7. Appendix

7.1. Proofs for Noiseless GRC Case

Proof of Lemma 1

Proof. First, $rank(X_2A^{(C)}) = rank(X_1A^{(C)}) = r$ and $rank(A^{(R)}X_2) = rank(A^{(R)}X_1) = r$. Since $span(X_1A^{(C)}), span(X_2A^{(C)})$ are subspaces of $span(X_1), span(X_2)$ respectively, and $dim(span(X_2)) \leq r$ we get $span(X_2) = span(X_2A^{(C)}) = span(X_1A^{(C)}) = span(X_1)$, and we define $U \in \mathcal{O}_{n_1\times r}$ a basis for this subspace. For X_1, X_2 there are $Y_1, Y_2 \in \mathbb{R}_{r\times n_2}$ such that $X_1 = UY_1, X_2 = UY_2$. Therefore $A^{(R)}UY_1 = A^{(R)}UY_2$. Since $rank(A^{(R)}UY_1) = r$ and $U \in \mathcal{O}_{n_1\times r}$ we get $rank(A^{(R)}U) = r$, hence the matrix $U^T{A^{(R)}}^TA^{(R)}U$ is invertible, which gives $Y_1 = Y_2$, and therefore $X_1 = UY_1 = UY_2 = X_2$. □

Proof of Lemma 2

Proof. $span(XA^{(C)}) \subseteq span(X)$ and $rank(XA^{(C)}) = rank(X) = r$, hence $span(XA^{(C)}) = span(X)$ and $\hat{U}$ from stage 1 in Algorithm 1 is a basis for $span(X)$. We can write $X = \hat{U}Y$ for some matrix $Y \in \mathbb{R}_{r\times n_2}$. Since $rank(A^{(R)}\hat{U}Y) = rank(\hat{U}) = r$, we have $rank(A^{(R)}\hat{U}) = r$. Thus eq. (6) gives $\hat{X}$ in closed form and we get:

$$\begin{aligned} A^{(R)}\hat{X} = A^{(R)}\hat{U}[\hat{U}^T{A^{(R)}}^TA^{(R)}\hat{U}]^{-1}\hat{U}^T{A^{(R)}}^TB^{(R,0)} = \\ A^{(R)}\hat{U}[\hat{U}^T{A^{(R)}}^TA^{(R)}\hat{U}]^{-1}\hat{U}^T{A^{(R)}}^TA^{(R)}\hat{U}Y = \\ A^{(R)}\hat{U}Y = A^{(R)}X. \end{aligned} \tag{16}$$

$$\begin{aligned} \hat{X}A^{(C)} = \hat{U}[\hat{U}^T{A^{(R)}}^TA^{(R)}\hat{U}]^{-1}\hat{U}^T{A^{(R)}}^TA^{(R)}XA^{(C)} = \\ \hat{U}[\hat{U}^T{A^{(R)}}^TA^{(R)}\hat{U}]^{-1}\hat{U}^T{A^{(R)}}^TA^{(R)}\hat{U}YA^{(C)} = \\ \hat{U}YA^{(C)} = XA^{(C)}. \end{aligned} \tag{17}$$

□

Lemma 3. *Let $V \in \mathcal{O}_{n\times r}$ and $A^{(C)} \in \mathbb{R}_{n\times k}$ be a random matrix $A^{(C)} \overset{i.i.d.}{\sim} N(0,\sigma^2)$. Then $V^TA^{(C)} \overset{i.i.d.}{\sim} N(0,\sigma^2)$.*

Proof. For any two matrices $A \in \mathbb{R}_{n_1\times n_2}$ and $B \in \mathbb{R}_{m_1\times m_2}$ we define their Kronecker product as a matrix in $\mathbb{R}_{n_1m_1\times n_2m_2}$:

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & . & . & a_{1n_2}B \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ a_{n_11}B & a_{n_12}B & . & . & a_{n_1n_2}B \end{pmatrix} \tag{18}$$

Now, we have $vec(V^TA^{(C)}) = (I_n \otimes V^T)vec(A^{(C)})$ and since $vec(A^{(C)}) \sim N(0, \sigma I_n)$ the vector $(I_n \otimes V^T)vec(A^{(C)})$ is also a multivariate Gaussian vector with zero mean and covariance matrix:

$$\begin{aligned} COV\left(V^TA^{(C)}\right) = COV\left((I_n \otimes V^T)vec(A^{(C)})\right) = \\ (I_n \otimes V^T)COV\left(vec(A^{(C)})\right)(I_n \otimes V^T)^T = \\ \sigma^2(I_n \otimes V^T)(I_n \otimes V^T)^T = \sigma^2 I_r \otimes I_n = \sigma^2 I_{nr}. \end{aligned} \tag{19}$$

□

Proof of Theorem 1

For the GRC model, Lemmas 1,2 and 3 can be used to prove exact recovery of X with the minimal possible number of measurements:

Proof. Let $U\Sigma V^T$ be the SVD of X. From Lemma 3 the elements of the matrix $V^T A^{(C)}$ have a continuous Gaussian distribution and since the measure of low rank matrices is zero and $k^{(C)} \geq r$ we get that $P(rank(V^T A^{(C)}) = r) = 1$. Since $B^{(C)} = U\Sigma V^T A^{(C)}$ we get $P(rank(B^{(C)}) = rank(U\Sigma V^T A^{(C)}) = r) = 1$. In the same way $P(rank(B^{(R)}) = r) = 1$. Combining Lemma 2 with Lemma 1 give us the required result. □

7.2. Gradient Descent

The gradient descent stage is performed directly in the space of rank r matrices, using the decomposition $\hat{X}$=WS where $W \in \mathbb{R}_{n_1 \times r}$ and $S \in \mathbb{R}_{r \times n_2}$ and computing the gradient of the loss as a function of W and S,

$$\mathcal{L}(W,S) = \mathcal{F}(WS) = ||A^{(R)}WS - B^{(R)}||_F^2 + ||WSA^{(C)} - B^{(C)}||_F^2. \tag{20}$$

We want to minimize eq. (20) but the loss $\mathcal{L}$ isn't convex and therefore gradient descent may fail to converge to a global optimum. We propose $\hat{X}$ (the output of SVLS) as a starting point which may be close enough to enable gradient descent to converge to the global optimum, and in addition may accelerate convergence.

