ECE598: Information-theoretic methods in high-dimensional statistics Spring 2016

Lecture 19: Denoising sparse vectors - Risk upper bound

Lecturer: Yihong Wu *Scribe: Ravi Kiran Raman, Apr 12, 2016*

This lecture focuses on the problem of denoising for a sparse vector and upper bound of the minimax risk corresponding to the problem.

Let $\theta \in \Theta = B_0(k) = \{\theta \in \mathbb{R}^p : \|\theta\|_0 \leq k\}$, be a sparse vector. We observe $Y = \theta + Z$, where $Z \sim \mathcal{N}(0, I_p)$. Recall that the last lecture obtains the upper bound on the minimax risk for the problem using the mutual information method as

$$R_n^*(\Theta) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta[\|\hat{\theta} - \theta\|_2^2] \gtrsim k \log\left(\frac{p}{k}\right).$$

This lecture focuses on obtaining the upper bound to the minimax error by analyzing the risk corresponding to the maximum likelihood estimator.

Remark 19.1. Estimators, $\hat{\theta}$ are typically efficiently computable for the denoising problem defined above. Further, adaptive estimators that function in the absence of knowledge of k can be defined.

19.1 Maximum Likelihood estimator and risk upper bound

19.1.1 MLE and Basic Inequality

The maximum likelihood estimator for the denoising problem under additive Gaussian noise is given by

$$\hat{\theta}_{\text{MLE}}(y) \in \arg\min_{\tilde{\theta} \in B_0(k)} \|y - \tilde{\theta}\|_2^2. \tag{19.1}$$

We now show that $\forall \theta \in B_0(k)$,

$$\|\hat{\theta}_{\text{MLE}} - \theta\|_2^2 \lesssim k \log\left(\frac{p}{k}\right)$$

holds both, under expectation and with high probability. For ease of notation, we shall henceforth refer to the ML estimator as $\hat{\theta}$.

We observe that the ground truth θ is a feasible solution of (19.1). Since the estimator minimizes the ℓ_2 distance, we have

$$\|Z - h\|_2^2 = \|y - \hat{\theta}\|_2^2 \leq \|y - \theta\|_2^2 = \|Z\|_2^2,$$

where $h = \hat{\theta} - \theta$. Thus $\|h\|_0 \leq 2k$. Hence we have

$$\begin{aligned} \|h\|_2^2 &\leq 2\langle h, Z\rangle = 2\|h\|_2 \left\langle Z, \frac{h}{\|h\|_2} \right\rangle \\ &\leq 2\|h\|_2 \sup_{u \in S^{p-1} \cap B_0(2k)} \langle Z, u\rangle \\ \Leftrightarrow \|h\|_2 &\leq 2 \sup_{u \in S^{p-1} \cap B_0(2k)} \langle Z, u\rangle. \end{aligned} \tag{19.2}$$

19.1.2 Risk upper bound through Gaussian width

Let $G = S^{p-1} \cap B_0(2k)$. Thus, from (19.2), we have

$$\mathbb{E}\left[\|h\|_2\right] \leq 2\mathbb{E}\left[\sup_{u \in G} \langle u, Z \rangle\right] = 2w(G),$$

where $w(\cdot)$ is the Gaussian width. We know that Sudakov minorization lower bounds the Gaussian width as

$$w(G) \gtrsim \epsilon \sqrt{\log\left(N(G, \|\cdot\|_2, \epsilon)\right)} \asymp \sqrt{k \log\left(\frac{ep}{k}\right)},$$

as long as $\epsilon \asymp 1$. The above result follows from the Gilbert-Varshamov lower bound via packing Hamming spheres.

However we are interesting in an upper bound for the Gaussian width here. One way to obtain this is using Dudley's entropy integral method [**?**],

$$\begin{aligned} w(G) &\lesssim \int_0^{\text{rad}(G)} \sqrt{\log(N(G, \|\cdot\|_2, \epsilon))} d\epsilon \\ &\lesssim \int_0^1 \sqrt{\log\left(\frac{1}{\epsilon}\right)^k \binom{p}{2k}} d\epsilon \qquad (19.3) \\ &\asymp \sqrt{k \log \frac{pe}{k}}, \end{aligned}$$

where (19.3) follows from the fact that the vectors projected onto the set of support vectors lie on S^{2k-1} and the fact that there are $\binom{p}{2k}$ possible support vector combinations.

