

Lecture 19: Denoising via Shrinkage

Lecturer: Yiqiao Zhong

Last update: 18:59:59, 2024-12-05

Disclaimer: materials in this pdf are used for STAT 709 “Mathematical Statistics” at the University of Wisconsin–Madison, but they are not guaranteed to be accurate. Consult textbooks whenever possible.

We have seen that Stein’s shrinkage may help improve estimation. However, the effect of shrinkage is most pronounced if we exploit the *sparsity* structure, which is common in many problems.

1 Sparsity in data analysis

1.1 Motivation

Q: Why is sparsity important?

- (1) Occam’s razor. Better interpretability. Less is more.
- (2) Variable selection. We collect more and more covariates or features in “Big Data Age”, but important ones are not many.
- (3) Biological mechanism. Nature favors sparsity. Sparse coding in primary visual cortex, genetics, etc.
- (4) Computation & Memory. Data compression.

1.2 Quantifying sparsity

For a vector $x \in \mathbb{R}^n$, we say that x is s -sparse if

$$\|x\|_0 := \sum_{j=1}^n \mathbb{1}(x_j \neq 0) \leq s.$$

In other words, the number of nonzeros in x is no larger than s . For this reason, it is often called exact sparsity.

We can relax this definition by allowing entries of x to have very small nonzero values. For $q \in (0, 2]$, we introduce “ ℓ_q ” norm

$$\|x\|_q := \left(\sum_{j=1}^n |x_j|^q \right)^{1/q} \leq s.$$

A few comments are in order. First, when $q < 1$, the above definition is only a quasinorm, since it does not satisfy the triangle inequality, but people often call it ℓ_q norm anyway. Second, as q gets smaller, the measurement is more stringent: the region of $\|x\|_q \leq 1$ is smaller (Fig. 1). Third, we have $\|x\|_0 = \lim_{q \rightarrow 0} \|x\|_q^q$.

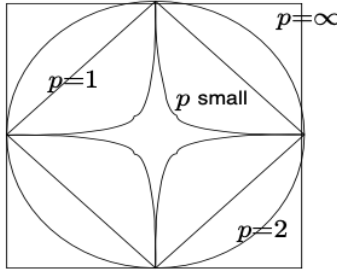


Figure 1.5 Contours of ℓ_p balls

Figure 1: A figure depicting the region $\{x : \|x\|_q \leq 1\}$ for different q .

1.3 Smoothness and sparsity are twins

Smoothness is an important property in many fields and arises naturally in physical systems. Let's consider a simple analysis to show that smoothness and sparsity are closely related.

Suppose that $f(x)$ is a function satisfying $f(-\pi) = f(\pi)$ (which can be extended to a periodic function on $\mathbb{R}$). Under some conditions, we can expand f into Fourier series.

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} [A_k \cos(kx) + B_k \sin(kx)]$$

The coefficients are determined by

$$C_k = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) e^{ikx} dx$$

and A_k, B_k are the real and imaginary part of C_k respectively. Suppose that f is smooth in the sense that bounded high-order derivatives: for some integer $m \geq 1$,

$$\sup_{x \in [-\pi, \pi]} \left| \frac{d^m}{dx^m} f(x) \right| \leq C.$$

Then by integration by parts,

$$C_k = \frac{(-1)^m}{(ik)^m} \cdot \frac{1}{\pi} \int_{-\pi}^{\pi} \frac{d^m}{dx^m} f(x) e^{ikx} dx \quad \implies \quad |C_k| \leq \frac{2}{k^m}$$

So the high-frequency coefficients decay very fast. (We can measure the sparsity using ℓ_q norm with $q \sim 1/m$).

2 Example: Wavelet transform*

We show an example of the NMR (nuclear magnetic resonance) data taken from [1].

