Persistence in High dimensional linear predictor-selection and the virtue of over-parametrization

Eitan Greenshtein*
Department of Statistics
Haifa University
Mount Carmel
31905 Haifa Israel.
geitan@stat.haifa.ac.il

Ya'acov Ritov
Department of Statistics
The Hebrew University of Jerusalem
91905 Jerusalem, Israel
yaacov.ritov@huji.ac.il

September 28, 2006

Abstract

Let $Z^i = (Y^i, X_1^i, ..., X_m^i)$, $i = 1, ..., n$, be i.i.d. random vectors, $Z^i \sim F$, $F \in \mathcal{F}$. It is desired to predict Y by $\sum \beta_j X_j$, where $(\beta_1, ..., \beta_m) \in B^n \subseteq R^m$, under a prediction loss. Suppose that $m = n^\alpha$, $\alpha > 1$, i.e., there are many more explanatory variables than observations. We consider sets B^n restricted by the maximal number of non-zero coefficients of their members, or by their l_1 radius. We study the following asymptotic question: How 'large' may the set B^n be, so that it is still possible to select empirically a predictor whose risk under F is close to that of the best predictor in the set. Sharp bounds for orders of magnitudes are given under various assumptions on $\mathcal{F}$. Algorithmic complexity of the ensuing procedures is also studied. The main message of this paper and the implications of the above derived orders are that under various sparsity assumptions on the optimal predictor there is "asymptotically no harm" in introducing many more explanatory variables than observations. Furthermore, such practice can be beneficial in comparison with a procedure that screens in advance a small subset of explanatory variables. Another main result is that 'Lasso'-type procedures, i.e., optimization under l_1 constraint, could be efficient in finding optimal sparse predictors in high dimensions.

Running head: Persistence and Predictor Selection. Key words: Consistency, Lasso, regression, variable selection.

1 Introduction

In practice, when modeling statistical phenomena, we tend to adopt more flexible models (e.g., with more parameters) as we get more observations. This practice suggests studying asymptotics of triangular arrays, i.e., when the model assumed for the observations $Z^1, ..., Z^n$ depends on n. Yet triangular array formulation is hardly studied in statistics. The standard mathematical statistical paradigm is the existence of a "true" model, and behavior of estimators is studied as the number of observations increases, while the model is kept fixed. We do not adopt this paradigm. We consider the problem of predictor selection in a given complex situation, and not that of estimation of a metaphysical unknown parameter, which may exist, but may not exist. In fact, the definitions of the predictor and of the parameter are intimately tied. Our parameter of interest is the best predictor out of a restricted class of potential predictors. A triangular array formulation is natural to our approach.

Consider now the setting of linear predictors in a triangular array. For simplicity, we will denote $Z_n^1, ..., Z_n^n$ of a triangular array simply as $Z^1, ..., Z^n$. Our study is dedicated to the case where the collection $\mathcal{F}^n$ is of distributions of $m+1$ dimensional i.i.d. vectors $Z^i = (Y^i, X_1^i, ..., X_m^i)$, $i = 1, ..., n$, where $m = n^\alpha$, $\alpha \geq 1$. The set of predictors, i.e., the set $\{g_\beta\ ; \beta \in R^m\}$ of functions of the explanatory variables, is of the form $g_\beta = g_\beta(X_1, ..., X_m) = \sum_{i=1}^m \beta_j X_j$ where β ranges over all m dimensional vectors. Denote:

$$L_F(\beta) = E_F(Y - \sum_{j=1}^m \beta_j X_j)^2. \tag{1}$$

The set of all possible predictors is too large for estimation. Minimization of the empirical analogue of (1) is essentially unrelated to the minimization of (1) itself. We will search for natural subsets $B^n \subset R^m$, so that the task of selecting (nearly) an optimal predictor from B^n is not too ambitious, and can be done empirically. It is, of course, desired that those sets will be as large as possible to include better predictors. Finally, the procedures that search for a predictor, i.e., the "estimation procedures", should be feasible in term of their algorithmic complexity.

In our setting a sequence of predictor selection procedures becomes a basic object. Given a set of predictors B^n and a distribution F_n, let $\beta^*_{F_n} = \arg\min_{\beta \in B^n} L_{F_n}(\beta)$.

Definition 1 *Given a sequence of sets of predictors B^n, the sequence of* *procedures $\hat{\beta}^n$, is called persistent, if for every sequence $F_n \in \mathcal{F}^n$:*

$$L_{F_n}(\hat{\beta}^n) - L_{F_n}(\beta^*_{F_n}) \xrightarrow{\text{p}} 0.$$

Remark 1 In the definition of persistence, we consider the distance between $L_{F_n}(\hat{\beta}^n)$ and $L_{F_n}(\beta^*_{F_n})$, rather than the more common l_2 distance between $\hat{\beta}^n$ and $\beta^*_{F_n}$. This is the more relevant distance to study in predictor selection. For example, we do not have to worry about colinearity. A consistent estimation of the parameter β is impossible, unless we assume that the matrix of the explanatory variables is not close to singularity.

Remark 2 The persistence criterion should have an appeal in particular in situations where $L_{F_n}(\beta^*_{F_n})$ does not approach 0. When $L_{F_n}(\beta^*_{F_n})$ could approach 0, a more delicate asymptotic study of rates of convergence, etc., becomes relevant. Yet in most situations and models (nearly) perfect prediction is impossible, thus convergence to 0 of $L_{F_n}(\beta^*_{F_n})$ does not hold.

A study of consistency (in the conventional sense of l_2 distance) in a triangular array setting in regression problems was conducted by Huber (1973) and by Portnoy (1984) (see also references there). They studied the problem of coefficients estimation under the setup $Y_i = \sum \beta_j X_{ij} + \epsilon_i$, where ϵ_i are i.i.d., $E\epsilon_i = 0$, and $m = m(n)$ increases with n. Their setup is more conventional than ours since they, unlike us, assume the linear model and study cases where $m(n) < n$. Major differences between our work and theirs is that they were concerned with robustness and M-estimators under heavy tailed distributions of ϵ_i, unlike us, and we consider random explanatory variables, unlike them. The motivation of these papers seems to be the same as ours, i.e., explore the limits to increasing the parameter set as the number of observations is increased. Yet by their approach, the number of explanatory variables m is taken, as is customary, to be less than n. In fact, it is shown in Huber (1973), under conditions on ϵ_i and the design matrix X, that consistency may be achieved as long as $m(n) = o(\sqrt{n})$. Portnoy, under further conditions, established consistency as long as $m(n) = o(n/log(n))$. Notice the huge gap!

