Calibrated Elastic Regularization in Matrix Completion

Authored by:

Cun-hui Zhang
Tingni Sun

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 Paper Body

Let ? ? IRd1 ?d2 be a matrix of interest and ?? = {1, . . . , d1 } ? {1, . . . , d2 }. Suppose we observe vectors (?i , yi), yi = ??i + ?i ,

i = 1, . . . , n,

(1)

where ?i ? ?? and ?i are random errors. We are interested in estimating ? when n is a small fraction of d1 d2 . A well-known application of matrix completion is the Netflix problem where yi is the rating of movie bj by user ai for ? = (ai , bj) ? ?? [1]. In such applications, the proportion of the observed entries is typically very small, so that the estimation or recovery of ? is impossible without a structure assumption on ?. In this paper, we assume that ? 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. ?i = 0. A direct approach is to minimize

rank(M) subject to M?i = yi . 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]: n o b (CR) = arg min kM k(N) : M? = yi ? i ? n . ? i M

We denote the nuclear norm by k ? k(N) here and throughout this paper. This procedure, analyzed in [2, 3, 4, 11] among others, is parallel to the replacement of the ‘0 penalty by the ‘1 penalty in solving the sparse recovery problem in a linear space. 1

In this paper, we focus on the problem of matrix completion with noisy observations (1) and take the exact recovery as a special case. P Since the exact constraint is no longer appropriate in the presence n of noise, penalized squared error i=1 (M?i ? yi)2 is considered. By reformulating the problem in Lagrange form, [8] proposed the spectrum Lasso n n nX o X b (MHT) = arg min ? M?2i /2 ? yi M?i + ?kM k(N) , (2) M

i=1

i=1

along with an iterative convex minimization algorithm. However, (2) is difficult to analyze the Pwhen n sample fraction ?0 = n/(d1 d2) is small, due to the ill-posedness of the quadratic term i=1 M?2i . This has led to two alternatives in [7] and [9]. While [9] proposed to minimize Pn (2) under an additional ‘? constraint on M , [7] modified (2) by replacing the quadratic term i=1 M?2i with ?0 kM k2(F) . Both [7, 9] provided nearly optimal error bounds when the noise level is of no smaller order than the ‘? norm of the target matrix ?, 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 ? and the error bound is optimal only when d1 and d2 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 n n nX o X e = arg min ? M?2i /2 ? yi M?i + ?1 kM k(N) + (?2 /2)kM k2(F) , (3) M

i=1

i=1

where k ? k(N) and k ? k(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 ‘1 and ‘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 ?0 /(?0 + ?2), we correct this bias by a calibration step, b = (1 + ?2 /?0)?. e ?

(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 ?1 and ?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 '? norm of ?, and than that of [6] when d2 /d1 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 min{(r log d)2 , r(log d)6 }d log d, where d = d1 + d2 . 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 r2 d log d, with an extra r in comparison to the ideal n 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 ? Rd1 ?d2 , kM k(N) is the nuclear norm (the sum of all singular values of M), kM k(S) is the spectrum norm (the largest 2

singular value), kM k(F) is the Frobenius norm (the '2 norm of vectorized M), and kM k? = maxjk —Mjk —. Linear mappings from Rd1 ?d2 to Rd1 ?d2 are denoted by the calligraphic letters. For a linear mapping Q, the operator norm is kQk(op) = supkM k(F) =1 kQM k(F) . We equip Rd1 ?d2 with the inner product hM1 , M2 i = trace(M1¿ M2) so that hM, M i = kM k2(F) . For projections P, P ? = I ? P with I being the identity. We denote by E? the unit matrix with 1 at ? ? {1, . . . , d1 } ? {1, . . . , d2 }, and by P? the projection to E? : M ? M? E? = hE? , M iE? .

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 o n arg min kZ ? W k2(F) /2 + ?1 kZk(N) + ?2 kZk2(F) /2 Z

is given by S(W ; ?1 , ?2) = U D?1 ,?2 V 0 with D?1 ,?2 = diag{(d1 ??1)+ , . . . , (dr ??1)+ }/(1+?2), where U DV 0 is the SVD of W , D = diag{d1 , . . . , dr } 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 {yi , ?i : i = 1, . . . , n}. Since ?i are allowed to have ties, we need the following notation. Let m? = #{i : ?i = ?, i ? n} be the multiplicity of observations at ? ? ?? and m? = max? m? be the maximum multiplicity. Suppose that the complete data is composed of m? observations at each ? for a certain integer m? . (com) (com) Let Y ? be the sample mean of the complete data at ? and Y be the matrix with components (com)

Y?