2.1 Discrete Fourier transformation

Given $X \in \mathbb{R}^N$, define

$$\hat{X}_k = \sum_{n=0}^{N-1} X_n \exp\left(-\frac{i2\pi kn}{N}\right), \quad k = 0, 1, \dots, N-1.$$

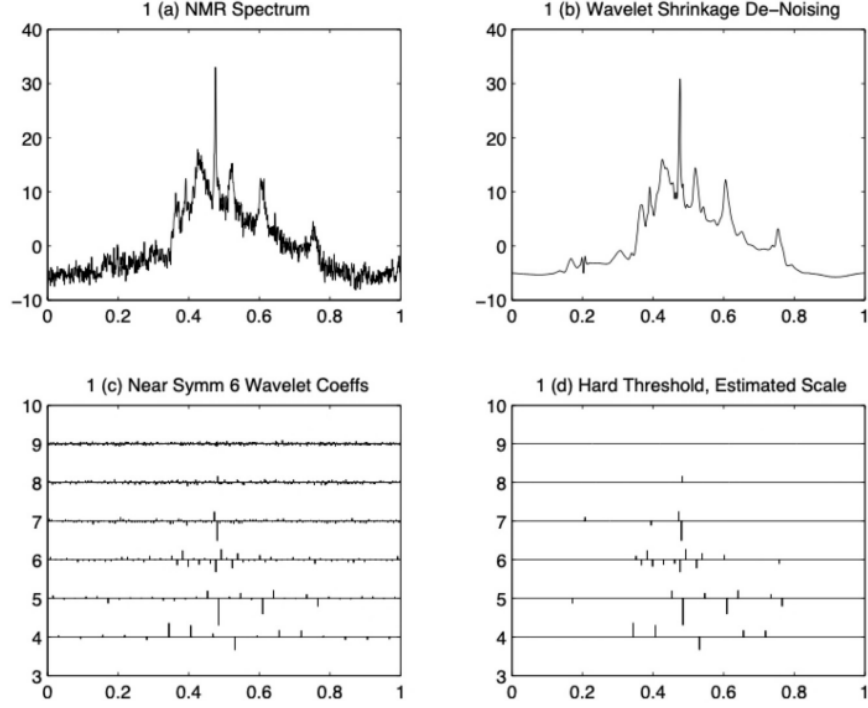


Figure 2: An example of wavelet transform where denoising is applied to estimate wavelet coefficients.

Let

$$F = \left(\exp\left(-\frac{iw\pi kn}{N}\right) \right)_{0 \leq k, n < N} \in \mathbb{C}^{N \times N},$$

be an orthogonal matrix. Then,

$$\hat{X} = FX,$$

is essentially a change of basis, where

$$F : \text{time domain} \longleftrightarrow \text{frequency domain}.$$

2.2 Wavelet transform

$$\begin{cases} \text{Fourier: localized in frequency.} \\ \text{Wavelet: localized in both time and frequency.} \end{cases}$$

A generic form:

$$\psi_{jk}(t) = 2^{j/2} \psi(2^j t - k),$$

where in the RHS, ψ is the mother wavelet, 2^j is the scale, and k is the shift. $\{\psi_{jk}\}$ forms an orthonormal basis for $L_2(\mathbb{R})$. Pair (j, k) indicates wavelet ψ_{jk} , at resolution j and approximate location $k2^{-j}$. For discrete