Given observations $Z^1, ..., Z^n$, denote their empirical distribution by $\hat{F}$ and let

$$L_{\hat{F}}(\beta) = n^{-1}\sum_{i=1}^{n}(Y^i - \sum_{j=1}^{m}\beta_j X^i_j)^2.$$

Consider predictor selection methods of the following type. For a $c = c(n)$, choose:

$$\hat{\beta}^n = \arg\min_{\beta} L_{\hat{F}}(\beta) + c(n)||\beta||_1^2$$

Here $||\beta||_1$ is the l_1 norm of β. A related type of method is: for a $b = b(n)$, let

$$\hat{\beta}^n = \underset{\{\beta \mid ||\beta||_1 \leq b(n)\}}{\arg\min} L_{\hat{F}}(\beta).$$

We consider these procedures as Lasso-like methods.

These procedures motivated the study in this paper. In Tibshirani (1996) those procedures were introduced and were named Lasso. In that paper a heuristic and numerical study is conducted to find the appropriate $c(n)$ and $b(n)$ for such procedures. In Juditsky and Nemirovski (2000), properties of such procedures with $b(n) \equiv 1$ are studied. Yet the value 1 for $b(n)$ is chosen somewhat arbitrarily. Lee, Bartlett, and Williamson (1996) studied similar procedures for estimating parameters in neural network, and they also concentrated on $b(n) = 1$. In Chen, Donoho and Saunders (2001), in the context of denoising a signal represented by an overcomplete wavelet system, an analog of a Lasso-like procedure is suggested, see their equation (5.1). They also discuss the choice of $c(n)$ (choice of λ in their setup). An overcomplete system defines over-parametrization in our terminology.

Two types of sets, $B^n \subset R^m$ of possible predictors, are studied in this paper.

1. B^n is the set of all vectors $(\beta_1, ..., \beta_m)$ having at most $k = k(n)$ non-zero entries. Those are 'model-selection' or 'variable-selection' type of procedures that choose k explanatory variables out of the initial set of m variables. These sets will be denoted B^n_k

2. B^n is the set of all vectors $(\beta_1, ..., \beta_m)$ having l_1 norm less than or equal to $b = b(n)$. These sets will be denoted B^n_b.

We will explore the interplay between B^n_k and B^n_b. The first type, 'model-selection' is of interest as problems of variable-selection have long been studied from various aspects in numerous papers. The second type is of interest because of its relation to Lasso-like methods.

In Section 2 of this paper we will motivate the Lasso-like procedures. We also present an argument that suggests that the proper values of $c(n)$ and $b(n)$ are $c(n) = o((log(n)/n)^{1/2})$ and $b(n) = o((n/log(n))^{1/4})$ respectively. A careful study reveals that, in settings like the multivariate normal, this is not the case. In fact, from the results in Section 4 it follows that, when Z^i are multivariate normal, the values of $c(n)$ and $b(n)$ should be of the order of $o(log(n)/n)$ and $o((n/log(n))^{1/2})$. In Section 3 we will show persistence with respect to $B^n_{k(n)}$, for $k(n) = o(n/\log(n))$. Optimality of the last rate is proved, i.e., there exist no persistent procedures with respect to sets $B^n_{k(n)}$, when $k(n) = O(n/\log(n))$.

The persistent procedures in Section 3 are algorithmically inefficient: they involve searching over all the subsets of size of order $n/\log(n)$ out of the m explanatory variables. Additional assumptions, in Section 4, yield persistent and algorithmically efficient procedures with respect to B^n_k for $k(n) = o(n/\log(n))$.

The implications of the study of the above rates is the following. Consider a triangular array; suppose it is known that $\beta^*_{F_n}$, the (nearly) optimal predictor under F_n, has less than $k'(n)$ non-zero coefficients, alternatively, suppose that it is known that $||\beta^*_{F_n}||_1 \leq b'(n)$. We will say that the k-sparsity rate and the b-sparsity rate are respectively $k'(n)$ and $b'(n)$. Suppose now that there exist persistent procedures with respect to sets $B^n_{k(n)}$ (sets $B^n_{b(n)}$), where $k(n) > k'(n)$ (where $b(n) > b'(n)$). Then, there is 'asymptotically' no virtue in screening in advance smaller subsets of explanatory variables. This follows since the "persistence rates" $k(n)$ and $b(n)$ imply that by doing so we will not (significantly) improve on procedures that search through the entire set of explanatory variables. Yet obviously, when screening a small subset in advance, we may do harm by dropping potentially important variables.

In practice persistence rates and sparsity rates are not known. The practical way to act is to test estimators resulting from various assumptions about the persistence rates (e.g., resulting from various constraints $b(n)$ in the Lasso procedure) on a test set.

Thus, the importance of our study stems from its suggestion to turn to high dimensions, and pointing out that often there is 'no harm' in doing so.

In many cases it turns out that persistence rates are $k(n) = o(n/\log(n))$ when $m = n^\alpha$. Such cases are presented in more general prediction problems in a subsequent paper Greenshtein (work in progress).

Finally, a practical implication of this paper is its recommendation of the Lasso procedure in high dimensions as an effective method to find optimal predictors under sparsity conditions.

2 Motivating and exploring the Lasso-like methods

Consider a triangular array, where $Z = (Y, X_1, ..., X_{m(n)}) \sim F$, $F \in \mathcal{F}^n$. Denote $X_0 = Y$. We think of Y as a response variable and of X_j as explanatory variables. For any linear predictor, associated with a vector $(\beta_1, ..., \beta_m)$, denote

$$\gamma' = (-1, \beta_1, ..., \beta_m) = (\beta_0, ..., \beta_m).$$

Denote

$$\begin{aligned} L_F(\beta) &= E_F(Y - \sum_{j=1}^{m} X_j \beta_j)^2 \\ &= \gamma' \Sigma_F \gamma \end{aligned}$$

Here $\Sigma_F = (\sigma_{ij})$, $\sigma_{ij} = E_F X_i X_j$ $0 \leq i, j \leq m$.

We think of a sequence of problems where n observations, $Z^1, ..., Z^n$, are given and $m = n^\alpha$, $\alpha > 1$.

Let $\hat{F}_n$ be the empirical distribution determined by the sample $Z_1, ..., Z_n$. Note that

$$L_{\hat{F}_n}(\beta) = \gamma' \Sigma_{\hat{F}_n} \gamma,$$

where $\Sigma_{\hat{F}_n} = (\hat{\sigma}_{ij})$ and $\hat{\sigma}_{ij} = n^{-1} \sum_{k=1}^{n} X_i^k X_j^k$.