. If the complete data are available, (3) is equivalent to n o (com) arg min (m? /2)kY ? M k2(F) + ?1 kM k(N) + (?2 /2)kM k2(F) . M

(obs)

Let Y ?

= m?1 ?

(obs) (Y ?)d1 ?d2 . In (obs) (m? /m?)Y ? +

Y

(imp)

?i =?

yi be the sample mean of the observations at ? and Y

the white noise model, the conditional expectation of Y

(com) ?

(obs)

given Y

=

(obs)

is

(1 ? m? /m?)?? for m? ? m? . This leads to a generalized E-step:

(imp)

= (Y ?

P

(imp)

)d1 ?d2 , Y ?

(obs)

= min{1, (m? /m?)}Y ?

+ (1 ? m? /m?)+ Z?(old) ,

(5)

where Z (old) is the estimation of ? 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. e in (3). We now present the EM-algorithm for the computation of the spectrum E-net ? Algorithm 1 Initialize with Z (0) and k = 0. Repeat the following steps: ? E-step: Compute Y

(imp)

in (5) with Z (old) = Z (k) and assign k ? k + 1,

? M-step: Compute Z (k) = S(Y

(imp)

; ?1 /m? , ?2 /m?),

until kZ (k) ? Z (k?1) k2(F) /kZ (k) k2(F) ? . Then, return Z (k) . The following theorem states the convergence of Algorithm 1. Theorem 1 As k ? ?, Z (k) converges to a limit Z (?) as a function of the data and (?1 , ?2 , m?), e for m? ? m? . and Z (?) = ? 3

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 (obs) multiplicity m? is small, we simply use Z (0) = Y 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 = rank(?), U DV ¿ be the SVD of ?, and s1 ? . . . ? sr be the nonzero singular values of ?. Let T be the tangent space with respect to U V ¿ , the space of all matrices of the form U U ¿ M1 + M2 V V ¿ . The orthogonal projection to T is given by PT M = U U ¿ M + M V V ¿ ? U U ¿ M V V ¿ . Pn Theorem 2 Let ? = 1 + ?2 /?0 and H = i=1 P?i . Define R

=

(H ? ?0)PT /(?0 + ?2),

?

=

R(?2 ? + ?1 U V ¿),

(6)

Q = I ? H(PT HPT + ?2 PT)?1 PT . Let ? =

Pn

i=1 ?i E?i .

Suppose

kPT Rk(op) ? 1/2, sr ? 5?1 /?2 ,

? kPT ?k(F) ? r?1 /8, k? ? R(PT R + PT)?1 PT ? (S) ? ?1 /4, ? kPT ?k(F) ? r?1 /8, kQ?k(S) ? 3?1 /4, kPT? ?k(S) ? ?1 .

(7) (8) (9)

Then the calibrate spectrum E-net (4) satisfies ? b ? ?k(F) ? 2 r?1 /?0 . k?

(10)

The proof of Theorem 2 is provided in Section 5. When ?i are random entries in ?? , EH = ?0 I, so that (8) and the first inequality of (7) are expected to hold under proper conditions. Since the rank of PT ? is no greater than 2r, (9) essentially requires k?k(S) ?1 . Our analysis allows ?2 to lie in a certain range [?? , ??], and ?? /?? is large under proper conditions. Still, the choice of ?2 is constrained by (7) and (8) since ? is linear in ?2 . When ?2 /?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 ? 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 p ?0 = max{(d1 /r)kU U ¿ k? , (d2 /r)kV V ¿ k? }, ?1 = d1 d2 /rkU V ¿ k? , (11) where U and V are matrices of singular vectors of ?. [9] considered a more general notation of spikiness of a matrix M , defined as the ratio between the '? and dimension-normalized '2 norms, p ?sp (M) = kM k? d1 d2 /kM k(F) . (12) Suppose in the rest of the section that ?i are iid points uniformly distributed in ?? and ?i are iid N (0, ? 2) variables independent of {?i }. The following theorem asserts that under certain coherence conditions on the matrices ?, U U ¿ , V V ¿ and U V ¿ , all conditions of Theorem 2 hold with large probability when the sample size n is of the order r2 d log d. Theorem 3 Let d = d1 + d2 . Consider ?1 and ?2 satisfying ?1 = ?

