Confidence Intervals and Hypothesis Testing for High-Dimensional Statistical Models

Authored by:

Andrea Montanari
Adel Javanmard

Abstract

Fitting high-dimensional statistical models often requires the use of non-linear parameter estimation procedures. As a consequence, it is generally impossible to obtain an exact characterization of the probability distribution of the parameter estimates. This in turn implies that it is extremely challenging to quantify the 'uncertainty' associated with a certain parameter estimate. Concretely, no commonly accepted procedure exists for computing classical measures of uncertainty and statistical significance as confidence intervals or p-values. We consider here a broad class of regression problems, and propose an efficient algorithm for constructing confidence intervals and p-values. The resulting confidence intervals have nearly optimal size. When testing for the null hypothesis that a certain parameter is vanishing, our method has nearly optimal power. Our approach is based on constructing a 'de-biased' version of regularized M-estimators. The new construction improves over recent work in the field in that it does not assume a special structure on the design matrix. Furthermore, proofs are remarkably simple. We test our method on a diabetes prediction problem.

1 Paper Body

It is widely recognized that modern statistical problems are increasingly high-dimensional, i.e. require estimation of more parameters than the number of observations/examples. Examples abound from signal processing [16], to genomics [21], collaborative filtering [12] and so on. A number of successful estimation techniques have been developed over the last ten years to tackle these problems. A widely applicable approach consists in optimizing a suitably regularized likelihood function. Such estimators are, by necessity, non-linear and non-explicit (they are solution of certain optimization problems). The use of non-linear parameter estimators comes at a price. In general, it is impossible to characterize the distribution of the estimator. This situation is very different from the one of classical statistics in which either exact characterizations

are available, or asymptotically exact ones can be derived from large sample theory [26]. This has an important and very concrete consequence. In classical statistics, generic and well accepted procedures are available for characterizing the uncertainty associated to a certain parameter estimate in terms of confidence intervals or p-values [28, 14]. However, no analogous procedures exist in high-dimensional statistics. In this paper we develop a computationally efficient procedure for constructing confidence intervals and p-values for a broad class of high-dimensional regression problems. The salient features of our procedure are: (i) Our approach guarantees nearly optimal confidence interval sizes and testing power. (ii) It is the first one that achieves this goal under essentially no assumptions on the population covariance matrix of the parameters, beyond the standard conditions for high-dimensional consistency. (iii) It allows for a streamlined analysis with respect to earlier work in the same area. 1

Table 1: Unbiased estimator for ?0 in high dimensional linear regression models Input: Measurement vector y, design matrix X, parameter ?. Output: Unbiased estimator ?bu . 1: Set ? = ??, and let ?bn be the Lasso estimator as per Eq. (3). b ? (XT X)/n. 2: Set ? 3: for i = 1, 2, . . . , p do 4: Let mi be a solution of the convex program: b mT ?m b ? ei k? ? ? subject to k?m

minimize

(4)

5: Set M = (m1 , . . . , mp)T . If any of the above problems is not feasible, then set M = Ip?p . 6: Define the estimator ?bu as follows:

1 ?bu = ?bn (?) + M XT (Y ? X?bn (?)) n

(5)

(iv) Our method has a natural generalization non-linear regression models (e.g. logistic regression, see Section 4). We provide heuristic and numerical evidence supporting this generalization, deferring a rigorous study to future work. For the sake of clarity, we will focus our presentation on the case of linear regression, deferring the generalization to Section 4. In the random design model, we are given n i.i.d. pairs (Y1 , X1), (Y2 , X2), . . . , (Yn , Xn), with vectors Xi ? Rp and response variables Yi given by Wi ? N(0, ? 2) .

Yi = h?0 , Xi i + Wi ,

(1)

Here h ? , ? i is the standard scalar product in Rp . In matrix form, letting Y = (Y1 , . . . , Yn)T and denoting by X the design matrix with rows X1T , . . . , XnT , we have Y = X ?0 + W ,

W ? N(0, ? 2 In?n) .