Denote $\hat{\sigma}_{ij} = \sigma_{ij} + \epsilon_{ij}^n$, then $\hat{\Sigma} = \Sigma_F + E$, where $E = (\epsilon_{ij}^n)$. Let $Y_{ij} = X_i X_j$. We assume the following condition:

A1: Under the distributions in $\mathcal{F}^n$, the random variables $Y_{ij} = X_i X_j$ have bounded variances and moment-generating functions with bounded third derivative in the neighborhood of 0.

Under the above condition, we have for large enough A, depending on the bounds in condition **A1**,

$$\sup_{F_n \in \mathcal{F}^n} P_{F_n}(-\sqrt{\frac{A \log(n)}{n}} \leq \epsilon_{ij}^n \leq \sqrt{\frac{A \log(n)}{n}} \; \forall i, j) \to 1; \qquad (2)$$

(2) follows by Bonferroni, since for large enough A, and for any pair i, j,

$$\sup_{F_n \in \mathcal{F}^n} P_{F_n}(-\sqrt{\frac{A \log(n)}{n}} \leq \epsilon_{ij}^n \leq \sqrt{\frac{A \log(n)}{n}}) = 1 - o(m^{-2}).$$

The last equality is by moderate deviation principle as in Billingsley (1995) p. 153. The uniformity in $\mathcal{F}^n$ follows from the uniform boundness of the third derivative; see such an argument in Lemma 2.2 of Breiman and Freedman (1983).

Denote by $\hat{E}$ the matrix with identical entries equal to $\sqrt{An^{-1} \log(n)}$. Then (2) implies:

$$\sup_{F_n \in \mathcal{F}^n} P_{F_n}(L_{F_n}(\beta) \leq \gamma' \Sigma_{\hat{F}_n} \gamma \; + \; |\gamma|' \hat{E} |\gamma| \; \forall \beta \in R^{m+1}) \to 1, \qquad (3)$$

where $|\gamma| = (1, |\beta_1|, ..., |\beta_m|)$.

Equation (3) suggests the following method for selecting a predictor. Select the predictor $\hat{\beta}$ where

$$(-1, \hat{\beta}) = \underset{(\gamma \in R^{m+1}; \beta_0 = -1)}{\arg\min} \gamma' \Sigma_{\hat{F}_n} \gamma + |\gamma|' \hat{E} |\gamma|.$$

Equivalently, write

$$(-1, \hat{\beta}) = \underset{(\gamma \in R^{m+1}; \beta_0 = -1)}{\arg\min} \gamma' \hat{\Sigma}_F \gamma + c(n) ||\gamma||_1^2, \quad (4)$$

which may be rephrased as optimization of a convex function in a convex domain. Note that in the last equation $c(n) = O(\sqrt{\log(n)/n})$.

Here and throughout, we consider procedures that use the appropriate values of $c(n)$, $b(n)$, etc. In practice the appropriate values are not known, and one should try various values and test the resulting estimators on a test set.

We will now summarize our findings on persistence of procedures of the type

$$\hat{\beta}^n = \underset{\{\beta:\ ||\beta||_1 \leq b(n)\}}{\arg\min} L_{\hat{F}}(\beta). \quad (5)$$

Theorem 1 *Under Assumption* **A1** *on $\mathcal{F}^n$, for any sequence $B_{b(n)}^n \subset R^m$, where $B_{b(n)}^n$ consists of all vectors with l_1 norm less than $b(n) = o((n/\log(n))^{1/4})$, there exists a persistent sequence of procedures. A concrete persistent sequence of procedures is given in (5).*

Proof. As in (3), $\sup_{F_n \in \mathcal{F}^n} \sup_{\beta \in B_{b(n)}^n} P_{F_n}(|L_{F_n}(\beta) - L_{\hat{F}_n}(\beta)| < |\gamma|' \hat{E} |\gamma|) \to 1$. Now, for sequences of vectors β of order $b(n) = o((n/\log(n))^{1/4})$, the corresponding sequence $|\gamma|' \hat{E} |\gamma|$ approaches 0. The result now follows immediately from the definition of persistence. □

Suppose in addition that the following condition holds.

A2: Let $B_{k(n)}^n$ be the set of all vectors with $k(n) = o((n/\log(n))^{1/2})$ non-zero entries. There exists a constant C, $C < \infty$, such that $|| \arg\min_{\beta \in B_{k(n)}^n} L_{F_n}(\beta) ||_2 < C$ for any sequence $F_1, F_2, \ldots, F_n \in \mathcal{F}$.

Remark 3 When $E_F Y^2$ is bounded, **A2** follows whenever the minimal eigenvalue of the covariance matrix of the explanatory variables is bounded from below. As pointed out in Remark 1, assumptions about minimal eigenvalues and near singularity of the random matrix X are essential when dealing with persistence in the conventional sense, i.e., when dealing with consistency.

Note that the range of the procedures achieving persistence need not be within the variable selection sets. It is a matter of formalism, but such a requirement was not part of the definition; i.e., an estimator $\tilde{\beta}_n$ may be persistent with respect to a set B^n, while $\tilde{\beta}_n \notin B^n$ for some n. We will use this fact in the proof of the following theorem, where sequences with range outside the B^n will be considered. Yet as shown in Section 4, these procedures may be adjusted so that their range will be within B^n_k.

Theorem 2 *Suppose Assumptions* **A1, A2** *hold. There exists a persistent sequence of procedures with respect to the sets* $B^n_{k(n)}$ *with* $k(n) = o((n/\log(n))^{1/2})$.

Proof. We consider the particular sequence of procedures which is defined by (5). By Assumption A2, we can consider only vectors β with l_2 norm bounded by, say, $C < \infty$. However, any vector with l_2 norm c and of dimension $k(n)$ has l_1 norm less than or equal to $b(n) = c\sqrt{k(n)}$. It follows from Theorem 1 that the estimator defined in (5), with $b(n)$ as above, is persistent with respect to the larger set $B^b_{b(n)}$ hence also for $B^n_{k(n)}$. □

The persistence rate in Theorem 1 is implied also by the following assumption **A3**, alternative to **A1**.

A3 There are finite constants C and L, such that under any $F \in \mathcal{F}^n$, $n = 1, 2, \ldots$: $E_F Y^2 < C$, and all $|X_j| < L$ w.p. 1, $j = 1, ..., m(n)$.

