Step into High-dimensional Statistics: Sparse Mean Estimation

Yajie Bao*

March 2, 2020

Classical Statistics always has a basic assumption: $n > p$, and many estimation methods with good asymptotic properties were build based on this assumption. Many multivariate statistical models (see Anderson [1958]) will fail when the number of variates is greater than sample size, such as linear regression, LDA, PCA... And in many situations, the dimension p will increase with the growth of sample size n.

Take the normal mean estimation as an example, $X_i,\ i = 1, 2, ..., n$ are i.i.d samples from multivariate normal distribution $N(\boldsymbol{\mu}, \sigma^2 I_p)$, and sample mean $\bar{\boldsymbol{X}}$ is the minimax estimator of $\boldsymbol{\mu}$. Note that the minimax error is

$$\mathbb{E}\left(\bar{\boldsymbol{X}} - \boldsymbol{\mu}\right)^2 = \sum_{j=1}^{p} \mathbb{E}\left(\bar{X}_j - \mu_j\right)^2 = \frac{p\sigma^2}{n},$$

and obviously $\bar{\boldsymbol{X}}$ is not a consistent estimator when $n = o(p)$, which is called the curse of dimensionality.

Another example is LDA, we need to compute the linear discriminant vector $\widehat{\boldsymbol{\Sigma}}\left(\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}}\right)$. The rank of sample covariance matrix is $\min\{n, p\} = n$, which means $\widehat{\boldsymbol{\Sigma}}$ is non-invertible. So we can't obtain $\widehat{\boldsymbol{\Sigma}}^{-1}$ directly.

1 Gaussian sequence model

A toy model in high-dimensional statistics is the Gaussian sequence model,

$$y_{ij} = \beta_j + z_{ij},\ i = 1, 2, ..., n;\ j = 1, 2, ..., p \tag{1.1}$$

where z_{ij} are i.i.d normal r.v with mean 0 and variance σ^2 for each j. Now we have n observations for each y_j to estimate $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$. To overcome this problem, we need to add some assumptions on high-dimensional parameter $\boldsymbol{\beta}$. A direct thought is sparsity, i.e., there are only few non-zero elements in $\boldsymbol{\beta}$. And this assumption can be written as

$$\sum_{j=1}^{p} 1(|\beta_j| \neq 0) \leq s_0. \tag{1.2}$$

*Department of Mathematics, Shanghai Jiao Tong University, Eamil: baoyajie2019stat@sjtu.edu.cn

Next step is to find the positions of non-zero parameter entries and obtain their estimation, and the first part of our goal is also called support recovery. If $\beta_j = 0$, then $\hat{\beta}_j = \bar{Y}_j$ will be quite small. Thus we can only keep $\hat{\beta}_j$ with large magnitude, which leads to the idea of thresholding.

There are many thresholding functions like hard thresholding, soft thresholding (see Donoho and Johnstone [1994]), SCAD (see Fan and Li [2001]) etc. Here we use hard thresholding method

$$\widehat{\beta}_j = \bar{Y}_j \mathbb{I}\left(\left|\bar{Y}_j\right| \geq t\right), \quad \forall j \in \{1, \ldots, p\}, \tag{1.3}$$

where $\bar{Y}_j = \sum_{i=1}^n y_{ij}/n$. Next we will give some theoretical results on estimation error and support recovery. Before this we need a lemma on the bound of $\max_j \left|\bar{Y}_j - \beta_j\right|$

Lemma 1.1 *For the sample mean $\bar{Y}_j,\ j = 1, 2, ..., p$*

$$\max_{i=1}^{p} \left|\bar{Y}_j - \beta_j\right| = O_p\left(\sqrt{\frac{\log p}{n}}\right). \tag{1.4}$$