The gradient of $\mathcal{L}$ is (using the chain rule)

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial W} &= 2\Big[A^{(R)^T}(A^{(R)}WS - B^{(R)})S^T + (WSA^{(C)} - B^{(C)})A^{(C)^T}S^T\Big] \\ \frac{\partial \mathcal{L}}{\partial S} &= 2\Big[W^T A^{(R)^T}(A^{(R)}WS - B^{(R)}) + W^T(WSA^{(C)} - B^{(C)})A^{(C)^T}\Big]\end{aligned} \tag{21}$$

7.3. Proofs for Noiseless RCMC Case

We prove that if $U \in \mathcal{O}_{n_1 \times r}$ is orthonormal then with high probability $p^{-1}||U^T A^{(R)^T} A^{(R)} U - pI_r||_2 < 1$. Because U is orthonormal, this is equivalent to

$$p-1||UU^T A^{(R)^T} A^{(R)} UU^T - pUU^T||_2 < 1 \Leftrightarrow p^{-1}||P_U P_{A^{(R)^T}} P_U - pP_U||_2 < 1 \tag{22}$$

where $P_U = UU^T, P_{A^{(R)^T}} = A^{(R)^T} A^{(R)}$ and $p^{(R)} = p$. We generalize Theorem 4.1 from (Candès & Recht, 2009).

Lemma 4. *Suppose $A^{(R)}$ as in the RCMC model with inclusion probability p, and $U \in \mathcal{O}_{n_1 \times r}$ with $\mu(U) = \frac{n_1}{r} max_i ||P_U(e_i)||^2 = \mu$. Then there is a numerical constant C_R such that for all $\beta > 1$, if $C_R\sqrt{\frac{\beta log(n_1) r\mu}{pn_1}} < 1$ then:*

$$P\left(p^{-1}||P_U P_{A^{(R)^T}} P_U - pP_U||_2 < C_R\sqrt{\frac{\beta log(n_1) r\mu}{pn_1}}\right) > 1 - 3n_1^{-\beta} \tag{23}$$

The proof of Lemma 4 builds upon (yet generalizes) the proof of Theorem 4.1 from (Candès & Recht, 2009). We next present a few lemmas which are required for the proof of Lemma 4. We start with a lemma from (Candès & Romberg, 2007).

Lemma 5. *If y_i is a family of vectors in $\mathbb{R}^d$ and r_i is a sequence of i.i.d. Bernoulli random variables with $P(r_i = 1) = p$, then*

$$E\big(p^{-1}||\Sigma_i(r_i - p)y_i \otimes y_i||\big) < C\sqrt{\frac{log(d)}{p}} max_i ||y_i|| \tag{24}$$

for some numerical constant C provided that the right hand side is less than 1.

We next use a result from large deviations theory (Talagrand, 1996):

Theorem 4. *Let $Y_1 ... Y_n$ be a sequence of independent random variables taking values in a Banach space and define*

$$Z = sup_{f \in F} \sum_{i=1}^{n} f(Y_i) \tag{25}$$

where F is a real countable set of functions such that if $f \in F$ then $-f \in F$.

Assume that $|f| \leq B$ and $E(f(Y_i)) = 0$ for every $f \in F$ and $i \in [n]$. Then there exists a constant C such that for every $t \geq 0$

$$P\big(|Z - E(Z)| \geq t\big) \leq 3exp\left(\frac{-t}{CB} log(1 + \frac{t}{\sigma + Br})\right) \tag{26}$$

where $\sigma = sup_{f \in F} \sum_{i=1}^{n} E(f^2(Y_i))$.

Theorem 4 is used in the proof of the next lemma which is taken from Theorem 4.2 in (Candès & Recht, 2009). We bring here the lemma and proof in our notations for convenience.

Lemma 6. *Let $U \in \mathcal{O}_{n \times r}$ with incoherence constant μ. Let r_i be i.i.d. Bernoulli random variables with $P(r_i = 1) = p$ and let $Y_i = p^{-1}(r_i - p)P_U(e_i) \otimes P_U(e_i)$ for $i = 1, .., n$. Let $Y = \sum_{i=1}^{n} Y_i$* and $Z = ||Y||_2$. Suppose $E(Z) \leq 1$. Then for every $\lambda > 0$ we have

$$P\Big(|Z - E(Z)| \geq \lambda\sqrt{\frac{\mu r log(n)}{pn}}\Big) \leq 3exp\Big(-\gamma min(\lambda^2 log(n), \lambda\sqrt{\frac{pnlog(n)}{\mu r}})\Big) \tag{27}$$

for some positive constant γ.