Theorem 3 *If Assumption* **A3** *holds, then for any sequence* $B^n_{b(n)} \subset R^m$, *where* $B^n_{b(n)}$ *consists of all vectors with* l_1 *norm less than* $b(n) = o((n/\log(n))^{1/4})$, *there exists a persistent sequence of procedures. A concrete persistent sequence of procedures is given in (5).*

Theorem 3 is implied by an adaptation of the results in Juditsky and Nemirovsky (2000). Condition **A3** is close to their setup. We will now describe their setup and explain their method and its adaptation to our purpose. We use their notations. Juditsky and Nemirovski study prediction, in a manner similar to ours, of a response variable y, based on a linear combination of given functions $f_1, ..., f_M$, where $f_j = f_j(x)$ are functions bounded by some L. They assume a model $y = f(x) + e$, where e and x are independent and $E(e) = 0$. Given n independent replicates (y_t, x_t), $t = 1, ..., n$, they study the problem of estimating the 'best' linear combination of $f_1, ..., f_m$ under the constraint that the l_1 norm of the vector of coefficients is 1. 'Best' is in terms of the L_2 distance between f and the function obtained by the linear combination. As in our problem, they study asymptotics when

$M = n^{\alpha},\ \alpha > 1$. This setting is very close to ours, when their $f_j(x)$ is identified with our X_j. As demonstrated in what follows, *their assumption, about independence of x_t and e_t, is not needed under our definition of persistence.* A definition of consistency according to their approach (consistency is not defined in their paper) would involve L_2 distance between f and the linear combination of f_j. Thus the class we handle in Theorem 3 is slightly larger than the class treated in their setup.

In the sequel we also formulate and prove the conclusion of Theorem 2 under such alternative conditions, stated as Theorem 4. The proofs of Theorems 3 and 4 are along the lines of the technique of Juditsky and Nemirovski.

A statement and a proof of the following key result that is needed may be found in Nemirovski (1998) p. 188.

Lemma 1 (Nemirovski's inequality:) *Let $\xi^t \in R^K$, $t = 1, ..., n$, be independent random vectors with zero means and finite variance, and $K \geq 3$. Then for every $p \in [2, \infty]$ one has:*

$$E||\sum_{t=1}^{n} \xi^t||_p^2 \ \leq \ O(1) \min[p, \log(K)] \sum_{t=1}^{n} E||\xi^t||_p^2,$$

where $||\ ||_p$ is the l_p norm.

We will use the inequality in the case $p = \infty$. There are related results in empirical processes which bound the expectation of the maximum of a finite sequence of random variables. Yet we do not know of a result that can replace the above inequality in the context that we need.

Consider the matrix $(\Sigma_{\hat{F}} - \Sigma_F)$ as an $(m+1)^2$ dimensional vector. Write $(\Sigma_{\hat{F}} - \Sigma_F)$ as $\sum_{t=1}^{n} \xi^t$, where

$$\xi^t = \frac{1}{n}(X_0^t X_0^t - E X_0^t X_0^t,\ X_0^t X_1^t - E X_0^t X_1^t, ...)$$

is an $(m+1)^2$ dimensional vector. Suppose there is an envelope function with respect to $X_i X_j$ $0 \leq i, j \leq m$ with a second moment, i.e., $E(\max_{i,j} X_i X_j)^2 < \infty$. Then, we get by Nemirovski's inequality that the expected value of the l_∞ norm of $(\Sigma_{\hat{F}} - \Sigma_F)$ satisfies:

$$E||\sum \xi^t||_\infty = O\Big(\sqrt{\frac{\log(n)}{n}}\Big).$$

Consider now B_b^n with $b = b(n) = o((n/\log(n))^{1/4})$. For $\beta \in B_b^n$, by the last inequality and by Markov inequality, for $\gamma^t = (-1, \beta_1, ..., \beta_m)$ we get: $|\gamma^t(\Sigma_{\hat{F}_n} - \Sigma_{F_n})\gamma| \xrightarrow{\text{p}} 0$; equivalently, we get $|L_{\hat{F}_n}(\beta) - L_{F_n}(\beta)| \xrightarrow{\text{p}} 0$.

Consequently, persistent procedures, relative to sets $B_{b(n)}^n$, of predictors β with l_1 norm less than $b(n) = o((n/\log(n))^{1/4})$ exist. Now under the assumption **A2** and by the Cauchy-Schwartz inequality, a persistent selection relative to sets $B_{k(n)}^n$ with $k(n) = o((n/\log(n))^{\frac{1}{2}})$ is also possible.

Remark 4 An envelope function with a second moment for the collection X_iX_j $0 \leq i, j \leq m$ exists in our triangular array setting if all but fixed number of X_j, $j = 0, ..., m$ are bounded by some L, and all of them have second moment. In particular when $X_0 \equiv Y$ has a bounded second moment and X_j, $j = 1, ..., m$ are bounded as in Theorem 3. Thus Theorem 3 is obtained as a corollary.

The following Theorem 4 is obtained from Theorem 3, in the same manner that Theorem 2 follows from Theorem 1.

Theorem 4 *Suppose that the set X_iX_j $0 \leq i, j \leq m$ has an envelope function with a bounded second moment under $F_n \in \mathcal{F}^n$ $n = 1, 2, ...$ Suppose that condition* **A2** *holds. Then there exists a method which is persistent with respect to 'variable-selection' sets, B_k^n, with $k(n) = o((n/\log(n))^{\frac{1}{2}})$.*

Theorems 2 and 4 are obtained as immediate corollaries of Theorems 1 and 3 respectively, when assuming boundedness of $||\beta_{F_n}^*||_2$. With some more effort Theorem 2 may be strengthened and a more flexible condition may replace the one in Theorem 4. In fact, under boundedness of $||\beta_{F_n}^*||_2$, a sufficient condition that implies the $k(n) = o((n/\log(n))^{1/2})$ rate is that $E_{F_n}X_j^{2+\delta}$ $n = 1, 2, ...,$ $j = 0, ..., m(n)$ is bounded for some $\delta > 0$. To show it, one should apply truncation and diagonalization, as in the Section 4 in the sequel.

Summary. We established the existence of persistent procedures, under various assumptions, when $b(n)$ and $k(n)$ are of orders $o((n/\log(n))^{1/4})$ and $o((n/\log(n))^{1/2})$ respectively. The methods of proofs were based on bounding l_∞ distance between Σ_F and $\Sigma_{\hat{F}}$.

In the sequel, using different methods, we will explore conditions under which $b(n)$ and $k(n)$ may be 'pushed' towards the rates $o((n/\log(n))^{1/2})$ and $o((n/\log(n)))$ respectively. Those rates are optimal in a sense, as will follow below. Compare the huge gap we get, under various conditions, for the rates of $k(n)$, with the differences, mentioned in the Introduction, between the rates derived by Huber and those derived by Portnoy.

As mentioned, in a subsequent paper by Greenshtein the $o(n/\log(n))$ rate for $k(n)$ is shown to hold in general triangular arrays, extending linear prediction under a squared prediction loss. Still, we do not know whether the lower rates, obtained in this section, may be improved even under the elementary assumption that the entries of Z^i are bounded. We state the problem in the following.