p
8?0 d log d,
1?
?2 k?k(F) ? 2. ?1 {n/(d log d)}1/4 4
(13)

Then, there exists a constant C such that n o 4/3 n ? C max ?20 r2 d log d, (?1 + r)?1 rd log d, (?sp ? ?4?)r2 d log d
(14)

implies b ? ?k2 /(d1 d2) ? 32(? 2 rd log d)/n k? (F) with probability at least 1 ? 1/d2 , where ?0 and ?1 are the coherence constants in (11), ?sp = ?sp (?) is the spikiness of ? and ?? = k?k(F) /(r1/2 sr). We require the knowledge of noise level ? to determine the penalty level that is usually considered as tuning parameter in practice. The Frobenius norm k?k(F) in (13) can be replaced by an estimate of the same magnitude in Theorem 3. In our simulation experiment, we use Pn ?2 = ?1 {n/(d log d)}1/4 /Fb with Fb = (i=1 yi2 /?0)1/2 . The Chebyshev inequality provides Fb/k?k(F) ? 1 when ?sp = O(1) and ? 2 k?k2? . 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 kR(PT R + PT)?1 (?2 ? + ?1 U V ¿)k(S) ¿ ?1 /4 . (15) This guarantees the existence of a primal dual certificate for the spectrum E-net penalty [14]. For ?2 = 0, a similar inequality was proved in [3], where the sample size requirement is n ? C0 min{?2 r2 (log d)2 d, ?2 r(log d)6 d} for a certain coherence factor ?. 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 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). Pn ? Lemma 2 Let H = i=1 P?i where ?i are iid points uniformly distributed in ? . Let R = (H ? ?0)PT /(?0 + ?2) and ? = 1 + ?2 /?0 . Let M be a deterministic matrix. Then, there exists a numerical constant C such that, for all k ? 1 and m ? 1, n okm 2m p ?2 2 2 ? 2km EkRk M k2m ? C? r dkm/n ? (d d /r)kM k . (16) 1 2 ? 0 0 (S) We use a different graphical approach than those in [3] to bound E trace({(Rk M)¿ (Rk 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 = ?2 ? + ?1 U

V ¿ and write R(PT R + PT)?1 M = ? ? RM ?R2 M +? ? ?+(?1)k ?1 Rk M +Rem. By?(16) with km log d for k ? 2 and an even simpler bound for k = 1 and Rem, (15) holds when (d1 d2 /r)kM k? ?1 ?, where ? r2 d(log d)/n. Since ?sp + ?1 + k?k2(F) /(rs2r) = O(1), this is equivalent to ?(sr ?2 /?1 + 1) . 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 r2 in (16) can be replaced by r(log d)? , e.g. ? = 5 in view of [3], the rest of the proof of Theorem 3 is intact with ? rd(log d)1+? /n and a proper adjustment of ?2 in (13). Compared with [7] and [9], the main advantage ofPTheorem 3 is the proportionality of its error n bound to the noise level. In [7], the quadratic term i=1 M?2i in (2) is replaced by its expectation 2 ?0 kM k(F) and the resulting minimizer is proved to satisfy b (KLT) ? ?k2 /(d1 d2) ? C max(? 2 , k?k2?)rd(log d)/n k? (F)

(17)

with large probability, where C is a numerical constant. This error bound achieves the squared error rate ? 2 rd(log d)/n as in Theorem 3 when the noise level ? is of no smaller order than k?k? , but not of smaller order. In particular, (17) does not imply exact recovery when ? = 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 kM k? ? ?? / d1 d2 . For k?k(F) ? 1 and ?sp (?) ? ?? , [9] proved b (NW) ? ?k2 ? C max(d1 d2 ? 2 , 1)(??)2 rd(log d)/n k? (F) 5