19.1.3 Risk upper bound through Covering argument

We now provide an alternate proof to show that the upper bound is held with high probability (consequently in expectation). Let J represent a set of indices. Let us partition G as

$$G = \cup_{|J|=2k} G_J = \cup_{|J|=2k} \left\{ x \in \mathbb{R}^p : \text{supp}(x) = J, x_J \in S^{2k-1} \right\}.$$

Hence, we have

$$\sup_{u \in G} \langle u, Z \rangle = \max_{|J|=2k} \sup_{u \in G_J} \langle u, Z \rangle = \max_{|J|=2k} \|Z_J\|_2.$$

Fix an index set J such that $|J| = 2k$. Let $\mathcal{U} = \{u_1, \ldots, u_N\}$ be an ϵ-net of G_J. Thus, the set of vectors form a cover of a $2k$ dimensional sphere. Thus,

$$N = N(S^{2k-1}, \|\cdot\|_2, \epsilon) \leq \left(\frac{3}{\epsilon}\right)^{2k}.$$

Now, $\forall u \in G_J$, $\exists i \in [N]$ such that $\|u - u_i\|_2 \leq \epsilon$. Thus, $\exists r \in \sqrt{2} G_J$ such that $u = u_i + r$. Thus we have

$$\sup_{u \in G_J} \langle u, Z \rangle \leq \max_{i \in [N]} \langle u_i, Z \rangle + \sup_{r \in \sqrt{2} G_J} \langle r, Z \rangle.$$

Now, we know that

$$\sup_{r\in\sqrt{2}G_J}\langle r,Z\rangle \le \sqrt{2}\epsilon \sup_{u\in G_J}\langle u,Z\rangle .$$

Using this, we have

$$\sup_{u\in G_J}\langle u,Z\rangle \lesssim \max_{i\in[N]}\langle u_i,Z\rangle ,$$

when ϵ is an appropriately chosen constant. Here $\langle u_i,Z\rangle \sim \mathcal{N}(0,1)$ as $\|u_i\|_2=1$.

Since $\binom{p}{2k}$ choices of index sets are possible, we bound the tail probability using union bound as

$$\begin{aligned}
\mathbb{P}\left[\sup_{u\in G}\langle u,Z\rangle > t\right] &\le \sum_{|J|=2k}\mathbb{P}\left[\sup_{u\in G_J}\langle u,Z\rangle > t\right] \\
&\le \sum_{|J|=2k}\sum_{i\in[N]}\mathbb{P}\left[\langle u_i,Z\rangle > t\right] \\
&\le \binom{p}{2k}\exp(ck)Q(t) \le \exp\left(2k\log\frac{p}{2k}\right)\exp(ck)\exp\left(-\frac{t^2}{2}\right),
\end{aligned}$$

where the last step follows from bounding the size of the ϵ-net and the Q-function. Thus, for $t \asymp \sqrt{k\log\frac{ep}{k}}$, (scaled by an appropriately large constant), the tail probability is arbitrarily low. Thus, with high probability,

$$\sup_{u\in G}\langle u,Z\rangle \lesssim \sqrt{k\log\frac{pe}{k}}.$$

19.1.4 Risk upper bound using tail bound for χ^2 distribution

As observed earlier,

$$\sup_{u\in G}\langle u,Z\rangle = \max_{|J|=2k}\|Z_J\|_2.$$

Since $Z\sim\mathcal{N}(0,I_p)$, $\|Z_J\|_2^2 \sim \chi^2_{2k}$ for a given J. We first study a few properties of the χ^2 random variable.

Let $L\sim\chi^2_m$. Then, $\mathbb{E}[L]=m$, $\mathrm{var}(L)\asymp m$ *i.e.*$\sigma_L \asymp \sqrt{m}$.

Theorem 19.1 ([**?**]). *If $L\sim\chi^2_m$, then*

$$\begin{aligned}
\mathbb{P}\left[L-m > S\sqrt{m}+S^2\right] &\le \exp\left(\frac{-S^2}{2}\right) \\
\mathbb{P}\left[L-m < -S\sqrt{m}\right] &\le \exp\left(\frac{-S^2}{2}\right).
\end{aligned}$$

