Calibrated Elastic Regularization in Matrix Completion

Tingni Sun
Statistics Department, The Wharton School
University of Pennsylvania
Philadelphia, Pennsylvania 19104
tingni@wharton.upenn.edu

Cun-Hui Zhang
Department of Statistics and Biostatistics
Rutgers University
Piscataway, New Jersey 08854
czhang@stat.rutgers.edu

Abstract

This paper concerns the problem of matrix completion, which is to estimate a matrix from observations in a small subset of indices. We propose a calibrated spectrum elastic net method with a sum of the nuclear and Frobenius penalties and develop an iterative algorithm to solve the convex minimization problem. The iterative algorithm alternates between imputing the missing entries in the incomplete matrix by the current guess and estimating the matrix by a scaled soft-thresholding singular value decomposition of the imputed matrix until the resulting matrix converges. A calibration step follows to correct the bias caused by the Frobenius penalty. Under proper coherence conditions and for suitable penalties levels, we prove that the proposed estimator achieves an error bound of nearly optimal order and in proportion to the noise level. This provides a unified analysis of the noisy and noiseless matrix completion problems. Simulation results are presented to compare our proposal with previous ones.

1 Introduction

Let $\Theta \in \mathbb{R}^{d_1 \times d_2}$ be a matrix of interest and $\Omega^* = \{1, \ldots, d_1\} \times \{1, \ldots, d_2\}$. Suppose we observe vectors (ω_i, y_i),

$$y_i = \Theta_{\omega_i} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where $\omega_i \in \Omega^*$ and ε_i are random errors. We are interested in estimating Θ when n is a small fraction of $d_1 d_2$. A well-known application of matrix completion is the Netflix problem where y_i is the rating of movie b_j by user a_i for $\omega = (a_i, b_j) \in \Omega^*$ [1]. In such applications, the proportion of the observed entries is typically very small, so that the estimation or recovery of Θ is impossible without a structure assumption on Θ. In this paper, we assume that Θ is of low rank.

A focus of recent studies of matrix completion has been on a simpler formulation, also known as exact recovery, where the observations are assumed to be uncorrupted, i.e. $\varepsilon_i = 0$. A direct approach is to minimize rank(M) subject to $M_{\omega_i} = y_i$. An iterative algorithm was proposed in [5] to project a trimmed SVD of the incomplete data matrix to the space of matrices of a fixed rank r. The nuclear norm was proposed as a surrogate for the rank, leading to the following convex minimization problem in a linear space [2]:

$$\widehat{\Theta}^{(\mathrm{CR})} = \arg\min_M \Big\{ \|M\|_{(N)} : M_{\omega_i} = y_i \, \forall\, i \le n \Big\}.$$

We denote the nuclear norm by $\|\cdot\|_{(N)}$ here and throughout this paper. This procedure, analyzed in [2, 3, 4, 11] among others, is parallel to the replacement of the ℓ_0 penalty by the ℓ_1 penalty in solving the sparse recovery problem in a linear space.

In this paper, we focus on the problem of matrix completion with noisy observations (1) and take the exact recovery as a special case. Since the exact constraint is no longer appropriate in the presence of noise, penalized squared error $\sum_{i=1}^{n}(M_{\omega_i}-y_i)^2$ is considered. By reformulating the problem in Lagrange form, [8] proposed the spectrum Lasso

$$\widehat{\Theta}^{(\mathrm{MHT})}=\arg\min_{M}\Big\{\sum_{i=1}^{n}M_{\omega_i}^2/2-\sum_{i=1}^{n}y_iM_{\omega_i}+\lambda\|M\|_{(N)}\Big\}, \tag{2}$$

along with an iterative convex minimization algorithm. However, (2) is difficult to analyze when the sample fraction $\pi_0=n/(d_1d_2)$ is small, due to the ill-posedness of the quadratic term $\sum_{i=1}^{n}M_{\omega_i}^2$. This has led to two alternatives in [7] and [9]. While [9] proposed to minimize (2) under an additional ℓ_∞ constraint on M, [7] modified (2) by replacing the quadratic term $\sum_{i=1}^{n}M_{\omega_i}^2$ with $\pi_0\|M\|_{(F)}^2$. Both [7, 9] provided nearly optimal error bounds when the noise level is of no smaller order than the ℓ_∞ norm of the target matrix Θ, but not of smaller order, especially not for exact recovery. In a different approach, [6] proposed a non-convex recursive algorithm and provided error bounds in proportion to the noise level. However, the procedure requires the knowledge of the rank r of the unknown Θ and the error bound is optimal only when d_1 and d_2 are of the same order.

Our goal is to develop an algorithm for matrix completion that can be as easily computed as the spectrum Lasso (2) and enjoys a nearly optimal error bound proportional to the noise level to continuously cover both the noisy and noiseless cases. We propose to use an elastic penalty, a linear combination of the nuclear and Frobenius norms, which leads to the estimator

$$\widetilde{\Theta}=\arg\min_{M}\Big\{\sum_{i=1}^{n}M_{\omega_i}^2/2-\sum_{i=1}^{n}y_iM_{\omega_i}+\lambda_1\|M\|_{(N)}+(\lambda_2/2)\|M\|_{(F)}^2\Big\}, \tag{3}$$

where $\|\cdot\|_{(N)}$ and $\|\cdot\|_{(F)}$ are the nuclear and Frobenius norms, respectively. We call (3) spectrum elastic net (E-net) since it is parallel to the E-net in linear regression, the least squares estimator with a sum of the ℓ_1 and ℓ_2 penalties, introduced in [15]. Here the nuclear penalty provides the sparsity in the spectrum, while the Frobenius penalty regularizes the inversion of the quadratic term. Meanwhile, since the Frobenius penalty roughly shrinks the estimator by a factor $\pi_0/(\pi_0+\lambda_2)$, we correct this bias by a calibration step,

$$\widehat{\Theta}=(1+\lambda_2/\pi_0)\widetilde{\Theta}. \tag{4}$$

We call this estimator calibrated spectrum E-net.

