approximation(least square) problem:

$$\min_{Y \in \mathcal{M}_r} \sum_{(i,j) \in \Omega} \left(M_{ij} - Y_{ij}\right)^2, \tag{1}$$

where

$$\mathcal{M}_r := \left\{Y \in \mathbb{R}^{n_1 \times n_2} : \operatorname{rank}(Y) = r\right\} \tag{2}$$

In this paper, the question aim to address is how to determine the rank r. To proceed we assume the following model with noisy and possibly biased observations of a subset of matrix entries. There is a (population) value Y^* of $n_1 \times n_2$ matrix of rank $r < \mathfrak{R}(n_1, n_2, m)$ [11] and M_{ij} are viewed as observed (estimated) values of Y^*_{ij}, $(i,j) \in \Omega$, based on a sample of size N.

The observed values are modeled as

$$M_{ij} = Y^*_{ij} + N^{-1/2}\Delta_{ij} + \varepsilon_{ij},\ (i,j) \in \Omega, \tag{3}$$

where $Y^* \in \mathcal{M}_r$ and Δ_{ij} are some (deterministic) numbers. The random errors ε_{ij} are assumed to be independent of each other and such that $N^{1/2}\varepsilon_{ij}$ converge in distribution to normal with mean zero and variance σ_{ij}^2, $(i,j) \in \Omega$. The additional terms $N^{-1/2}\Delta_{ij}$ in (3) represent a possible deviation of population values from the "true" model and are often referred to as the population drift or a sequence of local alternatives (we can refer to [12] for a historical overview of invention of the local alternatives setting). This is a reasonably realistic model motivated by many real applications.

3. STATISTICAL TEST PROCEDURE

Consider the following weighted least squares test statistic

$$T_N(r) := N \min_{Y \in \mathcal{M}_r} \sum_{(i,j)\in\Omega} w_{ij} \left(M_{ij} - Y_{ij}\right)^2, \qquad (4)$$

where $w_{ij} := 1/\hat{\sigma}_{ij}^2$ with $\hat{\sigma}_{ij}^2$ being consistent estimates of σ_{ij}^2 (i.e., $\hat{\sigma}_{ij}^2$ converge in probability to σ_{ij}^2 as $N \to \infty$). Recall that the well-posedness condition [11] is sufficient for local identifiability of Y^*. The following asymptotic results can be compared with similar results in the analysis of covariance structures (cf., [13]).

Proposition 3.1 (Asymptotic properties of test statistic)**.** *Consider the noisy observation model* (3). *Suppose that the model is globally identifiable at $Y^* \in \mathcal{M}_r$ and Y^* is well-posed for minimum rank matrix completion problem [11]. Then as $N \to \infty$, the test statistic $T_N(r)$ converges in distribution to noncentral χ^2 distribution with degrees of freedom $\mathrm{df}_r = m - r(n_1 + n_2 - r)$ and the noncentrality parameter*

$$\delta_r = \min_{H \in \mathcal{T}_{\mathcal{M}_r}(Y^*)} \sum_{(i,j)\in\Omega} \sigma_{ij}^{-2} \left(\Delta_{ij} - H_{ij}\right)^2, \qquad (5)$$

where $\mathcal{T}_{\mathcal{M}_r}(Y)$ is the tangent space to $\mathcal{M}_r$ at $Y \in \mathcal{M}_r$.

Remark 3.1. The above asymptotic results are formulated in terms of the "sample size N" suggesting that the observed values are estimated from some data. That is, the given values $\bar{M}_{ij}$, $(i,j) \in \Omega$, are obtained by averaging i.i.d. data points M_{ij}^{ℓ}, $\ell = 1, ..., N$. In that case asymptotic normality of $N^{1/2}\varepsilon_{ij}$ can be justified by application of the Central Limit Theorem, and the corresponding variances σ_{ij}^2 can be estimated from the data in the usual way $\hat{\sigma}_{ij}^2 = (N - 1)^{-1}\sum_{\ell=1}^{N}(M_{ij}^{\ell} - \bar{M}_{ij})^2$.