Proof: let $X_j = \bar{Y}_j - \beta_j = \frac{\sum_{i=1}^n z_{ji}}{n}$, where $z_{ji} \sim N(0, \sigma^2)$ and independent. Using the tail probability of normal random variables and the fact $X_j \sim N(0, \frac{\sigma^2}{n})$, we have

$$\begin{aligned} \mathbb{P}\left(\max_{j=1}^{p} |X_j| \geq t\right) &\leq \sum_{j=1}^{p} \mathbb{P}\left(|X_j| \geq t\right) \\ &\leq 2p \exp\left(-\frac{nt^2}{2\sigma^2}\right). \end{aligned}$$

Set $t = \lambda\sqrt{\frac{\log p}{n}}$ for sufficiently large λ and the result follows. ■

Theorem 1.2 (Support recovery) *Let $S(\beta) = \{j : |\beta_j| \neq 0\}$ and $S(\widehat{\beta}) = \{j : |\widehat{\beta}_j| \neq 0\}$, assume that $\min_{j\in S} |\beta_j| > \sigma\sqrt{\frac{2\log(2p/\delta)}{n}}$ and set $t = \sigma\sqrt{\frac{2\log(2p/\delta)}{n}}$ then with probability at least $1 - \delta$,*

$$S(\beta) = S(\widehat{\beta}). \tag{1.5}$$

Proof: According to the proof of Lemma 1.1, with probability at least $1 - \delta$,

$$\max_{i=1}^{p} \left|\bar{Y}_j - \beta_j\right| \leq \sigma\sqrt{\frac{2\log(2p/\delta)}{n}}.$$

If $j \in S$, then $|\widehat{\beta}_j| > 0$, otherwise the error will be great than $\sigma\sqrt{\frac{2\log(2p/\delta)}{n}}$. If $j \in S^c$, then

$$\mathbb{P}\left(|\widehat{\beta}_j| = 0\right) = \mathbb{P}\left(\left|\bar{Y}_j - \beta_j\right| \geq t\right) \leq 1 - \delta.$$

Then we have completed the proof. ■

Theorem 1.3 (ℓ_1 error bound) *Under the assumption (1.2) and set threshold* $t = \lambda\sqrt{\frac{\log p}{n}}$

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\widehat{\beta}_j - \beta_j| = O_p\left(s_0\sqrt{\frac{\log p}{n}}\right), \tag{1.6}$$

where $\lambda > \sqrt{2}\sigma$.

Proof: First using the assumption (1.2) and Lemma 1.1, we have

$$\begin{aligned}
\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|_1 &= \sum_{j=1}^{p} \left|\bar{Y}_j \mathbb{I}\left(\left|\bar{Y}_j\right| \geq t\right) - \beta_j\right| \\
&= \sum_{j\in S} \left|\bar{Y}_j \mathbb{I}\left(\left|\bar{Y}_j\right| \geq t\right) - \beta_j\right| + \sum_{j\in S^c} \left|\bar{Y}_j \mathbb{I}\left(\left|\bar{Y}_j\right| \geq t\right)\right| \\
&\leq \sum_{j\in S} \left|\bar{Y}_j - \beta_j\right| + \sum_{j\in S} \left|\bar{Y}_j\right| \mathbb{I}\left(\left|\bar{Y}_j\right| < t\right) + \sum_{j\in S^c} \left|\bar{Y}_j \mathbb{I}\left(\left|\bar{Y}_j\right| \geq t\right)\right| \\
&\leq s_0 \max_{i=1}^{p} \left|\bar{Y}_j - \beta_j\right| + s_0 t + \max_{i=1}^{p} \left|\bar{Y}_j - \beta_j\right| \sum_{j\in S^c} \mathbb{I}\left(\left|\bar{Y}_j\right| \geq t\right) \\
&= O_p\left(s_0\sqrt{\frac{\log p}{n}}\right) + I.
\end{aligned}$$

Then note that when $\beta_j = 0$, $\bar{Y}_j \sim N(0, \frac{\sigma^2}{n})$ and

$$\begin{aligned}
\mathbb{P}\left(\sum_{j\in S^c} \mathbb{I}\left(\left|\bar{Y}_j\right| \geq t\right) > 0\right) &= \mathbb{P}\left(\max_{j\in S^c} |\bar{Y}_j| \geq t\right) \\
&\leq 2p \exp\left(-\frac{\lambda^2}{2\sigma^2}\log p\right) \\
&= 2\exp\left(-\frac{\lambda^2}{2\sigma^2}\log p + \log p\right) \to 0.
\end{aligned}$$

Thus $I = o_p\left(s_0\sqrt{\frac{\log p}{n}}\right)$ and the result follows. ∎

Remark. Through the analysis above, under the sparsity assumption (1.2), if $s_0\sqrt{\frac{\log p}{n}} \to 0$ then hard thresholding estimator is still consistent.