Motivated by [8], we develop an EM algorithm to solve (3) for matrix completion. The algorithm iteratively replaces the missing entries with those obtained from a scaled soft-thresholding singular value decomposition (SVD) until the resulting matrix converges. This EM algorithm is guaranteed to converge to the solution of (3).

Under proper coherence conditions, we prove that for suitable penalty levels λ_1 and λ_2, the calibrated spectrum E-net (4) achieves a desired error bound in the Frobenius norm. Our error bound is of nearly optimal order and in proportion to the noise level. This provides a sharper result than those of [7, 9] when the noise level is of smaller order than the ℓ_∞ norm of Θ, and than that of [6] when d_2/d_1 is large. Our simulation results support the use of the calibrated spectrum E-net. They illustrate that (4) performs comparably to (2) and outperforms the modified method of [7].

Our analysis of the calibrated spectrum E-net uses an inequality similar to a duel certificate bound in [3]. The bound in [3] requires sample size $n\asymp\min\{(r\log d)^2,r(\log d)^6\}d\log d$, where $d=d_1+d_2$. We use the method of moments to remove a $\log d$ factor in the first component of their sample size requirement. This leads to a sample size requirement of $n\asymp r^2d\log d$, with an extra r in comparison to the ideal $n\asymp rd\log d$. Since the extra r does not appear in our error bound, its appearance in the sample size requirement seems to be a technicality.

The rest of the paper is organized as follows. In Section 2, we describe an iterative algorithm for the computation of the spectrum E-net and study its convergence. In Section 3, we derive error bounds for the calibrated spectrum E-net. Some simulation results are presented in Section 4. Section 5 provides the proof of our main result.

We use the following notation throughout this paper. For matrices $M\in\mathbb{R}^{d_1\times d_2}$, $\|M\|_{(N)}$ is the nuclear norm (the sum of all singular values of M), $\|M\|_{(S)}$ is the spectrum norm (the largest singular value), $\|M\|_{(F)}$ is the Frobenius norm (the ℓ_2 norm of vectorized M), and $\|M\|_\infty = \max_{jk}|M_{jk}|$. Linear mappings from $\mathbb{R}^{d_1\times d_2}$ to $\mathbb{R}^{d_1\times d_2}$ are denoted by the calligraphic letters. For a linear mapping $\mathcal{Q}$, the operator norm is $\|\mathcal{Q}\|_{(op)} = \sup_{\|M\|_{(F)}=1}\|\mathcal{Q}M\|_{(F)}$. We equip $\mathbb{R}^{d_1\times d_2}$ with the inner product $\langle M_1, M_2\rangle = \text{trace}(M_1^\top M_2)$ so that $\langle M, M\rangle = \|M\|^2_{(F)}$. For projections $\mathcal{P}$, $\mathcal{P}^\perp = \mathcal{I} - \mathcal{P}$ with $\mathcal{I}$ being the identity. We denote by E_ω the unit matrix with 1 at $\omega \in \{1,\dots,d_1\}\times\{1,\dots,d_2\}$, and by $\mathcal{P}_\omega$ the projection to E_ω: $M \to M_\omega E_\omega = \langle E_\omega, M\rangle E_\omega$.

2 An algorithm for spectrum elastic regularization

We first present a lemma for the M-step of our iterative algorithm.

Lemma 1 *Suppose the matrix Z has rank r. The solution to the optimization problem*

$$\arg\min_Z \Big\{\|Z - W\|^2_{(F)}/2 + \lambda_1\|Z\|_{(N)} + \lambda_2\|Z\|^2_{(F)}/2\Big\}$$

is given by $S(W;\lambda_1,\lambda_2) = UD_{\lambda_1,\lambda_2}V'$ with $D_{\lambda_1,\lambda_2} = diag\{(d_1-\lambda_1)_+,\dots,(d_r-\lambda_1)_+\}/(1+\lambda_2)$, where UDV' is the SVD of W, $D = diag\{d_1,\dots,d_r\}$ and $t_+ = max(t,0)$.

The minimization problem in Lemma 1 is solved by a scaled soft-thresholding SVD. This is parallel to Lemma 1 in [8] and justified by Remark 1 there. We use Lemma 1 to solve the M-step of the EM algorithm for the spectrum E-net (3).

We still need an E-step to impute a complete matrix given the observed data $\{y_i, \omega_i : i = 1,\dots,n\}$. Since ω_i are allowed to have ties, we need the following notation. Let $m_\omega = \#\{i : \omega_i = \omega, i \le n\}$ be the multiplicity of observations at $\omega \in \Omega^*$ and $m^* = \max_\omega m_\omega$ be the maximum multiplicity. Suppose that the complete data is composed of m_* observations at each ω for a certain integer m_*. Let $\overline{Y}^{(\text{com})}_\omega$ be the sample mean of the complete data at ω and $\overline{Y}^{(\text{com})}$ be the matrix with components $\overline{Y}^{(\text{com})}_\omega$. If the complete data are available, (3) is equivalent to

$$\arg\min_M \Big\{(m_*/2)\|\overline{Y}^{(\text{com})} - M\|^2_{(F)} + \lambda_1\|M\|_{(N)} + (\lambda_2/2)\|M\|^2_{(F)}\Big\}.$$

Let $\overline{Y}^{(\text{obs})}_\omega = m_\omega^{-1}\sum_{\omega_i=\omega} y_i$ be the sample mean of the observations at ω and $\overline{Y}^{(\text{obs})} = (\overline{Y}^{(\text{obs})}_\omega)_{d_1\times d_2}$. In the white noise model, the conditional expectation of $\overline{Y}^{(\text{com})}_\omega$ given $\overline{Y}^{(\text{obs})}$ is $(m_\omega/m_*)\overline{Y}^{(\text{obs})}_\omega + (1 - m_\omega/m_*)\Theta_\omega$ for $m_\omega \le m_*$. This leads to a generalized E-step:

$$\overline{Y}^{(\text{imp})} = (\overline{Y}^{(\text{imp})}_\omega)_{d_1\times d_2},\ \overline{Y}^{(\text{imp})}_\omega = \min\{1,(m_\omega/m_*)\}\overline{Y}^{(\text{obs})}_\omega + (1 - m_\omega/m_*)_+ Z^{(\text{old})}_\omega, \tag{5}$$

where $Z^{(\text{old})}$ is the estimation of Θ in the previous iteration. This is a genuine E-step when $m_* = m^*$ but also allows a smaller m_* to reduce the proportion of missing data.

