High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity

Authored by:

Martin J. Wainwright
Po-ling Loh

Abstract

Although the standard formulations of prediction problems involve fully-observed and noiseless data drawn in an i.i.d. manner, many applications involve noisy and/or missing data, possibly involving dependencies. We study these issues in the context of high-dimensional sparse linear regression, and propose novel estimators for the cases of noisy, missing, and/or dependent data. Many standard approaches to noisy or missing data, such as those using the EM algorithm, lead to optimization problems that are inherently non-convex, and it is difficult to establish theoretical guarantees on practical algorithms. While our approach also involves optimizing non-convex programs, we are able to both analyze the statistical error associated with any global optimum, and prove that a simple projected gradient descent algorithm will converge in polynomial time to a small neighborhood of the set of global minimizers. On the statistical side, we provide non-asymptotic bounds that hold with high probability for the cases of noisy, missing, and/or dependent data. On the computational side, we prove that under the same types of conditions required for statistical consistency, the projected gradient descent algorithm will converge at geometric rates to a near-global minimizer. We illustrate these theoretical predictions with simulations, showing agreement with the predicted scalings.

1 Paper Body

In standard formulations of prediction problems, it is assumed that the covariates are fully-observed and sampled independently from some underlying distribution. However, these assumptions are not realistic for many applications, in which covariates may be observed only partially, observed with corruption, or exhibit dependencies. Consider the problem of modeling the voting behavior of politicians: in this setting, votes may be missing due to abstentions, and temporally dependent due to collusion or ?tit-for-tat? behavior. Similarly, surveys

often suffer from the missing data problem, since users fail to respond to all questions. Sensor network data also tends to be both noisy due to measurement error, and partially missing due to failures or drop-outs of sensors. There are a variety of methods for dealing with noisy and/or missing data, including various heuristic methods, as well as likelihood-based methods involving the expectation-maximization (EM) algorithm (e.g., see the book [1] and references therein). A challenge in this context is the possible non-convexity of associated optimization problems. For instance, in applications of EM, problems in which the negative likelihood is a convex function often become non-convex with missing or noisy data. Consequently, although the EM algorithm will converge to a local minimum, it is difficult to guarantee that the local optimum is close to a global minimum. In this paper, we study these issues in the context of high-dimensional sparse linear regression, in the case when the predictors or covariates are noisy, missing, and/or dependent. Our main contribution is to develop and study some simple methods for handling these issues, and to prove theoretical results about both the associated statistical error and the optimization error. Like EM-based approaches, our estimators are based on solving optimization problems that may be non-convex; however, despite this non-convexity, we are still able to prove that a simple form of projected gradient descent will produce an output that is ?sufficiently close??meaning as 1

small as the statistical error?to any global optimum. As a second result, we bound the size of this statistical error, showing that it has the same scaling as the minimax rates for the classical cases of perfectly observed and independently sampled covariates. In this way, we obtain estimators for noisy, missing, and/or dependent data with guarantees similar to the usual fully-observed and independent case. The resulting estimators allow us to solve the problem of high-dimensional Gaussian graphical model selection with missing data. There is a large body of work on the problem of corrupted covariates or errors-in-variables for regression problems (see the papers and books [2, 3, 4, 5] and references therein). Much of the earlier theoretical work is classical in nature, where the sample size n diverges with the dimension p held fixed. Most relevant to this paper is more recent work that has examined issues of corrupted and/or missing data in the context of highdimensional sparse linear models, allowing for n p. St?adler and B?uhlmann [6] developed an EM-based method for sparse inverse covariance matrix estimation in the missing data regime, and used this result to derive an algorithm for sparse linear regression with missing data. As mentioned above, however, it is difficult to guarantee that EM will converge to a point close to a global optimum of the likelihood, in contrast to the methods studied here. Rosenbaum and Tsybakov [7] studied the sparse linear model when the covariates are corrupted by noise, and proposed a modified form of the Dantzig selector, involving a convex program. This convexity produces a computationally attractive method, but the statistical error bounds that they establish scale proportionally with the size of the additive perturbation, hence are often weaker than the bounds that can be proved using our methods. The remainder of this paper is organized as follows. We begin in Section 2 with background and a precise description of the problem. We then introduce the class

