Denoising using wavelets

WTBV

December 19, 2017

- Threshold functions $s_\lambda(t)$
  - are used to suppress parts of a signal with very low amplitudes and (usually) high frequencies ("noise")
  - examples are

$$\{\text{Hard}, \lambda\} \qquad \begin{cases} 0 & |x| \leq \lambda \\ x & |x| > \lambda \end{cases}$$

$$\{\text{Soft}, \lambda\} \qquad \begin{cases} 0 & |x| \leq \lambda \\ \operatorname{sgn}(x)(|x| - \lambda) & |x| > \lambda \end{cases}$$

$$\{\text{PiecewiseGarrote}, \lambda\} \qquad \begin{cases} 0 & |x| \leq \lambda \\ x - \frac{\lambda^2}{x} & |x| > \lambda \end{cases}$$

$$\{\text{SmoothGarrote}, \lambda, n\} \qquad \frac{x^{2n+1}}{x^{2n} + \lambda^{2n}}$$

$$\{\text{Hyperbola}, \lambda\} \qquad \begin{cases} 0 & |x| \leq \lambda \\ \operatorname{sgn}(x)\sqrt{x^2 - \lambda^2} & |x| > \lambda \end{cases}$$

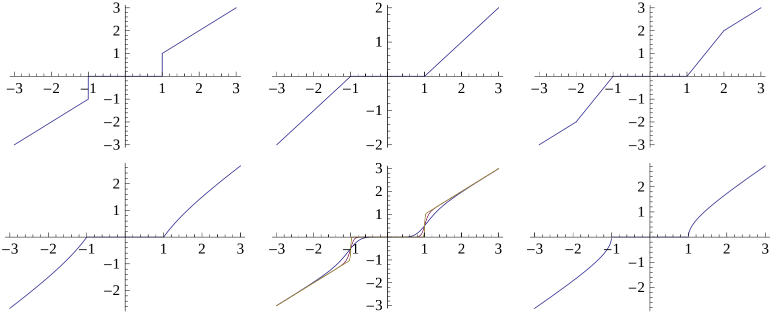

Figure: Examples of threshold functions $s_\lambda(t)$

- Setting and strategy
  - "true" signal : $\boldsymbol{v} = (v_1, v_2, \ldots, v_N)$
  - noise vector: $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N)$
  - noised signal $\boldsymbol{y} = (y_1, y_2, \ldots, y_N) = \boldsymbol{v} + \boldsymbol{\varepsilon}$
  - wavelet filtering (orthogonal transform!)

$$\boldsymbol{y} \overset{WT}{\longmapsto} \boldsymbol{z} = (\boldsymbol{a}, \boldsymbol{d}) = (H\boldsymbol{y}, G\boldsymbol{y})$$

  - applying thresholding with $s_\lambda(t)$ to the high-pass component

$$\boldsymbol{z} \longmapsto \widehat{\boldsymbol{z}} = (\boldsymbol{a}, \widehat{\boldsymbol{d}}) \ \text{ with } \ \widehat{\boldsymbol{d}} = s_\lambda(\boldsymbol{d})$$

  - inverse wavelet transform

$$\widehat{\boldsymbol{z}} \overset{WT^{-1}}{\longmapsto} \widehat{\boldsymbol{v}} = H^\dagger \boldsymbol{a} + G^\dagger \widehat{\boldsymbol{d}}$$

- Heuristic considerations
  - Noise modelled as *Gaussian white noise* with noise level σ: $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N)$ generated by independent and identically $\mathcal{N}(0, \sigma^2)$-distributed random variables
  - For a vector $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_N)$ of independent, $\mathcal{N}(0, \sigma^2)$-distributed random variables ε_i and an orthogonal $(N \times N)$ matrix U the components γ_i of the vector

$$\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_N) = U\,\boldsymbol{\varepsilon},$$

    are again independent $\mathcal{N}(0, \sigma^2)$-distributed random variables
  - Due to orthogonality of the wavelet transform the transformed noise term $WT(\boldsymbol{\varepsilon})$ in

$$\boldsymbol{y} = \boldsymbol{v} + \boldsymbol{\varepsilon} \longmapsto WT(\boldsymbol{y}) = WT(\boldsymbol{v}) + WT(\boldsymbol{\varepsilon})$$

    is still characterized by being white noise with noise level σ

- Heuristic considerations (contd.)
  - In wavelet transformations most energy goes into the approximation (low-pass) component $\boldsymbol{a}$
  - Noise of high frequency goes into the detail (high-pass) component $\boldsymbol{d}$
  - $\Longrightarrow$ the detail component mainly (but not exclusivlely) consists of noise (detail coefficients $\lesssim \sigma$) — that is where to attack!
  - The problem: the noise level σ is not known and has to be estimated from the data to be denoised themselves
  - How to choose λ (depending on the estimate for σ) ?
  - Measure of quality for denoising: *mean squared error* (MSE)

$$\mathrm{E}\left[\|\,\boldsymbol{v} - \widehat{\boldsymbol{v}}\|^2\,\right] = \mathrm{E}\left[\sum_{1 \leq j \leq N} (v_j - \widehat{v}_j)^2\right]$$