We now present the EM-algorithm for the computation of the spectrum E-net $\widetilde{\Theta}$ in (3).

Algorithm 1 *Initialize with $Z^{(0)}$ and $k = 0$. Repeat the following steps:*

- *E-step: Compute $\overline{Y}^{(\text{imp})}$ in (5) with $Z^{(\text{old})} = Z^{(k)}$ and assign $k \leftarrow k + 1$,*
- *M-step: Compute $Z^{(k)} = S(\overline{Y}^{(\text{imp})}; \lambda_1/m_*, \lambda_2/m_*)$,*

until $\|Z^{(k)} - Z^{(k-1)}\|^2_{(F)}/\|Z^{(k)}\|^2_{(F)} \le \epsilon$. Then, return $Z^{(k)}$.

The following theorem states the convergence of Algorithm 1.

Theorem 1 *As $k \to \infty$, $Z^{(k)}$ converges to a limit $Z^{(\infty)}$ as a function of the data and $(\lambda_1, \lambda_2, m_*)$, and $Z^{(\infty)} = \widetilde{\Theta}$ for $m_* \ge m^*$.*

Theorem 1 is a variation of a parallel result in [8] and follows from the same proof there. As [8] pointed out, a main advantage of Algorithm 1 is the speed of each iteration. When the maximum multiplicity m^* is small, we simply use $Z^{(0)} = \overline{Y}^{(\text{obs})}$ and $m_* = m^*$; Otherwise, we may first run the EM-algorithm for an $m_* < m^*$ and use the output as the initialization $Z^{(0)}$ for a second run of the EM-algorithm with $m_* = m^*$.

3 Analysis of estimation accuracy

In this section, we derive error bounds for the calibrated spectrum E-net. We need the following notation. Let $r = \text{rank}(\Theta)$, $UDV^\top$ be the SVD of Θ, and $s_1 \geq \ldots \geq s_r$ be the nonzero singular values of Θ. Let T be the tangent space with respect to $UV^\top$, the space of all matrices of the form $UU^\top M_1 + M_2 VV^\top$. The orthogonal projection to T is given by

$$\mathcal{P}_T M = UU^\top M + MVV^\top - UU^\top MVV^\top. \tag{6}$$

Theorem 2 *Let $\xi = 1 + \lambda_2/\pi_0$ and $\mathcal{H} = \sum_{i=1}^n \mathcal{P}_{\omega_i}$. Define*

$$\begin{aligned} \mathcal{R} &= (\mathcal{H} - \pi_0)\mathcal{P}_T/(\pi_0 + \lambda_2), \\ \overline{\Delta} &= \mathcal{R}(\lambda_2\Theta + \lambda_1 UV^\top), \\ \mathcal{Q} &= \mathcal{I} - \mathcal{H}(\mathcal{P}_T\mathcal{H}\mathcal{P}_T + \lambda_2\mathcal{P}_T)^{-1}\mathcal{P}_T. \end{aligned}$$

Let $\varepsilon = \sum_{i=1}^n \varepsilon_i E_{\omega_i}$. Suppose

$$\|\mathcal{P}_T\mathcal{R}\|_{(op)} \leq 1/2, \quad s_r \geq 5\lambda_1/\lambda_2, \tag{7}$$

$$\|\mathcal{P}_T\overline{\Delta}\|_{(F)} \leq \sqrt{r}\lambda_1/8, \quad \big\|\overline{\Delta} - \mathcal{R}(\mathcal{P}_T\mathcal{R} + \mathcal{P}_T)^{-1}\mathcal{P}_T\overline{\Delta}\big\|_{(S)} \leq \lambda_1/4, \tag{8}$$

$$\|\mathcal{P}_T\varepsilon\|_{(F)} \leq \sqrt{r}\lambda_1/8, \quad \|\mathcal{Q}\varepsilon\|_{(S)} \leq 3\lambda_1/4, \quad \|\mathcal{P}_T^\perp\varepsilon\|_{(S)} \leq \lambda_1. \tag{9}$$

Then the calibrate spectrum E-net (4) satisfies

$$\|\widehat{\Theta} - \Theta\|_{(F)} \leq 2\sqrt{r}\lambda_1/\pi_0. \tag{10}$$

The proof of Theorem 2 is provided in Section 5. When ω_i are random entries in Ω^*, $E\mathcal{H} = \pi_0\mathcal{I}$, so that (8) and the first inequality of (7) are expected to hold under proper conditions. Since the rank of $\mathcal{P}_T\varepsilon$ is no greater than $2r$, (9) essentially requires $\|\varepsilon\|_{(S)} \asymp \lambda_1$. Our analysis allows λ_2 to lie in a certain range $[\lambda_*, \lambda^*]$, and λ^*/λ_* is large under proper conditions. Still, the choice of λ_2 is constrained by (7) and (8) since $\overline{\Delta}$ is linear in λ_2. When λ_2/π_0 diverges to infinity, the calibrated spectrum E-net (4) becomes the modified spectrum Lasso of [7].