1. Consider the case where $\mathcal{F}^n$ consists of all the distributions under which the entries of $Z = (Y, X_1, ..., X_m)$ are bounded. Does a procedure exists that is persistent with respect to sets B_b^n, with l_1 radius $b(n)$ which is not $o(n/\log(n))^{1/4}$?

2. Assume that $\mathcal{F}^n$ consist of all distributions under which the entries of Z are bounded. Does a procedure exist which is persistent with respect to sets B_k^n, for $k(n)$ which is not $o(n/\log(n))^{1/2}$?

3 Persistence of model-selection type of procedures. The normal case.

In this section we will study persistence of model-selection type of procedures, assuming that $\mathcal{F}^n$ consist of multivariate normal distributions. Those procedures select at the first stage a model, i.e., a subset of $k(n)$ explanatory variables, and then choose a linear predictor based on those variables. Persistence of such procedures is studied with respect to the sets B_k^n that correspond to vectors that have at most $k(n)$ non-zero entries. The question is how far we may push $k(n)$ and still achieve persistence.

Prediction when the explanatory variables are multivariate normal, and there are many more explanatory variables than observations, was studied by Bickel and Levina (2003). Yet, they predict Y which is 0 or 1, i.e., they study classification in this setting.

Denote the collection of all subsets, of size $k = k(n)$, of explanatory variables by $\mathcal{K} = \mathcal{K}_n$; each of its members is denoted by K, $K \in \mathcal{K}$. Let $\hat{\beta}(K)$ be the least squares estimator based on the subset K of explanatory variables, and let

$$\hat{\beta} = \arg\min_{K \in \mathcal{K}} L_{\hat{F}}(\hat{\beta}(K)). \tag{6}$$

Similarly, let $\beta_F^*(K)$ be the best linear predictor based on the subset K of the explanatory variables, under F, and $\beta_F^* = \arg\min_{K \in \mathcal{K}} L_F(\beta_F^*(K))$.

The following is assumed throughout this section

B The sets $\mathcal{F}^n$ consist of all multivariate normal distributions with uniformly bounded variance of Y.

The main result of this section is

Theorem 5 *Suppose $k(n) = o(n/\log(n))$; then there exists a persistent sequence of procedures with respect to the corresponding B_k^n.*

The procedure presented in the proof of Theorem 5 involves searching over all the subsets of size $k(n)$ of the m explanatory variables. In Section 4 we will consider procedures with a lower complexity, which are persistent under a more restricted version of Assumption **B**.

The proof of the Theorem will preceded by the following lemmas and propositions.

Proposition 1 *Suppose $V_n \sim \chi^2_{k_n}$ where $k_n \le \alpha n$, $0 < \alpha < 1$. Then $P(V_n > n) = o(\exp(-\gamma n))$ for some $\gamma > 0$.*

Proof. Since V_n has $\Gamma(k_n/2, 2)$ distribution, its Lebesgue density is given by

$$f(x) = \frac{1}{\Gamma(k_n/2)2^{k_n/2}} x^{k_n/2-1} e^{-x/2}.$$

In particular, $f(x) = o(1)e^{-(1-\alpha')x/2}$ on (n, ∞) for any $1 > \alpha' > \alpha - \alpha \log \alpha$. The proposition follows. □

Let $A_n^\epsilon(K)$ be the event $|L_{\hat{F}_n}(\hat{\beta}(K)) - L_{\hat{F}_n}(\beta^*_{F_n}(K))| > \epsilon$, and denote by $B_n^\epsilon(K)$ the event $|L_{\hat{F}_n}(\beta^*_{F_n}(K)) - L_{F_n}(\beta^*_{F_n}(K))| > \epsilon$.

Lemma 2 *There is $\gamma_1 > 0$ such that for any non-random $K \in \mathcal{K}$, $\sup_{F_n \in \mathcal{F}^n} P_{F_n}(A_n^\epsilon(K) \cup B_n^\epsilon(K)) = o(\exp(-\gamma_1 n))$.*

Proof. The proof follows since the probability of both $A_n^\epsilon(K)$ and $B_n^\epsilon(K)$ approach 0 exponentially fast: for $A_n^\epsilon(K)$ observe that $n \times (L_{\hat{F}_n}(\hat{\beta}(K)) - L_{\hat{F}_n}(\beta^*_{F_n}(K)))$ is distributed χ^2 with k degrees of freedom and apply Proposition 1. For $B_n^\epsilon(K)$ apply large deviation principle for the difference between the random mean and its expectation. □

The number of elements in $\mathcal{K}$ is of order m^k, $k = k(n)$, hence if $m^k \exp(-\gamma_1 n) \to 0$ for some $\gamma_1 > 0$, then we get by Bonferroni

$$\sup_{F_n \in \mathcal{F}^n} P_{F_n}(\cup_{K \in \mathcal{K}_n}(A_n^\epsilon(K) \cup B_n^\epsilon(K)) \to 0.$$

If $k(n) = \delta n/\log(n)$ for δ small enough, then $m^k \exp(-\gamma_1 n) \to 0$. Thus we get

Corollary 1 *If $k(n) = \delta n/\log(n)$, then for small enough δ*

$$\sup_{F_n \in \mathcal{F}^n} P_{F_n}(|L_{\hat{F}_n}(\hat{\beta}) - L_{F_n}(\beta^*_{F_n})| > \epsilon) \to 0.$$

The last equation establishes that $L_{\hat{F}_n}(\hat{\beta})$ is a consistent estimator for $L_{F_n}(\beta^*_{F_n})$. It does not, however, imply that, $\hat{\beta}$ is a persistent estimator for $\beta^*_{F_n}$. Recall that for the latter it is necessary that for every $\epsilon > 0$,

$$\sup_{F_n \in \mathcal{F}^n} P_{F_n}(|L_{F_n}(\hat{\beta}) - L_{F_n}(\beta^*_{F_n})| > \epsilon) \to 0. \tag{7}$$

To obtain (7), hence to prove Theorem 5, we need the following lemma and its corollary.