Now, applying the above concentration inequality for $m=2k$, $S=\sqrt{ck\log\frac{p}{k}}$, we have

$$\mathbb{P}\left[\|Z_J\|_2^2 > 2k+k\sqrt{c\log\frac{p}{k}}+ck\log\frac{p}{k}\right] \le \mathbb{P}\left[\|Z_J\|_2^2 > k\log\frac{pe}{k}\right] \qquad \le \exp\left(\frac{-ck\log\frac{p}{k}}{2}\right).$$

Thus, with high probability,

$$\sup_{u\in G}\langle u,Z\rangle \lesssim \sqrt{k\log\frac{pe}{k}}.$$

19.2 Thresholding schemes and Risk upper bounds

19.2.1 Hard and Soft thresholding

For the denoising problem defined above, the *hard thresholding* estimate corresponding to the threshold τ is given by

$$\hat{\theta}_{\mathrm{HT}}(y)_i = \begin{cases} y_i, & \text{if } |y_i| > \tau \\ 0, & \text{if } |y_i| \leq \tau \end{cases}$$

Similarly, the *soft thresholding* estimate is given by

$$\hat{\theta}_{\mathrm{ST}}(y)_i = \begin{cases} y_i - \tau, & \text{if } y_i > \tau \\ y_i + \tau, & \text{if } y_i < -\tau \\ 0, & \text{if } |y_i| \leq \tau \end{cases}$$

The HT estimate is not continuous and the corresponding risk function does not vary monotonically. On the other hand, the soft thresholding avoids both these issues.

The hard and soft thresholding estimators can alternatively be written in the form of penalized objective functions. Consider the problem defined as follows:

$$\theta'(y) = \arg\min_{\tilde{\theta} \in \mathbb{R}^p} \|y - \tilde{\theta}\|_2^2 + \lambda \|\tilde{\theta}\|_0.$$

Then, for appropriately chosen penalty factor λ, $\theta'(y) = \hat{\theta}_{\mathrm{HT}}(y)$. Similarly, for the problem

$$\theta'(y) = \arg\min_{\tilde{\theta} \in \mathbb{R}^p} \|y - \tilde{\theta}\|_2^2 + \lambda \|\tilde{\theta}\|_1,$$

for appropriately chosen λ, $\theta'(y) = \hat{\theta}_{ST}(y)$.

Note: Under such thresholding schemes, we may not necessarily obtain a k-sparse vector as we desire. However, we shall ignore this fact as we are interested in only the risk upper bounds.

19.2.2 ℓ_∞-constrained procedure

Consider the following ℓ_∞-constrained formulation of the problem

$$\hat{\theta}(y) \in \arg \min_{\tilde{\theta} \in \mathbb{R}^p : \|y - \tilde{\theta}\|_\infty \leq \tau} \|\tilde{\theta}\|_0.$$

We observe that the hard thresholding estimate is a feasible solution to the above problem. (The set that minimizes the above objective function is in reality a continuum of points.) The constraint of interest is that $\|y - \tilde{\theta}\|_\infty \leq \tau$. Thus, setting $\tilde{\theta}_i = 0$ when $|y_i| \leq \tau$ and $\tilde{\theta}_i = y_i$ when $|y_i| > \tau$ satisfies the constraint. Further, this estimate also minimizes the ℓ_0 norm and thus $\hat{\theta}(y)$ is a feasible solution.

Theorem 19.2. *For all $\theta \in B_0(k)$, $\hat{\theta}$ a feasible solution to the above problem, for $\tau = \sqrt{2 \log p}$, with high probability,*

$$\|\hat{\theta} - \theta\|_2^2 \leq 16 k \log p.$$

Proof. We shall decompose the proof into three steps.

Step 1: Set τ to ensure feasibility of ground truth.

Since $Y = \theta + Z$,

$$\|y - \theta\|_\infty = \|Z\|_\infty \lesssim \sqrt{2 \log p} \quad \text{whp.}$$

Thus we observe that the ground truth is feasible high probability.

Step 2: Analyze structure of error.

The error is given by $h = \hat{\theta} - \theta$. Since θ is a feasible solution,

$$\|\hat{\theta}\|_0 \leq \|\theta\|_0 \leq k.$$

Thus, $\|h\|_0 \leq 2k$.

