Proceedings of the International Conference IMCS-60, MSU
October 10–13, 2024, Chisinau, Republic of Moldova

High Dimensional Inference – a Literature Review

Alexandru Lopotenco, Marcel Teleuca

Abstract

High dimensional inference has been a crucial topic in statistical learning, especially with the rise of data that has a lot of features yet is difficult to gather.

We analyze regressions on data in large dimensions p that are much greater than the samples available, n. The usual method in constructing confidence intervals used for OLS is not useful when $p \gg n$ since asymptotic normality fails. We survey literature that comes up with new method of constructing these intervals.

Keywords: statistical learning, high-dimensional inference, confidence intervals, LASSO

1 Introduction

We will start discussing the case of high-dimensional linear regression with normal errors. Namely, we are given I.I.D. pairs $(Y_1, X_1), (X_2, Y_2), \dots (Y_n, X_n)$ which satisfy

$$Y_i = X_i^T \theta_0 + W_i \tag{1}$$

for $W_i \sim N(0, \sigma^2 I)$ and $\theta_0, Y_i, X_i \in \mathbb{R}^p$ for all $1 \leq i \leq n$. We denote $Y = (Y_1, Y_2, \dots, Y_n)$ and X to be the design matrix with rows $X_1^T, X_2^T, \dots X_n^T$. For simplicity, we denote $Z_j = (X_j, Y_j)$ and $Z = (X, Y)$. To study the high-dimensional case when $p \gg n$, LASSO regression has been introduced by [5] which promotes the sparsity of

©2024 by Alexandru Lopotenco, Marcel Teleuca

the learned coefficients via L1 regularization. Formally, the estimator of this regression is defined as

$$\hat{\theta}^n(X, Y; \lambda) = \arg\min_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|Y - \theta X\|_2^2 + \lambda \|\theta\|_1 \right\}. \tag{2}$$

Notably, this estimator comes with two drawbacks, especially in regards to building confidence intervals. First, the estimator is asymptotically biased, hence it is unclear how to center the confidence intervals. Second, it is difficult to find a distribution for it, hence building confidence intervals is theoretically complicated.

Another pitfall in high-dimensional regression is that the usual procedure of obtaining a reliable standard deviation estimator is no longer viable. Thankfully, this concern is addressed in [4] by providing the joint convex optimization problem,

$$(\hat{\theta}^n(\lambda), \hat{\sigma}(\lambda)) = \arg\min_{\theta \in \mathbb{R}^p, \sigma > 0} \left\{ \frac{1}{2\sigma n} \|Y - \theta X\|_2^2 + \frac{\sigma}{2} + \lambda \|\theta\|_1 \right\}, \tag{3}$$

which produces $\hat{\sigma}(\lambda)$, a consistent estimator for σ when λ has suitable asymptotics that we will discuss later.

1.1 Understanding the bias

To address the first concern for $\hat{\theta}^n$ it is useful to study its relation to the true parameter, θ_0. As mentioned before, extracting a limiting distribution for $\hat{\theta}^n$ is not feasible, hence one has to look to concentration inequalities for a distance between $\hat{\theta}^n$ and θ_0. To do so, we first need to introduce some notation.

Definition 1. *We denote by $S \equiv supp(\theta_0)$ the support of θ_0, defined as $supp(\theta_0) = \{i \in [p] : \theta_{0,i} \neq 0 \ \}$ and let $s_0 = |S|$.*

Definition 2. *For a symmetric matrix $\hat{\Sigma} \in \mathbb{S}_p$ and a set $S \in [p]$ we define the compatibility constant as*

$$\phi^2\left(\hat{\Sigma}, S\right) \equiv \min_{\theta \in \mathbb{R}^p} \left\{ \frac{|S| \theta^T \hat{\Sigma} \theta}{\|\theta_S\|_1^2} : \theta \in \mathbb{R}^p, \|\theta_{S^c}\|_1 \leq 3\|\theta_S\|_1 \right\},$$

and we further say that $\hat{\Sigma}$ *satisfies the compatibility condition for a set* $S \subseteq [p]$ *with constant* ϕ_0 *if* $\phi^2\left(\hat{\Sigma}, S\right) \geq \phi_0$. *We say that it holds for design matrix* X *if it holds for* $\hat{\Sigma} = \frac{1}{n}X^TX$.

We will also require a few technical conditions for the design matrix.