(18)

with large probability. Scale change from the above error bound yields b (NW) ? ?k2 /(d1 d2) ? C max{? 2 , k?k2 /(d1 d2)}(??)2 rd(log d)/n. k? (F) (F) ? ? ? Since ? ? 1 and ? k?k(F) / d1 d2 ? k?k? , 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 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 d2 /d1 , where d2 ? d1 is assumed without loss of generality by symmetry. The error bound in [6] implies b (KMO) ? ?k2 /(d1 d2) ? C0 (s1 /sr)4 ? 2 rd(log d)/n k? (19) (F) p for sample size n ? C1? rd log d + C2? r2 d d2 /d1 , where {C1? , C2? } are constants depending on the same set of coherence factors as in (14) and s1 ¿ ? ?p ? ¿ sr are the singular values of ?. Therefore, Theorem 3 effectively replaces the root aspect ratio d2 /d1 in the sample size requirement of (19) with a log factor, and removes the coherence factor (s1 /sr)4 on the right-hand side of (19). We note that s1 /sr is a larger coherence factor than k?k(F) /(r1/2 sr) in the sample size requirement in Theorem 3. The root aspect ratio can be removed from the sample size requirement for (19) if ? 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 ? = U V ¿ , where Ud1 ?r and Vd2 ?r are random matrices with

independent standard normal entries. The sampling points ?i have no tie and ? = {?i : i = 1, . . . , n} is a uniformly distributed random subset of {1, . . . , d1 } ? {1, . . . , d2 }, where n is fixed. The P errors ? are iid N (0, ? 2) variables. Thus, the observed matrix is n Y = P? (? + ?) with ? P? = H = i=1 P?i being a projection. The signal to noise ratio (SNR) is defined as SNR = r/?. We compare the calibrated spectrum E-net (4) with the spectrum Lasso (2) and its modification b (KLT) of [7]. For all methods, we compute a series of estimators with 100 different penalty lev? els, 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 ?2 = Pn ?1 {n/(d log d)}1/4 /Fb, where Fb = (i=1 yi2 /?0)1/2 is an estimator for k?k(F) . We plot the training errors and test errors as functions of estimated ranks, where the training and test errors are defined as b ? ?)k2 b ? Y)k2 kP?? (? kP? (? (F) (F) . , Test error = Training error = ? 2 kP? Y k2(F) kP? ?k(F) In Figure 1, we report the estimation performance of three methods. The rank of ? is 10 but {?, ?, ?} 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. b (NW) would have the same performance as We note that the constrained spectrum Lasso estimator ? b ? ?? is set with a sufficiently high ?? . However, the spectrum Lasso when the constraint ?sp (?) 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 6

?0=0.2, SNR=1

?0=0.2, SNR=10

1 Error

Error

1 0.5 0

0

10

20

30

0.5 0

40

0

10

Rank ?0=0.5, SNR=1

0.5 0

20
40
0.5 0
60
0
10
Rank ?0=0.8, SNR=1
30
40
1 Error
Error
20 Rank ?0=0.8, SNR=10
1 0.5 0
30
1 Error
Error
1
0
20
Rank ?0=0.5, SNR=10
0
20
40
60
0.5 0
80
Rank
0
5
10
15
20
25
Rank

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; d1 = d2 = 100, rank(?) = 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 ? = U DV ¿ be the SVD and M be another matrix. Then, 0

? kM k(N) ? k?k(N) ? kPT? M k(N) ? hU V ¿ , M ? ?i ? k(PT M ? ?)V D?1/2 k2(F) + kD?1/2 U ¿ (PT M ? ?)k2(F) .

Proof of Theorem 2. Define ?? = (PT HPT + ?2 PT)?1 (PT ? + PT H? ? ?1 U V ¿), ? = (?0 + ?2)?1 (?0 ? ? ?1 U V ¿), e ? ?? , ?? = ?? ? ?, ?? = ? e ? ?. ?=? b = ?? e and ?? ? ? = ?(?1 /?0)U V ¿ , Since ? b ? ?k(F) k?

? ?k?? k(F) + k?? ? ?k(F) ? = ?k?? k(F) + r?1 /?0 ? ? ?k?k(F) + ?k?? k(F) + r?1 /?0 .

(20) (21)

We consider two cases by comparing ?2 and ?0 . Case 1: ?2 ? ?0 . By algebra ??? = ?0?1 (PT R + PT)?1 PT (? + ?), so that ? ?k?? k(F) ? ?0?1 k(PT R + PT)?1 k(op) kPT ? + PT ?k(F) ? r?1 /(2?0).