Theorem 1.4 (ℓ_∞ error bound) *Under the assumption (1.2) and set threshold* $t = M_0\sqrt{\frac{\log p}{n}}$ *for some* $M_0 > 0$, *then we have*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_\infty = O_p\left(\sqrt{\frac{\log p}{n}}\right). \tag{1.7}$$

Proof: Note that there exists some C_0 such that,

$$\begin{aligned}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_\infty &\le \max_{j=1}^{p}\left|\bar{Y}_j-\beta_j\right| + \max_{j=1}^{p}\left|\bar{Y}_j\right| \mathbb{I}\left(|\bar{Y}_j|<t\right)\\ &\le C_0\sqrt{\frac{\log p}{n}}+t.\end{aligned}$$

■

Remark. Using the simple norm inequality and Theorem 1.2, we have ℓ_2 error bound

$$\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2 \le \sqrt{s}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_\infty = O_p\left(\sqrt{\frac{s\log p}{n}}\right). \tag{1.8}$$

And according to Johnstone [1986], (1.8) is statistical minimax lower bound of sparse mean estimation.

2 New tail bound assumption

Note that the assumption of normality is used to construct tail bound (1.4), and this assumption can be substituted by the following condition:

Assumption 2.1 (Exponential-type tails) *Suppose that there exists some $\gamma>0$ such that*

$$\mathrm{E}\exp\left(tz_{ij}^2\right)\le K_1<\infty \quad \textit{for all } |t|\le\gamma \textit{ and } i,j \tag{2.1}$$

Here we use a lemma in Cai and Liu [2011] as following:

Lemma 2.2 *Let $\xi_1,\ldots,\xi_n$ be independent random variables with mean 0. Suppose that there exists some $\eta>0$ and $\bar{B}_n^2$ such that $\sum_{k=1}^n \mathrm{E}\xi_k^2 e^{\eta|\xi_k|}\le\bar{B}_n^2$. Then for $0<x\le\bar{B}_n$,*

$$\mathrm{P}\left(\sum_{k=1}^n\xi_k\ge C_\eta\bar{B}_n x\right)\le\exp\left(-x^2\right), \tag{2.2}$$

where $C_\eta=\eta+\eta^{-1}$.

Proof: By the inequality $|e^s-1-s|\le s^2e^{|s|}$, we have for any $t\ge 0$,

$$\begin{aligned}\mathrm{P}\left(\sum_{k=1}^n\xi_k\ge C_n\bar{B}_n x\right) &\le \exp\left(-tC_\eta\bar{B}_n x\right)\prod_{k=1}^n\mathrm{E}\exp\left(t\xi_k\right)\\ &\le \exp\left(-tC_\eta\bar{B}_n x\right)\prod_{k=1}^n\left(1+t^2\mathrm{E}\xi_k^2e^{t|\xi_k|}\right)\\ &\le \exp\left(-tC_\eta\bar{B}_n x+\sum_{k=1}^n t^2\mathrm{E}\xi_k^2e^{t|\xi_k|}\right).\end{aligned}$$

Take $t = \eta \left(x/\bar{B}_n \right)$, it follows that

$$\mathrm{P}\left(\sum_{k=1}^{n} \xi_k \geq C_\eta \bar{B}_n x\right) \leq \exp\left(-\eta C_\eta x^2 + \eta^2 x^2\right) = \exp\left(-x^2\right).$$

■

Theorem 2.3 *Assume that the noise z_{ij} satisfying Assumption 2.1, we have*

$$\max_{i=1}^{p} \left| \bar{Y}_j - \beta_j \right| = O_p\left(\sqrt{\frac{\log p}{n}}\right). \tag{2.3}$$