Assumption 1. *The design matrix* X *satisfies* $C^{-1} \leq \lambda_{\min}(X) \leq \lambda_{\max}(X) \leq C$ *for some constant, where* $\lambda_{\min}$ *and* $\lambda_{\max}$ *are the minimum and maximum eigenvalues of the sample covariance matrix* $\hat{\Sigma} = \frac{1}{n}X^TX$.

Assumption 2. *The matrix* $X\mathbb{E}[X^TX]^{-\frac{1}{2}}$ *has independent sub-Gaussian rows with mean* 0 *and sub-Gaussian norm equal to some* κ

Assumption 3. *The maximum diagonal entry of* $\hat{\Sigma}$ *is at most* K, *formally* $\max_i \frac{1}{n}(X^TX)_{ii} \leq K$.

Definition 2 is rather technical, yet it enables us to introduce the following concentration inequality for an $\hat{\theta}^n$ inferred from a design matrix X that satisfies the compatibility condition for a set S with constant ϕ_0 and $|S| = S_0$ due to [1]

$$P\left(\|\hat{\theta}^n - \theta_0\|_1 \geq t^2 s_0 \sqrt{\frac{\log(p)}{n}}\right) \leq 2p^{-C_0 t^2/2}, \tag{4}$$

for some constant C_0 dependent on p, σ and K. While the imposed compatibility conditions on X seem restrictive, [2] proved in Theorem 7 that under Assumptions 1, 2 and 3 these conditions hold with probability at least $1 - e^{-C^3 n}$ for some constant C_3 dependent on K, κ and C when $n \geq Ls_0 \log(p/s_0)$ for a large enough constant L.

1.2 Debiasing the estimator

Since we can control the quantity $\|\hat{\theta}^n - \theta_0\|_1$ we want to build a better estimator for θ_0 via adding an asymptotically debiasing term. Formally, we want an $\hat{\theta}^u = \hat{\theta}^n + V$ for a term V that makes $\hat{\theta}^u - \theta_0$ to control.

Hence it makes sense to write $V = UX\hat{\theta}^n - UY$ since it helps us factorize as follows

$$\hat{\theta}^u - \theta_0 = \hat{\theta}^n - \theta_0 + UX(\hat{\theta}^n - \theta_0) - UW = (I + UX)(\hat{\theta}^n - \theta_0) - UW \tag{5}$$

Clearly, UW is an error term with mean 0. Hence, we want to minimize the quantity $\|I + UX\|_\infty$, which can attain 0 by setting $U = -\hat{\Sigma}^{-1}X^T$ if $\hat{\Sigma}$ is non-singular. However, this is not always the case when $p \gg n$, so [2] propose to set $U = -\frac{1}{n}MX^T$ where

$$M = \arg\min_{M \in \mathbb{R}^{p \times p}} \|M\hat{\Sigma} - I\|_\infty. \tag{6}$$

In this setting, M serves as a pseudo-inverse for $\hat{\Sigma}$ with respect to the infinity norm. [2] call $\mu_*(X) = \min_{M \in \mathbb{R}^{p \times p}} \|M\hat{\Sigma} - I\|_\infty$ the generalized coherence parameter of X and they prove in Lemma 23 that under Assumptions 1 and 2 we have the following concentration inequality

$$P\left(\mu_*(X) \geq a\sqrt{\frac{\log p}{n}}\right) \leq 2p^{-a^2 C_1}, \tag{7}$$

for a constant C_1 dependent on κ and C. Now, notice that due to Equations 4 and 7, under Assumptions 1, 2 and 3 we can deduce the following inequality on the residual term $\Delta := (M\hat{\Sigma} - I)(\theta_0 - \hat{\theta}^n)$ in 5:

$$P\left(\|\Delta\|_\infty \geq ac\frac{s_0 \log p}{n}\right) \leq 4e^{-C_3 n} + 4p^{-C_4 a^2 \wedge C_5 c^2}, \tag{8}$$

for some constants C_4 and C_5 dependent on κ and C. Notice that $4e^{-C_3 n}$ is the probability that X does not satisfies the compatibility conditions. This implies that $\|\Delta\|_\infty = O\left(\frac{s_0 \log p}{n}\right)$. Hence, as per the definition of U we have

$$\hat{\theta}^u = \hat{\theta}^n + \frac{1}{n}MX^T(Y - X^T\hat{\theta}^n), \tag{9}$$

and equation 5 can be rewritten as

$$\sqrt{n}(\hat{\theta}^u - \theta_0) = \sqrt{n}(M\hat{\Sigma} - I)(\theta_0 - \hat{\theta}^n) + \frac{1}{n}MX^TW, \tag{10}$$