Theorem 2 provides sufficient conditions on the target matrix and the noise for achieving a certain level of estimation error. Intuitively, these conditions on the target matrix Θ must imply a certain level of coherence (or flatness) of the unknown matrix since it is impossible to distinguish the unknown from zero when the observations are completely outside its support. In [2, 3, 4, 11], coherence conditions are imposed on

$$\mu_0 = \max\{(d_1/r)\|UU^\top\|_\infty, (d_2/r)\|VV^\top\|_\infty\}, \quad \mu_1 = \sqrt{d_1 d_2/r}\|UV^\top\|_\infty, \tag{11}$$

where U and V are matrices of singular vectors of Θ. [9] considered a more general notation of spikiness of a matrix M, defined as the ratio between the ℓ_∞ and dimension-normalized ℓ_2 norms,

$$\alpha_{sp}(M) = \|M\|_\infty\sqrt{d_1 d_2}/\|M\|_{(F)}. \tag{12}$$

Suppose in the rest of the section that ω_i are iid points uniformly distributed in Ω^* and ε_i are iid $N(0, \sigma^2)$ variables independent of $\{\omega_i\}$. The following theorem asserts that under certain coherence conditions on the matrices Θ, $UU^\top$, $VV^\top$ and $UV^\top$, all conditions of Theorem 2 hold with large probability when the sample size n is of the order $r^2 d\log d$.

Theorem 3 *Let $d = d_1 + d_2$. Consider λ_1 and λ_2 satisfying*

$$\lambda_1 = \sigma\sqrt{8\pi_0 d\log d}, \quad 1 \leq \frac{\lambda_2\|\Theta\|_{(F)}}{\lambda_1\{n/(d\log d)\}^{1/4}} \leq 2. \tag{13}$$

Then, there exists a constant C such that

$$n \geq C \max\Big\{\mu_0^2 r^2 d \log d, (\mu_1 + r)\mu_1 r d \log d, (\alpha_{sp}^{4/3} \vee \kappa_*^4) r^2 d \log d\Big\} \tag{14}$$

implies

$$\|\widehat{\Theta} - \Theta\|_{(F)}^2 / (d_1 d_2) \leq 32(\sigma^2 r d \log d)/n$$

with probability at least $1 - 1/d^2$, where μ_0 and μ_1 are the coherence constants in (11), $\alpha_{sp} = \alpha_{sp}(\Theta)$ is the spikiness of Θ and $\kappa_* = \|\Theta\|_{(F)}/(r^{1/2} s_r)$.

We require the knowledge of noise level σ to determine the penalty level that is usually considered as tuning parameter in practice. The Frobenius norm $\|\Theta\|_{(F)}$ in (13) can be replaced by an estimate of the same magnitude in Theorem 3. In our simulation experiment, we use $\lambda_2 = \lambda_1\{n/(d \log d)\}^{1/4}/\widehat{F}$ with $\widehat{F} = (\sum_{i=1}^n y_i^2/\pi_0)^{1/2}$. The Chebyshev inequality provides $\widehat{F}/\|\Theta\|_{(F)} \to 1$ when $\alpha_{sp} = O(1)$ and $\sigma^2 \ll \|\Theta\|_\infty^2$.

A key element in our analysis is to find a probabilistic bound for the second inequality of (8), or equivalently an upper bound for

$$P\big\{\|\mathcal{R}(\mathcal{P}_T\mathcal{R} + \mathcal{P}_T)^{-1}(\lambda_2\Theta + \lambda_1 UV^\top)\|_{(S)} > \lambda_1/4\big\}. \tag{15}$$

This guarantees the existence of a primal dual certificate for the spectrum E-net penalty [14]. For $\lambda_2 = 0$, a similar inequality was proved in [3], where the sample size requirement is $n \geq C_0 \min\{\mu^2 r^2 (\log d)^2 d, \mu^2 r (\log d)^6 d\}$ for a certain coherence factor μ. We remove a log factor in the first bound, resulting in the sample size requirement in (14), which is optimal when $r = O(1)$. For exact recovery in the noiseless case, the sample size $n \asymp rd(\log d)^2$ is sufficient if a golfing scheme is used to construct an approximate dual certificate [4, 11]. We use the following lemma to bound (15).

Lemma 2 *Let $\mathcal{H} = \sum_{i=1}^n \mathcal{P}_{\omega_i}$ where ω_i are iid points uniformly distributed in Ω^*. Let $\mathcal{R} = (\mathcal{H} - \pi_0)\mathcal{P}_T/(\pi_0 + \lambda_2)$ and $\xi = 1 + \lambda_2/\pi_0$. Let M be a deterministic matrix. Then, there exists a numerical constant C such that, for all $k \geq 1$ and $m \geq 1$,*

$$\xi^{2km} E\|\mathcal{R}^k M\|_{(S)}^{2m} \leq \Big\{C\mu_0^2 r^2 dkm/n\Big\}^{km}\Big(\mu_0^{-2}(\sqrt{d_1 d_2}/r)\|M\|_\infty\Big)^{2m}. \tag{16}$$

We use a different graphical approach than those in [3] to bound $E\,\mathrm{trace}(\{(\mathcal{R}^k M)^\top(\mathcal{R}^k M)\}^m)$ in the proof of Lemma 2. The rest of the proof of Theorem 3 can be outlined as follows. Assume that all coherence factors are $O(1)$. Let $M = \lambda_2\Theta + \lambda_1 UV^\top$ and write $\mathcal{R}(\mathcal{P}_T\mathcal{R} + \mathcal{P}_T)^{-1}M = \mathcal{R}M - \mathcal{R}^2 M + \cdots + (-1)^{k^*-1}\mathcal{R}^{k^*}M + \text{Rem}$. By (16) with $km \asymp \log d$ for $k \geq 2$ and an even simpler bound for $k = 1$ and Rem, (15) holds when $(\sqrt{d_1 d_2}/r)\|M\|_\infty \asymp \lambda_1\eta$, where $\eta \asymp r^2 d(\log d)/n$. Since $\alpha_{sp} + \mu_1 + \|\Theta\|_{(F)}^2/(r s_r^2) = O(1)$, this is equivalent to $\eta(s_r\lambda_2/\lambda_1 + 1) \lesssim 1$. Finally, we use matrix exponential inequalities [10, 12] to verify other conditions of Theorem 2. We omit technical details of the proof of Lemma 2 and Theorem 3. We would like to point out that if the r^2 in (16) can be replaced by $r(\log d)^\gamma$, e.g. $\gamma = 5$ in view of [3], the rest of the proof of Theorem 3 is intact with $\eta \asymp rd(\log d)^{1+\gamma}/n$ and a proper adjustment of λ_2 in (13).

