ECE 830 Fall 2010 Statistical Signal Processing

instructor: R. Nowak , **scribe:** Haiyan Xu

Lecture 20: Signal Subspaces and Sparsity

1 Signal Subspaces and Sparsity

Recall the classical linear signal model:

$$X = H\theta + w, \quad w \sim N(0, \sigma_w^2 I)$$

Where $S = H\theta$, is a linear-parametric model for the signal and w is noise. Here H is a known $n \times k$ matrix, whose columns span the signal subspace, and $\theta \in \mathbb{R}^k$ are the signal parameters.

MLE:

The MLE of θ is:

$$\hat{\theta}_{MLE} = (H^T H)^{-1} H^T X$$

and the MLE of the signal is:

$$\hat{S} = H\hat{\theta} = \underbrace{H(H^T H)^{-1} H^T}_{P_H} X$$

Where $P_H = H(H^H H)^{-1} H^T$

Wiener filter:

The Bayes MMSE estimator based on a prior $\theta \sim N(0, \sigma_\theta^2 I)$ is the Wiener filter:

$$\hat{\theta}_{wiener} = H^T \Big(HH^T + \frac{\sigma_w^2}{\sigma_\theta^2} I \Big)^{-1} X = \Big(\frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma^2} \Big) \hat{\theta}_{MLE}$$

This follows directly from the Gauss-Markov Theorem, also

$$H^T \Big(HH^T + \frac{\sigma_w^2}{\sigma_\theta^2} I \Big)^{-1} \longrightarrow (H^T H)^{-1} H^T \qquad \text{as} \qquad \frac{\sigma_w^2}{\sigma_\theta^2} \longrightarrow \infty$$

So in the high SNR situation, the Wiener filter acts essentially the same as the MLE; it projects X onto the signal subspace. At low SNR the Wiener filter "shrinks" the MLE to balance the tradeoff between $bias^2$ and variance.

2 Sparsity

In the classic set-up:

$$X = H\theta + w$$

we assume that we know the low-dimensional signal subspace. In many problems we may not have this information, but we might know that the signal lies in one of many subspaces in a certain trasform domain.

Example 1 *Narrowband communications*

The communication signal lies in one of many narrow frequency bands, but we may not know which band it will be in (e.g. frequency hopping communication).

Example 2 *Wavelet-based Image Processing*

The discrete wavelet trasform (DWT) is very effective at compressing natural images. In fact it is the basis of the JPEG 2000 Standard. The DWT 06 images tends to be "sparse" in the following sense.

If x is an image and w denotes the DWT, then the DWT coefficients $\theta = wX$ tend to be mostly zero (or very nearly zero). The locations of the relatively few non-zero (or significant) coefficients in the vector θ depend in a complicated way on x itself.

So, while image do approximately lie in a subspace of the wavelet domain, the subspace is different for each different image.

3 Sparse Signal Models

Let u be an $n \times n$ matrix whoes columns form an orthobasis for $\mathbb{R}$. For example, U could be the DFT or DWT. Consider the denoising problem:

$$X = H\theta + w, w \sim N(0, \sigma_w^2 I)$$

An equivalent observation model is:

$$\begin{aligned} u^T X &= u^T u\theta + u^T w \\ &= \theta + w' \end{aligned}$$

Where $w' \sim N(0, \sigma^2 u^T I u)$

Consider this model:

$$y = \theta + w, w \sim N(0, \sigma_w^2 I)$$

Where $y = u^T X$

If we make no assumption about the θ, then we might use the MLE:

$$\begin{aligned} \hat{\theta}_{MLE} &= u^T X = y, \\ \hat{S}_{MLE} &= X \end{aligned}$$

If we suppose that the coefficients tend to have a certain energy, then we could use the prior: $\theta \sim N(0, \sigma_\theta^2 I)$ and the wiener filter:

$$\hat{\theta}_{wiener} = u^T \Big(uu^T + \frac{\sigma_w^2}{\sigma_\theta^2} I \Big)^{-1} X$$