of estimators we will consider and the form of the projected gradient descent algorithm. Section 3 is devoted to a description of our main results, including a pair of general theorems on the statistical and optimization error, and then a series of corollaries applying our results to the cases of noisy, missing, and dependent data. In Section 4, we demonstrate simulations to confirm that our methods work in practice. For detailed proofs, we refer the reader to the technical report [8]. Notation. For a matrix M , we write kM kmax := maxi,j —mij — to be the elementwise ‘? -norm of M . Furthermore, ———M ———1 denotes the induced ‘1 -operator norm (maximum absolute column sum) of M , and ———M ———op is the (M) , the condition number of M . induced ‘2 -operator norm (spectral norm) of M . We write ?(M) := ??max min (M)

2

Background and problem set-up

In this section, we provide a formal description of the problem and motivate the class of estimators studied in the paper. We then describe a class of projected gradient descent algorithms to be used in the sequel. 2.1

Observation model and high-dimensional framework

Suppose we observe a response variable yi ? R that is linked to a covariate vector xi ? Rp via the linear model yi = hxi , ? ? i + i , ?

for i = 1, 2, . . . , n.

(1)

p

Here, the regression vector ? ? R is unknown, and i ? R is observation noise, independent of xi . Rather than directly observing each xi ? Rp , we observe a vector zi ? Rp linked to xi via some conditional distribution: zi ? Q(? — xi),

for i = 1, 2, . . . , n.

(2)

This setup allows us to model various types of disturbances to the covariates, including (a) Additive noise: We observe zi = xi + wi , where wi ? Rp is a random vector independent of xi , say zero-mean with known covariance matrix ?w . (b) Missing data: For a fraction ? ? [0, 1), we observe a random vector zi ? Rp such that independently for each component j, we observe zij = xij with probability 1 ? ?, and zij = ? with probability ?. This model can also be generalized to allow for different missing probabilities for each covariate. Our first set of results is deterministic, depending on specific instantiations of the observed variables {(yi , zi)}ni=1 . However, we are also interested in proving results that hold with high probability when the xi ?s and zi ?s are drawn at random from some distribution. We develop results for both the i.i.d. setting and the case of dependent covariates, where the xi ?s are generated according to a stationary vector autoregressive (VAR) process. Furthermore, we work within a high-dimensional framework where n p, and assume ? ? has at most k non-zero parameters, where the sparsity k is also allowed to increase to infinity with the sample size n. We assume the scaling k? ? k2 = O(1), which is reasonable in order to have a non-diverging signal-to-noise ratio. 2

2.2

M -estimators for noisy and missing covariates

We begin by examining a simple deterministic problem. Let Cov(X) = ?x 0, and consider the program

1 T ? ?x ? ? h?x ? ? , ?i . (3) ?b ? arg min k?k1 ?R 2 As long as the constraint radius R is at least k? ? k1 , the unique solution to this convex program is ?b = ? ? . This idealization suggests various estimators based on the plug-in principle. We form unbiased estimates of ?x and b and ? ?x ? ? , denoted by ? b, respectively, and consider the modified program and its regularized version: 1 T