Compared with [7] and [9], the main advantage of Theorem 3 is the proportionality of its error bound to the noise level. In [7], the quadratic term $\sum_{i=1}^n M_{\omega_i}^2$ in (2) is replaced by its expectation $\pi_0\|M\|_{(F)}^2$ and the resulting minimizer is proved to satisfy

$$\|\widehat{\Theta}^{(\mathrm{KLT})} - \Theta\|_{(F)}^2/(d_1 d_2) \leq C \max(\sigma^2, \|\Theta\|_\infty^2) rd(\log d)/n \tag{17}$$

with large probability, where C is a numerical constant. This error bound achieves the squared error rate $\sigma^2 rd(\log d)/n$ as in Theorem 3 when the noise level σ is of no smaller order than $\|\Theta\|_\infty$, but not of smaller order. In particular, (17) does not imply exact recovery when $\sigma = 0$. In Theorem 3, the error bound converges to zero as the noise level diminishes, implying exact recovery in the noiseless case. In [9], a constrained spectrum Lasso was proposed that minimizes (2) subject to $\|M\|_\infty \leq \alpha^*/\sqrt{d_1 d_2}$. For $\|\Theta\|_{(F)} \leq 1$ and $\alpha_{sp}(\Theta) \leq \alpha^*$, [9] proved

$$\|\widehat{\Theta}^{(\mathrm{NW})} - \Theta\|_{(F)}^2 \leq C \max(d_1 d_2 \sigma^2, 1)(\alpha^*)^2 rd(\log d)/n \tag{18}$$

with large probability. Scale change from the above error bound yields

$$\|\widehat{\Theta}^{(\mathrm{NW})} - \Theta\|_{(F)}^2/(d_1 d_2) \le C \max\{\sigma^2, \|\Theta\|_{(F)}^2/(d_1 d_2)\}(\alpha^*)^2 rd(\log d)/n.$$

Since $\alpha^* \ge 1$ and $\alpha^* \|\Theta\|_{(F)}/\sqrt{d_1 d_2} \ge \|\Theta\|_\infty$, the right-hand side of (18) is of no smaller order than that of (17). We shall point out that (17) and (18) only require sample size $n \asymp rd \log d$. In addition, [9] allows more practical weighted sampling models.

Compared with [6], the main advantage of Theorem 3 is the independence of its sample size requirement on the aspect ratio d_2/d_1, where $d_2 \ge d_1$ is assumed without loss of generality by symmetry. The error bound in [6] implies

$$\|\widehat{\Theta}^{(\mathrm{KMO})} - \Theta\|_{(F)}^2/(d_1 d_2) \le C_0 (s_1/s_r)^4 \sigma^2 rd(\log d)/n \tag{19}$$

for sample size $n \ge C_1^* rd \log d + C_2^* r^2 d \sqrt{d_2/d_1}$, where $\{C_1^*, C_2^*\}$ are constants depending on the same set of coherence factors as in (14) and $s_1 > \cdots > s_r$ are the singular values of Θ. Therefore, Theorem 3 effectively replaces the root aspect ratio $\sqrt{d_2/d_1}$ in the sample size requirement of (19) with a log factor, and removes the coherence factor $(s_1/s_r)^4$ on the right-hand side of (19). We note that s_1/s_r is a larger coherence factor than $\|\Theta\|_{(F)}/(r^{1/2}s_r)$ in the sample size requirement in Theorem 3. The root aspect ratio can be removed from the sample size requirement for (19) if Θ can be divided into square blocks uniformly satisfying the coherence conditions.

4 Simulation study

This experiment has the same setting as in Section 9 of [8]. We provide the description of the simulation settings in our notation as follows: The target matrix is $\Theta = UV^\top$, where $U_{d_1 \times r}$ and $V_{d_2 \times r}$ are random matrices with independent standard normal entries. The sampling points ω_i have no tie and $\Omega = \{\omega_i : i = 1, \ldots, n\}$ is a uniformly distributed random subset of $\{1, \ldots, d_1\} \times \{1, \ldots, d_2\}$, where n is fixed. The errors ε are iid $N(0, \sigma^2)$ variables. Thus, the observed matrix is $Y = \mathcal{P}_\Omega(\Theta + \varepsilon)$ with $\mathcal{P}_\Omega = \mathcal{H} = \sum_{i=1}^n \mathcal{P}_{\omega_i}$ being a projection. The signal to noise ratio (SNR) is defined as SNR $= \sqrt{r}/\sigma$.

We compare the calibrated spectrum E-net (4) with the spectrum Lasso (2) and its modification $\widehat{\Theta}^{(\mathrm{KLT})}$ of [7]. For all methods, we compute a series of estimators with 100 different penalty levels, where the smallest penalty level corresponds to a full-rank solution and the largest penalty level corresponds to a zero solution. For the calibrated spectrum E-net, we always use $\lambda_2 = \lambda_1 \{n/(d \log d)\}^{1/4}/\widehat{F}$, where $\widehat{F} = (\sum_{i=1}^n y_i^2/\pi_0)^{1/2}$ is an estimator for $\|\Theta\|_{(F)}$. We plot the training errors and test errors as functions of estimated ranks, where the training and test errors are defined as

$$\text{Training error} = \frac{\|\mathcal{P}_\Omega(\widehat{\Theta} - Y)\|_{(F)}^2}{\|\mathcal{P}_\Omega Y\|_{(F)}^2}, \quad \text{Test error} = \frac{\|\mathcal{P}_\Omega^\perp(\widehat{\Theta} - \Theta)\|_{(F)}^2}{\|\mathcal{P}_\Omega^\perp \Theta\|_{(F)}^2}.$$

In Figure 1, we report the estimation performance of three methods. The rank of Θ is 10 but $\{\Theta, \Omega, \varepsilon\}$ are regenerated in each replication. Different noise levels and proportions of the observed entries are considered. All the results are averaged over 50 replications. In this experiment, the calibrated spectrum E-net and the spectrum Lasso estimator have very close testing and training errors, and both of them significantly outperform the modified Lasso. Figure 1 also illustrates that in most cases, the calibrated spectrum E-net and spectrum Lasso achieve the optimal test error when the estimated rank is around the true rank.