Proof. We know that $Z = ||Y||_2 = sup_{f_1, f_2} \langle f_1, Y f_2\rangle = sup_{f_1, f_2} \sum_{i=1}^{n} \langle f_1, Y_i f_2\rangle$, where the supremum is taken over a countable set of unit vectors $f_1, f_2 \in F_V$. Let F be the set of all functions f such that $f(Y) = \langle f_1, Y f_2\rangle$ for some unit vectors $f_1, f_2 \in F_V$. For every $f \in F$ and $i \in [n]$ we have $E(f(Y_i)) = 0$. From the incoherence of U we conclude that

$$|f(Y_i)| = p^{-1}|r_i - p| \times |\langle f_1, P_U(e_i)\rangle| \times |\langle P_U(e_i), f_2\rangle| \leq p^{-1}||P_U(e_i)||^2 \leq p^{-1}\frac{r}{n}\mu. \tag{28}$$

In addition

$$E(f^2(Y_i)) = p^{-1}(1-p)\langle f_1, P_U(e_i)\rangle^2 \langle P_U(e_i), f_2\rangle^2 \leq$$
$$p^{-1}||P_U(e_i)||^2 |\langle P_U(e_i), f_2\rangle^2| \leq p^{-1}\frac{r}{n}\mu |\langle P_U(e_i), f_2\rangle|^2. \tag{29}$$

Since $\sum_{i=1}^{n} |\langle P_U(e_i), f_2\rangle|^2 = \sum_{i=1}^{n} |\langle e_i, P_U(f_2)\rangle|^2 = ||P_U(f_2)||^2 \leq 1$, we get $\sum_{i=1}^{n} E(f^2(Y_i)) \leq p^{-1}\frac{r}{n}\mu$.

We can take $B = 2p^{-1}\frac{r}{n}\mu$ and $t = \lambda\sqrt{\frac{\mu r log(n)}{pn}}$ and from Theorem 4:

$$P(|Z - E(Z)| \geq t) \leq 3exp\left(\frac{-t}{KB} log(1 + \frac{t}{2})\right) \leq 3exp\left(\frac{-tlog(2)}{KB} min(1, \frac{t}{2})\right) \tag{30}$$

where the last inequality is due to the fact that for every $u > 0$ we have $log(1 + u) \geq log(2)min(1, u)$. Taking $\gamma = -log(2)/K$ finishes our proof. □

We are now ready to prove Lemma 4

Proof. (Lemma 4) Represent any vector $w \in R^{n_1}$ in the standard basis as $w = \sum_{i=1}^{n_1} \langle w, e_i\rangle e_i$. Therefore $P_U(w) = \sum_{i=1}^{n_1} \langle P_U(w), e_i\rangle e_i = \sum_{i=1}^{n_1} \langle w, P_U(e_i)\rangle e_i$. Recall the r_i Bernoulli variables which determine if e_i is included as a row of $A^{(R)}$ as in Section 2 and define Y_i and Z as in Lemma 6. We get

$$P_{A^{(R)^T}} P_U(w) = \sum_{i=1}^{n_1} r_i \langle w, P_U(e_i)\rangle e_i \Longrightarrow P_U P_{A^{(R)^T}} P_U(w) = \sum_{i=1}^{n_1} r_i \langle w, P_U(e_i)\rangle P_U(e_i) \tag{31}$$

In other words the matrix $P_U P_{A^{(R)^T}} P_U$ is given by

$$P_U P_{A^{(R)^T}} P_U = \sum_{i=1}^{n_1} r_i P_U(e_i) \otimes P_U(e_i) \tag{32}$$

U is $\mu-$incoherent, thus $max_{i\in[n_1]}||P_U(e_i)|| \leq \sqrt{\frac{r\mu}{n_1}}$, hence from Lemma 5 we have for p large enough:

$$E(p^{-1}||P_U P_{A^{(R)^T}} P_U - pP_U||_2) < C\sqrt{\frac{log(n_1)r\mu}{pn_1}} \leq 1. \tag{33}$$

For $\beta > 1$ which satisfy the lemma's requirement, take $\lambda = \sqrt{\frac{\beta}{\gamma}}$ where γ as in Theorem 4. We get that if $p > \frac{\mu log(n_1) r\beta}{n_1\gamma}$ then from Lemma 6 with probability of at least $1 - 3n_1^{-\beta}$ we have $Z \leq C\sqrt{\frac{log(n_1)r\mu}{pn_1}} + \frac{1}{\sqrt{\gamma}}\sqrt{\frac{log(n_1)r\mu\beta}{pn_1}}$. Taking $C_R = C + \frac{1}{\sqrt{\gamma}}$ finishes our proof. □

Proof of Theorem 2

Proof. From Lemma 4 and using a union bound we have that with probability $> 1 - 6min(n_1, n_2)^{-\beta}$, ${p^{(R)}}^{-1}||p^{(R)}I_r - U^T{A^{(R)}}^T A^{(R)}U||_2 < 1$ and ${p^{(C)}}^{-1}||p^{(C)}I_r - V^T A^{(C)}{A^{(C)}}^T V||_2 < 1$. Since the singular values of $p^{(R)}I_r - U^T{A^{(R)}}^T A^{(R)}U$ are $|p^{(R)} - \sigma_i(U^T{A^{(R)}}^T A^{(R)}U)|$ for $1 \leq i \leq r$, we have

$$p^{(R)} - \sigma_r(U^T{A^{(R)}}^T A^{(R)}U) \leq \sigma_1(p^{(R)}I_r - U^T{A^{(R)}}^T A^{(R)}U) < p^{(R)} \Rightarrow 0 < \sigma_r(U^T{A^{(R)}}^T A^{(R)}U) \tag{34}$$

and similarly for $V^T A^{(C)}{A^{(C)}}^T V$. Therefore $rank(A^{(R)}U) = rank(V^T A^{(C)}) = r$ and $rank(A^{(R)}X) = rank(XA^{(C)}) = r$ with probability $> 1 - 6min(n_1, n_2)^{-\beta}$. From Lemma 2 we get $A^{(R)}X = A^{(R)}\hat{X} \quad XA^{(C)} = \hat{X}A^{(C)}$ and from Lemma 1 we get $X = \hat{X}$ with probability $> 1 - 6min(n_1, n_2)^{-\beta}$. □

7.4. Proofs for Noisy GRC Case

The proof of Theorem 3 is using strong concentration results on the largest and smallest singular values of $n \times k$ matrix with i.i.d Gaussian entries:

Theorem 5. *(Szarek, 1991) Let $A \in \mathbb{R}_{n\times k}$ be a random matrix $A \overset{i.i.d.}{\sim} N(0, \frac{1}{n})$. Then, its largest and smallest singular values obey:*

$$P\Big(\sigma_1(A) > 1 + \frac{\sqrt{k}}{\sqrt{n}} + t\Big) \leq e^{-nt^2/2}$$
$$P\Big(\sigma_k(A) \leq 1 - \frac{\sqrt{k}}{\sqrt{n}} - t\Big) \leq e^{-nt^2/2}. \tag{35}$$

Corollary 2. *Let $A \in \mathbb{R}_{n\times k}$ be a random matrix $A \overset{i.i.d.}{\sim} N(0, 1)$ where $n \geq 4k$, and let $A^\dagger$ be the Moore-Penrose pseudoinverse of A. Then*

$$P\left(||A^\dagger||_2 \leq \frac{6}{\sqrt{n}}\right) > 1 - e^{-n/18} \tag{36}$$

Proof. Since $A^\dagger$ is the pseudoinverse of A, $||A^\dagger||_2 = \frac{1}{\sigma_k(A)}$ and from Theorem 5 we get $\sigma_k(A) \geq \sqrt{n} - \sqrt{k} - t\sqrt{n}$ with probability $\geq 1 - e^{nt^2/2}$ (notice the scaling by $\sqrt{n}$ of the entries of A compared to Theorem 5). Therefore, if we take $n \geq 4k$ and $t = \frac{1}{3}$ we get

$$P\left(||A^\dagger||_2 \leq \frac{6}{\sqrt{n}}\right) = P\left(\sigma_k(A) \geq \frac{\sqrt{n}}{6}\right) \geq 1 - e^{-n/18}. \tag{37}$$

□

We also use the following lemma from (Shalev-Shwartz & Ben-David, 2014):

Lemma 7. *Let Q to be a finite set of vectors in $\mathbb{R}^n$, let $\delta \in (0, 1)$ and k be an integer such that*

$$\epsilon \equiv \sqrt{\frac{6log(2|Q|/\delta)}{k}} \leq 3. \tag{38}$$

Let $A \in \mathbb{R}_{k\times n}$ be a random matrix with $A \overset{i.i.d.}{\sim} N(0, \frac{1}{k})$. Then,

$$P\left(max_{x\in Q}\left|\frac{||Ax||^2}{||x||^2} - 1\right| \leq \epsilon\right) > 1 - \delta. \tag{39}$$

Lemma 7 is a direct result of the Johnson-Lindenstrauss lemma (Dasgupta & Gupta, 2003) applied to each vector in Q and using the union bound. Representing the vectors in Q as a matrix, Lemma 7 shows that $A^{(R)}, A^{(C)}$ preserve matrix Frobenius norm with high probability - a weaker property than the RIP which holds for *any* low-rank matrix.

To prove Theorem 3, we first represent $||X - \hat{X}||_F$ as a sum three parts (Lemma 8), then give probabilistic upper bounds to each of the parts and finally use union bound. We define $A^{(R)}_{\hat{U}} = A^{(R)}\hat{U}$ and $A^{(C)}_{V^T} = V^T A^{(C)}$. From Lemma 3 $A^{(R)}_{\hat{U}}, A^{(C)}_{V^T} \overset{i.i.d.}{\sim} N(0,1)$, hence $rank(A^{(R)}_{\hat{U}}) = rank(A^{(C)}_{V^T}) = r$ with probability 1. We assume w.l.o.g that $\hat{X} = \hat{X}^{(R)}$ (see SVLS description). Therefore, from eq. (9) we have $\hat{X} = \hat{U}(A^{(R)^T}_{\hat{U}} A^{(R)}_{\hat{U}})^{-1} A^{(R)^T}_{\hat{U}} B^{(R)}$.

We denote by $A^{(R)^\dagger}_{\hat{U}} = (A^{(R)^T}_{\hat{U}} A^{(R)}_{\hat{U}})^{-1} A^{(R)^T}_{\hat{U}}$ and $A^{(C)^\dagger}_{V^T} = A^{(C)^T}_{V^T}(A^{(C)}_{V^T} A^{(C)^T}_{V^T})^{-1}$ the Moore-Penrose pseudoinverse of $A^{(R)}_{\hat{U}}$ and $A^{(C)}_{V^T}$, respectively. We next prove the following lemma

Lemma 8. *Let $A^{(R)}$ and $A^{(C)}$ be as in the GRC model and $Z^{(R)}, Z^{(C)}$ be noise matrices. Let $\hat{X}$ be the output of SVLS. Then:*

$$||X - \hat{X}||_F \leq \mathbf{I} + \mathbf{II} + \mathbf{III}$$

where:

$$\mathbf{I} \equiv ||(B^{(C,0)} - B^{(C)}_{(r)})A^{(C)^\dagger}_{V^T}||_F \tag{40}$$

$$\mathbf{II} \equiv ||\hat{U}A^{(R)^\dagger}_{\hat{U}} A^{(R)}(B^{(C,0)} - B^{(C)}_{(r)})A^{(C)^\dagger}_{V^T}||_F \tag{41}$$

$$\mathbf{III} \equiv ||\hat{U}A^{(R)^\dagger}_{\hat{U}} Z^{(R)}||_F. \tag{42}$$

Proof. We represent $||X - \hat{X}||_F$ as follows

$$\begin{aligned} ||X - \hat{X}||_F = \\ ||X - \hat{U}(A^{(R)^T}_{\hat{U}} A^{(R)}_{\hat{U}})^{-1} A^{(R)^T}_{\hat{U}}(A^{(R)}X + Z^{(R)})||_F = \\ ||X - \hat{U}A^{(R)^\dagger}_{\hat{U}} A^{(R)}X \ - \hat{U}A^{(R)^\dagger}_{\hat{U}} Z^{(R)}||_F \leq \\ ||X - \hat{U}A^{(R)^\dagger}_{\hat{U}} A^{(R)}X||_F + \mathbf{III} \end{aligned} \tag{43}$$