Proof: Using the simple inequality

$$s^2 e^s \leq e^{2s} \leq e^{s^2+1},$$

we have for each $i,\ j$

$$\mathrm{E}\left(z_{ij}^2 e^{\eta |z_{ij}|}\right) \leq \mathrm{E}\left(\eta^{-2} \exp(2\eta |z_i j|)\right) \leq e\mathrm{E}\left(\eta^{-2} \exp(\eta^2 |z_i j|^2)\right).$$

By Assumption 2.1, we can set

$$\bar{B}_n^2 = ne\eta^{-2} K_1,$$

where $0 < \eta < \sqrt{\gamma}$. Then for sufficiently large η

$$\begin{aligned}
\mathrm{P}\left(\max_{i=1}^{p} \left| \bar{Y}_j - \beta_j \right| > C\sqrt{\frac{\log p}{n}}\right) &\leq \sum_{j=1}^{p} \mathrm{P}\left(\sum_{i=1}^{n} |z_{ij}| > C\sqrt{n \log p}\right) \\
&= p\mathrm{P}\left(\sum_{i=1}^{n} |z_{ij}| > C\bar{B}_n e^{-1} \eta K_1^{-\frac{1}{2}} \sqrt{\log p}\right) \\
&\to 0,
\end{aligned}$$

whcih completes the proof. ■

Remark. Assumption 2.1 is very similar to sub-Gaussian (see Vershynin [2018]), which has tail

$$\mathbb{P}\{|X| \geq t\} \leq 2\exp\left(-t^2/K_1^2\right) \quad \text{for all } t \geq 0. \tag{2.4}$$

And there is concentration inequality about sum of independent sub-Gaussian random variables.

Theorem 2.4 (General Hoeffding's inequality) *Let $X_i,\ i = 1, 2, ..., N$ be be independent, mean zero, sub-gaussian random variables with parameter σ_i, then for every $t \geq 0$,*

$$\mathbb{P}\left\{\left|\sum_{i=1}^{N} X_i\right| \geq t\right\} \leq 2\exp\left(-\frac{ct^2}{\sum_{i=1}^{N} \sigma_i^2}\right). \tag{2.5}$$

Besides Exponential-type tails, there is another common tail called Polynomial-type tails.

Assumption 2.5 (Polynomial-type tails) *Suppose that for some $\gamma > 0$,*

$$E\left|z_{ij}\right|^{2(1+\gamma)} \leq K \quad \text{for all } i, j. \tag{2.6}$$

Theorem 2.6 *Under the Assumption (2.5), we have*

$$\max_{i=1}^{p}\left|\bar{Y}_j - \beta_j\right| = O_p\left(\frac{p^{1/2(1+\gamma)}}{n^{1/2}}\right). \tag{2.7}$$

Proof: We use a moment inequality in Shao [2003], for $q > 0$

$$E\left|\sum_{i=1}^{n} z_{ij}\right|^q \leq \frac{C_q}{n^{1-q/2}}\sum_{i=1}^{n} E\left|X_i\right|^q. \tag{2.8}$$

By Markov inequality,

$$\begin{aligned}
\mathrm{P}\left(\max_{i=1}^{p}\left|\bar{Y}_j - \beta_j\right| > t\right) &\leq p\frac{\mathrm{E}\left|\sum_{i=1}^{n} z_{ij}\right|^{2(1+\gamma)}}{(nt)^{2(1+\gamma)}} \\
&\leq p\frac{Cn^{1+\gamma}K_2}{(nt)^{2(1+\gamma)}} \\
&= pC_pK_2n^{-(1+\gamma)}t^{-2(1+\gamma)}.
\end{aligned}$$

Let $t = M\frac{p^{1/2(1+\gamma)}}{n^{1/2}}$ for sufficiently large M, then we complete the proof. ■

Remark. If we take threshold $t = M\frac{p^{1/2(1+\gamma)}}{n^{1/2}}$, then the convergence rate of ℓ_1 error will be $O_p(s_0\frac{p^{1/2(1+\gamma)}}{n^{1/2}})$.