b ? hb ?b ? arg min ? , ?i , (4) ? ?? k?k1 ?R 2

1 T b ? hb ? , ?i + ?n k?k1 , (5) ?b ? arg minp ? ?? ??R 2 where ?n ¿ 0 is the regularization parameter. The Lasso [9, 10] is a special case of these programs, where 1 b Las := 1 X T X and ? bLas := X T y, (6) ? n n and we have introduced the shorthand y = (y1 , . . . , yn)T ? Rn , and X ? Rn?p , with xTi as its ith row. In this paper, we focus on more general instantiations of the programs (4) and (5), involving different choices of b ? b Las is positive the pair (?, b) that are adapted to the cases of noisy and/or missing data. Note that the matrix ? semidefinite, so the Lasso program is convex. In sharp contrast, for the cases of noisy or missing data, the most b is not positive semidefinite, hence the loss functions appearing in the problems (4) natural choice of the matrix ? and (5) are non-convex. It is generally impossible to provide a polynomial-time algorithm that converges to a (near) global optimum of a non-convex problem. Remarkably, we prove that a simple projected gradient descent algorithm still converges with high probability to a vector close to any global optimum in our setting. Let us illustrate these ideas with some examples: Example 1 (Additive noise). Suppose we observe the n ? p matrix Z = X + W , where W is a random matrix independent of X, with rows wi drawn i.i.d. from a zero-mean distribution with known covariance ?w . Consider the pair 1 b add := 1 Z T Z ? ?w and ? ? badd := Z T y, (7) n n ? which correspond to unbiased estimators of ?x and ?x ? , respectively. Note that when ?w = 0 (corresponding b add is to the noiseless case), the estimators reduce to the standard Lasso. However, when ?w 6= 0, the matrix ? not positive semidefinite in the high-dimensional regime (n p) of interest. Indeed, since the matrix n1 Z T Z b add to have a large number of negative eigenvalues. has rank at most n, the subtracted matrix ?w may cause ? Example 2 (Missing data). Suppose each entry of X is missing independently with probability ? ? [0, 1), and we observe the matrix Z ? Rn?p with entries

Xij with probability 1 ? ?, Zij = 0 otherwise. Given the observed matrix Z ? Rn?p , consider an estimator of the general form (4), based on the choices eT e eT e 1 eT b mis := Z Z ? ? diag Z Z ? and ? bmis := Z y, (8) n n n eij = Zij /(1 ? ?). It is easy to see that the pair (? b mis , ? b Las , ? where Z bmis) reduces to the pair (? bLas) for the standard Lasso when ? = 0, corresponding to no missing data. In the more interesting case when ? ? (0, 1), the matrix eT Z e Z b n in equation (8) has rank at most n, so the subtracted diagonal matrix may cause the matrix ?mis to have a large number of negative eigenvalues when n p, and the associated quadratic function is not convex. 2.3

Restricted eigenvalue conditions

b there are various ways to assess its closeness to ? ? . We focus on the ‘2 -norm k?b ?? ? k2 , as Given an estimate ?, well as the closely related ‘1 -norm k?b ? ? ? k1 . When the covariate matrix X is fully observed (so that the Lasso b Las = 1 X T X can be applied), it is well understood that a sufficient condition for ‘2 -recovery is that the matrix ? n satisfy a restricted eigenvalue (RE) condition (e.g., [11, 12, 13]). In this paper, we use the following condition: 3

b satisfies a lower restricted eigenvalue condition with curvaDefinition 1 (Lower-RE condition). The matrix ? ture ¿ ¿ 0 and tolerance ¿ (n, p) ¿ 0 if b ? ¿ k?k22 ? ¿ (n, p)k?k21 ?T ??

for all ? ? Rp .

(9)

b Las = 1 X T X satisfies this RE condition (9), the Lasso estimate It can be shown that when the Lasso matrix ? n 1 has low ‘2 -error for any vector ? ? supported on any subset of size at most k . ¿ (n,p) . Moreover, it is known b Las will satisfy such an RE condition that for various random choices of the design matrix X, the Lasso matrix ? with high probability (e.g., [14]). We also make use of the analogous upper restricted eigenvalue condition: b satisfies an upper restricted eigenvalue condition with Definition 2 (Upper-RE condition). The matrix ? smoothness ?u ¿ 0 and tolerance ?u (n, p) ¿ 0 if b ? ?u k?k22 + ?u (n, p)k?k21 ?T ??

for all ? ? Rp .

(10)

In recent work on high-dimensional projected gradient descent, Agarwal et al. [15] use a more general form of bounds (9) and (10), called the restricted strong convexity (RSC) and restricted smoothness (RSM) conditions. 2.4

Projected gradient descent

In addition to proving results about the global minima of programs (4) and (5), we are also interested in polynomial-time procedures for approximating such optima. We show that the simple projected gradient descent algorithm can be used to solve the program (4). The algorithm generates a sequence of iterates ? t according to

1 b t ? t+1 = ? ? t ? (?? ?? b) , (11) ? where ? ¿ 0 is a stepsize parameter, and ? denotes the ‘2 -projection onto the ‘1 -ball of radius R. This projection can be computed rapidly in O(p) time, for instance using a procedure due to Duchi et al. [16]. Our analysis shows that under a reasonable set of conditions, the iterates for the family of programs (4) converges to a point extremely close to any global optimum in both ‘1 -norm and ‘2 -norm, even for the non-convex program.