(2) p

The goal is estimate the unknown (but fixed) vector of parameters ?0 ? R . In the classic setting, n p and the estimation method of choice is ordinary least squares yielding ?bOLS = (XT X)?1 XT Y . In particular ?b is Gaussian with mean ?0 and covariance ? 2 (XT X)?1 . This directly allows to construct confidence intervals1 . In the high-dimensional setting where p ¿ n, the matrix (XT X) is rank deficient and one has to resort to biased estimators. A particularly successful approach is the Lasso [24, 7] which promotes sparse reconstructions through an ‘1 penalty. o n 1 kY ? X?k22 + ?k?k1 . (3) ?bn (Y, X; ?) ? arg minp

??R 2n In case the right hand side has more than one minimizer, one of them can be selected arbitrarily for our purposes. We will often omit the arguments Y , X, as they are clear from the context. We denote by S ? supp(?0) ? [p] the support of ?0 , and let s0 ? —S—. A copious theoretical literature [6, 2, 4] shows that, under suitable assumptions on X, the Lasso is nearly as accurate as if the support S was known a priori. Namely, for n = ?(s0 log p), we have k?bn ? ?0 k22 = O(s0 ? 2 (log p)/n). These remarkable properties come at a price. Deriving an exact characterization for the distribution of ?bn is not tractable in general, and hence there is no simple procedure to construct confidence intervals and p-values. In order to overcome this challenge, we construct a de-biased estimator from the Lasso solution. The de-biased estimator is given by the simple formula ?bu = ?bn +(1/n) M XT (Y ?X?bn), as in Eq. (5). The basic intuition is that XT (Y ? X?bn)/(n?) is a subgradient of the ‘1 norm at the Lasso solution ?bn . By adding a term proportional to this subgradient, our procedure compensates the bias introduced by the ‘1 penalty in the Lasso. For instance, letting Q ? (XT X/n)?1 , ?biOLS ? 1.96? dence interval [28]. 1

2

p

Qii /n, ?biOLS + 1.96?

p Qii /n] is a 95% confi-

We will prove in Section 2 that ?bu is approximately Gaussian, with mean ?0 and covariance b)/n, where ? b = (XT X/n) is the empirical covariance of the feature vectors. This result ? 2 (M ?M allows to construct confidence intervals and p-values in complete analogy withp classical statistics p u u b b b procedures. For instance, letting Q ? M ?M , [?i ? 1.96? ? Qii /n, ?i + 1.96? Qii /n] is a 95% confidence interval. The size of this interval is of order ?/ n, which is the optimal (minimum) one, i.e. the same that would have been obtained by knowing a priori the support of ?0 . In practice the noise standard deviation is not known, but ? can be replaced by any consistent estimator ? b. A key role is played by the matrix M ? Rp?p whose function is to ?decorrelate? the columns of X. We propose here to construct M by solving a convex program that aims at optimizing two objectives. b ? I—? (here and below — ? —? denotes the entrywise ‘? norm) One one hand, we try to control —M ? which ?as shown in Theorem 2.1? controls the non-Gaussianity and bias of ?bu . On the other, we b]i,i , for each i ? [p], which controls the variance of ?bu . minimize [M ?M i The idea of constructing a de-biased estimator of the form ?bu = ?bn + (1/n) M XT (Y ? X?bn) was used by Javanmard and Montanari in [10], that suggested the choice M = c??1 , with ? = E{X1 X1T } the population covariance matrix and c a positive constant. A simple estimator for ? was proposed for sparse covariances, but asymptotic validity and optimality were proven only for uncorrelated Gaussian designs (i.e. Gaussian X with ? = I). Van de Geer, B?ulhmann and Ritov [25] used the same construction with M an estimate of ??1 which is appropriate for sparse inverse covariances. These authors prove semi-parametric optimality in a non-asymptotic setting, provided the sample size is at least n = ?(s20 log p). In this paper, we do not assume any sparsity constraint on ??1 , but still require the sample size scaling n = ?(s20 log