We note that the constrained spectrum Lasso estimator $\widehat{\Theta}^{(\mathrm{NW})}$ would have the same performance as the spectrum Lasso when the constraint $\alpha_{sp}(\widehat{\Theta}) \le \alpha^*$ is set with a sufficiently high α^*. However, analytic properties of the spectrum Lasso is unclear without constraint or modification.

5 Proof of Theorem 2

The proof of Theorem 2 requires the following proposition that controls the approximation error of the Taylor expansion of the nuclear norm with subdifferentiation. The result, closely related to those

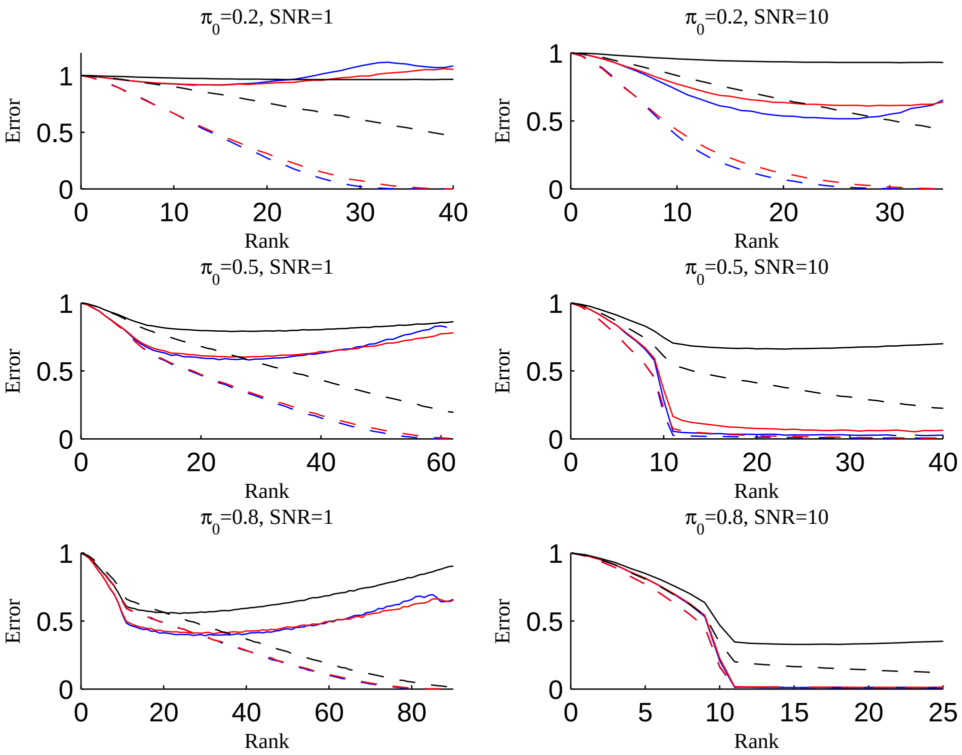

Figure 1: Plots of training and testing errors against the estimated rank: testing error with solid lines; training error with dashed lines; spectrum Lasso in blue, calibrated spectrum E-net in red; modified spectrum Lasso in black; $d_1 = d_2 = 100$, rank$(\Theta) = 10$.

in [13], is used to control the variation of the tangent space of the spectrum E-net estimator. We omit its proof.

Proposition 1 *Let $\Theta = UDV^\top$ be the SVD and M be another matrix. Then,*

$$\begin{aligned} 0 &\le \|M\|_{(N)} - \|\Theta\|_{(N)} - \|\mathcal{P}_T^\perp M\|_{(N)} - \langle UV^\top, M - \Theta\rangle \\ &\le \|(\mathcal{P}_T M - \Theta)VD^{-1/2}\|_{(F)}^2 + \|D^{-1/2}U^\top(\mathcal{P}_T M - \Theta)\|_{(F)}^2. \end{aligned}$$

Proof of Theorem 2. Define

$$\begin{aligned} &\Theta^* = (\mathcal{P}_T\mathcal{H}\mathcal{P}_T + \lambda_2\mathcal{P}_T)^{-1}(\mathcal{P}_T\varepsilon + \mathcal{P}_T\mathcal{H}\Theta - \lambda_1 UV^\top), \\ &\overline{\Theta} = (\pi_0 + \lambda_2)^{-1}(\pi_0\Theta - \lambda_1 UV^\top), \\ &\Delta = \widetilde{\Theta} - \Theta^*, \quad \Delta^* = \Theta^* - \overline{\Theta}, \quad \Delta_* = \widetilde{\Theta} - \overline{\Theta}. \end{aligned}$$

Since $\widehat{\Theta} = \xi\widetilde{\Theta}$ and $\xi\overline{\Theta} - \Theta = -(\lambda_1/\pi_0)UV^\top$,

$$\begin{aligned} \|\widehat{\Theta} - \Theta\|_{(F)} &\le \xi\|\Delta_*\|_{(F)} + \|\xi\overline{\Theta} - \Theta\|_{(F)} \\ &= \xi\|\Delta_*\|_{(F)} + \sqrt{r}\lambda_1/\pi_0 \qquad (20) \\ &\le \xi\|\Delta\|_{(F)} + \xi\|\Delta^*\|_{(F)} + \sqrt{r}\lambda_1/\pi_0. \qquad (21) \end{aligned}$$

We consider two cases by comparing λ_2 and π_0.