3

Main results and consequences

We provide theoretical guarantees for both the constrained estimator (4) and the regularized variant 1 T

b ? hb ?b ? arg min ? ? ?? ? , ?i + ?n k?k1 , 2 k?k1 ?b0 k

(12)

for a constant b0 ? k? ? k2 , which is a hybrid between the constrained (4) and regularized (5) programs. 3.1

Statistical error

b satisfies a lower-RE condition with curvature In controlling the statistical error, we assume that the matrix ? b and vector ? ¿ and tolerance ¿ (n, p), as previously defined (9). In addition, recall that the matrix ? b serve as surrogates to the deterministic quantities ?x ? Rp?p and ?x ? ? ? Rp , respectively. We assume there is a function ?(Q, ?), depending on the standard deviation ? of the observation noise vector from equation (1) and the conditional distribution Q from equation (2), such that the following deviation conditions are satisfied: r r log p b ? ?x)? ? k? ? ?(Q, ?) log p . kb ? ? ?x ? ? k? ? ?(Q, ?) and k(? (13) n n q The following result applies to any global optimum ?b of the program (12) with ?n ? 4 ?(Q, ?) logn p : b ? Theorem 1 (Statistical error). Suppose the surrogates (?, b) satisfy the deviation bounds (13), and the matrix b ? satisfies the lower-RE condition (9) with parameters (¿ , ¿) such that r ? ¿ ?(Q, ?) log p ? , k ¿ (n, p) ? min . (14) 2 b0 n 128 k 4

Then for any vector ? ? with sparsity at most k, there is a universal positive constant c0 such that any global optimum ?b satisfies the bounds r ?

k log p c 0 ? max ?(Q, ?) , ?n , and (15a) k?b ? ? k2 ? ¿ n r

log p 8 c0 k ? b max ?(Q, ?) , ?n . (15b) k? ? ? k1 ? ¿ n The same bounds (without ?n) also apply to the constrained program (4) with radius choice R = k? ? k1 . b Las , ? Remarks: Note that for the standard Lasso pair (? bLas), bounds of the form (15) for sub-Gaussian noise are well-known from past work (e.g., [12, 17, 18, 19]). The novelty of Theorem 1 is in allowing for general pairs of such surrogates, which can lead to non-convexity in the underlying M -estimator. 3.2

Optimization error

Although Theorem 1 provides guarantees that hold uniformly for any choice of global minimizer, it does not provide any guidance on how to approximate such a global minimizer using a polynomial-time algorithm. b satNonetheless, we are able to show that for the family of programs (4), under reasonable conditions on ? isfied in various settings, a simple projected gradient method will converge geometrically fast to a very good approximation of any global optimum. Theorem 2 (Optimization error). Consider the program (4) with any choice of radius R for which the constraint b satisfies the lower-RE (9) and upper-RE (10) conditions with is active. Suppose that the surrogate matrix ? log p ?u , ?l n , and that we apply projected gradient descent (11) with constant stepsize ? = 2?u . Then as long as n % k log p, there is a contraction coefficient ? ? (0, 1) independent of (n, p, k) and universal positive b the gradient descent iterates {? t }? satisfy the bound constants (c1 , c2) such that for any global optimum ?, t=0 log p b 2 ? ? t k? 0 ? ?k b 2 + c1 k? t ? ?k k?b ? ? ? k21 + c2 k?b ? ? ? k22 for all t = 0, 1, 2, ... (16) 2 2 n In addition, we have the ‘1 -bound ? ? b 1 ? 2 k k? t ? ?k b 2 + 2 k k?b ? ? ? k2 + 2 k?b ? ? ? k1 for all t = 0, 1, 2, ... k? t ? ?k (17) Note that the bound (16) controls the ‘2 -distance between the iterate ? t at time t, which is easily computed in polynomial-time, and any global optimum ?b of the program (4), which may be

difficult to compute. Since ? ? (0, 1), the first term in the bound vanishes as t increases. Together with Theorem q 1, equations (16) and (17) k log p imply that the ‘2 - and ‘1 -optimization error are bounded as O(n) and O k logn p , respectively. 3.3