Lemma 3 *Suppose $k(n) = o(n)$. Then for any fixed $K \in \mathcal{K}$ and $\epsilon > 0$, there exists $\gamma > 0$ such that $\sup_{F_n \in \mathcal{F}^n} P_{F_n}\Big(L_{F_n}(\hat{\beta}(K)) - L_{F_n}(\beta^*_{F_n}(K)) > \epsilon\Big) = o(\exp(-\gamma n))$*

Proof. We consider a concrete subset K with indices (say) 1,2,...,k, and a concrete F_n. We will omit the index n when there is no ambiguity. Note that for such a concrete subset we may assume, w.l.o.g., that (i) $\beta^*_F(K) = 0$, (ii) the r.v. $X_1, ..., X_k$ are i.i.d distributed $N(0,1)$. Assumption (ii) is possible thanks to our definition of persistence in which we consider $L_F(\hat{\beta}) - L_F(\beta^*_F)$ rather than $||\hat{\beta} - \beta^*_F||_2^2$, so the problem is invariant under linear transformation of the explanantory variables. Now $L_F(\hat{\beta}(K)) - L_F(\beta^*_F(K)) = E((W\hat{\beta}(K))^2|\hat{\beta}(K))$; the random vector W is k dimensional and consists of i.i.d. $N(0,1)$ entries which are independent of $\hat{\beta}(K)$; W may be thought of as the explanatory variables in the subset of a future observation. Thus $E((W\hat{\beta}(K))^2|\hat{\beta}(K)) = ||\hat{\beta}(K)||^2$. Let $X(K)$ be the random design matrix, corresponding to the subset of explanatory variables, obtained by the n observations. Then $\hat{\beta}(K) \sim N(0, \sigma_K^2(X(K)'X(K))^{-1})$, w.l.o.g. $\sigma_K^2 = 1$. Hence $\hat{\beta}'(K)(X(K)'X(K))\hat{\beta}(K) \equiv V \sim \chi^2_{(k)}$. Let λ be the (random) minimal eigenvalue of $X(K)'X(K)$, then $V > ||\hat{\beta}(K)||^2\lambda$. Hence:

$$P(||\hat{\beta}(K)||^2 > \epsilon) \le P(\frac{V}{\lambda} > \epsilon) = P(\frac{V}{\lambda/n} > \epsilon n).$$

Now from Silverstein's (1985) proof about a.s. convergence of the minimal eigenvalue of a Wishart matrix, for any $0 < a < 1$ there exists $\gamma > 0$ such that

$$P(\frac{\lambda}{n} < a) = o(\exp(-\gamma n)).$$

Also, since $k = o(n)$ and $V \sim \chi^2_{(k)}$ as in Proposition 1 we have

$$P(V > a\epsilon n) = o(\exp(-\gamma n)),$$

for some $\gamma > 0$. Combining the last two equations we obtain

$$P(||\hat{\beta}(K)||^2 > \epsilon) = o(\exp(-\gamma n))$$

for $\gamma > 0$. The proof now follows. □

Corollary 2 *Suppose $k(n) = o(n/\log(n))$, then*

$$\sup_{F_n \in \mathcal{F}^n} P_{F_n}\Big(\cup_{K \in \mathcal{K}} [L_{F_n}(\hat{\beta}^n(K)) - L_{F_n}(\beta^*_{F_n}(K)) > \epsilon]\Big) \to 0.$$

Proof of Theorem 5: The proof follows from corollaries 1 and 2.

We show now an optimality property of the suggested procedure. It is shown that persistence cannot be achieved under **B** if $k(n)$ is of order $n/\log(n)$.

Theorem 6 *Suppose that $m = n^\alpha,\ \alpha > 1$. If $k(n) > c(n/\log(n))\ \ c > 0$, then there exists no procedure which is persistent with respect to the corresponding B^n_k.*

Proof. We start the proof by stating Fano inequality (See LeCam and Yang (1990) p. 128). Let $K(P, Q)$ be the Kullback Leibler distance between P and Q and let $J(P, Q) = K(P, Q)+K(Q, P)$. Suppose $X \sim F$, $F \in \{F_1, ..., F_M\}$, and $M > 2$; consider the problem of estimating F, based on X, under a 0-1 loss function. Then the minimax risk is at least

$$1 - \frac{1}{\log(M-1)}[\log(2) + \frac{1}{2}\max_{i,j} J(F_i, F_j)].$$

Let $Z = (Y, X_1, ..., X_m)$, where X_i are i.i.d. $N(0, 1)$. For any subset $X_{i_1}, ..., X_{i_k}$ of size k, of the explanatory variables, consider the joint distribution of Z determined by $Y = \frac{c_1}{\sqrt{k}} \sum_{j=1}^k X_{i_j} + \epsilon$; here $\epsilon \sim N(0, 1)$ is independent of $X_j,\ i = 1, ..., m$ and c_1 is a small enough properly chosen constant.

Among all subsets of size k, choose M such subsets in the following way. At each stage after choosing a subset, 'delete' all the "neighboring" subsets, having more than $k/2$ common indices with that subset, then choose the next subset from the remaining ones; keep on selecting subsets according to that procedure until all the subsets of size k, are either deleted or chosen. There are M chosen subsets at the end of the process, with corresponding M distributions. Denote the distributions by $F_1, ..., F_M$. Given n i.i.d., observations $Z^1, ..., Z^n$, the relevant distributions are the product measures $F_1^{(n)}, ..., F_M^{(n)}$. Now, note that for the distributions $F_i,\ i = 1, ..., M,\ J(F_i, F_j) = O(1)$, which may be made arbitrarily small by choosing small enough c_1; thus, $J(F_i^{(n)}, F_j^{(n)}) < c_3 n$, for c_3 that may be made arbitrarily small when choosing small enough c_1. By construction $L_{F_i}(\beta^*_{F_j}) > L_{F_i}(\beta^*_{F_i}) + c_2$ for a small enough constant c_2 when $i \neq j$.

We now approximate the term $\log(M-1)$ that appears in Fano inequality. At each stage we delete "neighboring" subsets, having at least $k/2$ common indices with the subset that was chosen at this stage, until all subsets are either deleted or chosen. The number of subsets of size k is of order m^k. The number of deleted subsets at each stage is of order $n^{\alpha' k},\ \alpha' < \alpha$. Thus, the number of stages, or, equivalently the number of chosen subsets, M, is:

$$M \approx m^k / n^{\alpha' k} \approx n^{(\alpha - \alpha')k} = \exp(\log(n)[\alpha - \alpha'] \frac{n}{\log(n)}).$$

Thus $\log(M) > c_4 n$ for small enough c_4. Applying Fano inequality, we get the desired result. □

Remark 5 For the case where the explanatory variables are non-random, related results are the 'Oracle inequality', Theorem 3 of Donoho and Johnstone (1994), and Lemma A.2 by Foster and George (1994). Those results give finer inequalities than needed for the proof of Theorem 6, in the case of orthogonal and non-random explanatory variables.

It seems that those results may be adjusted for our case of random explanatory variables, and yield the conclusion of Theorem 6 even for the case $\alpha = 1$. Yet the main interest in this paper is the case $\alpha > 1$, i.e., more explanatory variables than observations. Thus, our relatively simple argument, using Fano inequality seems worthwhile. Another advantage of our proof is that it does not rely on normality; it uses general properties of K-L numbers. Thus, this method of proof indicates that the $k(n) = o(n/\log(n))$ cannot be improved in typical situations.

