A Technical Assumptions

Before stating the assumptions, we first define the ψ_2-norm.

Definition A.1 (ψ_2-norm)**.** *For a real valued random variable A, its ψ_2 norm is defined by*

$$\|A\|_{\psi_2} = \inf\{u > 0 \ : \ \mathbb{E}\exp(A^2/u^2) \leq 2\}.$$

Definition A.2 (sub-Gaussian)**.** *We say that a real random variable A is 1-sub-Gaussian if $\|A\|_{\psi_2} < 1$. We say that a random variable B with values in $\mathbb{R}^N$ is 1-sub-Gaussian if $\langle B, v\rangle$ is 1-sub-Gaussian for all $v \in \mathbb{R}^N$ with $\|v\| = 1$.*

As mentioned previously, we need to have assumptions that control the growth of W_t to be not too large and not too small. Because we have two phases the algorithm, initialization and iteration, we require two forms of these bounds. For initialization, our assumption is essentially the same as Assumption 1 of Lounici et al. (2014).

Assumption A.3 (sub-Gaussian W_t)**.** *For each $m \leq t$, each column $w_m \in \mathbb{R}^r$ of W_t satisfies:*

1. *w_m is drawn independently (for each m) from a 1-sub-Gaussian distribution;*

2. *there exists a numerical constant c_1 with $0 < c_1 \leq 1$ such that*

$$\mathbb{E}(\langle w_m, u\rangle) \geq c_1 \|\langle w_m, u\rangle\|_{\psi_2} \ \forall u \in \mathbb{R}^r.$$

For iteration, we also need non-asymptotic bounds on the singular values, which would hold if W_t were i.i.d. Gaussian from results from random matrix theory (see Corollary 5.35 of Vershynin (2010)).

Assumption A.4 (Growth of Singular Values)**.** *We assume that $\sigma_r(\mathring{X}) > 0$, and that there exists a C_{sv} large enough that for every $t \geq C_{\text{sv}}$, $\mathring{X}W_t$ satisfies*

$$\sigma_r(\mathring{X}W_t^T) \geq \frac{3}{4}\sigma_r(\mathring{X})\sqrt{t}, \|\mathring{X}W_t^T\| \leq \frac{3}{2}\sigma_1(\mathring{X})\sqrt{t} \tag{8}$$

with probability at least $1 - t^{-2}$ for $t \geq C_{\text{sv}}$.

For matrix completion, we need an incoherence assumption as in Candès and Tao (2009), Candès and Recht (2009), and Recht (2011). There are many ways of interpreting this parameter, but intuitively, it says that observing an entry actually gives information about other entries. It turns out that generating i.i.d. Gaussians for each entry of W_M will produce right singular vectors that are incoherent: with $W_M = U_{W_M}\Sigma_{W_M}V_{W_M}^T$ the SVD, for some constants C, c, with probability at least $1 - cM^{-3}\log M$, $\max_i \|P_{V_{W_M}} e_i\| \leq \sqrt{C\max\{r, \log M\}/M}$ (See Lemma 2.2 of (Candès and Recht, 2009)). Here P_V denotes projection to the column space of V. This metric is equivalent to the coherence definition given below, which leads to Assumption A.6.

Definition A.5. *The coherence of an $M \times r$ matrix V is $\mu(V) := \max_{m\in[M]}(M/r)\|e_m^T V\|_2^2$.*

Assumption A.6 (Incoherence)**.** *There exists some C_{inc} such that for large enough M, for any subset of $[t]$ of size M, with probability at least $1 - M^{-3}\log M$, $\mu(V_{W_M}) \leq C_{\text{inc}}\log M$.*

Note we do not assume incoherence of the column space of $\mathring{X}$. In practice, having incoherent column space is probably helpful. But for our theoretical results, because N is fixed as the number of columns t is growing, incoherence of $\mathring{X}$, which provides high probability bounds with respect to N (not t), are not as useful.