where we have used the triangle inequality. We next use the following equality

$$XA^{(C)}A^{(C)^\dagger}_{V^T} V^T = U\Sigma V^T A^{(C)} A^{(C)^\dagger}_{V^T} V^T = U\Sigma V^T = X \tag{44}$$

to obtain:

$$\begin{aligned} ||X - \hat{U}A^{(R)^\dagger}_{\hat{U}} A^{(R)}X||_F = \\ ||(I_n - \hat{U}A^{(R)^\dagger}_{\hat{U}} A^{(R)})X||_F = \\ ||(I_n - \hat{U}A^{(R)^\dagger}_{\hat{U}} A^{(R)})XA^{(C)}A^{(C)^\dagger}_{V^T} V^T||_F = \\ ||(I_n - \hat{U}A^{(R)^\dagger}_{\hat{U}} A^{(R)})B^{(C,0)}A^{(C)^\dagger}_{V^T}||_F \end{aligned} \tag{45}$$

where the last equality is true because V is orthogonal.

Since $\hat{U}$ is a basis for $span(B_{(r)}^{(C)})$ there exists a matrix Y such that $\hat{U}Y = B_{(r)}^{(C)}$ and we get:

$$(I_n - \hat{U}A_{\hat{U}}^{(R)\dagger}A^{(R)})B_{(r)}^{(C)} = B_{(r)}^{(C)} - \hat{U}A_{\hat{U}}^{(R)\dagger}A^{(R)}\hat{U}Y = B_{(r)}^{(C)} - \hat{U}Y = 0. \tag{46}$$

Therefore

$$\begin{aligned} ||(I_n - \hat{U}A_{\hat{U}}^{(R)\dagger}A^{(R)})B^{(C,0)}A_{V^T}^{(C)\dagger}||_F = \\ ||(I_n - \hat{U}A_{\hat{U}}^{(R)\dagger}A^{(R)})(B^{(C,0)} - B_{(r)}^{(C)})A_{V^T}^{(C)\dagger}||_F \leq \\ ||(B^{(C,0)} - B_{(r)}^{(C)})A_{V^T}^{(C)\dagger}||_F + ||\hat{U}A_{\hat{U}}^{(R)\dagger}A^{(R)}(B^{(C,0)} - B_{(r)}^{(C)})A_{V^T}^{(C)\dagger}||_F = \mathbf{I} + \mathbf{II} \end{aligned} \tag{47}$$

Combining eq. (43) and eq. (47) gives the required result. □

We next bound each of the three parts in the formula of Lemma 8. We use the following claim:

Claim 1. $||B^{(C,0)} - B_{(r)}^{(C)}||_2 \leq 2||Z^{(C)}||_2$

Proof. We know that $||B^{(C)} - B_{(r)}^{(C)}||_2 \leq ||B^{(C)} - B^{(C,0)}||_2$ since $rank(B_{(r)}^{(C)}) = rank(B^{(C,0)}) = r$ with probability 1, and by definition $B_{(r)}^{(C)}$ is the closest rank-r matrix to $B^{(C)}$ in Frobenius norm. Therefore from the triangle inequality

$$||(B^{(C,0)} - B_{(r)}^{(C)})||_2 \leq ||B^{(C)} - B_{(r)}^{(C)}||_2 + ||B^{(C)} - B^{(C,0)}||_2 \leq 2||B^{(C,0)} - B^{(C)}||_2 = 2||Z^{(C)}||_2. \tag{48}$$

□

Now we are ready to prove Theorem 3. The proof uses the following inequalities for matrix norms for any two matrices A, B:

$$\begin{aligned} ||AB||_2 \leq ||A||_2||B||_2 \\ ||AB||_F \leq ||A||_F||B||_2 \\ rank(A) \leqslant r \Rightarrow ||A||_F \leq \sqrt{r}||A||_2. \end{aligned} \tag{49}$$

Proof. (Theorem 3) We prove (probabilistic) upper bounds on the three terms appearing in Lemma 8.

1. We have
$$rank\left((B^{(C,0)} - B_{(r)}^{(C)})A_{V^T}^{(C)\dagger}\right) \leqslant rank\left(A_{V^T}^{(C)\dagger}\right) \leqslant r. \tag{50}$$
Therefore
$$\mathbf{I} = ||(B^{(C,0)} - B_{(r)}^{(C)})A_{V^T}^{(C)\dagger}||_F \leq \sqrt{r}||(B^{(C,0)} - B_{(r)}^{(C)})||_2||A_{V^T}^{(C)\dagger}||_2 \tag{51}$$
Since $A_{V^T}^{(C)} \overset{i.i.d.}{\sim} N(0,1)$, from Corollary 2 we get $P\left(||A_{V^T}^{(C)\dagger}||_2 \leq \frac{6}{\sqrt{k}}\right) \geq 1 - e^{-k/18}$ for $k \geq 4r$, hence with probability $\geq 1 - e^{-k/18}$,
$$\mathbf{I} \leq 6\sqrt{\frac{r}{k}}||(B^{(C,0)} - B_{(r)}^{(C)})||_2. \tag{52}$$
From Claim 1 and eq. (40) we get a bound on $\mathbf{I}$ for some absolute constants C_1, c_1:
$$P\left(\mathbf{I} \leq C_1\sqrt{\frac{r}{k}}||Z^{(C)}||_2\right) > 1 - e^{-c_1 k}. \tag{53}$$

2. $\hat{U}$ is orthogonal and can be omitted from **II** without changing the norm. Applying the second inequality in eq. (49) twice, we get the inequality:

$$\mathbf{II} = ||\hat{U} A_{\hat{U}}^{(R)^\dagger} A^{(R)}(B^{(C,0)} - B_{(r)}^{(C)}) A_{V^T}^{(C)^\dagger}||_F \leq ||A_{\hat{U}}^{(R)^\dagger}||_2 ||A^{(R)}(B^{(C,0)} - B_{(r)}^{(C)})||_F ||A_{V^T}^{(C)^\dagger}||_2. \tag{54}$$