Since u is an orthonormal transform $uu^T = I$ and the wiener filter simplifies to:

$$\begin{aligned} \hat{\theta}_{wiener} &= u^T \Big(1 + \frac{\sigma_w^2}{\sigma_\theta^2} \Big)^{-1} IX \\ &= \Big(1 + \frac{\sigma_w^2}{\sigma_\theta^2} \Big)^{-1} u^T X \\ &= \Big(\frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_w^2} \Big) \cdot u^T X \\ &= \Big(\frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_w^2} \Big) \cdot \hat{\theta}_{MLE} \end{aligned}$$

But now suppose our prior knowledge about θ is that it is sparse; i.e. many or most of the coefficients are zero (or near zero). This is not captured by the Gaussian prior, which models every coefficient as a Gaussian random variable with power σ_θ^2. If many coefficients are zero, then many should have approximately zero power!

So we would like to design a prior probability density that reflects our belief that most of the coefficients are zero or near zero in magnitude.

Example 3 *Gaussian mixture*

Let $\theta_1, \ldots, \theta_n$ denote the coefficients and model them as follows:

$$\theta_i \overset{\text{iid}}{\sim} p \cdot N(0, \sigma_0^2) + (1-p) \cdot N(0, \sigma_1^2)$$

where i=1,. . . , n, with $\sigma_0^2 << \sigma_1^2$.

In words this prior is saying that a fraction p of the coefficients tend to be very small in magnitude (i.e. $|\theta_i| \sim \sigma_0$) and 1-p tend to be large.

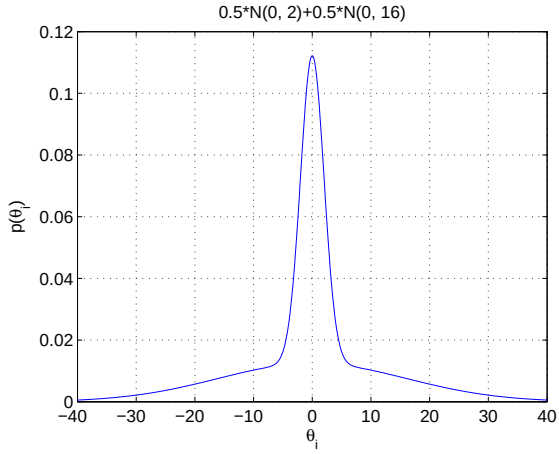

Figure 1: Example of Gaussian Mixture

Example 4 *Laplacian prior:*

$$\theta_i \overset{\text{iid}}{\sim} \frac{\lambda}{2} e^{-\lambda|\theta_i|} \quad , \qquad i = 1, \ldots, n$$

We will focus on the laplacian prior because it leads to very simple and intuitive solutions to the donoising problem and it is log-concave, which makes it computationablly tractable when used in inverse problems such as deconvolution.

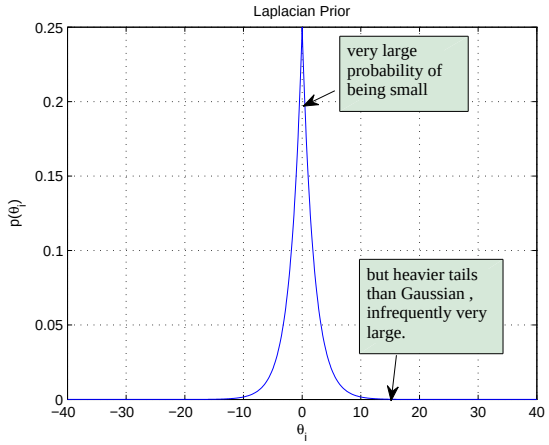

Figure 2: Example of Laplacian prior

4 Laplacian priors for sparsity

Assume the prior as:

$$p(\theta) = \prod_{i=1}^{n} p(\theta_i) = \prod_{i=1}^{n} \frac{\lambda}{2} e^{-\lambda|\theta_i|}$$