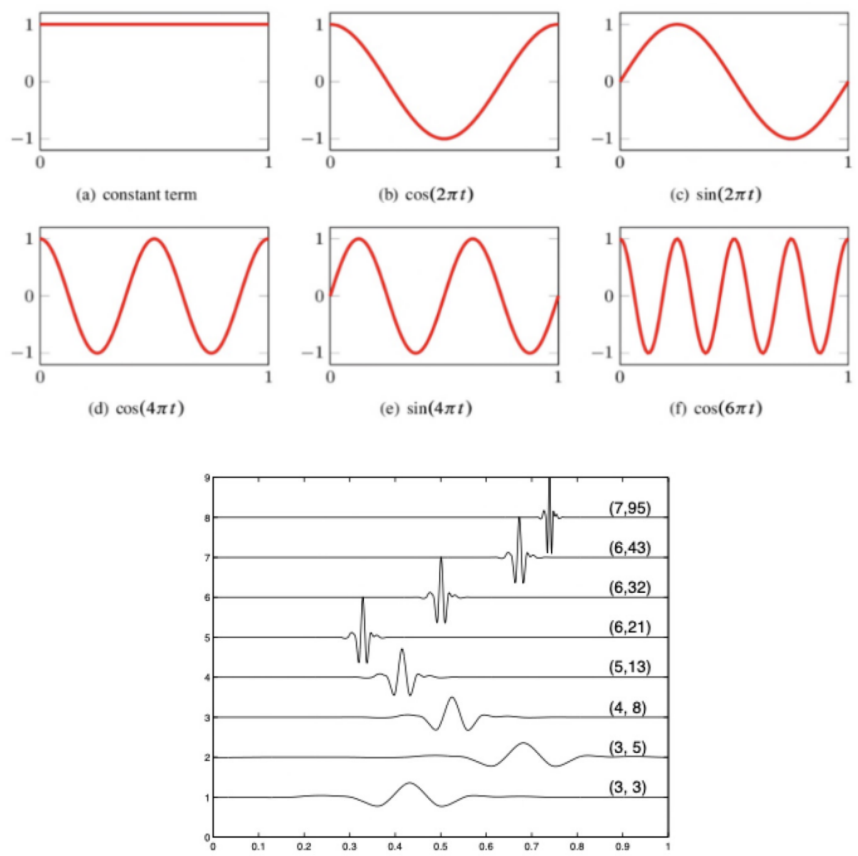


Figure 3: Selected vectors in the wavelet basis, which has a frequency index and a location index.

data $X = (t_l)_{l \leq N}$, define $\hat{X} = WX$, where $W = O(N)^1$ is discrete wavelet matrix. Often $\hat{X}$ is sparse or approximately sparse. $X = W^\top \hat{X}$ is sparse representation.

$$\begin{cases} \text{Smoothness of functions} \implies \text{Fourier coefficients decay fast.} \\ \text{(Piecewise) smoothness of functions} \implies \text{Wavelet coefficients sparse / decay fast.} \end{cases}$$

3 Thresholding

3.1 Gaussian sequence model

We assume that our observations follow the model.

$$Y_k = \theta_k + \sigma Z_k, \quad \text{where } Z_k \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1). \quad (1)$$

The observations (Y_k) can be regarded as transformed or preprocessed measurements. We have raw measurements $X \in \mathbb{R}^n$, target (unknown) discretized function $(f(t_k))_{k \leq n}$,

$$X_k = f(t_k) + \sigma \omega_k, \quad \omega_k \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1).$$

¹ W can be computed in time $O(N)$.

Apply orthogonal transform $Y = WX$ where W is an orthogonal matrix (constructed from discrete Fourier transform or wavelet transform). Then we introduce

$$(\theta_1, \dots, \theta_n)^\top = W(f(t_1), \dots, f(t_n))^\top.$$

Since standard Gaussian random vector remains standard Gaussian after orthogonal transformation, we obtain (1). In vector form, That is,

$$Y \sim \mathcal{N}(\theta, \sigma^2 I_n). \quad (2)$$

Assumption. We assume that $\theta \in \mathbb{R}^n$ is sparse, such that

$$(\text{exact sparsity}) \quad \|\theta\|_0 = \sum_{j=1}^n \mathbb{1}(\theta_j \neq 0) \leq S.$$

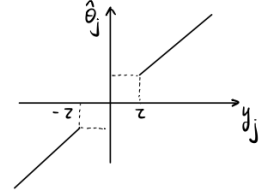
3.2 Thresholding functions

Without the sparsity assumption, we may estimate $\theta \in \mathbb{R}^n$ via either the unbiased natural estimator Y or James-Stein estimator $(1 - \frac{(p-2)\sigma^2}{\|Y\|^2})Y$ (assuming σ^2 is known).

We exploit the sparsity assumption by estimation via two typical thresholding functions $\hat{\theta}_j = T_\tau(Y_j)$.