- $\boldsymbol{v} = (v_1, \ldots, v_N) \in \mathbb{R}^N$
- $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_N)$ white noise with variance σ^2, $\boldsymbol{y} = \boldsymbol{v} + \boldsymbol{\varepsilon}$
- $\widehat{\boldsymbol{v}} = (\widehat{v}_1, \ldots, \widehat{v}_N)$ estimate for $\boldsymbol{v}$ with $A \subseteq \{1, 2, \ldots, N\}$ and

$$\widehat{v}_j = \begin{cases} y_j & \text{if } j \in A \\ 0 & \text{if } j \notin A \end{cases}$$

- In this case the MSE equals

$$\mathrm{E}\left[\|\boldsymbol{v} - \widehat{\boldsymbol{v}}\|^2\right] = \mathrm{E}\left[\textstyle\sum_{1 \leq j \leq N}(v_j - \widehat{v}_j)^2\right] = \sum_{j \in A} \mathrm{E}\left[\varepsilon_j^2\right] + \sum_{j \notin A} \mathrm{E}\left[v_j^2\right]$$

  and this is minimized by setting $\quad j \in A \Longleftrightarrow v_j^2 > \sigma^2$
- so that the ideal MSE is

$$\mathrm{E}\left[\|\boldsymbol{v} - \widehat{\boldsymbol{v}}\|^2\right] = \sum_{1 \leq j \leq N} \min(v_j^2, \sigma^2)$$

- VisuShrink :
  - is Wavelet shrinkage with
  $$\lambda = \lambda^{\text{univ}} = \sigma \cdot \sqrt{2 \ln N} \qquad (\text{“universal tolerance”})$$
  - Theorem [Donoho-Johnstone, 1995]
  For $\boldsymbol{v} \in \mathbb{R}^N$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_N)$ white noise with noise level σ, for $\widehat{\boldsymbol{v}} = s_\lambda(\boldsymbol{v} + \boldsymbol{\varepsilon})$ (soft threshold) one gets
  $$\mathrm{E}\left[\|\boldsymbol{v} - \widehat{\boldsymbol{v}}\|^2\right] \leq (2 \ln N + 1) \cdot \left(\sigma^2 + \sum_{1 \leq j \leq N} \min(v_j^2, \sigma^2)\right)$$
  - σ will be estimated (on the first high-pass component!) using the *mean absolute deviation* (MAD):
    - $\boldsymbol{w} = (w_1, \ldots, w_N)$
    - $\widetilde{\boldsymbol{w}} =$ Median of $\boldsymbol{w}$
    - $\boldsymbol{v} = (|w_1 - \widetilde{\boldsymbol{w}}|, \ldots, |w_N - \widetilde{\boldsymbol{w}}|)$
    - $\mathrm{MAD}(\boldsymbol{w}) =$ Median of $\boldsymbol{v} = \widetilde{\boldsymbol{v}}$
  - Theorem [Hampel, 1974]
  $$\mathrm{MAD}(\boldsymbol{w}) \approx 0.6745 \cdot \sigma$$

- SURE method (STEINs unbiased risk estimator, 1981)
- Goal: choice of the λ parameter for soft-shrinking methods

$$s_\lambda(x) = \begin{cases} x-\lambda & \text{if } x > \lambda \\ 0 & \text{if } |x| \leq \lambda \\ x+\lambda & \text{if } x < -\lambda \end{cases}$$

- C.M. STEIN, Estimation of the mean of a multivariate normal distribution, *Ann. Stat.* 1981.
  D. DONOHO, I. JOHNSTONE, Adapting to unknown smoothness via wavelet shrinkage, *J. Amer. Stat. Assoc.* 1995.
  P. VAN FLEET, *Discrete Wavelet Transformations*, Wiley, 2008 (ch. 9).