From Corollary 2 we know that for $k > 4r$ we have $||A_{\hat{U}}^{(R)^\dagger}||_2 \leq \frac{6}{\sqrt{k}}$ and $||A_{V^T}^{(C)^\dagger}||_2 \leq \frac{6}{\sqrt{k}}$, each with probability $> 1 - e^{-k/18}$. Therefore,

$$P\Big(\mathbf{II} \leq \frac{36}{k} ||A^{(R)}(B^{(C,0)} - B_{(r)}^{(C)})||_F\Big) > 1 - 2e^{-k/18}. \tag{55}$$

$A^{(R)}$ and $B^{(C,0)} - B_{(r)}^{(C)}$ are independent and $rank(B^{(C,0)} - B_{(r)}^{(C)}) \leq 2r$. Therefore we can apply Lemma 7 with k such that $\frac{k}{6} > log(2k) + \frac{k}{18}$ (this holds for $k \geq 40$) to get with probability $> 1 - 2e^{-k/18}$:

$$\mathbf{II} \leq \frac{36}{k} ||A^{(R)}(B^{(C,0)} - B_{(r)}^{(C)})||_F \leq \frac{36\sqrt{2k}}{k} ||(B^{(C,0)} - B_{(r)}^{(C)})||_F \leq 36\sqrt{4\frac{r}{k}} ||(B^{(C,0)} - B_{(r)}^{(C)})||_2. \tag{56}$$

From eq. (55) and (56) together with Claim 1 we have constants C_2 and c_2 such that,

$$P\Big(\mathbf{II} \leq C_2 ||Z^{(C)}||_2\Big) > 1 - 3e^{-c_2 k}. \tag{57}$$

3. $rank(A_{\hat{U}}^{(R)^\dagger}) \leq r$ and from Corollary 2 we get $P\Big(||A_{\hat{U}}^{(R)^\dagger}||_2 \leq \frac{6}{\sqrt{k}}\Big) > 1 - e^{-k/18}$ for $k > 4r$. Therefore, with probability $> 1 - e^{-k/18}$:

$$\mathbf{III} = ||\hat{U} A_{\hat{U}}^{(R)^\dagger} Z^{(R)}||_F = ||A_{\hat{U}}^{(R)^\dagger} Z^{(R)}||_F \leq \sqrt{r} ||A_{\hat{U}}^{(R)^\dagger} Z^{(R)}||_2 \leq \sqrt{r} ||A_{\hat{U}}^{(R)^\dagger}||_2 ||Z^{(R)}||_2 \leq \frac{6\sqrt{r}}{\sqrt{k}} ||Z^{(R)}||_2. \tag{58}$$

Hence we have constants C_3 and c_3 such that, $> 1 - e^{-c_3 k}$.

$$P\Big(\mathbf{III} \leq C_3 ||Z^{(R)}||_2\Big) > 1 - e^{-c_3 k}. \tag{59}$$

Combining equations (53,57,59) with Lemma 8 and taking the union bound while setting $c^{(C)} = C_1 + C_2$, $c^{(R)} = C_3$ with $c = min(c_1, c_2, c_3)$ concludes our proof. □

7.5. Simulations for Large Values of n

We varied n between 10 and 1000, with results averaged over 100 different matrices of rank 3 at each point, and tried to recover them using $k = 20$ row and column measurements. Measurement matrices were $A^{(R)}, A^{(C)} \overset{i.i.d.}{\sim} \frac{1}{n}$ to allow similar norms for each measurement vector for different values of n. Recovery performance was insensitive to n. if we take $A^{(R)}, A^{(C)} \overset{i.i.d.}{\sim} N(0,1)$ instead of $N(0, \frac{1}{n})$, the scaling of our results is in agreement with Theorem 3.

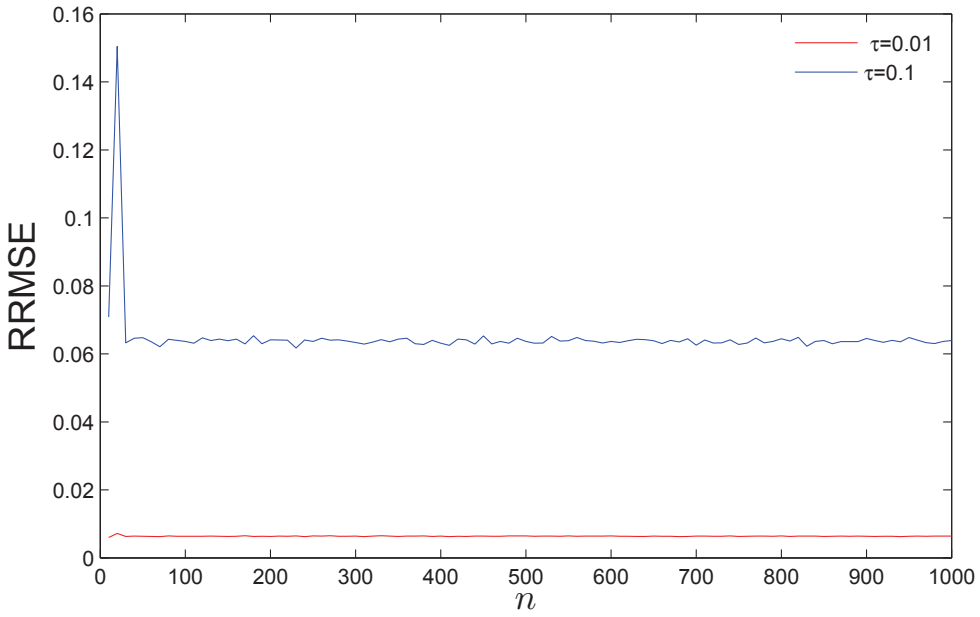

Figure 5. Reconstruction error for $n \times n$ matrix where n is varied between 10 and 1000, $k = 20$ and $r = 3$ and two different noise levels: $\tau = 0.1$ (blue) and $\tau = 0.01$ (red). Each point represents average performance over 100 random matrices.