(1) **Hard Thresholding.**

$$\hat{\theta}_j = \begin{cases} Y_j, & \text{if } |Y_j| > \tau \\ 0, & \text{if } |Y_j| \leq \tau. \end{cases}$$



(2) **Soft Thresholding.**

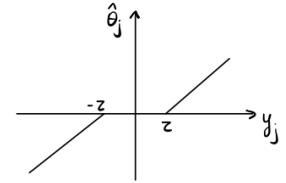
$$T_\tau(u) = \text{sgn}(u)(|u| - \tau)_+ = \arg \min_{v \in \mathbb{R}} \left\{ \frac{1}{2}(v - u)^2 + \tau|v| \right\}.$$

Determining thresholding level τ . From lectures on convergence modes, we know that

$$\|Z\|_\infty \xrightarrow{p} \sqrt{2\sigma^2 \log n} \quad \text{as } n \rightarrow \infty.$$

So we choose (assuming we know σ^2 , which is estimated separately)

$$\tau = \sqrt{2\sigma^2 \log n}.$$



We can have more accurate results for $\|Z\|_\infty$. By Gaussian concentration inequality, $\max_{1 \leq i \leq n} Z_i$ is close to $\mathbb{E}[\max_{1 \leq i \leq n} Z_i]$: the difference is bounded by $O(1)$ with high probability. Moreover, it is known that

$$\sqrt{2 \log n} - 1 - \sqrt{2 \log 2} \leq \frac{1}{\sigma} \mathbb{E} \max_{1 \leq i \leq n} Z_i \leq \sqrt{2 \log n}.$$

So choosing $\tau = (1 + 0.01)\sqrt{2\sigma^2 \log n}$ is guaranteed to remove all noise with high probability.

Summary. The transform-shrinkage approach is

$$(X_k) \xrightarrow{W} Y \xrightarrow[\text{Shrinkage}]{T_\tau} \hat{\theta} \xrightarrow{W^{-1}} (\hat{f}(t_k)).$$

4 Denoising theory

4.1 General oracle inequalities

Let us evaluate how well thresholding estimators perform. Consider the MSE as risk function $R(\theta) = \mathbb{E}[(\hat{\theta} - \theta)^2]$ for an estimator $\hat{\theta}$. In this section, our results hold *without any sparsity assumption*.

Suppose there is an oracle, who tells us the locations (indices) of large signals θ_j : we estimate θ_j by $\hat{\theta}_j$ and incur MSE σ^2 , and by 0 and incur MSE θ_j^2 otherwise.

We define the (oracle-aided) ideal risk as follows.

$$R(\theta, \sigma) = \sum_{i=1}^n \min\{\theta_i^2, \sigma^2\}.$$

In particular, if θ is S -sparse, then

$$R(\theta, \sigma) = s\sigma^2.$$

Theorem 1. *Suppose $Y \sim \mathcal{N}(\theta, \sigma^2 I_n)$. For soft-thresholding estimator $\hat{\theta}_{\tau_n} = T_{\tau_n}(Y)$ with the level $\tau_n = \sqrt{2\sigma \log n}$,*

$$\mathbb{E}[\|\hat{\theta}_{\tau_n} - \theta\|_2^2] \leq (2 \log n + 1)[\sigma^2 + R(\theta, \sigma)].$$

Similar results hold for other (e.g., hard) thresholding. One striking aspect of this result is that we don't know the indices of large signals among $\theta_1, \dots, \theta_n$. Compared with the ideal risk, the risk is increased only by a log factor.