Step 3: Bound ℓ_2 norm.

$$\begin{aligned} \|h\|_2^2 &\leq \|h\|_\infty^2 \|h\|_0 \\ &\leq 2k \|\hat{\theta} - \theta\|_\infty^2 \\ &\leq 2k (\|\hat{\theta} - y\|_\infty + \|y - \theta\|_\infty)^2 \qquad (19.4) \\ &\leq 8k\tau^2 = 16k \log p, \end{aligned}$$

where (19.4) follows from the triangle inequality. We note that all the above statements hold with high probability following the statement of feasibility. □

Similarly, consider the problem

$$\hat{\theta}(y) \in \arg \min_{\tilde{\theta} \in \mathbb{R}^p : \|y - \tilde{\theta}\|_\infty \leq \tau} \|\tilde{\theta}\|_1.$$

We observe here that for any $\tilde{\theta}$ satisfying the constraint, $\|\tilde{\theta}\|_1 \geq \sum_{i=1}^p (|y| - \tau) \mathbf{1}\{|y| > \tau\}$. The soft thresholding estimate satisfies the above bound and the constraint and is thus a feasible solution to the problem.

Theorem 19.3. *For all $\theta \in B_0(k)$, $\hat{\theta}$ a feasible solution to the above problem, for $\tau = \sqrt{2 \log p}$, with high probability,*

$$\|\hat{\theta} - \theta\|_2^2 \leq 32k \log p.$$

Proof. We proceed in similar fashion to the earlier proof.

Step 1: Set τ to ensure feasibility of ground truth.

Since $Y = \theta + Z$,

$$\|y - \theta\|_\infty = \|Z\|_\infty \lesssim \sqrt{2 \log p} \quad \text{whp.}$$

Thus we observe that the ground truth is feasible with high probability.

Step 2: Analyze structure of error.

The error is given by $h = \hat{\theta} - \theta$. Thus $\|h\|_\infty \le 2\tau$. Let $J = \text{supp}(\theta)$. Define the cone

$$\mathcal{C} = \{x \in \mathbb{R}^p : \|x_{J^c}\|_1 \le \|x_J\|_1\}.$$

We now have

$$\|h_J\|_1 - \|h_{J^c}\|_1 = \sum_{i \in J} |\hat{\theta}_i - \theta_i| - \sum_{i \in J^c} |\hat{\theta}_i| \ge \|\theta\|_1 - \|\hat{\theta}\|_1 \ge 0,$$

which follows from the triangle inequality and the feasibility of θ. Thus $h \in \mathcal{C}$.

Step 3: Bound ℓ_2 norm.

$$\begin{aligned} \|h\|_2^2 &\le \|h\|_1 \|h\|_\infty && (19.5) \\ &\le 4\tau \|h_J\|_1 \\ &\le 4\tau\sqrt{k}\|h_J\|_2 && (19.6) \\ &\le 4\tau\sqrt{k}\|h\|_2 \\ \Leftrightarrow \|h\|_2^2 &\le 32k \log p, \end{aligned}$$

where (19.5) and (19.6) follow from Holder's inequality and Cauchy-Schwarz inequality respectively. □

Remark 19.2 (Approximate Sparsity)**.** Let J be a set of indices of size k. Let $h \in \mathcal{C} = \{x \in \mathbb{R}^p : \|x_{J^c}\|_1 \le \|x_J\|_1\}$. Consider the set of k largest elements in h_{J^c} indexed by the set K. Then,

$$\|h_{(J \cup K)^c}\|_2^2 \ge \frac{1}{2}\|h\|_2^2.$$

Proof. For every element, we have

$$|h_{J^c}^{(i)}| \le \frac{1}{i}\|h_{J^c}\|_1.$$

Thus,

$$\begin{aligned} \|h_{J^c}\|_2^2 &\le \sum_{i=k+1}^{p-k} |h_{J^c}^{(i)}|^2 \le \sum_{i=k+1}^{p-k} \frac{1}{i^2}\|h_{J^c}\|_1^2 \\ &\le \frac{1}{k} k \|h_{J^c}\|_2^2 \le \|h_{J \cup K}\|_2^2, \end{aligned}$$

which follows from Cauchy-Schwarz inequality and the fact that $h \in \mathcal{C}$. □

Remark 19.3.

1. When the vector is sufficiently sparse, specifically $k = o(p)$,

$$R^* \le (2 + o(1))k \log \frac{p}{k}.$$

   Further, the bound can be achieved in the adaptive case too.

2. If $k = \Theta(p)$, i.e, $\frac{k}{p} \to \alpha \in (0, 1]$ as $p \to \infty$, then,

$$R^* \asymp p\left(\beta(\alpha) + o(1)\right),$$

   where $\beta(\alpha)$ is a constant dependent on α.

References