Next, we take $n, k, r \to \infty$ while the ratios $\frac{n}{k} = 5$ and $\frac{k}{r} = 4$ are kept constant, and compute the relative error for different noise level. Again, the relative error converges rapidly to constant, independent of n, k, r .

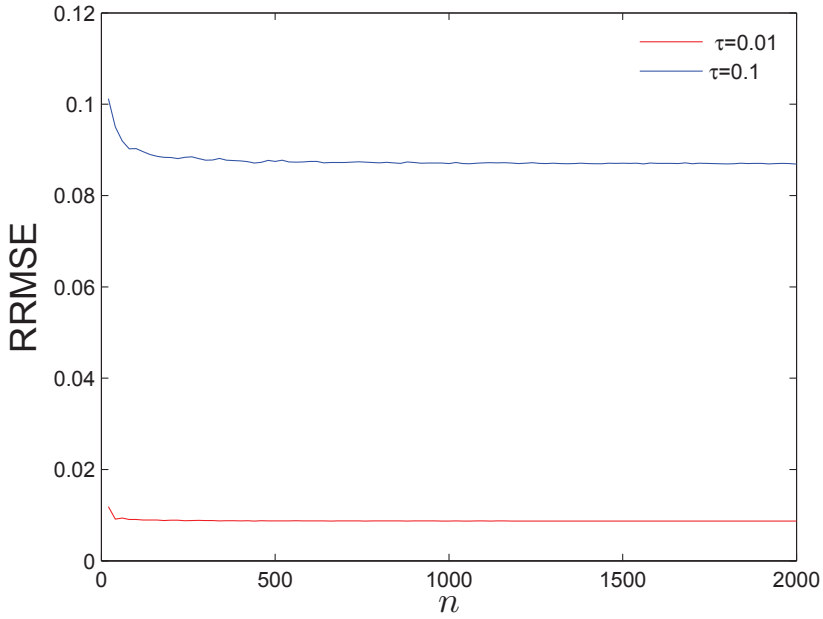

Figure 6. Reconstruction error for $n \times n$ matrix X with rank r varying from 1 to 50 and with $n = 20r, k = 4r$. Two different noise level are shown: $\tau = 0.1$ (blue) and $\tau = 0.01$ (red). Each point represents average performance over 100 random matrices.

7.6. Low Rank matrix Approximation

We bring here the one pass algorithm to approximate X from (Halko et al., 2011) for the convenience of the reader. The output of this algorithm isn't low rank if $k > r$. This algorithm is different from $SVLS_P$ and its purpose is to approximate a (possibly full rank) matrix by low rank matrix. We adjusted Algorithm 3 to our purpose with some changes. First, we estimate the rank of X using the elbow method from Section 3.3 and instead of calculating the QR decomposition of $B^{(C)}$ and ${B^{(R)}}^T$ we find their $\hat{r}$ largest singular vectors. Furthermore, we repeat part two in algorithm 3 while replacing the roles of columns and rows as in SVLS . This variation gives our modified algorithm $SVLS_P$ as described in Section 3.4.

Algorithm 3

Input: $A^{(R)}, A^{(C)}, B^{(R)}, B^{(C)}$

1. compute $Q^{(C)}R^{(C)}$ the QR decomposition of $B^{(C)}$, and $Q^{(R)}R^{(R)}$ the QR decomposition for ${B^{(R)}}^T$
2. Find the least-squares solution $Y = argmin_C ||Q^{(C)}B^{(C)} - CQ^{(R)^T}B^{(R)^T}||_F$.
3. Return the estimate $\hat{X} = Q^{(C)}YQ^{(R)^T}$.

We compared our SVLS to $SVLS_P$ which is presented in Section 3.4. We took $X \in \mathcal{M}^{(10)}_{1000\times 1000}$ and $\sigma = 1$. We tried to recover X in the GRC model with $k = 12$ for 100 different matrices. For each matrix, we compared the $RRMSE$ obtained for the outputs of SVLS and $SVLS_P$. The $RRMSE$ for $SVLS_P$ was lower than the $RRMSE$ for SVLS in most cases but the differences were very small and negligible.