Remark 1. *Theorem 1 holds with no assumption on θ . Without assumption on θ , this is not improvable due to the following minimax result [1, Proposition 8.8]. As $n \rightarrow \infty$,*

$$\inf_{\hat{\theta}} \sup_{\theta \in \mathbb{R}^n} \frac{\mathbb{E}[\|\hat{\theta} - \theta\|^2]}{\sigma^2 + \sum_{i=1}^n \min\{\theta_i^2, \sigma^2\}} \geq (2 \log n) \cdot (1 + o(1)).$$

4.2 Oracle Inequalities for S -sparse Signals

Denote the space of S -sparse vectors as

$$\Theta_{n,s} = \{\theta \in \mathbb{R}^n : \|\theta\|_0 \leq s\}.$$

Theorem 2. *Suppose $S = S_n$ satisfies*

$$\frac{S_n}{n} \rightarrow 0, \quad n \rightarrow \infty.$$

Then, as $n \rightarrow \infty$,

(1) *The soft-thresholding estimator $\hat{\theta}_{\tau_n}$ with $\tau_n = \sqrt{2\sigma^2 \log(n/S_n)}$,*

$$\sup_{\theta \in \Theta_{n,s}} \mathbb{E}[\|\hat{\theta}_{\tau_n} - \theta\|^2] \leq (1 + o(1)) \cdot 2S_n \sigma^2 \log\left(\frac{n}{S_n}\right).$$

(2) *The soft-thresholding estimator matches the minimax lower bound:*

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta_{n,s}} \mathbb{E}[\|\hat{\theta}_{\tau_n} - \theta\|^2] \geq (1 + o(1)) \cdot 2S_n \sigma^2 \log\left(\frac{n}{S_n}\right).$$

Proof of (1). Let us simply write $\widehat{\theta}$ instead of $\widehat{\theta}_{\tau_n}$ and define

$$\mathcal{S} = \{i \leq n : \theta_i \neq 0\}.$$

The high-level idea is the following: if $i \notin \mathcal{S}$, most noise Y_i will be denoised by $\widehat{\theta}$. If $i \in \mathcal{S}$, the estimator incurs additional bias. We will conduct coordinate-wise analysis since

$$\|\widehat{\theta} - \theta\|^2 = \sum_{j=1}^n (\widehat{\theta}_j - \theta_j)^2$$

and divide our analysis into two cases.

Case I: If $j \notin \mathcal{S}$, namely $\theta_j = 0$,

$$\begin{aligned} \mathbb{E}[(\widehat{\theta}_j - \theta_j)^2] &= \mathbb{E}[(\sigma|Z_j| - \tau)_+^2] \\ &= \sigma^2 \int_0^\infty \mathbb{P}((|Z_j| - \frac{\tau}{\sigma})_+^2 > t) dt \quad (\text{by Fubini's theorem}) \\ &= \sigma^2 \int_0^\infty \mathbb{P}(|Z_j| - \frac{\tau}{\sigma} > \sqrt{t}) dt \\ &\leq 2\sigma^2 \int_{\tau/\sigma}^\infty (u - \frac{\tau}{\sigma}) \mathbb{P}(|Z_j| \geq u) du \quad (\text{change of variables } u := \sqrt{t} + \frac{\tau}{\sigma}) \\ &\leq 4\sigma^2 \int_{\tau/\sigma}^\infty u \exp\left(-\frac{u^2}{2}\right) du \quad (\text{by Gaussian tail probability bound}) \\ &= 4\sigma^2 \exp\left(-\frac{\tau^2}{2\sigma^2}\right). \end{aligned}$$

Case II: If $j \in \mathcal{S}$ namely $\theta_j \neq 0$,

$$\begin{aligned} \mathbb{E}[(\widehat{\theta}_j - \theta_j)^2] &= \mathbb{E}[(\widehat{\theta}_j - T_\tau(\theta_j) + T_\tau(\theta_j) - \theta_j)^2] \\ &\leq \mathbb{E}[2(T_\tau(Y_j) - T_\tau(\theta_j))^2 + 2(T_\tau(\theta_j) - \theta_j)^2] \quad (\text{since } (a+b)^2 \leq 2a^2 + 2b^2) \\ &= \underbrace{2\mathbb{E}[(Y_j - \theta_j)^2]}_{\text{estimation error}} + \underbrace{2\tau^2}_{\text{bias}} \quad (\text{since } T_\tau \text{ is 1-Lipschitz}) \\ &= 2\sigma^2 + 2\tau^2. \end{aligned} \tag{3}$$