B Algorithm for Two Block Sizes and Uniformly Random Sampling Theorems

Algorithm 2 DoubleColumnSpaceEstimate: column space estimation with two block sizes

Input: Partially observable $Y_t \in \mathbb{R}^{N\times t}$; $k^{(1)}, k^{(2)} \in \mathbb{N}$, such that the total number of samples per column is $k^{(1)} + k^{(2)}$; $M_{\text{init}} \in \mathbb{N}$ the number of columns for initialization; $M_1, M_2 \in \mathbb{N}$, the sizes of blocks of columns for least squares; $s_1, s_2 \in \mathbb{N}$ the numbers of blocks; ϵ, the desired accuracy; a, a boolean indicator of active sampling

function DoubleColumnSpaceEstimate(Y_t, $k^{(1)}$, $k^{(2)}$, M_{init}, $M_1, M_2, s_1, s_2, \epsilon, a$)
  ▷ Spectral initialization with uniformly random sampling
  $\Omega_{M_{\text{init}}} \leftarrow \emptyset$
  for $m = 1, \dots, M_{\text{init}}$ **do**
    $S \sim \text{Unif}(\mathcal{C}(N, k^{(1)} + k^{(2)}))$
    $\Omega \leftarrow \Omega \cup (S \times \{m\})$
  end for
  $\hat{X} \leftarrow$ScaledPCA $(\mathcal{P}_\Omega(Y_t), k^{(1)} + k^{(2)}, N)$
  ▷ Least squares iteration
  $L_1 \leftarrow C^{\text{med}} \lceil \log M_1 \rceil$
  for $i = 1, \dots, s_1$ **do**
    $m \leftarrow M^{\text{init}} + (i-1)L_1 M_1 + 1$
    $I \leftarrow [m : (m + L_1 M_1 - 1)]$
    $\Omega^{(1)}, \Omega^{(2)} \leftarrow$Sample$(\hat{X}, k^{(1)}, k^{(2)}, I, a)$
    $\hat{X} \leftarrow$ MedianLS $(\hat{X}, Y_t, \Omega^{(1)}, \Omega^{(2)}$, M_1, m, $\epsilon)$
    $\Omega \leftarrow \Omega \cup \Omega^{(1)} \cup \Omega^{(2)}$
  end for
  $L_2 \leftarrow C^{\text{med}} \lceil \log M_2 \rceil$
  for $i = 1, \dots, s_2$ **do**
    $m \leftarrow M^{\text{init}} + s_1 L_1 M_1 + (i-1)L_2 M_2 + 1$
    $I \leftarrow [m : (m + L_2 M_2 - 1)]$
    $\Omega^{(1)}, \Omega^{(2)} \leftarrow$Sample$(\hat{X}, k^{(1)}, k^{(2)}, I, a)$
    $\hat{X} \leftarrow$ MedianLS $(\hat{X}, Y_t, \Omega^{(1)}, \Omega^{(2)}$, M_2, m, $\epsilon)$
    $\Omega \leftarrow \Omega \cup \Omega^{(1)} \cup \Omega^{(2)}$
  end for
  return $\hat{X}, \Omega$
end function

Theorem B.1 (Noisy observations, random sampling, for small σ_Z/ϵ)**.** *Suppose that U, the orthonormal part of* $\text{QR}(\mathring{X})$*, is $k^{(1)}$-isomeric. Suppose further that Assumptions 2.1, 4.5, A.3, A.4, A.6 hold, and $N/2 \geq k^{(1)} \geq r, k^{(2)} \geq 1$, $1 \geq \epsilon \geq e^{-M}M$, and Equation* (2) *hold. Then there exists constants $C^{\text{init}}_{B.1}$, $C^{\text{iter}}_{B.1}$, $C^{\text{prob}}_{B.1}$ such that for all $\epsilon > 0$, if we initialize with M_{init} columns, where*

$$M_{\text{init}} \geq C^{\text{init}}_{B.1} \frac{\sigma_1(\mathring{X})^6 N^2 (\log M_{\text{init}})^3 r^2}{\sigma_r(\mathring{X})^6 (k^{(1)}+k^{(2)})^2 \sigma_*(U;k^{(1)})^2},$$

and we use s blocks, where $\text{s} \geq \log\left(\frac{\sigma_r \sigma_*(U;k^{(1)})}{48\sqrt{r}\epsilon}\right)$, *and each block has size M, with*

$$M \geq C^{\text{iter}}_{B.1} \frac{\sigma_1(\mathring{X})^6 r^3 N (\log M)^2}{\sigma_r(\mathring{X})^6 k^{(2)} \sigma_*(U;k^{(1)})^2} + \log\left(\frac{1}{\epsilon}\right),$$

then ColumnSpaceEstimate*(Y_t, $k^{(1)}, k^{(2)}$, M_{init}, M, s, ϵ, True) returns an $\hat{X}$ such that $\sin\theta(U, \hat{X}) \leq \epsilon$ with probability at least $1 - 2M_{\text{init}}^{-2} - C^{\text{prob}}_{B.1} s M^{-2}$.*