3 New sparsity assumption

Sparsity assumption (1.2) is actually an ℓ_0 ball in $\mathbb{R}^p$, which can be genlized to ℓ_q ball in $\mathbb{R}^p$, i.e., for $0 \leq q < 1$

$$\mathcal{U}\left(q, s_q\right) = \left\{\boldsymbol{\beta} \in \mathbb{R}^p : \sum_{j=1}^{p}|\beta_j|^q \leq s_q\right\}. \tag{3.1}$$

Next we will build convergence rate of ℓ_q, and the proof is very similar to the Theorem 1 in Bickel and Levina [2008].

Theorem 3.1 (ℓ_1 error bound) *If $\beta \in \mathcal{U}\left(q, s_q\right)$ and set threshold $t_n = M\sqrt{\frac{\log p}{n}}$ for sufficiently large M. Suppose thta noise z_{ij} are sub-Gaussian random variables with same parameter σ, then*

$$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|_1 = \sum_{j=1}^{p}|\widehat{\beta}_j - \beta_j| = O_p\left(s_q\left(\frac{\log p}{n}\right)^{(1-q)/2}\right). \tag{3.2}$$

Proof: Let T_{t_n} be hard thresholding function with threshold t_n, then note that

$$\left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right\|_1 \le \left\|T_{t_n}\left(\bar{\boldsymbol{Y}}\right)-T_{t_n}\left(\boldsymbol{\beta}\right)\right\|_1+\left\|\boldsymbol{\beta}-T_{t_n}\left(\boldsymbol{\beta}\right)\right\|_1. \tag{3.3}$$

By $\beta \in \mathcal{U}\left(q, s_q\right)$ we have

$$\begin{aligned}
\left\|\boldsymbol{\beta}-T_{t_n}\left(\boldsymbol{\beta}\right)\right\|_1 &= \sum_{j=1}^{p}\left|\beta_j-\beta_j \mathbb{I}\left(\left|\beta_j\right| \ge t_n\right)\right| \\
&= \sum_{j=1}^{p}\left|\beta_j\right| \mathbb{I}\left(\left|\beta_j\right| < t_n\right) \\
&\le \sum_{j=1}^{p}\left|\beta_j\right|^q t_n^{1-q} \mathbb{I}\left(\left|\beta_j\right| < t_n\right) \\
&\le s_q t_n^{1-q}.
\end{aligned}$$

Next we will bound the first term of (3.3),

$$\begin{aligned}
\left\|T_{t_n}\left(\bar{\boldsymbol{Y}}\right)-T_{t_n}\left(\boldsymbol{\beta}\right)\right\|_1 \le& \sum_{j=1}^{p}\left|\bar{Y}_j\right| \mathbb{I}\left(\left|\bar{Y}_j\right| \ge t_n,\ \left|\beta_j\right| < t_n\right) \\
&+\sum_{j=1}^{p}\left|\bar{Y}_j-\beta_j\right| \mathbb{I}\left(\left|\bar{Y}_j\right| \ge t_n,\ \left|\beta_j\right| \ge t_n\right) \\
&+\sum_{j=1}^{p}\left|\beta_j\right| \mathbb{I}\left(\left|\bar{Y}_j\right| < t_n,\ \left|\beta_j\right| \ge t_n\right) \\
=& \mathrm{I}+\mathrm{II}+\mathrm{III}.
\end{aligned}$$

For the second term, there exists some $C_1>0$ such that,

$$\begin{aligned}
\mathrm{II} &\le \sum_{j=1}^{p}\left|\bar{Y}_j-\beta_j\right| \mathbb{I}\left(\ \left|\beta_j\right| \ge t_n\right) \\
&\le \max_{j=1}^{p}\left|\bar{Y}_j-\beta_j\right| \sum_{j=1}^{p} \mathbb{I}\left(\ \left|\beta_j\right| \ge t_n\right) \\
&\le C_1 \sqrt{\frac{\log p}{n}} s_q t_n^{-q}.
\end{aligned}$$

