Wavelets, sparse representations, and denoising

Mladen Victor Wickerhauser
Washington University in St. Louis, Missouri
victor@math.wustl.edu
http://www.math.wustl.edu/~victor

PMF, University of Zagreb
February 25th, 2009

Bases

$L^2(\mathbf{R})$ is the complete inner product space, or Hilbert space, of measurable square-integrable functions, with

norm: $\|f\| \stackrel{\text{def}}{=} \left(\int_{\mathbf{R}} |f(x)|^2\,dx\right)^{1/2}$

inner product: $\langle f, g\rangle \stackrel{\text{def}}{=} \int_{\mathbf{R}} f(x)\bar{g}(x)\,dx$, so $\|f\| = \sqrt{\langle f, f\rangle}$

$L^2(\mathbf{R})$ has many *countable dense spanning sets* of the form $\mathbf{B} = \{b_i : i \in I\}$ with countable index set I. For example, take $I = \{[a,b] : a, b \in \mathbf{Q},\ a < b\}$ and define $b_i = \mathbf{1}_{[a,b]}$ for $i = [a,b] \in I$. Then $\overline{\text{span}}\,\mathbf{B} = L^2(\mathbf{R})$.

$\mathbf{B}$ is an *orthonormal basis* (ONB) if $\langle b_i, b_j\rangle = \delta_{ij} \stackrel{\text{def}}{=} \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{otherwise;} \end{cases}$

(δ is called the *Kronecker symbol*.) For example, take $I = \{(j,k) : j, k \in \mathbf{Z}\}$ and define

$$b_i = 2^{j/2}\left(\mathbf{1}_{2^j[k,k+\frac{1}{2})} - \mathbf{1}_{2^j[k+\frac{1}{2},k+1)}\right), \qquad \text{for } i = (j,k) \in I.$$

This gives the *Haar ONB*.

Wavelet orthonormal bases

In this talk, a *wavelet* $w \in L^2(\mathbf{R})$ will be a nice function for which

$$\mathbf{B} = \{w_{jk}(x) \stackrel{\text{def}}{=} 2^{j/2} w(2^j x + k) : j, k \in \mathbf{Z}\}$$

is a dense spanning set for $L^2(\mathbf{R})$. Notice that $\|w_{jk}\| = \|w\| = 1$ for all $j, k \in \mathbf{Z}$.

If $\mathbf{B}$ is an ONB, say that w is an *orthonormal wavelet*. An example is the *Haar wavelet*

$$w \stackrel{\text{def}}{=} \mathbf{1}_{[0,\frac{1}{2})} - \mathbf{1}_{[\frac{1}{2},1)}.$$

The *wavelet transform* of a function $u \in L^2(\mathbf{R})$ is its expansion coefficients in $\mathbf{B}$:

$$Wu \stackrel{\text{def}}{=} \{Wu(j,k) = \langle u, w_{jk} \rangle : j, k \in \mathbf{Z}\}.$$

When $\mathbf{B}$ is an ONB, **Parseval's theorem** implies $u \in L^2(\mathbf{R}) \iff Wu \in \ell^2(\mathbf{Z}^2)$, the square-summable double sequences. Equivalently,

$$c \in \ell^2(\mathbf{Z}^2) \iff W^{-1}c \in L^2(\mathbf{R}), \qquad W^{-1}c \stackrel{\text{def}}{=} \sum_{j,k} c(j,k) w_{jk}.$$

Nice wavelets

Various hypotheses are added to get "nice" wavelets w, such as

smoothness: $w \in C^d(\mathbf{R})$, with given $d \in \mathbf{Z}^+$, and $w^{(d)} \in L^2(\mathbf{R})$;

rapid decay: $x^p\, w(x) \to 0$ as $|x| \to \infty$, all $p \in \mathbf{Z}^+$;

compact support: $w(x) = 0$ for all sufficiently large $|x|$;

vanishing moments: $x^p w(x)$ is integrable and $\int_{\mathbf{R}} x^p w(x)\, dx = 0$ for $0 \le p < P$, with given $P \in \mathbf{Z}^+$.