Case 1: $\lambda_2 \le \pi_0$. By algebra $\xi\Delta^* = \pi_0^{-1}(\mathcal{P}_T\mathcal{R} + \mathcal{P}_T)^{-1}\mathcal{P}_T(\varepsilon + \overline{\Delta})$, so that

$$\xi\|\Delta^*\|_{(F)} \le \pi_0^{-1}\|(\mathcal{P}_T\mathcal{R} + \mathcal{P}_T)^{-1}\|_{(op)}\|\mathcal{P}_T\overline{\Delta} + \mathcal{P}_T\varepsilon\|_{(F)} \le \sqrt{r}\lambda_1/(2\pi_0). \qquad (22)$$

The last inequality above follows from the first inequalities in (7), (8) and (9). It remains to bound $\|\Delta\|_{(F)}$. Let $Y = \sum_{i=1}^n y_i E_{\omega_i}$. We write the spectrum E-net estimator (3) as

$$\widetilde{\Theta} = \arg\min_M \Big\{\langle\mathcal{H}M, M\rangle/2 - \langle Y, M\rangle + \lambda_1\|M\|_{(N)} + (\lambda_2/2)\|M\|_{(F)}^2\Big\}.$$

It follows that for a certain member $\widehat{G}$ in the sub-differential of $\|M\|_{(N)}$ at $M = \widetilde{\Theta}$,

$$0 = \partial L_{\lambda_1,\lambda_2}(\widetilde{\Theta}) = \mathcal{H}\widetilde{\Theta} - Y + \lambda_2\widetilde{\Theta} + \lambda_1\widehat{G} = (\mathcal{H}+\lambda_2)\Delta + (\mathcal{H}+\lambda_2)\Theta^* - Y + \lambda_1\widehat{G}.$$

Let $\mathrm{Rem}_1 = \|\Theta^*\|_{(N)} - \langle UV^\top, \Theta^*\rangle$. Since $\|\Theta^*\|_{(N)} - \|\widetilde{\Theta}\|_{(N)} \ge -\langle \Delta, \widehat{G}\rangle$, we have

$$\begin{array}{rcl}
\langle(\mathcal{H}+\lambda_2)\Delta,\Delta\rangle & \le & \langle \mathcal{H}\Theta + \varepsilon - (\mathcal{H}+\lambda_2)\Theta^*, \Delta\rangle + \lambda_1\|\Theta^*\|_{(N)} - \lambda_1\|\widetilde{\Theta}\|_{(N)} \\
& = & \langle \mathcal{H}(\Theta-\Theta^*) + \varepsilon - \lambda_2\Theta^*, \Delta\rangle + \lambda_1\mathrm{Rem}_1 + \lambda_1\langle UV^\top, \Theta^*\rangle - \lambda_1\|\widetilde{\Theta}\|_{(N)} \\
& \le & \lambda_1\mathrm{Rem}_1 + \langle \varepsilon + \mathcal{H}(\Theta-\Theta^*) - \lambda_2\Theta^* - \lambda_1 UV^\top, \Delta\rangle - \lambda_1\|\mathcal{P}_T^\perp\Delta\|_{(N)} \\
& = & \lambda_1\mathrm{Rem}_1 + \langle \varepsilon + \mathcal{H}(\Theta-\Theta^*), \mathcal{P}_T^\perp\Delta\rangle - \lambda_1\|\mathcal{P}_T^\perp\Delta\|_{(N)}.
\end{array} \tag{23}$$

The second inequality in (23) is due to $\|\widetilde{\Theta}\|_{(N)} \ge \|\mathcal{P}_T^\perp\widetilde{\Theta}\|_{(N)} + \langle UV^\top, \widetilde{\Theta}\rangle$ and $\mathcal{P}_T^\perp\widetilde{\Theta} = \mathcal{P}_T^\perp\Delta$. The last equality in (23) follows from the definition of $\Theta^* \in T$, since it gives $\mathcal{P}_T\varepsilon + \mathcal{P}_T\mathcal{H}(\Theta-\Theta^*) - \lambda_2\Theta^* - \lambda_1 UV^\top = -(\mathcal{P}_T\mathcal{H}\mathcal{P}_T + \lambda_2\mathcal{P}_T)\Theta^* + \mathcal{P}_T\varepsilon + \mathcal{P}_T\mathcal{H}\Theta - \lambda_1 UV^\top = 0$. By the definitions of $\mathcal{Q}$, Θ^* and $\overline{\Delta}$, $\varepsilon + \mathcal{H}(\Theta-\Theta^*) = \mathcal{Q}\varepsilon + \mathcal{H}(\Theta-\overline{\Theta}) - \mathcal{H}(\mathcal{P}_T\mathcal{H}\mathcal{P}_T + \lambda_2\mathcal{P}_T)^{-1}\mathcal{P}_T\overline{\Delta}$. Since $\mathcal{P}_T^\perp\mathcal{H}\mathcal{P}_T = \mathcal{P}_T^\perp(\mathcal{H}-\pi_0)\mathcal{P}_T = \mathcal{P}_T^\perp\mathcal{R}(\pi_0+\lambda_2)$ and $(\mathcal{H}-\pi_0)(\Theta-\overline{\Theta}) = \overline{\Delta}$, we find

$$\begin{array}{rl}
& \langle \varepsilon + \mathcal{H}(\Theta-\Theta^*), \mathcal{P}_T^\perp\Delta\rangle \\
= & \langle \mathcal{Q}\varepsilon + (\mathcal{H}-\pi_0)\{\Theta - \overline{\Theta} - (\mathcal{P}_T\mathcal{H}\mathcal{P}_T + \lambda_2\mathcal{P}_T)^{-1}\mathcal{P}_T\overline{\Delta}\}, \mathcal{P}_T^\perp\Delta\rangle \\
= & \langle \mathcal{Q}\varepsilon + \overline{\Delta} - \mathcal{R}(\mathcal{P}_T\mathcal{R} + \mathcal{P}_T)^{-1}\mathcal{P}_T\overline{\Delta}, \mathcal{P}_T^\perp\Delta\rangle.
\end{array}$$