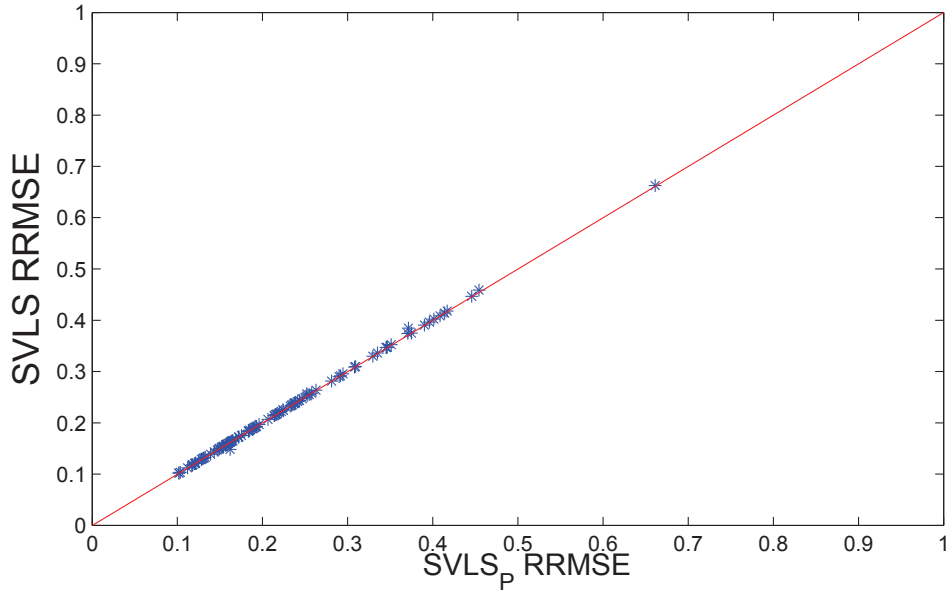

Figure 7. We recover a matrix X from 24000 measurements as in the GRC model 100 times. Figure shows average $RRMSE$ over 100 simulations for SVLS (Y axis) and $SVLS_P$ (X axis). The red linear line $Y = X$ was drawn for comparing those two algorithm, every dot that under the red line is a simulation that SVLS was better than $SVLS_P$ and every dot above the line tells the opposite

7.7. Rank Estimation

We test the elbow method for estimating the rank of X (see eq. (10)). We take a matrix X of size 400×400 and different ranks. We add Gaussian noise with $\sigma = 0.25$ while the measurements are sampled as in the RCMC model. For each number of measurements we sampled 100 matrices and took the average estimated rank. We compute the estimator $\hat{r}$ for different values of d, the number of measurements. We compare our method to the rank estimation which appears in

OptSpace (Keshavan et al., 2009) for the standard MC problem. Our simulation results, shown in Figure 8, indicate that the RCMC model with the elbow method is a much better design for rank estimation of X.

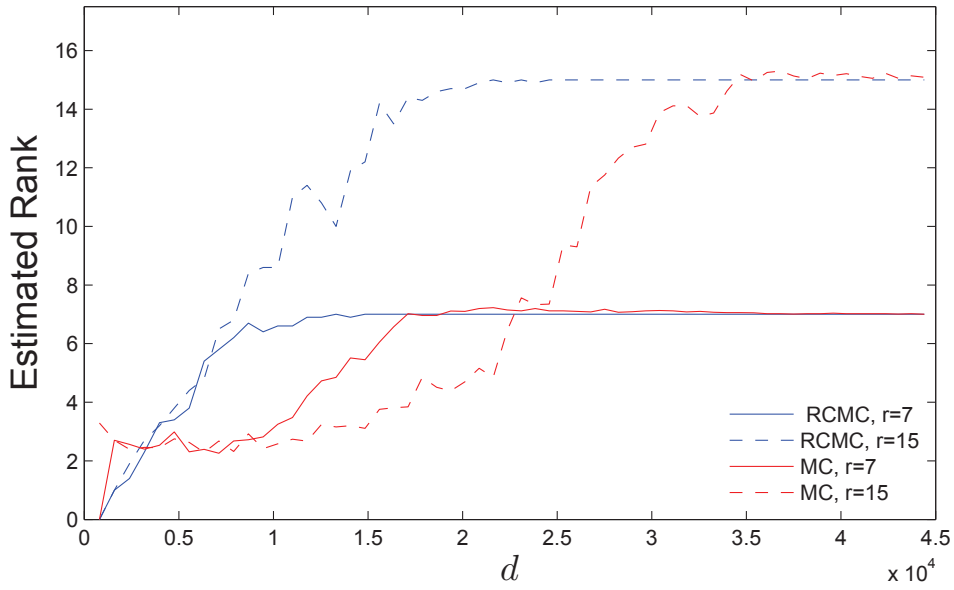

Figure 8. Estimation of $rank(X)$ vs. d, the number of measurements, $d = k(2n - k)$ where k is the number of columns in $B^{(C)}$ and number of rows in $B^{(R)}$. For each d we sampled 100 different matrices. Estimation was performed by the elbow method for RCMC model, as in eq. (10) in the main text, and for the MC model we used the method described in (Keshavan et al., 2009). RCMC recovers the correct rank with smaller number of measurements.

7.8. Test Error

In matrix completion with MC and RCMC ensembles the $RRMSE$ loss function measures the loss on both the observed and unobserved entries. This loss may be too optimistic when considering our prediction error only on unobserved entries. Thus, instead of including all measurements in calculation of the $RRMSE$ we compute a different measure of prediction error, given by the $RRMSE$ only on the unobserved entries. For each single-entry measurements operator $\mathcal{A}$ define $E(\mathcal{A})$ the set of measured entries and $\bar{E}$ it's complement, i.e. the set of unmeasured entries $(i, j) \in [n_1] \times [n_2]$. We define $X^{\bar{E}}$ to be a matrix such that $X^{\bar{E}}_{ij} = X_{ij}$ if $(i, j) \in \bar{E}$ and 0 otherwise. Instead of $RRMSE(X, \hat{X})$ we now calculate $RRMSE(X^{\bar{E}}, \hat{X}^{\bar{E}})$. This quantity measures our reconstruction only on the unseen matrix entries X_{ij}, and is thus not influenced by overfitting. In Table 2 we performed exactly the same simulation as in Table 1 but with $RRMSE(X^{\bar{E}}, \hat{X}^{\bar{E}})$. The results of OptSpace, SVT and SVLS stay similar to the results in Table 1 and our $RRMSE$ loss function does not show overfitting.

Table 2. $RRMSE$ only on the unknown measurements. for SVLS applied to RCMC, and OptSpace and SVT applied to the standard MC. Results represent average of 5 different random matrices. The results in the are parentheses the standard $RRMSE$ in Table 1.

NR	d	r	SVLS		OptSpace		SVT	
10^{-2}	120156	10	0.006	(0.006)	0.004	(0.004)	0.0074	(0.0073)
10^{-1}	120156	10	0.065	(0.064)	0.045	(0.044)	0.051	(0.05)
1	120156	10	0.619	(0.612)	0.49	(0.49)	0.52	(0.51)