Remark 3.2. For given index set Ω and observed (estimated) values M_{ij}, $(i,j) \in \Omega$, the statistic $T_N(r)$ can be used for testing the (null) hypothesis that the "true" rank is r (i.e. $Y^* \in \mathcal{M}_r$). That is the null hypothesis is rejected if $T_N(r)$ is large enough on the scale of the χ^2 distribution with the respective df_r degrees of freedom. In other words, our procedure chooses the smallest r, such that $T_N(r) \le q_{r,1-\alpha}$, where $q_{r,1-\alpha}$ is the $1-\alpha$ quantile of χ^2 distribution with degrees of freedom df_r. The role of values Δ_{ij} in the model is to suggest that the "true" model is true only approximately, and the corresponding noncentrality parameter δ_r gives an indication of the deviation from the exact rank r model. In practice, if Δ_{ij} is known, we can subtract it from the observation and then apply the rank test. If σ_{ij} is unknown and we have repetitive observations, we can estimate σ_{ij}^2 as shown in remark 3.1.

The asymptotics of the test statistic $T_N(r)$ depends on r and also on the cardinality m of the index set Ω. Suppose now that more observations become available at additional entries of the matrix. That is we are testing now the model with a larger index set Ω', of cardinality m', such that $\Omega \subset \Omega'$. In order to emphasize that the test statistic also depends on the corresponding index set we add the index set in the respective notations. Note that if Y^* is a solution of rank r for both sets Ω and Ω' and the model is globally (locally) identifiable (or the well-posedness condition holds [11]) at Y^* for the set Ω, then the model is globally (locally) identifiable(or the well-posedness condition holds) at Y^* for the set Ω'. The following result can be proved in the same way as Theorem 3.1 (cf., [13]).

Proposition 3.2. *Consider index sets $\Omega \subset \Omega'$ of cardinality $m = |\Omega|$ and $m' = |\Omega'|$, and the noisy observation model* (3). *Suppose that the model is globally identifiable at $Y^* \in \mathcal{M}_r$ and well-posedness condition [11] holds at Y^* for the smaller model (and hence for both models). Then the statistic $T_N(r,\Omega') - T_N(r,\Omega)$ converges in distribution to noncentral χ^2 with $\mathrm{df}_{r,\Omega'} - \mathrm{df}_{r,\Omega} = m' - m$ degrees of freedom and the noncentrality parameter $\delta_{r,\Omega'} - \delta_{r,\Omega}$, and $T_N(r,\Omega') - T_N(r,\Omega)$ is asymptotically independent of $T_N(r,\Omega)$.*

Corollary 3.1. *Suppose $N = 1$, $\Delta_{ij} = 0$ and $\varepsilon_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma^2)$. Consider a sequence of index set $\Omega_0 \supset \Omega_1 \supset \Omega_2 \supset \cdots \supset \Omega_K$, $|\Omega_{k-1}| - |\Omega_k| = L$, $\forall k = 1 \cdots K$. The conditions in proposition 3.2 hold. Let*

$$X_i = \min_{Y \in \mathcal{M}_r} \sum_{(i,j)\in\Omega_i} \left(M_{ij} - Y_{ij}\right)^2,$$
$$Z_i = (X_{i-1} - X_i)/L.$$

Apply the result in proposition 3.2 and Delta method, we can conclude that $\sqrt{K}(\bar{Z} - \sigma^2)$ converge in distribution to $\mathcal{N}(0, 2\sigma^4/L)$

Remark 3.3. In practise, the common situation is that we don't have repetitive observations. Then, we can assume $\Delta_{ij} = 0$ and the random errors are identically independent distributed. Then by applying corollary 3.1, with proper K and L, σ^2 can be estimated by $\bar{Z}$. Since it is only true when $Y^* \in \mathcal{M}_r$, we can use this in rank selection. More detail is shown in next section.

4. NUMERICAL EXPERIMENTS

We present some numerical experiments to illustrate our theory[1]. In this section, without further notification, nuclear norm minimization is solved by TFOCS [14] in Matlab and LRMA problem is solved by 'SoftImpute' [6](regularization parameter equals to 0) in R.

4.1. Testing for true rank

Asymptotic distribution of test statistic. In Section 3 (see (3)), we show that the asymptotical distribution of the test statistic for the "true" rank is χ^2 distribution, which we will verify numerically here. We generate the true matrix Y^*, an $n_1 \times n_2$ matrix of rank r^*, by uniformly generated an $n_1 \times r^*$ matrix V, an $n_2 \times r^*$ matrix W, and an $r^* \times r^*$ diagonal matrix D and setting $Y^* = \tilde{V} D \tilde{W}^\top$, where $\tilde{V}$ and $\tilde{W}$ are orthonormalization of V, W, respectively. We sample Ω uniformly random, where $|\Omega| = m$. In all the simulation Ω is sampled until well-posedness condition holds. The noisy and repeated observation matrices are generated by $M_{ij}^{(k)} = Y_{ij}^* + \varepsilon_{ij}^{(k)}, (i,j) \in \Omega$, where $\varepsilon_{ij}^{(k)} \sim \mathcal{N}(0, \sigma^2 N^{-1})$. In computing the test statistic $T_N^{(k)}(r)$ (4), the least square approximation is solved by a soft-threshholded SVD solver. The algorithm stops when either relative change in the Frobenius norm between two successive estimates, is less than some tolerance, denoted as tol or the number of iterations reaches the maximum, denoted as it.