4 Complexity of persistent procedures

The persistent procedure suggested in Section 3 has high complexity. It involves searching through all subsets of size $k(n) = o(n/\log(n))$ out of the m explanatory variables. Under further restriction on the triangular array, we will show, in this section, the existence of 'low complexity' procedures. The complexity of these procedures is essentially the same as that of solving a Lasso-like problem. The Lasso method involves optimization of a convex target function subject to convex constraints. Such convex optimization problems have efficient algorithms in general, see Nemirovski and Yudin (1983). For the particular Lasso method, an efficient computation algorithm was recently developed by Efron, Johnstone, Hastie, and Tibshirani (2003).

A key Lemma is the following Lemma 4. A proof under a slightly different setting is given in Juditsky and Nemirovski in their Proposition 2.2, and is attributed to B. Maurey. We give the proof here since there is a slight difference in the formulation, but mainly for being self contained.

Lemma 4 *Let $Z = (Y, X_1, ..., X_m)$, $Z \sim F$, be a random vector. Suppose $E_F Y^2 < \infty$; suppose further that $|X_j| < c$, $j = 1, ..., m$ w.p.1. Then for any predictor β with l_1 norm ν, there exists a corresponding predictor β' such that β' has k non-zero entries or less, and $L_F(\beta') < L_F(\beta) + \frac{c^2\nu^2}{k}$.*

Proof. Assume first that the entries β_j of β are positive. Denote $p_j = \beta_j/\nu$, $j = 1, ..., m$. Now consider a randomization of k trials in a multinomial setting with m categories, where the probability of category j is p_j, $j = 1, ..., m$. Let $\hat{P}^j$ be the fraction of the k trials whose outcome is in category j, $j = 1, ..., m$. Denote $\hat{P} = (\hat{P}_1, ..., \hat{P}_m)$. Note that the vector $\hat{P}$ has at most k non-zero entries. We will show that:

$$E_F L_F(\nu\hat{P}) \leq L_F(\beta) + \frac{\nu^2 c^2}{k},$$

the proof then follows.

Let $Z = (Y, X_1, ..., X_m)$ be independent of $\hat{P}$, in the following the expectation operator E is taken with respect to both $\hat{P}$ and Z.

$$\begin{aligned}
EL_F(\nu\hat{P}) &= E(Y - \sum \nu \hat{P}_j X_j)^2 \\
&= E(Y - \sum \nu p_j X_j + \sum \nu p_j X_j - \sum \nu \hat{P}_j X_j)^2 \\
&= E(Y - \sum \nu p_j X_j)^2 + E[\sum \nu X_j (p_j - \hat{P}_j)]^2 \\
&\qquad + 2E(Y - \sum \nu p_j X_j)(\sum X_j \nu (p_j - \hat{P}_j)) \\
&= L_F(\beta) + E[\sum \nu X_j (p_j - \hat{P}_j)]^2.
\end{aligned}$$

The last equality follows since $E(\hat{P}_j - p_j) = 0$ and since $\hat{P}$ and Z are independent. Now note that $cov(\hat{P}_l, \hat{P}_k) < 0$ for $k \neq l$ to obtain

$$\begin{aligned}
L_F(\beta) + E[\sum \nu X_j (p_j - \hat{P}_j]^2 &\leq L_F(\beta) + \nu^2 c^2 \sum var(\hat{P}_j) \\
&= L_F(\beta) + \nu^2 c^2 \sum \frac{p_j(1-p_j)}{k} \\
&\leq L_F(\beta) + \frac{\nu^2 c^2}{k}
\end{aligned}$$

The adaptation of the proof to the case where β_j may also be negative is straightforward.

□

Corollary 3 *Let $Z \sim F$, $E_F Y^2 < \infty$. Given $\epsilon > 0$ and β, let $c = c(\epsilon)$ be such that $E_F(Y - \sum \beta_j X_j)^2 - E_F(Y - \sum \beta_j \tilde{X}_j)^2 < \epsilon$, where $\tilde{X}_j = \max\{-c(\epsilon), \min\{X_j, c(\epsilon)\}\}$ is a truncation of X_j. Then there exists a corresponding predictor β' such that β' has k non-zero entries or less, and $L_F(\beta') < L_F(\beta) + \epsilon + c^2||\beta||_1^2/k$.*

For our main result in this section we will assume the following assumption about $\mathcal{F}^n$, which is more restrictive than condition **B**.

- **C:** Consider variable selection subsets B_k^n, with $k(n) = o(n/\log(n))$. Let $\kappa, C < \infty$. Assume for every n, $F \in \mathcal{F}^n$ if and only if F is multivariate normal distribution with second moments bounded by C, and $||\beta_F||_2 \leq \kappa$.

Theorem 7 *Suppose $\mathcal{F}^n$ satisfy condition* **C**. *Let B_k^n be the set of predictors with $k(n) = o(n/\log(n))$ non-zero entries. Then, there exists a sequence of procedures $\beta_n \in B_k^n$, $n = 1, 2, \ldots$, such that $\{\beta_n\}$ is persistent with respect to B_k^n, and the numerical complexity of calculating β_n is no more than the numerical complexity of Lasso plus an $O_p(m)$ term.*

The $O_p(m)$ term in the statement of Theorem 7, comes from extracting a vector β' with $o(n/\log(n))$ non-zero entries, from a vector β obtained by solving a Lasso problem. The extraction is in the manner described in the proof of Lemma 3. The strength of Theorem 5, compared with the results in Section 3, is in the lower complexity of the persistent procedures.

Proof. First we will show that for every $\epsilon > 0$ and n, there exists a $\tilde{\beta}_n = \tilde{\beta}_n(\epsilon)$, such that $\sup_{F_n \in \mathcal{F}^n} P_{F_n}(|L_{F_n}(\tilde{\beta}_n) - L_{F_n}(\beta^*_{F_n})| > \epsilon) \to 0$, where $\tilde{\beta}_n$ has $o(n/\log(n))$ non-zero coefficients. The result will then follow by a diagonalization argument: $\tilde{\beta}_n(\epsilon_n)$ will satisfy the theorem for $\epsilon_n \to 0$ slowly enough.