Combining the two cases: there are $n - S_n$ (bounded trivially by n) indices in Case I and S_n indices in Case II, so

$$\begin{aligned} \mathbb{E}[\|\widehat{\theta} - \theta\|^2] &\leq 2S_n(\sigma^2 + \tau_n^2) + 4n\sigma^2 \exp\left(-\frac{\tau_n^2}{2\sigma^2}\right) \\ &= 2S_n(\sigma^2 + 2\sigma^2 \log\left(\frac{n}{S_n}\right)) + 4S_n\sigma^2 \\ &= 6S_n\sigma^2 + 4S_n\sigma^2 \log\left(\frac{n}{S_n}\right) \end{aligned}$$

Inspecting the proof, we can improve the upper bound to $(1 + o(1)) \cdot 2S_n\sigma^2 \log(n/S_n)$ with a more careful inequality in (3) ([exercise](#)). $\square$

Remark 2 (Comparison with James-Stein estimator). *The James-Stein estimator fails on sparse signal, since it does not exploit the sparsity structure of the signals. Consider for simplicity $\sigma = 1$, we know that²*

$$\mathbb{E}[\|\widehat{\theta}_{\text{JS}} - \theta\|^2] \leq n - (n-2)^2 \mathbb{E}\left[\frac{1}{\|X\|^2}\right] \geq \frac{n\|\theta\|^2}{n + \|\theta\|^2}.$$

²Result taken from [1][Page 36], but I cannot find a proof.

So, we have risk lower bound in simple scenarios: θ is S -sparse with S signals having equal magnitude n/S with $\|\theta\|^2 = n$, then

$$\mathbb{E}[\|\hat{\theta}_{\text{JS}} - \theta\|^2] \geq \frac{n}{2}.$$

Remark 3 (Bias-variance tradeoff). Note that our thresholding level is $\sqrt{2\sigma^2 \log(n/S_n)}$ instead of $\sqrt{2\sigma^2 \log(n)}$. It has a smaller thresholding level, so many noises are not completely removed (reaching 0). This smaller thresholding level introduces a bit more variances from the noises, but incurs much smaller biases on large signals, thus leading to the minimax risk.

5 LASSO with orthogonal design

Suppose we have data $(x_i, y_i)_{1 \leq i \leq n}$, where $X_i \in \mathbb{R}^p$, $Y_i \in \mathbb{R}$. As usual, we write our data in matrix form

$$\begin{cases} X = [x_1, \dots, x_n]^\top \in \mathbb{R}^{n \times p}, \\ Y = [y_1, \dots, y_n]^\top \in \mathbb{R}^n. \end{cases}$$

The celebrated LASSO, an abbreviation for *least absolute shrinkage and selection operator* (who remembers the full name?), solves the following optimization problem. Given $\lambda \geq 0$, solve

$$\min_{\theta \in \mathbb{R}^p} \frac{1}{2n} \|X\theta - Y\|_2^2 + \lambda \|\theta\|_1 =: \min_{\theta \in \mathbb{R}^p} \ell_{n,\lambda}(\theta).$$

5.1 Connecting LASSO with denoising

Consider the orthogonal design case, which means X is an orthogonal matrix. For now, assume $n = p$, and $X \in O(n)$. By orthogonality,

$$\|Xu\| = \|X^\top u\| = \|u\|, \quad \forall u \in \mathbb{R}^n,$$

which implies that

$$\ell_{n,\lambda}(\theta) = \frac{1}{2n} \|\theta - X^\top y\|_2^2 + \lambda \|\theta\|_1 = \frac{1}{n} \sum_{j=1}^n [(\theta_j - Z_j)^2 + \lambda |\theta_j|],$$

where $z := X^\top y \in \mathbb{R}^n$. The loss $\ell_{n,\lambda}(\theta)$ is called “separable”, which can be solved coordinate-wisely (why?):

$$\arg \min_{\theta \in \mathbb{R}^p} \ell_{n,\lambda}(\theta) = (T_\lambda(z_1), T_\lambda(z_2), \dots, T_\lambda(z_p))^\top.$$

We can analyze LASSO with an orthogonal design using our results from denoising.