Figure 1 shows the Q-Q plot of $\{T_N^{(k)}(r)\}_{k=1}^{200}$ against the corresponding χ^2 distribution. In this experiment, $n_1 = 40$, $n_2 = 50$, $r^* = 11$, $m = 1000$, $\sigma = 5$, $N = 400$ and Ω is sampled until well-posedness condition is satisfied. The parameters $tol = 10^{-20}$ and $it = 50000$. From the result, we can see $T_N(r)$ follows the central χ^2 distribution with a degree of freedom $\mathrm{df}_r = m - r(n_1 + n_2 - r) = 131$, which is consistent with proposition 3.1.

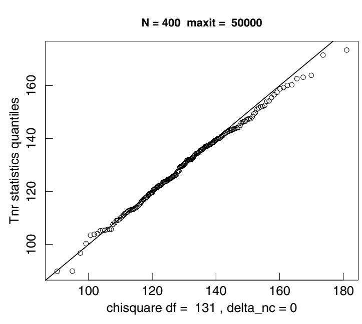


Fig. 1: Q-Q plot of $T_N(r)$ against quantiles of χ^2 distribution with degree of freedom 131: the observation matrix M is generated 200 times. $T_N^{(k)}(r)$ is computed as equation 4.

[1]More discussions can be found in a supplementary material at https://www2.isye.gatech.edu/∼yxie77/Experiment.pdf.

Table 1: sequential rank test(σ^2 is known).

rank	p-value	rank	p-value
2	0.00	7	0.00
3	0.00	8	0.00
4	0.00	**9**	**0.94**
5	0.00	10	0.69
6	0.00	11	0.41

Test for true rank. When σ^2 is known or there are repetitive observations, as discussed in Section 3, we can determine the true rank r^* by sequential χ^2 tests. That is, for r ranging from 1 to $\lceil \mathfrak{R}(n_1, n_2, m) \rceil$, we solve the least square approximations and compute $T_N(r)$. According to $T_N(r)$ we can determine which rank can be accepted for a predefined significant level. Table 1 shows a result of sequential rank test on a simulated data set with known σ^2. In this experiment, $n_1 = 40$, $n_2 = 50$, $r^* = 9$, $m = 1000$, $\sigma = 5$, $N = 100$. The true rank 9, is the first one accepted for 0.05 significant level.

When σ^2 is unknown and there is no repetitive observations, we can apply corollary 3.1 to estimate it as mentioned in remark 3.3. Recall that corollary 3.1 holds only when we do the estimation on the true rank (i.e. $r = r^*$). When $r < r^*$, $\bar{Z}$ largely overestimates the σ^2 and decreases hugely as r increases because part of the signal is treated as noise. When $r > r^*$, $\bar{Z}$ underestimates σ^2 and decreases slowly as r increases because part of noise is treated as signal. Therefore, we can select the first rank that $\hat{\sigma}^2$ is stable. Table 2 shows a result of estimating σ^2 for each rank r. In this experiment, $n_1 = n_2 = 100$, $r^* = 6$, $m = 8000$, $\sigma = 10$, $N = 1$. For each rank r ranging from 1 to 10, we estimate σ^2 by $\bar{Z}$ with $K = 20$ and $L = 300$. Our procedure chooses the true rank 6. Note that in this case, p-value can not be trusted since we don't have consistent estimate of σ^2 when $r \neq r^*$.

Figure 2 shows the comparison of rank selection between our sequential rank test (with known σ^2), nuclear norm minimization and the method suggested in [15] (we refer it as M^E method in the following). Since the nuclear norm minimization and M^E method can't give us the exact rank, we choose the rank by thresholding the percentage of the singular value of the recovered matrix in this two methods, i.e. $\hat{r} = \mathrm{argmin}_r \sum_{i=1}^r \lambda_{(i)} / \sum_{i=1}^{min(n_1,n_2)} \lambda_{(i)} > b$, where b is some threshold. In this experiment, $n_1 = 100$, $n_2 = 1000$, $\sigma = 5$, $N = 50$ and the sampling probability $p = 0.3$. For each true rank, we generate 100 instances of (Y^*, Ω, M), complete the rank selection with these three methods and compute the median of the error of estimated rank of each method. For the sequential rank test, we choose the first rank accepted with 0.05 significant level. For nuclear norm minimization and M^E method, we choose the threshold that gives us the best results for these two methods. It shows that selection by sequential χ^2 test outperforms the other two methods.