p). We refer to a forthcoming publication wherein the condition on the sample size scaling is relaxed [11]. From a technical point of view, our proof starts from a simple decomposition of the de-biased estimator ?bu into a Gaussian part and an error term, already used in [25]. However ?departing radically from earlier work? we realize that M need not be a good estimator of ??1 in order for the de-biasing procedure to work. We instead set M as to minimize the error term and the variance of the Gaussian term. As a consequence of this choice, our approach applies to general covariance structures ?. By contrast, earlier approaches applied only to sparse ?, as in [10], or sparse ??1 as in [25]. The only assumptions we make on ? are the standard compatibility conditions required for high-dimensional consistency [4]. We refer the reader to the long version of the paper [9] for the proofs of our main results and the technical steps. 1.1

Further related work

The theoretical literature on high-dimensional statistical models is vast and rapidly growing. Restricting ourselves to linear regression, earlier work investigated prediction error [8], model selection properties [17, 31, 27, 5], ‘2 consistency [6, 2]. Of necessity, we do not provide a complete set of references, and instead refer the reader to [4] for an in-depth introduction to this area. The problem of quantifying statistical significance in high-dimensional parameter estimation is, by comparison, far less understood. Zhang and Zhang [30], and B?uhlmann [3] proposed hypothesis testing procedures under restricted eigenvalue or compatibility conditions [4]. These methods are however effective only ? for detecting very ? large coefficients. Namely, they both require —?0,i — ? c max{?s0 log p/ n, ?/ n}, which is s0 larger than the ideal detection level [10]. In other words, in order for the coefficient ?0,i to be detectable with appreciable probability, it needs to be larger than the overall ‘2 error, rather than the ‘2 error per coordinate. Lockart et al. [15] develop a test for the hypothesis that a newly added coefficient along the Lasso regularization path is irrelevant. This however does not allow to test arbitrary coefficients at a given value of ?, which is instead the problem addressed in this paper. It further assumes that the current Lasso support contains the actual support supp(?0) and that the latter has bounded size. Finally, resampling methods for hypothesis testing were studied in [29, 18, 19]. 1.2

Preliminaries and notations

b ? XT X/n be the sample covariance matrix. For p ¿ n, ? b is always singular. However, We let ? b we may require ? to be nonsingular for a restricted set of directions. 3

b and a set S of size s0 , the compatibility condition is met, if for some Definition 1.1. For a matrix ? ?0 ¿ 0, and all ? satisfying k?S c k1 ? 3k?S k1 , it holds that k?S k21 ?

s0 T b ? ?? . ?20

Definition 1.2. The sub-gaussian norm of a random variable X, denoted by kXk?2 , is defined as kXk?2 = sup p?1/2 (E—X—p)1/p . p?1

The sub-gaussian norm of a random vector X ? Rn is defined as kXk?2 = supx?S n?1 khX, xik?2 . Further, for a random variable X, its sub-exponential norm, denoted by kXk?1 , is defined as kXk?1 = sup p?1 (E—X—p)1/p . p?1

For a matrix A and set of indices I, J, we let AI,J denote the submatrix formed by the rows in I and columns in J. Also, AI,? (resp. A?,I) denotes the submatrix containing just the rows (reps. columns) in I. Likewise, for a vector v, vI is the restriction of v to indices in I. We use the shorthand ?1 ?1 A?1)I,J . In particular, A?1)i,i . The maximum and the minimum singular values i,i = (A I,J = (A of A are respectively denoted by ?max (A) and ?min (A). We write kvkp for the standard ‘p norm of a vector v and kvk0 for the number of nonzero entries of v. For a matrix A, kAkp is the ‘p operator P norm, and —A—p is the elementwise ‘p norm, i.e., —A—p = (i,j —Aij —p)1/p . For an integer p ? 1, we let [p] ? {1, . . . , p}. For a vector v, supp(v) represents the positions of nonzero entries of v. Throughout, (w.h.p) means with probability converging to one as n ? ?, and ? R x with2high probability ?(x) ? ?? e?t /2 dt/ 2? denotes the CDF of the standard normal distribution.