Observation model:

$$X = u\theta + w \quad , \quad w \sim N((0, \sigma^2 I)$$

equivalently

$$u^T x = \theta + u^T w$$

Note $u^T w \sim N(0, u^T(\sigma_w^2 I)u) \equiv N(0, \sigma^2 I)$

Let's define $y = u^T X$, then we have the model:

$$y = \theta + w \quad , \quad w \sim N(0, \sigma^2 I)$$

The likelihood of θ given y is :

$$p(y|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\theta_i)^2}{2\sigma^2}}$$

The posterior distribution of θ is:

$$\begin{aligned} p(\theta|y) &\propto p(y|\theta)p(\theta) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\theta_i)^2}{2\sigma^2}} \cdot \frac{\lambda}{2} e^{-\lambda|\theta_i|} \end{aligned}$$

Consider the MAP estimator:

$$\begin{aligned} \hat{\theta} &= \underset{\theta}{argmax} \quad p(\theta|x) \\ &= \underset{\theta}{argmax} \quad log(p(\theta|x)) \\ &= \underset{\theta}{argmax} \sum_{i=1}^{n} \Big[-\frac{(y_i-\theta_i)^2}{2\sigma^2} - \lambda|\theta_i| \Big] + constant \\ &= \underset{\theta}{argmin} \sum_{i=1}^{n} \Big[\frac{(y_i-\theta_i)^2}{2\sigma^2} + \lambda|\theta_i| \Big] \end{aligned}$$

If $|\theta_i| \neq 0$, then we can differentiate to obtain:

$$\begin{aligned} -\frac{(y_i-\theta_i)}{\sigma^2} + \lambda sign(\theta_i) = 0 \\ \longrightarrow \theta_i = y_i - \lambda\sigma^2 sign(\theta_i) \end{aligned}$$

and clearly the minimizer must have the same sign as y_i, and so:

$$\hat{\theta}_i = y_i - \lambda\sigma^2 sign(\theta_i)$$

Plugging this into the argument of the minimization yields:

$$\begin{aligned} \frac{(y_i-\hat{\theta}_i)^2}{2\sigma^2} + \lambda|\theta_i| &= \frac{\lambda^2\sigma^4}{2\sigma^4} + \lambda|y_i - \lambda\sigma^2 sign(y_i)| \\ &= \frac{\lambda^2\sigma^2}{2} + \lambda|y_i - \lambda\sigma^2 sign(y_i)| \end{aligned} \tag{1}$$

On the other hand if $\hat{\theta}_i = 0$, then the argument of the minimization is:

$$\frac{(y_i - \hat{\theta}_i)^2}{2\sigma^2} + \lambda|\theta_i| = \frac{y_i^2}{2\sigma^2} \tag{2}$$

Observe that:

$$(1) < (2), \quad \text{if } |y_i| > \lambda\sigma^2;$$
$$(1) > (2), \quad \text{if } |y_i| \leq \lambda\sigma^2$$

Therefore, the optimal solution is:

$$\hat{\theta}_i = \begin{cases} 0, & \text{if } |y_i| \leq \lambda\sigma^2 \\ y_i - \lambda\sigma^2 sign(y_i), & \text{if } |y_i| > \lambda\sigma^2 \end{cases}$$

Graphically, the operation is this:

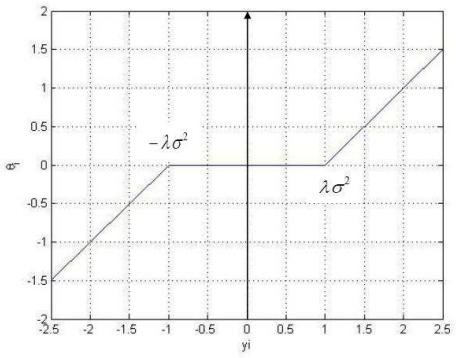

Figure 3: Plot of "soft-threshold"

This is called a "soft-threshold" function. It can be written compactly as:

$$\hat{\theta}_i = sign(y_i) \cdot max(|y_i| - \lambda\sigma^2, 0)$$

The "soft-threshold" estimator is:

$$\hat{\theta} = \begin{bmatrix} \hat{\theta}_1 \\ \vdots \\ \hat{\theta}_n \end{bmatrix} \qquad , \qquad \hat{s} = u\hat{\theta} = \sum_{i:|\hat{\theta}_i| \neq 0} \hat{\theta}_i u_i$$

Note that the soft-threshold estimator automatically selects a signal subspace based on the magnitude/energy of the observed data in each 1-dimension subspace.