Some consequences

Both Theorems 1 and 2 are deterministic results; applying them to specific models requires additional work to establish the stated conditions. We turn to the statements of some consequences of these theorems for different cases of noisy, missing, and dependent data. A zero-mean random variable Z is sub-Gaussian with parameter ? ¿ 0 if E(e?Z) ? exp(?2 ? 2 /2) for all ? ? R. We say that a random matrix X ? Rn?p is sub-Gaussian with parameters (?, ? 2) if each row xTi ? Rp is sampled independently from a zero-mean distribution with covariance ?, and for any unit vector u ? Rp , the random variable uT xi is sub-Gaussian with parameter at most ?. We begin with the case of i.i.d. samples with additive noise, as described in Example 1. Corollary 1. Suppose we observe Z = X + W , where the random matrices X, W ? Rn?p are sub2 Gaussian with parameters (?x , ?x2) and (?w , ?w), respectively, and the sample size is lower-bounded as ?2 +?2 2 x w b add , ? n % max ?min , 1 k log p. Then for the M -estimator based on the surrogates (? badd), the results of (?x) Theorems 1 and 2 hold with parameters p

1 2 2 , ¿ = ?min (?x) and ?(Q, ?) = c0 ?x2 + ?w + ? ?x2 + ?w 2 with probability at least 1 ? c1 exp(?c2 log p). 5

For i.i.d. samples with missing data, we have the following: Corollary 2. Suppose X ? Rn?p is a sub-Gaussian matrix with parameters (?x , ?x2), and Z is the missing 4

?x 1 , 1 k log p, then Theorems 1 and 2 hold with data matrix with parameter ?. If n % max (1??) 4 ?2 (?x) min

probability at least 1 ? c1 exp(?c2 log p) for ¿ = 12 ?min (?x) and ?(Q, ?) = c0

?x ?x ? + . 1?? 1??

Now consider the case where the rows of X are drawn from a vector autoregressive (VAR) process according to xi+1 = Axi + vi ,

for i = 1, 2, . . . , n ? 1,

(18)

where vi ? Rp is a zero-mean noise vector with covariance matrix ?v , and A ? Rp?p is a driving matrix with spectral norm ———A———2 ¡ 1. We assume the rows of X are drawn from a Gaussian distribution with covariance ?x , such that ?x = A?x AT + ?v , so the process is stationary. Corollary 3 corresponds to the case of additive noise for a Gaussian VAR process. A similar result can be derived in the missing data setting. Corollary 3. Suppose the rows of X are drawn according to a Gaussian VAR process with driving matrix A.

4 Suppose the additive noise matrix W is i.i.d. with Gaussian rows. If n % max ?2 ? (?x) , 1 k log p, with min

? 2 = ———?w ———op +

2———?x ———op , 1 ? ———A———op

then Theorems 1 and 2 hold with probability at least 1 ? c1 exp(?c2 log p) for ¿ = 12 ?min (?x) and ?(Q, ?) = c0 (? ? + ? 2). 3.4

Application to graphical model inverse covariance estimation

The problem of inverse covariance estimation for a Gaussian graphical model is closely related to the Lasso. Meinshausen and B?uhlmann [20] prescribed a way to recover the support of the precision matrix ? when each column of ? is k-sparse, via linear regression and the Lasso. More recently, Yuan [21] proposed a method for b ? ?———1 when estimating ? using linear regression and the Dantzig selector, and obtained error bounds on ———? the columns of ? are bounded in ‘1 . Both of these results assume the rows of X are observed noiselessly and independently. Suppose we are given a matrix X ? Rn?p of samples from a multivariate Gaussian distribution, where each row is distributed according to N (0, ?). We assume the rows of X are either i.i.d. or sampled from a Gaussian VAR process (18). Based on the modified Lasso, we devise a method to estimate ? based on a corrupted observation matrix Z. Let X j denote the j th column of X, and let X ?j denote the matrix X with j th column removed. By standard results on Gaussian graphical models, there exists a vector ?j ? Rp?1 such that X j = X ?j ?j + j , j