2

An de-biased estimator for ?0

Theorem 2.1. Consider the linear model (1) and let ?bu be defined as per Eq. (5). Then, ? u ? b . b T) , ? = n(M ? b ? I)(?0 ? ?) n(?b ? ?0) = Z + ? , Z—X ? N(0, ? 2 M ?M Further, suppose that ?min (?) = ?(1), and ?max (?) = O(1). In addition assume the rows of the whitened matrix X??1/2 are sub-gaussian, i.e., k??1/2 X1 k?2 = O(1). Let E be the event that the p b i,i = O(1). Then, using ? = O((log p)/n) b and maxi?[p] ? compatibility condition holds for ?, ? (see inputs in Table 1), the following holds true. On the event E, w.h.p, k?k? = O(s0 log p/ n). Note that compatibility condition (and hence the event E) holds w.h.p. for random design matrices of a general nature. In fact [22] shows that under some general assumptions, the compatibility b w.h.p., when n is sufficiently large. Bounds on condition on ? implies a similar condition on ?, T b the variances [M ?M]ii will be given in Section 3.2. Finally, the claim of Theorem 2.1 does not rely on the specific choice of the objective function in optimization problem (4) and only uses the optimization constraints. Remark 2.2. Theorem 2.1 does not make any assumption ? about the parameter vector ?0 . If we further assume that the support size s0 satisfies s0 = o(n/ log p), then we have k?k? = o(1), w.h.p. Hence, ?bu is an asymptotically unbiased estimator for ?0 .

3

Statistical inference

A direct application of Theorem 2.1 is to derive confidence intervals and statistical hypothesis tests ? for high dimensional models. Throughout, we make the sparsity assumption s0 = o(n/ log p). 3.1

Confidence intervals

We first show that the variances of variables Zj —X are ?(1). 4

Lemma 3.1. Let M = (m1 , . . . , mp)T be the matrix with rows mT i obtained by solving convex T 2 b b program (4). Then for all i ? [p], [M ?M]i,i ? (1 ? ?) /?i,i . By Remark 2.2 and Lemma 3.1, we have o n ?n(?bu ? ?)

0,i i P ? xX = ?(x) + o(1) , 1/2 T b ?[M ?M]i,i

?x ? R .

(6)

Since the limiting probability is independent of X, Eq. (6) also holds unconditionally for random design X. For constructing confidence intervals, a consistent estimate of ? is needed. To this end, we use the scaled Lasso [23] given by n 1 o ? {?bn (?), ? b} ? arg min kY ? X?k22 + + ?k?k1 . 2 ??Rp ,?¿0 2?n This is a joint convex minimization which provides an estimate of the noise level in addition to an p b, under the assumptions estimate of ?0 . We use ? = c1 (log p)/n that yields a consistent estimate ? of Theorem 2.1 (cf. [23]). We hence obtain the following. Corollary 3.2. Let q ?1 ?1/2 b T]i,i . ?(?, n) = ? (1 ? ?/2)b ?n [M ?M (7) Then Ii = [?biu ? ?(?, n), ?biu + ?(?, n)] is an asymptotic two-sided confidence interval for ?0,i with significance ?. Notice that the same corollary applies to any other consistent estimator ? b of the noise standard deviation. 3.2

Hypothesis testing

An important advantage of sparse linear regression models is that they provide parsimonious explanations of the data in terms of a small number of covariates. The easiest way to select the ?active? covariates is to choose the indexes i for which ?bin 6= 0. This approach however does not provide a measure of statistical significance for the finding that the coefficient is non-zero. More precisely, we are interested in testing an individual null hypothesis H0,i : ?0,i = 0 versus the alternative HA,i : ?0,i 6= 0, and assigning p-values for these tests. We construct a p-value Pi for the test H0,i as follows:

? bu n —?i — . (8) Pi = 2 1 ? ? b T]1/2 ? b[M ?M i,i The decision rule is then based on the p-value Pi :

1 if Pi ? ? Ti,X (y) = 0 otherwise

(reject H0,i) , (accept H0,i) .