Thus, by the second inequalities of (8) and (9),

$$\langle \varepsilon + \mathcal{H}(\Theta-\Theta^*), \mathcal{P}_T^\perp\Delta\rangle \le \lambda_1\|\mathcal{P}_T^\perp\Delta\|_{(N)}. \tag{24}$$

Since $\Theta^* = \Delta^* - \overline{\Theta} \in T$ and the singular values of $\overline{\Theta}$ is no smaller than $(\pi_0 s_r - \lambda_1)/(\pi_0+\lambda_2) \ge (s_r - \lambda_1/\lambda_2)/\xi \ge 4\lambda_1/(\lambda_2\xi)$ by the second inequality in (7), Proposition 1 and (22) imply

$$\mathrm{Rem}_1 \le 2\|\Theta^* - \overline{\Theta}\|_{(F)}^2/\{(\pi_0 s_r - \lambda_1)/(\pi_0+\lambda_2)\} \le r(\lambda_1/\pi_0)^2/(8\xi\lambda_1/\lambda_2). \tag{25}$$

It follows from (23), (24) and (25) that

$$\xi^2\|\Delta\|_{(F)}^2 \le \xi^2\langle(\mathcal{H}+\lambda_2)\Delta,\Delta\rangle/\lambda_2 \le \xi^2(\lambda_1/\lambda_2)\mathrm{Rem}_1 \le r\lambda_1^2/(4\pi_0^2). \tag{26}$$

Therefore, the error bound (10) follows from (21), (22) and (26).

Case 2: $\lambda_2 \ge \pi_0$. By applying the derivation of (23) to $\overline{\Theta}$ instead of Θ^*, we find

$$\begin{array}{rl}
& \langle(\mathcal{H}+\lambda_2)\Delta_*,\Delta_*\rangle + \lambda_1\|\mathcal{P}_T^\perp\Delta_*\|_{(N)} \\
\le & \lambda_1\big(\|\overline{\Theta}\|_{(N)} - \langle UV^\top, \overline{\Theta}\rangle\big) + \langle \varepsilon + \mathcal{H}(\Theta-\overline{\Theta}) - \lambda_2\overline{\Theta} - \lambda_1 UV^\top, \Delta_*\rangle.
\end{array}$$

By the definitions of $\overline{\Delta}$, $\mathcal{R}$, and $\overline{\Theta}$, $\overline{\Delta} = (\mathcal{H}-\pi_0)(\Theta-\overline{\Theta}) = \mathcal{H}(\Theta-\overline{\Theta}) - \lambda_2\overline{\Theta} - \lambda_1 UV^\top$. This and $\|\overline{\Theta}\|_{(N)} = \langle UV^\top, \overline{\Theta}\rangle$ gives

$$\langle(\mathcal{H}+\lambda_2)\Delta_*,\Delta_*\rangle + \lambda_1\|\mathcal{P}_T^\perp\Delta_*\|_{(N)} \le \langle \varepsilon + \overline{\Delta}, \Delta_*\rangle. \tag{27}$$

Since $\|\mathcal{P}_T^\perp(\varepsilon + \overline{\Delta})\|_{(S)} = \|\mathcal{P}_T^\perp\varepsilon\|_{(S)} \le \lambda_1$ by the third inequality in (9), we have

$$\langle \mathcal{P}_T^\perp(\varepsilon + \overline{\Delta}), \Delta_*\rangle \le \lambda_1\|\mathcal{P}_T^\perp\Delta_*\|_{(N)}. \tag{28}$$

It follows from (27), (28) and the first inequalities of (8) and (9) that

$$\lambda_2\|\Delta_*\|_{(F)}^2 \le \langle \mathcal{P}_T(\varepsilon + \overline{\Delta}), \Delta_*\rangle \le \Big\{\|\mathcal{P}_T\varepsilon\|_{(F)} + \|\mathcal{P}_T\overline{\Delta}\|_{(F)}\Big\}\|\Delta_*\|_{(F)} \le \sqrt{r}\lambda_1\|\Delta_*\|_{(F)}/2.$$

Thus, due to $\lambda_2 \ge \pi_0$,

$$\xi\|\Delta_*\|_{(F)} \le (\xi/\lambda_2)\sqrt{r}\lambda_1/2 \le \sqrt{r}\lambda_1/\pi_0. \tag{29}$$

Therefore, the error bound (10) follows from (20) and (29). □

Acknowledgments

This research is partially supported by the NSF Grants DMS 0906420, DMS-11-06753 and DMS-12-09014, and NSA Grant H98230-11-1-0205.

References

[1] ACM SIGKDD and Netflix. Proceedings of KDD Cup and workshop. 2007.

[2] E. Candes and B. Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9:717–772, 2009.

[3] E. J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inform. Theory*, 56(5):2053–2080, 2009.

[4] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *CoRR*, abs/0910.1879, 2009.

[5] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

[6] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from noisy entries. *Journal of Machine Learning Research*, 11:2057–2078, 2010.

[7] V. Koltchinskii, K. Lounici, and A. B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39:2302–2329, 2011.

[8] R. Mazumder, T. Hastie, and R. Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11:2287–2322, 2010.

[9] S. Negahban and M. J. Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. 2010.

[10] R. I. Oliveira. Concentration of the adjacency matrix and of the laplacian in random graphs with independent edges. Technical Report arXiv:0911.0600, arXiv, 2010.

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

[12] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math. doi:10.1007/s10208-011-9099-z.*, 2011.

[13] P.-A. Wedin. Perturbation bounds in connection with singular value decomposition. *BIT*, 12:99–111, 1972.

[14] C.-H. Zhang and T. Zhang. A general framework of dual certificate analysis for structured sparse recovery problems. Technical report, arXiv: 1201.3302v1, 2012.

[15] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B*, 67:301–320, 2005.