j

?j

(19) j ?1

where is a vector of i.i.d. Gaussians and ?? X . Defining aj := ?(?jj ? ?j,?j ?) , we have b j,?j = b ?j,?j = aj ?j . Our algorithm estimates ?bj and b aj for each j and combines the estimates to obtain ? aj ?bj . In the additive noise case, we observe Z = X + W . The equations (19) yield Z j = X ?j ?j + (j + W j). Note that ? j = j + W j is a vector of i.i.d. Gaussians, and since X ?? W , we have ? j ?? X ?j . Hence, our results on covariates with additive noise produce an estimate of ?j by solving the program (4) or (12) with the b = 1 Z T Z ? ?w . When Z is a missing-data version of X, b (j) , ? b ?j,?j , 1 Z ?jT Z j), where ? pair (? b(j)) = (? n n j we similarly estimate the vectors ? with suitable corrections. We arrive at the following algorithm: Algorithm 3.1. (1) Perform p linear regressions of the variables Z j upon the remaining variables Z ?j , b (j) , ? using the modified Lasso program (4) or (12) with the estimators (? b(j)), to obtain estimates ?bj . b jj ? ? b j,?j ?bj)?1 . Set ? e j,?j = b e jj = ?b (2) Estimate the scalars aj using b aj := ?(? aj ?bj and ? aj . b = arg min ———? ? ?——— e 1 , where S p is the set of symmetric p ? p matrices. (3) Construct the matrix ? p ??S

Note that the minimization in step (3) is a linear program, so is easily solved with standard methods. We have: 6

Corollary 4. Suppose the columns of the matrix ? are k-sparse, and suppose the condition number ?(?) is nonzero and finite. Suppose the deviation conditions r r log p log p (j) j (j) j b kb ? ? ??j,?j ? k? ? ?(Q, ?) and k(? ? ??j,?j)? k? ? ?(Q, ?) (20) n n b hold for all j, and suppose we have the following additional deviation condition on ?: r b ? ?kmax ? c?(Q, ?) log p . k? (21) n b (j) with the scaling (14). Then Finally, suppose the lower-RE condition holds uniformly over the matrices ? under the estimation procedure of Algorithm 3.1,

there exists a universal constant c0 such that r c0 ?2 (?) ?(Q, ?) ?(Q, ?) log p b ———? ? ?———op ? + k . ?min (?) ?min (?) ¿ n

4

Simulations

In this section, we provide simulation results to confirm that the scalings predicted by our theory are sharp. In Figure 1, we plot the results of simulations under the additive noise model described in Example 1, using 2 ?x = I and ?w = ?w I with ?w = 0.2. Panel (a) provides plots of ‘2 -error versus the sample size n, for p ? {128, 256, 512}. For all three choices of dimensions, the error decreases to zero as the sample size n increases, showing consistency of the method. If we plot the ‘2 -error versus the rescaled sample size n/(k log p), as depicted in panel (b), the curves roughly align for different values of p, agreeing with Theorem 1. Panel (c) shows analogous curves for VAR data with additive noise, using a driving matrix A with ———A———op = 0.2. Additive noise

Additive noise

0.28

0.55 p=128 p=256 p=512

0.26

0.22

0.22

0.4

0.2

0.2

0.35

0.16

0.45

l2 norm error

0.24

0.18

0.18 0.16

0.3 0.25

0.14

0.14

0.2

0.12

0.12

0.15

0.1 0.08

0.1

0

500

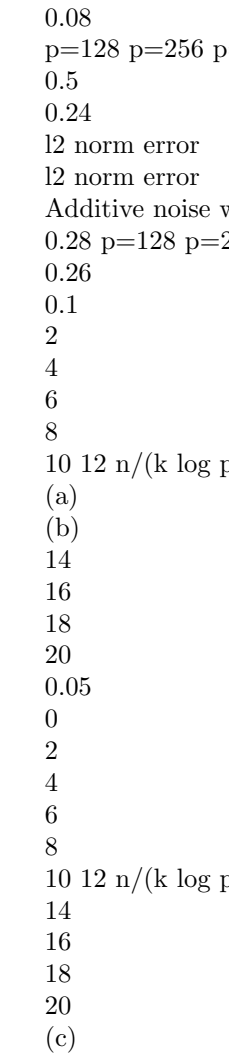
1000

1500 n

2000

2500

3000



? b Figure 1. Plots ? of the error k? ? ? k2 after running projected gradient descent on the non-convex objective, with sparsity k ? p. Plot (a) is an error plot for i.i.d. data with additive noise, and plot (b) shows ‘2 -error versus the n rescaled sample size k log . Plot (c) depicts a similar (rescaled) plot for VAR data with additive noise. As predicted p by Theorem 1, the curves align for different values of p in the rescaled plot.