For the third term,

$$\begin{aligned}
\mathrm{II} &\le \sum_{j=1}^{p}\left|\beta_j-\bar{Y}_j\right| \mathbb{I}\left(\left|\beta_j\right| \ge t_n\right)+t_n \sum_{j=1}^{p} \mathbb{I}\left(\left|\beta_j\right| \ge t_n\right) \\
&\le C_1 \sqrt{\frac{\log p}{n}} s_q t_n^{-q}+s_q t_n^{1-q}.
\end{aligned}$$

For the first term,

$$
\begin{aligned}
\mathrm{I} &\le \sum_{j=1}^{p} \left|\bar{Y}_j - \beta_j\right| \mathbb{I}\left(|\bar{Y}_j| \ge t_n,\ |\beta_j| < t_n\right) + \sum_{j=1}^{p} |\beta_j| \mathbb{I}\left(|\bar{Y}_j| \ge t_n,\ |\beta_j| < t_n\right) \\
&\le \sum_{j=1}^{p} \left|\bar{Y}_j - \beta_j\right| \mathbb{I}\left(|\bar{Y}_j| \ge t_n,\ |\beta_j| < t_n\right) + s_q t_n^{1-q} \\
&= \mathrm{IV} + s_q t_n^{1-q}.
\end{aligned}
$$

Now take $\gamma \in (0,1)$,

$$
\begin{aligned}
\mathrm{IV} &= \sum_{j=1}^{p} \left|\bar{Y}_j - \beta_j\right| \mathbb{I}\left(|\bar{Y}_j| \ge t_n,\ |\beta_j| < \gamma t_n\right) + \sum_{j=1}^{p} \left|\bar{Y}_j - \beta_j\right| \mathbb{I}\left(|\bar{Y}_j| \ge t_n,\ \gamma t_n \le |\beta_j| \le t_n\right) \\
&\le \sum_{j=1}^{p} \left|\bar{Y}_j - \beta_j\right| \mathbb{I}\left(|\bar{Y}_j| \ge t_n,\ |\beta_j| < \gamma t_n\right) + \sum_{j=1}^{p} \left|\bar{Y}_j - \beta_j\right| \mathbb{I}\left(|\beta_j| \ge \gamma t_n\right) \\
&\le C_1 \sqrt{\frac{\log p}{n}} \sum_{j=1}^{p} \mathbb{I}\left(\left|\bar{Y}_j - \beta_j\right| > (1-\gamma) t_n\right) + C_1 \sqrt{\frac{\log p}{n}} s_q (\gamma t_n)^{-q},
\end{aligned}
$$

moreover using (2.4) and make $(1-\gamma)^2 M > 2\sigma^2$ we have

$$
\begin{aligned}
\mathrm{P}\left(\sum_{j=1}^{p} \mathbb{I}\left(\left|\bar{Y}_j - \beta_j\right| > (1-\gamma) t_n\right) > 0\right) &= \mathrm{P}\left(\max_{j=1}^{p} |\bar{Y}_j - \beta_j| > (1-\gamma) t_n\right) \\
&\le p \exp\left(-\frac{(1-\gamma)^2 M \log p}{2\sigma^2}\right) \\
&= \exp\left(\log p - \frac{(1-\gamma)^2 M}{2\sigma^2} \log p\right) \\
&\to 0.
\end{aligned}
$$

Combining the inequalities above, (3.2) is proved. ■

References

T. W. Anderson. *An introduction to multivariate statistical analysis*. Wiley, New York, 1958.

P. J. Bickel and E. Levina. Covariance regularization by thresholding. *Annals of Statistics*, 36(6):2577–2604, 2008.

T. Cai and W. Liu. Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, 106(494):672–684, 2011.

D. L. Donoho and J. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

Johnstone. On minimax estimation of sparse normal mean vector. *Annals of Statistics*, 14 (2):590–606, 1986.

J. Shao. *Mathematical statistics.* Springer, 2003.

R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.