(9)

We measure the quality of the test Ti,X (y) in terms of its significance level ?i and statistical power 1 ? ?i . Here ?i is the probability of type I error (i.e. of a false positive at i) and ?i is the probability of type II error (i.e. of a false negative at i). Note that it is important to consider the tradeoff between statistical significance and power. Indeed any significance level ? can be achieved by randomly rejecting H0,i with probability ?. This test achieves power 1 ? ? = ?. Further note that, without further assumption, no nontrivial power can be achieved. In fact, choosing ?0,i 6= 0 arbitrarily close to zero, H0,i becomes indistinguishable from its alternative. We will therefore assume that, whenever ?0,i 6= 0, we have —?0,i — ¿ ? as well. We take a minimax perspective and require the test to behave uniformly well over s0 -sparse vectors. Formally, for ? ¿ 0 and i ? [p], define n o ?i (n) ? sup P?0 (Ti,X (y) = 1) : ?0 ? Rp , k?0 k0 ? s0 (n), ?0,i = 0 . n o ?i (n; ?) ? sup P?0 (Ti,X (y) = 0) : ?0 ? Rp , k?0 k0 ? s0 (n), —?0,i — ? ? .

5

Here, we made dependence on n explicit. Also, P? (?) is the induced probability for random design X and noise realization w, given the fixed parameter vector ?. Our next theorem establishes bounds on ?i (n) and ?i (n; ?). Theorem 3.3. Consider a random? design model that satisfies the conditions of Theorem 2.1. Under the sparsity assumption s0 = o(n/ log p), the following holds true

for any fixed sequence of integers i = i(n): lim ?i (n) ? ? .

(10)

n??

lim

n??

1 ? ?i (?; n) ? 1, 1 ? ?i? (?; n)

1 ? ?i? (?; n) ? G ?,

?

n? 1/2 ?[??1 i,i]

,

(11)

where, for ? ? [0, 1] and u ? R+ , the function G(?, u) is defined as follows: ? ? G(?, u) = 2 ? ?(??1 (1 ?) + u) ? ?(??1 (1 ?) ? u) . 2 2 It is easy to see that, for any ? ¿ 0, u 7? G(?, u) is continuous and monotone increasing. Moreover, G(?, 0) = ? which is the trivial power obtained by randomly rejecting H0,i with probability ?. As ? deviates from zero, we obtain nontrivial to achieve a specific power ? power. Notice that in order ?1 ? ¿ ?, our scheme requires ? = O(?/ n), since ??1) ? (?min (?))?1 = O(1). i,i ? ?max (? 3.2.1

Minimax optimality

The authors of [10] prove an upper bound for the minimax power of tests with a given significance level ?, under the Gaussian random design models (see Theorem 2.6 therein). This bound is obtained by considering an oracle test that knows all the active parameters except i, i.e., S{i}. To state the bound formally, for a set S ? [p] and i ? S c , define ?i—S ? ?i,i ? ?i,S (?S,S)?1 ?S,i , and let n o ??,s0 ? min ?i—S : S ? [p]{i}, —S— ¡ s0 . i?[p],S

? In asymptotic regime and under our sparsity assumption s0 = o(n/ log p), the bound of [10] simplifies to lim

n??

1 ? ?iopt (?; ?) ? 1, G(?, ?/?eff)

?eff = ?