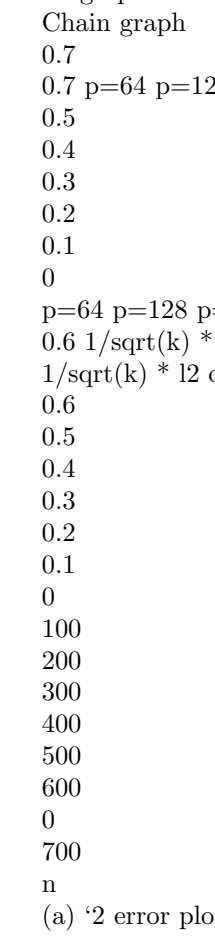
We also verified the results of Theorem 2 empirically. Figure 2 shows the results of applying projected gradient descent to solve the optimization problem (4) in the cases of additive noise and missing data. We first applied b then reapplied projected gradient descent 10 times, tracking projected gradient to obtain an initial estimate ?, t b the optimization error k? ? ?k2 (in blue) and statistical error k? t ? ? ? k2 (in red). As predicted by Theorem 2, the iterates exhibit geometric convergence to roughly the same fixed point, regardless of

starting point. Finally, we simulated the inverse covariance matrix estimation algorithm on three types of graphical models: (a) Chain-structured graphs. In this case, all nodes of are arranged in a line. The diagonal entries of ? equal 1, and entries corresponding to links in the chain equal 0.1. Then ? is rescaled so ———?———op = 1. (b) Star-structured graphs. In this case, all nodes are connected to a central node, which has degree k ? 0.1p. All other nodes have degree 1. The diagonal entries of ? are set equal to 1, and all entries corresponding to edges in the graph are set equal to 0.1. Then ? is rescaled so ———?———op = 1. (c) Erd?os-Renyi graphs. As in Rothman et al. [22], we first generate a matrix B with diagonal entries 0, and all other entries independently equal to 0.5 with probability k/p, and 0 otherwise. Then ? is chosen so ? = B + ?I has condition number p, and ? is rescaled so ———?———op = 1. 7

Log error plot: additive noise case
Log error plot: missing data case
0.5
0.5 Stat error Opt error
0
?0.5
?1
log(——‘t ? ‘——2)
log(——‘t ? ‘——2)
?0.5
?1.5 ?2 ?2.5
?1 ?1.5 ?2 ?2.5
?3 ?3.5
Stat error Opt error
0
?3
0
20
40 60 Iteration count
80
?3.5
100
0
20
40 60 Iteration count
(a)
80
100
(b)

b 2) and statistical error log(k? t ? ? ? k2) versus iteration Figure 2. Plots of the optimization error log(k? t ? ?k number t, generated by running projected gradient descent on the non-convex objective. As predicted by Theorem 2, the optimization error decreases geometrically.

After generating the matrix X of n i.i.d. samples from the appropriate graphical model, with covariance matrix ?x = ??1 , we generated the corrupted matrix Z = X + W with ?w = (0.2)2 I. Figure 3 shows the rescaled b ? ?———op plotted against the sample size n for a chain-structured graph, with panel (a) showing the ‘2 -error ?1k ———? original plot and panel (b) plotting against the rescaled sample size. We obtained qualitatively similar results for the star and Erd?os-Renyi graphs, in the presence of missing and/or dependent data. Chain graph



b Figure 3. (a) Plots of the rescaled error ?1k ———???——— op after running projected gradient descent for a chain-structured Gaussian graphical model with additive noise. As predicted by Theorems 1 and 2, all curves align

when the rescaled n , as shown in (b). Each point represents the average over 50 trials. error is plotted against the ratio k log p