5.2 Improvement of LASSO*

Since

$$X^\top y \sim \mathcal{N}(\theta^*, \sigma^2 I_n).$$

Even when z_j is large (strong signal), we shrink our estimate by λ because $T_\lambda(z_j)$ always applies shrinkage. So, the bias of soft-thresholding/LASSO is not asymptotically zero for large signals.

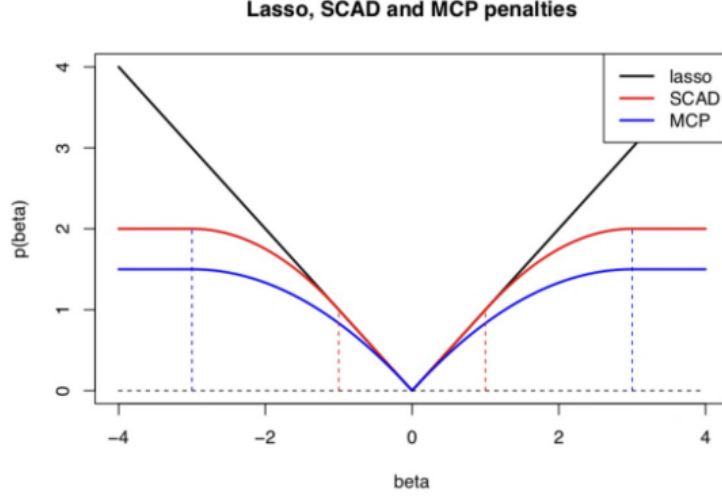


Figure 4: Three popular regularizers for sparse linear regression.

Two improvements: SCAD (smoothly clipped absolute deviation) and MCP (minimax concave penalty). Consider the scalar case $\ell_{1,\lambda}(\theta)$ where $\theta \in \mathbb{R}$:

$$\ell_{1,\lambda}(\theta) = \frac{1}{2} \|\theta - z\|^2 + P_\lambda(\theta)$$

where the regularizer $P_\lambda(\theta)$ is depicted in Figure 4.

- Instead of the ℓ_1 regularizer, we use a different regularizer P_λ that is flat after a threshold.
- For larger signals, it applies small shrinkage to incur less bias.

Similar to LASSO, both SCAD and MCP have a non-differential point at $\theta = 0$. But if a minimizer is $\hat{\theta} > 0$, Recall that a minimizer is necessarily the first-order stationary condition: if $\theta > 0$,

$$\ell'_{1,\lambda}(\theta)|_{\theta=\hat{\theta}} = 0. \quad \Longleftrightarrow \quad \hat{\theta} - z + P'_\lambda(\hat{\theta}) = 0.$$

If $P'_\lambda(\hat{\theta}) = 0$, then there is no bias in $\hat{\theta}$. We note that

$$\begin{cases} P'_\lambda(\hat{\theta}) \equiv \lambda, & \text{in LASSO, inducing a bias regardless} \\ P'_\lambda(\hat{\theta}) = 0, & \text{if } \hat{\theta} \text{ large in SCAD / MCP, usually inducing no bias when true signal } \theta \text{ is large.} \end{cases}$$

Pros and Cons. Compared with LASSO, MCP / SCAD both tend to denoise small signals and induce no bias for large signal, thus achieving better statistical accuracy. However, unlike LASSO, they involve a nonconvex optimization problem! It is harder (though still possible under additional conditions) to solve the following nonconvex problem under a general design matrix.

$$\frac{1}{2} \|y - X\theta\|_2^2 + \sum_{j=1}^p P_\lambda(\theta_j).$$

6 Further reading

The main reference of shrinkage theory is [1]. The theorem in Section 4.2 can be found in Chapter 8.4 of [1].

References

- [1] Iain Johnstone. *Gaussian estimation: Sequence and wavelet models*. 2017.