For a given ϵ, we will obtain such a $\tilde{\beta} = \tilde{\beta}(\epsilon)$ in a few stages. At the first stage, we obtain $\tilde{\beta}_1$ as follows. W.l.o.g. $\kappa = 1$ in condition **C**. Let

$$\tilde{\beta}_1 = \underset{\{\beta \mid ||\beta||_1 \leq \sqrt{k(n)}\}}{\arg\min} L_{\hat{F}}(\beta). \tag{8}$$

Note that by the Cauchy-Schwartz inequality and since $\kappa = 1$, the l_1 norm of of β^*_F is less than $\sqrt{k(n)}$, hence:

$$L_{\hat{F}_n}(\tilde{\beta}_1) \leq L_{\hat{F}_n}(\beta^*_{F_n}). \tag{9}$$

One may check that Corollary 3 may be applied on $\hat{F}_n$ with $\epsilon > 0$ and $c = c_n(\epsilon) = O_p(1)$. Thus we may extract a vector $\tilde{\beta}'_1$ from $\tilde{\beta}$, having $k_1 = k_1(n)$ non-zero coefficients, that satisfy

$$L_{\hat{F}_n}(\tilde{\beta}'_1) \leq L_{\hat{F}_n}(\tilde{\beta}_1) + \epsilon + \frac{c^2 k}{k_1}. \tag{10}$$

The extraction is through the multinomial simulation method, described in the proof of Lemma 4.

Choose $k_1(n) = o(n/\log(n))$ that satisfy $\frac{k(n)}{k_1(n)} \to 0$. Let $\tilde{\beta}$ be the least squares, with respect to the subset on which $\tilde{\beta}'_1$ has non-zero coefficients. Since this subset is chosen to be of order $o(n/\log(n))$, we may apply the reasoning and arguments of Section 3 , which together with the above imply that

$$\sup_{F_n \in \mathcal{F}^n} P_{F_n}(L_{F_n}(\tilde{\beta}) - L_{F_n}(\beta^*_{F_n}) > 2\epsilon) \to 0. \tag{11}$$

The above constructed $\tilde{\beta}$ is not persistent, since the last equation should hold for every ϵ. The latter is now easy to achieve using the diagonalization described above.

□

5 Concluding Remarks

We demonstrated for the case of multivariate normal Z^i that by increasing the number of explanatory variables from $o(n)$ (for which persistence may be achieved) to n^α, $\alpha > 1$, we can still achieve persistence with respect to all subsets of size $k(n) = o(n/\log(n))$. In cases where there are no clear favorite explanatory variables or physical understanding of a phenomena (a 'black box' situation), such a practice merits recommendation. This is especially true since we demonstrated the existence of algorithmically effective, persistent procedures. In more general situations (more general than the normal case) our results and techniques of proof also indicate that there is almost no loss, but a lot to be gained when increasing the number of explanatory variables. Thus we recommend an *inverse of Occam's razor principle.* Occam's razor principle does not seem relevant for prediction.

The various theorems we proved show that we may expect persistence for $k(n)$ of an order between $o((n/\log(n))^{\frac{1}{2}})$ and $o(n/\log(n))$. Consequently the l_1 constraint, $b(n)$, in the Lasso procedure should be of an order between $o(n/\log(n))^{\frac{1}{4}})$ and $o(n/\log(n))^{\frac{1}{2}})$.

In practice we do not know what is the right value for $b(n)$. Thus, we might want to use cross validation in order to try various points in that range. It might be helpful to try, through cross validation, even values of $b(n)$ that are larger than those suggested by our theory, for example, values for which there is still a unique solution to the Lasso optimization.

Finally, methods that use many more parameters than observations have recently been employed, and the fact that they do not get poor results due to overfitting is something of a mystery; see Breiman (2001). We demonstrated that methods that use many more parameters than observations may give good results as long as some restraint is kept (e.g., optimization under l_1 constraint). It might give some insight into the mystery of not getting poor results due to overfit.

We speculate that in the more general framework of predictor selection from a parametrized set of predictors $\{g_\beta\ ; \beta \in B\}$, under appropriate conditions, empirical minimization subject to l_1 constraints might have good properties, as explored here in the case of linear predictors. This is a subject that is studied in the work in progress mentioned earlier.

Acknowledgment: Helpful discussions with Larry Brown, Arkadi Nemirovski, and Tsachy Weissman are gratefully acknowledged. This work was presented in the workshop 'On high-dimensional data' at the Lorentz Center at the University of Leiden. The lovely and generous hospitality and the stimulating discussions with the participants are gratefully acknowledged. Y. Ritov was supported by part by Israeli Science Foundation grant

References

Bai, Z. D. and Yin, Y. Q. (1993). Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. Ann.Prob. **21**, 3, pp 1275-1294.

Bickel, P. and Levina, E. (2003). Some theory for Fisher's linear discriminant functions "naive Bayes", and some alternatives where there are more variables than observations. To appear in Bernoulli.

Billingsley, P. (1995). Probability and measure. Wiley, third edition.

Breiman, L. (2001). Statistical modeling: the two cultures. Statistical Science **16**, 3, pp 199-231.

Breiman, L. and Freedman, D. (1983). How many variables should be entered in a regression equation? JASA, **78**, No 381, pp 131-136.

Chen, S., Donoho, D., and Saunders, M. (2001). Atomic decomposition by basis pursuit. Siam Review. **43**. No. 1 pp 129-159.

Donoho, D.L. and Johnstone, I.M. (1994). Ideal spatial adaptation by wavelet shrinkage. Biometrika **81**, 3, pp 425-455.

Efron, B., Johnstone, I., Hastie, T., and Tibshirani, R. (2003). The angle regression (LAR) algorithm for solving the Lasso. Ann.Stat (to appear)

Emery, M., Nemirovski, A. and Voiculescu, D. (1998). Lectures on probability theory and statistics. Lecture notes in Mathematics. Springer.

Foster, D.P. and George, E.L. (1994). The risk inflation criterion for multiple regression. Ann.Stat. **22** 1947-1975.

Greenshtein, E. (work in progress). Best subset selection, consistency in high dimensional statistical learning and optimization under l_1 constraint.

Huber, P. (1973). Robust regression: asymptotics, conjectures, and Monte Carlo. Ann.Stat. **1**. No 5, pp 799-821.

Juditsky, A. and Nemirovski, A. (2000). Functional aggregation for nonparametric regression. Ann.Stat. **28**, No 3, pp 681-712.

Le Cam, L. and Yang, G. L. (1990). Asymptotics in Statistics. Springer-Verlag.

Lee, W. S., Bartlett, P. L., and Williamson, R. C. (1996). Efficient agnostics learning of neural networks with bounded fan in. IEEE trans. inform. theory. **42**, 2118-2132.

Nemirovski, A. and Yudin, D. (1983). Problems complexity and method efficiency in optimization. Wiley, New York.

Portnoy, S. (1984). Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large, I. Consistency. Ann. Stat. **12**, No 4, pp 1298-1309.

Silverstein, J. W. (1985). The smallest eigen value of a large dimensional Wishart matrix. Ann.Prob. **13** No 1, pp 1364-1368.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. JRSS B, **58** No 1, pp 267-288.