- Lemma:
  For a $\mathcal{N}(0, \sigma^2)$-distributed random variable ε, any $z \in \mathbb{R}$ and any piecewise differentiable function $g : \mathbb{R} \to \mathbb{R}$ one has

$$\mathrm{E}\left[\varepsilon \cdot g(z+\varepsilon)\right] = \sigma^2 \cdot \mathrm{E}\left[g'(z+\varepsilon)\right]$$

- This follows from partial integration:

$$\begin{aligned}
\mathrm{E}\left[\varepsilon \cdot g(z+\varepsilon)\right] &= \frac{1}{\sqrt{2\pi\sigma^2}} \int x \cdot g(z+x) \cdot e^{-\frac{x^2}{2\sigma^2}}\, dx \\
&= \frac{-1}{\sqrt{2\pi\sigma^2}} \int \sigma^2 \cdot g(y) \cdot \frac{d}{dy} e^{-\frac{(y-z)^2}{2\sigma^2}}\, dy \\
&= \sigma^2 \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \int \frac{d}{dy} g(y) \cdot e^{-\frac{(y-z)^2}{2\sigma^2}}\, dy \\
&= \sigma^2 \cdot \mathrm{E}\left[g'(z+\varepsilon)\right]
\end{aligned}$$

- Consequence:
  With z, ε, g as in the Lemma, one gets for the MSE of the estimate $\widehat{z} = w + g(w)$ of the random variable $w = z + \varepsilon$:

$$\begin{aligned} \mathsf{E}\left[(\widehat{z} - z)^2\right] &= \mathsf{E}\left[(\varepsilon + g(z+\varepsilon))^2\right] \\ &= \mathsf{E}\left[(\varepsilon^2 + 2\varepsilon \cdot g(z+\varepsilon) + g(z+\varepsilon)^2\right] \\ &= \mathsf{E}\left[\sigma^2 + 2\sigma^2 \cdot g'(w) + g(w)^2\right] \end{aligned}$$

- Special case: soft-shrinking with threshold value λ
  - The function is

$$g(z) = \begin{cases} -z & \text{if } |z| < \lambda \\ -\lambda \operatorname{sgn}(z) & \text{if } |z| \geq \lambda \end{cases}$$

    and thus

$$\frac{d}{dz} g(z) = \begin{cases} -1 & \text{if } |z| < \lambda \\ 0 & \text{if } |z| \geq \lambda \end{cases}$$

  - Therefore

$$\sigma^2 + 2\,\sigma^2 g'(w) + g(w)^2 = \begin{cases} w^2 - \sigma^2 & \text{if } |w| < \lambda \\ \sigma^2 + \lambda^2 & \text{if } |w] \geq \lambda \end{cases}$$

- General situation:
  - $\boldsymbol{z} = (z_1, z_2, \ldots, z_N) \in \mathbb{R}^N$
  - $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N)$ vector of $\mathcal{N}(0, \sigma_k^2)$-distributed random variables (noise, not necessarily independent and identically distributed)
    $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_N)$
  - $\boldsymbol{w} = \boldsymbol{z} + \boldsymbol{\varepsilon}$ noised vector
  - $\boldsymbol{g} = (g_1, g_2, \ldots, g_N)$ with $g_k : \mathbb{R}^N \to \mathbb{R}$ correcting functions
  - $\widehat{\boldsymbol{z}} = \boldsymbol{w} + \boldsymbol{g}(\boldsymbol{w})$ estimate for $\boldsymbol{z}$
- Theorem (Stein)
  With the notions just introduced, the MSE can be written as

$$
\begin{aligned}
\mathrm{E}\left[\|\widehat{\boldsymbol{z}} - \boldsymbol{z}\|^2\right] &= \mathrm{E}\left[\sum_{1 \le j \le N} \left(\sigma_j^2 + 2\sigma_j^2 \frac{\partial}{\partial w_j} g_j(\boldsymbol{w}) + g_j(\boldsymbol{w})^2\right)\right] \\
&= \|\boldsymbol{\sigma}\|^2 + 2 \sum_{1 \le j \le N} \sigma_j^2 \,\mathrm{E}\left[\frac{\partial}{\partial w_j} g_j(\boldsymbol{w})\right] + \mathrm{E}\left[\|\boldsymbol{g}(\boldsymbol{w})\|^2\right]
\end{aligned}
$$

- Special case: soft-shrinking with threshold value λ

$$\mathrm{E}\left[\|\widehat{\boldsymbol{z}} - \boldsymbol{z}\|^2\right] = \mathrm{E}\left[\sum_{1\le j\le N}\left(w_j^2 - \sigma_j^2 + \left(2\sigma_j^2 - w_j^2 + \lambda^2\right)\chi_{|w_j|\ge\lambda}\right)\right]$$

$$= \mathrm{E}\left[\|\boldsymbol{w} - \boldsymbol{\sigma}\|^2\right] + \mathrm{E}\left[\sum_{1\le j\le N}\left(2\sigma_j^2 - w_j^2 + \lambda^2\right)\chi_{|w_j|\ge\lambda}\right]$$