Theorem B.2 (Noisy observations, random sampling, for large $\frac{\sigma_Z}{\epsilon}$)**.** *Suppose Assumptions 2.1, 4.1, A.3, A.4, A.6 hold, and $N/2 \geq k^{(1)} \geq r, k^{(2)} \geq 1$, $1 \geq \epsilon \geq e^{-M}M$. Let ϵ satisfy equation* (3)*. Then there exist constants $C^{\text{init}}_{B.2}, C^{\text{iter}}_{B.2}, C^{\text{prob}}_{B.2}$ such that for every $\epsilon > 0$, if we initialize with M_{init} columns, where*

$$M_{\text{init}} \geq C^{\text{init}}_{B.2} \frac{\sigma_1(\mathring{X})^6 N^2 (\log M_{\text{init}})^3 r^2}{\sigma_r(\mathring{X})^6 (k^{(1)}+k^{(2)})^2 \sigma_*(U;k^{(1)})^2}$$

and perform alternating minimization with $s_1 = \log\left(\frac{\sigma_r \sigma_*(U;k^{(1)})}{48\sigma_Z\sqrt{k^{(1)}}}\right)$ *blocks of size*

$$M_1 \geq C^{\text{iter}}_{B.1} \frac{\sigma_1(\mathring{X})^6 r^3 N(\log M)^2}{\sigma_r(\mathring{X})^6 k^{(2)} \sigma_*(U;k^{(1)})^2} + \log\left(\tfrac{1}{\epsilon}\right),$$

followed by alternating minimization with $s_2 = 1$ *block of size*

$$M_2 \geq C^{\text{iter}}_{B.2} \frac{r^2 \sigma_Z^2 \sigma_1(\mathring{X})^4 N k^{(1)} (\log M)^2}{\sigma_r(\mathring{X})^6 k^{(2)} \sigma_*(U;k^{(1)})^2 \epsilon^2} + \log\left(\tfrac{1}{\epsilon}\right),$$

then DoubleColumnSpaceEstimate*(*Y_t*,* $k^{(1)}$*,* $k^{(2)}$*,* M_{init}*,* M_1*,* M_2*,* s_1, s_2, ϵ*, True) returns an* $\hat{X}$ *such that* $\sin\theta(U, \hat{X}) \leq \epsilon$ *with probability at least* $1 - 2M_{\text{init}}^{-2} - C^{\text{prob}}_{B.1} s_1 M_1^{-2} - C^{\text{prob}}_{B.2} M_2^{-2}$.

C SmoothQR Hardt (2014)

SmoothQR (Hardt, 2014): Smooth Orthonormalization

function SmoothQR($\tilde{W}, \epsilon, \mu$)
  $W \leftarrow \text{QR}(\tilde{W}), G_H \leftarrow 0, \sigma \leftarrow \epsilon\|\tilde{W}\|/M$
  while $\mu(W) > \mu$ and $\sigma \leq \|\tilde{W}\|$ **do**
    $\tilde{W} \leftarrow \text{GS}(\tilde{W} + G_H)$ where $G_H \sim \mathcal{N}(0, \sigma^2/M)$
    $\sigma \leftarrow 2\sigma$
  end while
return (W, G_H)
end function

D Scaled PCA Estimator

Here $\mathbf{1}_N \in \mathbb{R}^{N\times N}$ is the matrix with each entry equal to 1, and $I_N \in \mathbb{R}^{N\times N}$ is the identity matrix.

ScaledPCA

Input: Partially observed $\mathcal{P}_\Omega(Y) \in \mathbb{R}^{N\times M}$; k, the number of entries per column, N the number of rows of $\mathcal{P}_\Omega(Y)$

function ScaledPCA($\mathcal{P}_\Omega(Y)$, k, N)
  $C \leftarrow \mathcal{P}_\Omega(Y)\mathcal{P}_\Omega(Y)^T$
  ▷ We denote by $\circ$ the Hadamard (elementwise) product
  $C_{\text{scaled}} \leftarrow \left(\frac{N^2}{k(k-1)}\mathbf{1}_N\right) \circ C + \left(\left(\frac{N}{k} - \frac{N^2}{k(k-1)}\right) I_N\right) \circ C$
  $\hat{X} \leftarrow \text{QR}(C_{\text{scaled}})$
  return $\hat{X}$
end function