? , n ??,s0

(12)

Using the bound of (12) and specializing the result of Theorem 3.3 to Gaussian design X, we obtain that our scheme achieves a near optimal minimax power for a broad class of covariance matrices. We can compare our test to the optimal test by computing how much ? must be increased in order to achieve the minimax optimal power. It follows from the above that ? must be increased to ? ?, with the two differing by a factor: q q p ? ?/? = ??1 ? ? ??1 ?max (?)/?min (?) , ?,s 0 ii i,i ?i,i ? ?1 since ??1 , and ?i—S ? ?i,i ? ?max (?) due to ?S,S 0. ii ? (?min (?))

4

General regularized maximum likelihood

In this section, we generalize our results beyond the linear regression model to general regularized maximum likelihood. Here, we only describe the debiasing method. Formal guarantees can be obtained under suitable restricted

strong convexity assumptions [20] and will be the object of a forthcoming publication. For univariate Y , and vector X ? Rp , we let {f? (Y —X)}??Rp be a family of conditional probability densities parameterized by ?, that are absolutely continuous with respect to a common measure ?(dy), and suppose that the gradient ?? f? (Y —X) exists and is square integrable. As in for linear regression, we assume that the data is given by n i.i.d. pairs (X1 , Y1), . . . (Xn , Yn), where conditional on Xi , the response variable Yi is distributed as Yi ? f?0 (? —Xi) . 6

for some parameter vector ?0 ? Rp . Let Li (?) = ? log f? (Yi —Xi) be the normalized negative Pn log-likelihood corresponding to the observed pair (Yi , Xi), and define L(?) = n1 i=1 Li (?) . We consider the following regularized estimator:

(13) ?b ? arg minp L(?) + ?R(?) , ??R

where ? is a regularization parameter and R : Rp ? R+ is a norm. b Let Ii (?) be the Fisher information of f? (Y —Xi), defined as We next generalize the definition of ?. h i h

T i

Ii (?) ? E ?? log f? (Y —Xi) ?? log f? (Y —Xi) Xi = ?E ?2? log f (Y —Xi , ?) Xi , where the second identity holds under suitable regularity conditions [13], and ?2? denotes the Hesb ? Rp?p as follows: sian operator. We assume E[Ii (?)] 0 define ? n

X b . b? 1 ? Ii (?) n i=1

(14)

b Finally, the de-biased estimator ?bu is defined by ?bu ? b depends on ?. Note that (in general) ? b b ? ? M ?? L(?) , with M given again by the solution of the convex program (4), and the definition of b provided here. Notice that this construction is analogous to the one in [25] (although the present ? setting is somewhat more general) with the crucial difference of the construction of M . b A a simple heuristic derivation of this method is the following. By Taylor expansion of L(?) u 2 2 b around ?0 we get ?b ? ?b ? M ?? L(?0) ? M ?? L(?0)(?b ? ?0) . Approximating ?? L(?0) ? ? u b (which amounts to taking expectation with respect to the response variables yi), we get ? ? ?0 ? b ? I](?b ? ?b0). Conditionally on {Xi }1?i?n , the first term has zero expectation ?M ?? L(?0) ? [M ? b]. Further, by central limit theorem, its low-dimensional marginals are apand covariance [M ?M b ? I](?b ? ?b0) can be bounded as in the linear regression proximately Gaussian. The bias term ?[M ? b ? I—? ? ?. case, building on the fact that M is chosen such that —M ? Similar to the linear case, an asymptotic two-sided confidence interval for ?0,i (with significance ?) is given by Ii = [?biu ? ?(?, n), ?biu + ?(?, n)], where b T]1/2 . ?(?, n) = ??1 (1 ? ?/2)n?1/2 [M ?M i,i Moreover, an asymptotically valid p-value Pi for testing null hypothesis H0,i is constructed as:

? bu n—?i — . Pi = 2 1 ? ? b T]1/2 [M ?M i,i In the next section, we shall apply the general approach presented here to L1 -regularized logistic regression. In this case, the binary response Yi ? {0, 1} is distributed as Yi ? f?0 (? —Xi) where f?0 (1—x) = (1 + e?hx,?0 i)?1 and f?0 (0—x) = (1 + ehx,?0 i)?1 . It is easy to see that in this case b i i ?1 b = qbi (1 ? qbi)Xi X T , with qbi = (1