- The left summand is independent of λ.
  Minimizing the MSE can be achieved by choosing λ depending on the sample vector $\boldsymbol{w}$ so that the integrand in the second summand is minimal!

$$f(\lambda) = \sum_{1\le j\le N}\left(2\sigma_j^2 - w_j^2 + \lambda^2\right)\chi_{|w_j|\ge\lambda}$$

- Instead of dealing with $f(\lambda)$ it is more convenient to consider

$$\widetilde{f}(\lambda) = \sum_{1 \le j \le N} \left(2\sigma_j^2 - w_j^2 + \lambda^2\right) \chi_{|w_j| > \lambda}$$

which changes nothing as far as the expectation is concerned

- Assume that the components of the vector $\boldsymbol{w} = (w_1, \ldots, w_N)$ are ordered by increasing absolute value

$$|w_0| = 0 \le |w_1| \le |w_2| \le \cdots \le |w_N|$$

- The function $\widetilde{f}(\lambda)$ is continuous from the right for $\lambda \in \mathbb{R}_+$, and if $|w_j| < |w_{j+1}|$ holds, then $\widetilde{f}(\lambda)$ is strictly increasing on the half-open interval $[|w_j|, |w_{j+1}|)$, so that it takes its minimum at $|w_j|$:

$$\widetilde{f}(|w_j|) = (N - j)w_j^2 + \sum_{j+1 \le k \le N} (2\sigma_k^2 - w_k^2)$$

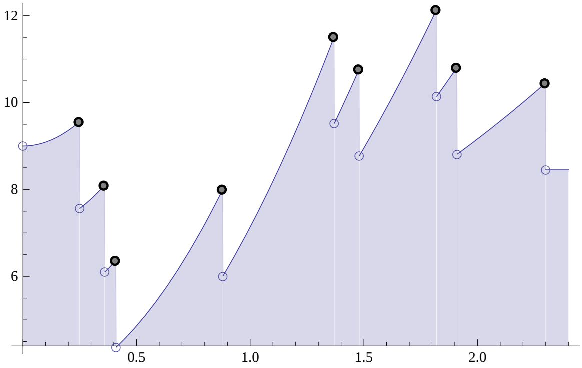

Figure: Example for the computation of the minimum value of $f(\lambda)$ for the sequence $\{0.25, 0.36, 0.41, 0.88, 1.37, 1.48, 1.82, 1.91, 2.3\}$

- From this one gets

$$\min_{\lambda\in\mathbb{R}_+} \widetilde{f}(\lambda) = \min_{0\leq j\leq N} \widetilde{f}(|w_j|) \qquad \lambda^{sure} := \mathrm{argmin}_{|w_j|}\widetilde{f}(|w_j|)$$

- The definition of $\widetilde{f}$ yields a (downward) recursion

$$\widetilde{f}(|w_j|) = \widetilde{f}(|w_{j+1}|) + 2\sigma_{j+1}^2 + (N-j)(w_j^2 - w_{j+1}^2)$$

  which starting from

$$\widetilde{f}(w_N) = 0$$

  gives a fast computation of the minimum!

- The usual assumption $\sigma_1 = \ldots = \sigma_N$ somewhat simplifies the formulas and the computations

SURE in the context of the wavelet transform

- Consider a one-level WT

$$\boldsymbol{z} \longmapsto \boldsymbol{y} = \boldsymbol{z} + \varepsilon \overset{WT}{\longmapsto} (\boldsymbol{a}, \boldsymbol{d}) \longmapsto (\boldsymbol{a}, \widehat{\boldsymbol{d}}) \overset{WT^{-1}}{\longmapsto} \widehat{\boldsymbol{z}}$$

where

$$\boldsymbol{d} = G(\boldsymbol{z} + \boldsymbol{\varepsilon}) = G\boldsymbol{z} + G\boldsymbol{\varepsilon} \overset{s_\lambda}{\longmapsto} \widehat{\boldsymbol{d}}$$

Note: the high-pass component $G\boldsymbol{\varepsilon}$ has the same noise as ε

- σ must be estimated beforehand (as in the VISUSHRINK method)
- One has $\lambda^{sure} \leq \lambda^{univ}$. Recommendation: If $\boldsymbol{y}$ is sparse, it is better to use λ^{univ} instead of λ^{sure}, based on the criterion

$$\frac{1}{N} \sum_{1 \leq j \leq n} (y_j^2 - \sigma^2) \leq \frac{3}{2\sqrt{N}} \log_2(N)$$

(DONOHOE, JOHNSTONE)