5 Summary

$$y = u^T x + w, \qquad w \sim N(0, \sigma^2 I)$$

MLE:

$$\begin{aligned}\hat{\theta}_{MLE} &= \underset{\theta}{argmin}\frac{||x-u\theta||_2^2}{2\sigma^2}\\ &= \underset{\theta}{argmin}\frac{||x-u\theta||^2}{2\sigma^2}\\ &= u^T x\\ &= y\end{aligned}$$

Wiener filter:

$$\hat{\theta}_{wiener} = \left(\frac{\sigma_\theta^2}{\sigma_\theta^2+\sigma^2}\right)\cdot y \quad , \qquad \theta \sim N(0,\sigma_\theta^2 I)$$

$$\text{or} \qquad \hat{\theta}_{wiener,i} = \left(\frac{\sigma_{\theta i}^2}{\sigma_{\theta i}^2+\sigma^2}\right)\cdot y_i \quad , \qquad \theta_i \sim N(0,\sigma_{\theta i}^2 I)$$

$\longrightarrow$ Fixed, non-adaptive trade off.

Soft-threshod: $\theta_i \overset{\text{iid}}{\sim} \frac{\lambda}{2}e^{-\lambda|\theta_i|}$

$$\hat{\theta} = \underset{\theta}{argmin}\frac{||y-\theta||_2^2}{2\sigma^2} + \lambda||\theta||_1$$
$$\hat{\theta}_i = sign(y_i)\cdot max(|y_i|-\lambda\sigma^2, 0)$$

$\longrightarrow$ Data-adaptive shrinkage to trade off bias and variance.

Example 5

$$\sigma^2 = 1, \qquad y = \begin{bmatrix}10\\1\end{bmatrix} \swarrow \textit{ probably just noise}$$

MLE:

$$\hat{\theta}_{MLE} = y = \begin{bmatrix}10\\1\end{bmatrix} \qquad \textit{full dimension}$$

Wiener filter:

$$\hat{\theta}_{wiener} = \left(\frac{\sigma_\theta^2}{\sigma_\theta^2+1}\right)\begin{bmatrix}10\\1\end{bmatrix} \quad \propto \quad \begin{bmatrix}10\\1\end{bmatrix} \qquad \textit{full dimension}$$

Soft-threshod:

$$\hat{\theta}_{wiener} = \begin{bmatrix}max(10-\lambda,0)\\max(1-\lambda,0)\end{bmatrix} \quad \overset{\lambda=1}{=} \quad \begin{bmatrix}9\\0\end{bmatrix} \qquad \textit{shrink to 1-dimension}$$

6 Inverse problems

Suppose we observe a distorted signal s in noise:

$$\begin{aligned}x &= As + w\\ &= Au\theta + w \quad , \quad w \sim N(0,\sigma^2 I)\end{aligned}$$

A is a known matrix, suppose s is sparse in basis u, and write $s = u\theta$.

Wiener Filter (with Gaussian Prior): $\theta \sim N(0, \lambda I)$

$$\hat{\theta}_{wiener} = \underset{\theta}{argmin}\Big(\frac{||x - Au\theta||_2^2}{2\sigma^2} + \lambda||\theta||_2^2\Big)$$

$\longrightarrow$ linear, non-adaptive.

Sparse Solution (Laplacian Prior): "LASSO"

$$\hat{\theta}_L = \underset{\theta}{argmin} \underbrace{\Big(\frac{||x - Au\theta||_2^2}{2\sigma^2} + \lambda||\theta||_1^2\Big)}_{\text{"LASSO"}}$$

$\longrightarrow$ non-linear, adaptive.

Both are convex optimizations.