(22)

The last inequalityP above follows from the first inequalities in (7), (8) and (9). It remains to bound n k?k(F) . Let Y = i=1 yi E?i . We write the spectrum E-net estimator (3) as n o e = arg min hHM, M i/2 ? hY, M i + ?1 kM k(N) + (?2 /2)kM k2 ? (F) . M

7

b in the sub-differential of kM k(N) at M = ?, e It follows that for a certain member G e = H? e ? Y + ?2 ? e + ?1 G b = (H + ?2)? + (H + ?2)?? ? Y + ?1 G. b 0 = ?L?1 ,?2 (?) e (N) ? ?h?, Gi, b we have Let Rem1 = k?? k(N) ? hU V ¿ , ?? i. Since k?? k(N) ? k?k h(H + ?2)?, ?i

e (N) ? hH? + ? ? (H + ?2)?? , ?i + ?1 k?? k(N) ? ?1 k?k ? ? e (N) = hH(? ? ?) + ? ? ?2 ? , ?i + ?1 Rem1 + ?1 hU V ¿ , ?? i ? ?1 k?k ? ? ¿ ? ? ?1 Rem1 + h? + H(? ? ?) ? ?2 ? ? ?1 U V , ?i ? ?1 kPT ?k(N) = ?1 Rem1 + h? + H(? ? ??), PT? ?i ? ?1 kPT? ?k(N) . (23)

e (N) ? kP ? ?k e (N) + hU V ¿ , ?i e and P ? ? e = P ? ?. The The second inequality in (23) is due to k?k T T T ? last equality in (23) follows from the definition of ? ? T , since it gives PT ? + PT H(? ? ??) ? ?2 ?? ? ?1 U V ¿ = ?(PT HPT + ?2 PT)?? + PT ? + PT H? ? ?1 U V ¿ = 0. By the definitions of Q, ?? and ?, ? + H(? ? ??) = Q? + H(? ? ?) ? H(PT HPT + ?2 PT)?1 PT ?. Since PT? HPT = PT? (H ? ?0)PT = PT? R(?0 + ?2) and (H ? ?0)(? ? ?) = ?, we find h? + H(? ? ??), PT? ?i = hQ? + (H ? ?0){? ? ? ? (PT HPT + ?2 PT)?1 PT ?}, PT? ?i = hQ? + ? ? R(PT R + PT)?1 PT ?, PT? ?i. Thus, by the second inequalities of (8) and (9), h? + H(? ? ??), PT? ?i ? ?1 kPT? ?k(N) .

(24)

Since ?? = ?? ? ? ? T and the singular values of ? is no smaller than (?0 sr ? ?1)/(?0 + ?2) ? (sr ? ?1 /?2)/? ? 4?1 /(?2 ?) by the second inequality in (7), Proposition 1 and (22) imply Rem1 ? 2k?? ? ?k2(F) /{(?0 sr ? ?1)/(?0 + ?2)} ? r(?1 /?0)2 /(8??1 /?2).

(25)

It follows from (23), (24) and (25) that ? 2 k?k2(F) ? ? 2 h(H + ?2)?, ?i/?2 ? ? 2 (?1 /?2)Rem1 ? r?21 /(4?02).

(26)

Therefore, the error bound (10) follows from (21), (22) and (26). Case 2: ?2 ? ?0 . By applying the derivation of (23) to ? instead of ?? , we find h(H + ?2)?? , ?? i + ?1 kPT? ?? k(N)

? ?1 k?k(N) ? hU V ¿ , ?i + h? + H(? ? ?) ? ?2 ? ? ?1 U V ¿ , ?? i. By the definitions of ?, R, and ?, ? = (H ? ?0)(? ? ?) = H(? ? ?) ? ?2 ? ? ?1 U V ¿ . This and k?k(N) = hU V ¿ , ?i gives h(H + ?2)?? , ?? i + ?1 kPT? ?? k(N) ? h? + ?, ?? i.

(27)

Since kPT? (? + ?)k(S) = kPT? ?k(S) ? ?1 by the third inequality in (9), we have hPT? (? + ?), ?? i ? ?1 kPT? ?? k(N) .

(28)

It follows from (27), (28) and the first inequalities of (8) and (9) that n o ? ?2 k?? k2(F) ? hPT (? + ?), ?? i ? kPT ?k(F) + kPT ?k(F) k?? k(F) ? r?1 k?? k(F) /2. Thus, due to ?2 ? ?0 ,

? ? ?k?? k(F) ? (?/?2) r?1 /2 ? r?1 /?0 .

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

(29)

Acknowledgments This research is partially supported by the NSF Grants DMS 0906420, DMS-11-06753 and DMS12-09014, and NSA Grant H98230-11-1-0205. 8

2 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. Cand‘es 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. abs/0910.1879, 2009.

CoRR,

[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.

9