+ e?h?,X Ii (?)) , and thus i n
X b= 1 ? qbi (1 ? qbi)Xi XiT . n i=1
5
Diabetes data example

We consider the problem of estimating relevant attributes in predicting type-2 diabetes. We evaluate the performance of our hypothesis testing procedure on the Practice Fusion Diabetes dataset [1]. This dataset contains de-identified medical records of 10000 patients, including information on diagnoses, medications, lab results, allergies, immunizations, and vital signs. From this dataset, we extract p numerical attributes resulting in a sparse design matrix Xtot ? Rntot ?p , with ntot = 10000, 7
0.3
ZSc
ZS N(0, 1)
0.1
0.2
Density
2 0 -2
0.0
-4
Sample Quantiles of Z
4
0.4
Histograms of Z
-3
-2
-1
0
1
2
3
-10
-5
0
5
10
Standard normal quantiles
(a)
(b)
Q-Q plot of Z
? for one realization. Normalized histograms of Z

Figure 1: Q-Q plot of Z and normalized histograms of Z?S (in red) and Z?S c (in blue) for one realization. No fitting of the Gaussian mean and variance was done in panel (b).

and p = 805 (only 5.9% entries of Xtot are non-zero). Next, we standardize the columns of X to have mean 0 and variance 1. The attributes consist of:

(i)Transcript records: year of birth, gender and BMI; (ii)Diagnoses informations: 80 binary attributes corresponding to different ICD-9 codes. (iii)Medications: 80 binary attributes indicating the use of different medications. (iv) Lab results: For 70 lab test observations, we include attributes indicating patients tested, abnormality flags, and the observed values. We also bin the observed values into 10 quantiles and make 10 binary attributes indicating the bin of the corresponding observed value. We consider logistic model as described in the previous section with a binary response identifying the patients diagnosed with type-2 diabetes. For the sake of performance evaluation, we need to know the true significant attributes. Letting L(?) be the logistic loss corresponding to the design Xtot and response vector Y ? Rntot , we take ?0 as the minimizer of L(?). Notice that here, we are in the low dimensional regime (ntot ¿ p) and no regularization is needed. Next, we take random subsamples of size n = 500 from the patients, and examine the performance of our testing procedure. The experiment is done using glmnet-package in R that fits the entire path of the regularized logistic estimator. We then choose the value of ? that yields maximum AUC (area under ROC curve), approximated by a 5-fold cross validation. Results: Type I errors and powers of our decision rule (9) are computed by comparing to ?0 . The average error and power (over 20 random subsamples) and significance level ? = 0.05 are respec? b]1/2 . tively, 0.0319 and 0.818. Let Z = (zi)pi=1 denote the vector with zi ? n(?biu ? ?0,i)/[M ?M i,i In Fig. 1(a), sample quantiles of Z are depicted versus the quantiles of a standard normal distribution. The plot clearly corroborates our theoretical result regarding the limiting distribution of Z. In order to build further intuition about the proposed p-values, let Z? = (? zi)pi=1 be the vector with ? bu b]1/2 . In Fig. 1(b), we plot the normalized histograms of Z?S (in red) and Z?S c (in z?i ? n?i /[M ?M i,i blue). As the plot showcases, Z?S c has roughly standard normal distribution, and the entries of Z?S appear as distinguishable spikes. The entries of Z?S with larger magnitudes are easier to be marked off from the normal distribution tail.

2 References

[1] Practice Fusion Diabetes Classification. http://www.kaggle.com/c/pf2012-diabetes, 2012. Kaggle competition dataset.