Table 2: sequential rank test(σ^2 is unknown)

rank	p-value	$\hat{\sigma}^2(=\bar{Z})$	rank	p-value	$\hat{\sigma}^2(=\bar{Z})$
1	0.82	34995.5	5	0.84	5050.63
2	0.86	26751.3	**6**	**0.43**	**97.7**
3	0.92	18719.6	7	0.76	96.6
4	0.62	11231.8	8	0.96	96.7

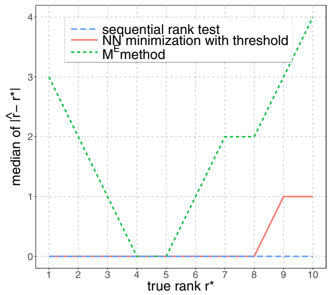


Fig. 2: Comparison of rank selection between sequential χ^2 test, nuclear norm minimization and M^E method, sampling probability p=0.3. For each true rank, we compute the median of rank error for 100 experiments. Threshold $b_{nm} = 0.25$, $b_{ME} = 0.13$ for nuclear norm minimization and M^E method, respectively.

4.2. Application: Source identification.

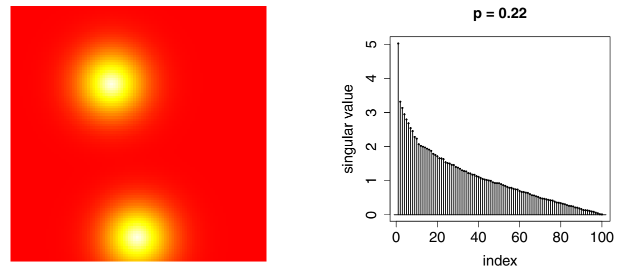
In recent years, matrix completion is used in source identification and localization [8, 9]. In this problem, limited data of a field generated by the source is collected. Then, matrix completion is used to recover the entire filed and identify the source, from the sparse samples. To solve this problem, the very first step is to know the number of sources. The general assumption is that the energy decays as the distance from the source increases (i.e. unimodal structure). With this assumption, the true rank of the matrix is closely related to the number of sources.
We simulate an energy field generate by two sources, and the energy of each location $\mathbf{x} \in \mathbb{R}^2$ is:

$$g(\mathbf{x}) = \frac{\|\mathbf{x} - \mu_2\| f_{\mu_1,\sigma_1}(\mathbf{x}) + \|\mathbf{x} - \mu_1\| f_{\mu_2,\sigma_2}(\mathbf{x})}{\|\mathbf{x} - \mu_1\|_2 + \|\mathbf{x} - \mu_2\|_2},$$

where $f_{\mu,\sigma}$ is PDF of $\mathcal{N}(\mu, \sigma^2)$. Then, the observation matrix: $M(\mathbf{x}) = \tilde{g}(\mathbf{x}) + \varepsilon_{\mathbf{x}}, \forall \mathbf{x} \in \Omega$, where $\tilde{g}(\mathbf{x}) = \frac{g(\mathbf{x})}{\max_{\mathbf{x}} g(\mathbf{x})}$ and $\varepsilon_{\mathbf{x}}$ is random error. In figure 3(a), it is the standardized energy field, $\tilde{g}(x)$, where the white spots are the sources. With sampling probability $p = 0.22$, the singular value of M is shown in figure 3(b) and there is only one outstanding singular value. Table 3 is the result of rank test, we can see it identify the rank as 2 which is the same as the number of sources.

4.3. Application: Delay matrix completion

Suppose we have a network of sensors distributed around the source(s) of signal. Each sensor detects the signals with different delays due to the difference of the locations. Denote the signal detected by ith sensor as $x_i(t)$ and the signal from the jth source as $s_j(t)$, then $x_i(t) = \sum_{j=1}^{J} s_j(t - \tau_i^j)$, where τ_i^j is the delay for jth signal detected by ith sensor and J is the number of sources. A delay matrix M is defined as: $M_{ij} = \text{argmax}_s x_i(s) \star x_j(s)$, where $\star$ denotes cross-correlation [16]. The rank of M is related to the number of sources, e.g. when $J = 1$ the rank of M is 2.
In this experiment, we perform our rank selection procedure on a real dataset collected from 130 sensors around Old Faithful geyser in Yellowstone National Park [17]. $M \in \mathbb{R}^{130\times130}$ and $|\Omega| = 15359$. We estimate σ^2 by $\bar{Z}$ with $L = 140$ and $K = 50$. The result is shown in table 4. The estimates of σ^2 is stable when $r = 4$, which suggests there might be more than one source of the signal.