where clearly $\frac{1}{n}MX^TW \sim N(0, \sigma^2 M\hat{\Sigma}M^T)$, so

$$\sqrt{n}(\hat{\theta}^u - \theta_0) - O\left(\frac{s_0 \log p}{\sqrt{n}}\right) \sim N(0, \sigma^2 M\hat{\Sigma}M^T), \tag{11}$$

which means that in the sparse regime $s_0 = o\left(\frac{\sqrt{n}}{\log p}\right)$ the estimator $\hat{\theta}^u$ introduced by [2] is indeed asymptotically unbiased.

1.3 Confidence intervals

We only need $\mu_*(X) = O\left(\sqrt{\frac{\log p}{n}}\right)$ to achieve the asymptotics in 11, hence in practice it is enough to impose that $\|\hat{\Sigma}m_i - e_i\|_\infty \leq a\sqrt{\frac{\log p}{n}} =: \mu$ for some parameter a. Recall that in 3 we introduced a consistent estimator for σ, so we can now build a $(1-\alpha)$ confidence interval for θ_0.

Algorithm 1 Computation of matrix the debiased coefficient - Algorithm 1 of [2]

Comment 1. **Input:** Design matrix X, parameters λ, μ and labels Y.
Initialize: $\hat{\Sigma} = \frac{1}{n}X^TX$ and $\hat{\theta}^n$ the solution of 2
for $i = 1$ **to** p **do**
Let m_i be the solution of
Minimize $m^T\hat{\Sigma}m$ subject to $\|\hat{\Sigma}m - e_i\|_\infty \leq 10\sqrt{\log p/n}$
end for
Set $M = (m_1, m_2, \dots m_p)^T$ if all the problems are feasible and I, the identity matrix, otherwise.
Output: $\hat{\theta}^u + \frac{1}{n}MX^T(Y - X^T\hat{\theta}^n)$

Theorem 1. *(Theorem 15 in [2]) Construct* $\hat{\theta}^u$ *as per Algorithm* 1*, with a large enough and let* $\hat{\sigma}$ *be the estimator of the standard deviation obtained from solving 3 with* $\lambda = 10\sqrt{2\log p/n}$*. Then, under Assumptions 1, 2, 3 and that* $s_0 = o\left(\frac{\sqrt{n}}{\log p}\right)$ *we have*

$$J_{i,\alpha}(Z, \theta_0) := \hat{\theta}^u \pm z_{1-\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}[M\hat{\Sigma}M^T]^{\frac{1}{2}}_{i,i} \tag{12}$$

satisfies $\lim_{n\to\infty} P(\theta_{0,i} \in J_{\alpha,i}(Z, \theta_0)) = 1 - \alpha$.

As a closing remark for future readings, the authors of [2] discussed improved asymptotics towards their initial discoveries. In particular, they discuss the case when s_0 is of the order $o(n/(\log p)^2$ in the scenario when the covariance matrix is known apriori or can be efficiently estimated in their further work [3].

Acknowledgments. This literature review is based on a project done for STAT9700 class while first author was at University of Pennsylvania. Second author contributed to formatting for this primer to be included in the conference handbook.

References

[1] P. Bühlmann and S. van de Geer. *Statistics for high-dimensional data.* Springer Series in Statistics. Springer, Heidelberg, 2011.

[2] A. Javanmard and A. Montanari. *Confidence intervals and hypothesis testing for high-dimensional regression.* 2014.

[3] A. Javanmard and A. Montanari. *Debiasing the lasso: Optimal sample size for Gaussian designs.* The Annals of Statistics, vol. 46(6A), pp. 2593–2622, 2018.

[4] Sun and C.-H. Zhang. *Scaled sparse linear regression.* Biometrika, vol. 99(4), pp. 879–898, 2012.

[5] R. Tibshirani. *Regression shrinkage and selection via the lasso.* Journal of the Royal Statistical Society. Series B (Methodological), vol. 58(1), pp. 267–288, 1996.

Alexandru Lopotenco[1], Marcel Teleuca[2]

[1] Stanford University
E–mail: alopotenco@stanford.edu
ORCID: https://orcid.org/0009-0005-8477-3602

[2] Vladimir Andrunachievici Institute of Mathematics and Computer Science, Moldova State University
E–mail: mteleuca@gmail.com
ORCID: https://orcid.org/0000-0003-1730-5284