8

[2] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. Amer. J. of Mathematics, 37:1705?1732, 2009. [3] P. B?uhlmann. Statistical significance in high-dimensional linear models. arXiv:1202.1377, 2012. [4] P. B?uhlmann and S. van de Geer. Statistics for high-dimensional data. Springer-Verlag, 2011. [5] E. Cand‘es and Y. Plan. Near-ideal model selection by ‘1 minimization. The Annals of Statistics, 37(5A):2145?2177, 2009. [6] E. J. Cand?es and T. Tao. Decoding by linear programming. IEEE Trans. on Inform. Theory, 51:4203? 4215, 2005. [7] S. Chen and D. Donoho. Examples of basis pursuit. In Proceedings of Wavelet Applications in Signal and Image Processing III, San Diego, CA, 1995. [8] E. Greenshtein and Y.

Ritov. Persistence in high-dimensional predictor selection and the virtue of overparametrization. Bernoulli, 10:971?988, 2004. [9] A. Javanmard and A. Montanari. Confidence Intervals and Hypothesis Testing for High-Dimensional Regression. arXiv:1306.3171, 2013. [10] A. Javanmard and A. Montanari. Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. arXiv:1301.4240, 2013. [11] A. Javanmard and A. Montanari. Nearly Optimal Sample Size in Hypothesis Testing for High-Dimensional Regression. arXiv:1311.0274, 2013. [12] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. Computer, 42(8):30?37, August 2009. [13] E. Lehmann and G. Casella. Theory of point estimation. Springer, 2 edition, 1998. [14] E. Lehmann and J. Romano. Testing statistical hypotheses. Springer, 2005. [15] R. Lockhart, J. Taylor, R. Tibshirani, and R. Tibshirani. A significance test for the lasso. arXiv preprint arXiv:1301.7161, 2013. [16] M. Lustig, D. Donoho, J. Santos, and J. Pauly. Compressed sensing mri. IEEE Signal Processing Magazine, 25:72?82, 2008. [17] N. Meinshausen and P. B?uhlmann. High-dimensional graphs and variable selection with the lasso. Ann. Statist., 34:1436?1462, 2006. [18] N. Meinshausen and P. B?uhlmann. Stability selection. J. R. Statist. Soc. B, 72:417?473, 2010. [19] J. Minnier, L. Tian, and T. Cai. A perturbation method for inference on regularized regression estimates. Journal of the American Statistical Association, 106(496), 2011. [20] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. Statistical Science, 27(4):538?557, 2012. [21] J. Peng, J. Zhu, A. Bergamaschi, W. Han, D.-Y. Noh, J. R. Pollack, and P. Wang. Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. The Annals of Applied Statistics, 4(1):53?77, 2010. [22] M. Rudelson and S. Zhou. Reconstruction from anisotropic random measurements. IEEE Transactions on Information Theory, 59(6):3434?3447, 2013. [23] T. Sun and C.-H. Zhang. Scaled sparse linear regression. Biometrika, 99(4):879?898, 2012. [24] R. Tibshirani. Regression shrinkage and selection with the Lasso. J. Royal. Statist. Soc B, 58:267?288, 1996. [25] S. van de Geer, P. B?uhlmann, and Y. Ritov. On asymptotically optimal confidence regions and tests for high-dimensional models. arXiv:1303.0518, 2013. [26] A. W. Van der Vaart. Asymptotic statistics, volume 3. Cambridge university press, 2000. [27] M. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ‘1 -constrained quadratic programming. IEEE Trans. on Inform. Theory, 55:2183?2202, 2009. [28] L. Wasserman. All of statistics: a concise course in statistical inference. Springer Verlag, 2004. [29] L. Wasserman and K. Roeder. High dimensional variable selection. Annals of statistics, 37(5A):2178, 2009. [30] C.-H. Zhang and S. Zhang. Confidence Intervals for Low-Dimensional Parameters in High-Dimensional Linear Models. arXiv:1110.2563, 2011. [31] P. Zhao and B. Yu. On model selection consistency of Lasso. The Journal of Machine Learning Research, 7:2541?2563, 2006.

9