(a) Energy field of two sources (b) SVD of M

Fig. 3: (a) Size of energy field is $100 * 100$, $\mu_1 = (10, 50)$, $\mu_2 = (70, 40)$, $\sigma_1 = \sigma_2 = 10$. (b) From singular value, we can't clearly identify whether the rank should be larger than 1. $\varepsilon_{\mathbf{x}} \sim \mathcal{N}(0, 0.04)$.

Table 3: p-value for sequential rank test of observed energy field.

rank	p-value	rank	p-value
1	0.00	4	1.00
2	**0.15**	5	1.00
3	0.98	6	1.00

Table 4: sequential rank test for delay matrix(σ^2 is unknown)

rank	p-value	$\hat{\sigma}^2(=\bar{Z})$	rank	p-value	$\hat{\sigma}^2(=\bar{Z})$
1	0.03	0.70	6	1.00	0.19
2	1.00	0.66	7	1.00	0.18
3	1.00	0.49	8	1.00	0.18
4	**1.00**	**0.20**	9	1.00	0.17
5	1.00	0.20	10	1.00	0.16

5. CONCLUSION

In this paper, we propose a general statistical model for noisy observations in matrix completion problem. With this model, a test statistics can be computed and used to do the statistical inference of rank selection. Numerical experiments illustrate our theoretical results and show the strength of proposed rank selection procedure in the real-world data.

6. REFERENCES

[1] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational mathematics*, vol. 9, no. 6, p. 717, 2009.

[2] A. M.-C. So and Y. Ye, "Theory of semidefinite programming for sensor network localization," *Mathematical Programming*, vol. 109, no. 2-3, pp. 367–384, 2007.

[3] M. Davenport and J. Romberg, "An overview of low-rank matrix recovery from incomplete observations," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 4, pp. 608–622, 2016.

[4] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[5] R. Sun and Z.-Q. Luo, "Guaranteed matrix completion via non-convex factorization," *IEEE Transactions on Information Theory*, vol. 62, no. 11, pp. 6535–6579, 2016.

[6] R. Mazumder, T. Hastie, and R. Tibshirani, "Spectral regularization algorithms for learning large incomplete matrices," *Journal of machine learning research*, vol. 11, no. Aug, pp. 2287–2322, 2010.

[7] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, no. 8, pp. 30–37, 2009.

[8] S. Choudhary and U. Mitra, "Analysis of target detection via matrix completion," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 3771–3775.

[9] J. Chen and U. Mitra, "Underwater acoustic source localization using unimodal-constrained matrix factorization," in *Signals, Systems, and Computers, 2017 51st Asilomar Conference on*. IEEE, 2017, pp. 2002–2006.

[10] R. H. Keshavan, S. Oh, and A. Montanari, "Matrix completion from a few entries," in *Information Theory, 2009. ISIT 2009. IEEE International Symposium on*. IEEE, 2009, pp. 324–328.

[11] A. Shapiro, Y. Xie, and R. Zhang, "Matrix completion with deterministic pattern-a geometric perspective," *arXiv preprint arXiv:1802.00047*, 2018.

[12] D. McManus, "Who invented local power analysis?" *Econometric Theory*, vol. 7, pp. 265–268, 1991.

[13] J. Steiger, A. Shapiro, and M. Browne, "On the multivariate asymptotic distribution of sequential chi-square statistics," *Psychometrika*, vol. 50, pp. 253–254, 1985.

[14] S. R. Becker, E. J. Candès, and M. C. Grant, "Templates for convex cone problems with applications to sparse signal recovery," *Mathematical programming computation*, vol. 3, no. 3, p. 165, 2011.

[15] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from a few entries," *IEEE Transactions on Information Theory*, vol. 56, no. 6, pp. 2980–2998, 2010.

[16] L. Xie, Y. Xie, S.-M. Wu, F.-C. Lin, and W. Song, "Communication efficient signal detection for distributed ambient noise imaging."

[17] S.-M. Wu, K. M. Ward, J. Farrell, F.-C. Lin, M. Karplus, and R. B. Smith, "Anatomy of old faithful from subsurface seismic imaging of the yellowstone upper geyser basin," *Geophysical Research Letters*, vol. 44, no. 20, 2017.