5

Discussion

In this paper, we formulated an ‘1 -constrained minimization problem for sparse linear regression on corrupted data. The source of corruption may be additive noise or missing data, and although the resulting objective is not generally convex, we showed that projected gradient descent is guaranteed to converge to a point within statistical precision of the optimum. In addition, we established ‘1 - and ‘2 -error bounds that hold with high probability when the data are drawn i.i.d. from a sub-Gaussian distribution, or drawn from a Gaussian VAR process. Finally, we used our results on linear regression to perform sparse inverse covariance estimation for a Gaussian graphical model, based on corrupted data. The bounds we obtain for the spectral norm of the error are of the same order as existing bounds for inverse covariance matrix estimation with uncorrupted, i.i.d. data. Acknowledgments PL acknowledges support from a Hertz Foundation Fellowship and an NDSEG Fellowship; MJW and PL were also partially supported by grants NSF-DMS-0907632 and AFOSR-09NL184. The authors thank Alekh Agarwal, Sahand Negahban, and John Duchi for discussions and guidance.

8

2 References

[1] R. Little and D. B. Rubin. Statistical analysis with missing data. Wiley, New York, 1987. [2] J. T. Hwang. Multiplicative errors-in-variables models with applications to recent data released by the U.S. Department of Energy. Journal of the American Statistical Association, 81(395):pp. 680?688, 1986. [3] R. J. Carroll, D. Ruppert, and L. A. Stefanski. Measurement Error in Nonlinear Models. Chapman and Hall, 1995. [4] S. J. Iturria, R. J. Carroll, and D. Firth. Polynomial regression and estimating functions in the presence of multiplicative measurement error. Journal of the Royal Statistical Society Series B - Statistical Methodology, 61:547?561, 1999. [5] Q. Xu and J. You. Covariate selection for linear errors-in-variables regression models. Communications in Statistics - Theory and Methods, 36(2):375?386, 2007. [6] N. St?adler and P. B?uhlmann. Missing values: Sparse inverse covariance estimation and an extension to sparse regression. Statistics and Computing, pages 1?17, 2010. [7] M. Rosenbaum and A. B. Tsybakov. Sparse recovery under matrix uncertainty. Annals of Statistics, 38:2620?2651, 2010. [8] P. Loh and M.J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. Technical report, UC Berkeley, September 2011. Available at http: //arxiv.org/abs/1109.3714. [9] R. Tibshirani. Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, Series B, 58(1):267?288, 1996. [10] S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. SIAM Journal on Sci-

entific Computing, 20(1):33?61, 1998. [11] S. van de Geer. The deterministic Lasso. In Proceedings of Joint Statistical Meeting, 2007. [12] P. J. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. Annals of Statistics, 37(4):1705?1732, 2009. [13] S. van de Geer and P. Buhlmann. On the conditions used to prove oracle results for the Lasso. Electronic Journal of Statistics, 3:1360?1392, 2009. [14] G. Raskutti, M. J. Wainwright, and B. Yu. Restricted eigenvalue properties for correlated Gaussian designs. Journal of Machine Learning Research, 11:2241?2259, 2010. [15] A. Agarwal, S. Negahban, and M.J. Wainwright. Fast global convergence of gradient methods for highdimensional statistical recovery. Technical report, UC Berkeley, April 2011. Available at http:// arxiv.org/abs/1104.4824. [16] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the ‘1 -ball for learning in high dimensions. In International Conference on Machine Learning, pages 272?279, 2008. [17] C. H. Zhang and J. Huang. The sparsity and bias of the Lasso selection in high-dimensional linear regression. Annals of Statistics, 36(4):1567?1594, 2008. [18] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. Annals of Statistics, 37(1):246?270, 2009. [19] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for the analysis of regularized M -estimators. In Advances in Neural Information Processing Systems, 2009. [20] N. Meinshausen and P. B?uhlmann. High-dimensional graphs and variable selection with the Lasso. Annals of Statistics, 34:1436?1462, 2006. [21] M. Yuan. High-dimensional inverse covariance matrix estimation via linear programming. Journal of Machine Learning Research, 99:2261?2286, August 2010. [22] A. J. Rothman, P. J. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. Electronic Journal of Statistics, 2:494?515, 2008.

9