The orthonormal Haar wavelet has compact support and $P = 1$ vanishing moment, but is not smooth (or even continuous).

The *derivative Gaussian* (dG) wavelet, defined for fixed $L \in \mathbf{Z}^+$ by

$$w(t) \stackrel{\text{def}}{=} \frac{d^L}{dt^L} \exp(-t^2),$$

is smooth, has rapid decrease and $P = L$ vanishing moments, but does not have compact support. It is not an orthogonal wavelet for any L.

Approximations

Suppose $\mathbf{B}_0 \subset \mathbf{B}$ is a finite subset of the ONB $\mathbf{B}$; relabel $\mathbf{B}_0 = \{b_i : i = 1, \ldots, M\}$.

Rank-M approximation of $u \in L^2(\mathbf{R})$ in $\operatorname{span} \mathbf{B}_0$ is given by the *orthogonal projection*

$$u_0 = P_0 u \stackrel{\text{def}}{=} \sum_{i=1}^{M} \langle u, b_i \rangle b_i \in \operatorname{span} \mathbf{B}_0$$

By **Bessel's inequality**,

$$\|u - u_0\| = \min_{v \in \operatorname{span} \mathbf{B}_0} \|u - v\|,$$

so u_0 is the best approximation to u available in $\operatorname{span} \mathbf{B}_0$.

Encode the approximation by $u \approx P_0 u = u_0 \longleftrightarrow \{c_i = \langle u, b_i \rangle = \langle u_0, b_i \rangle : i = 1, \ldots, M\}$.

Sparse approximation

If $u \in C^d(\mathbf{R})$ and w is a compactly-supported wavelet with $P > d$ vanishing moments, then by **Taylor's theorem**,

$$\langle w_{jk}, u \rangle = O(2^{[\frac{1}{2}-P]j}), \qquad \text{as } j \to +\infty$$

In addition, if u has compact support, then $\langle w_{jk}, u \rangle \neq 0$ for only $O(2^j)$ values of k. Hence the *decreasing rearrangement* of the wavelet coefficients $Wu(i)$ of u will decrease like $O(1/i^P)$.

Moral: wavelets with many vanishing moments represent smooth functions efficiently.

Probability spaces and random variables

A *probability space* X is a set equipped with

- a topology T consisting of subsets $Y \subset X$ called *events*;
- a nonnegative function $\mathrm{Pr} : T \to \mathbf{R}$ called the *probability measure*.

Pr satisfies $\mathrm{Pr}(X) = 1$, $\mathrm{Pr}(\emptyset) = 0$, and *countable additivity*: if K is a countable index set and the collection $\{Y_k \subset X : k \in I\}$ is pairwise disjoint, then

$$\mathrm{Pr}(\bigcup_{k \in K} Y_k) = \sum_{k \in K} \mathrm{Pr}(Y_k)$$

A *random variable* $v : X \to \mathbf{R}$ is a function nice enough so that for every interval $I \subset \mathbf{R}$, the pre-image of I is an event in the topology:

$$v^{-1}(I) \stackrel{\text{def}}{=} \{\xi \in X : v(\xi) \in I\} \in T.$$

(This is not a troublesome requirement.) Then we say that the *probability* that v takes values in I is given by $\mathrm{Pr}(v^{-1}(I))$, usually written as $\mathrm{Pr}(v \in I)$.

Distributions and densities

Let $v : X \to \mathbf{R}$ be a random variable.

The *distribution function* v is defined for each $x \in \mathbf{R}$ by:

$$\mathsf{Pr}(v \leq x) = \mathsf{Pr}(v \in (-\infty, x]) = \mathsf{Pr}(\{\xi \in X : v(\xi) \leq x\}).$$

The *probability density function* $\rho(x)$ of v is the derivative of this distribution function:

$$\rho(x) \stackrel{\text{def}}{=} \frac{d}{dx}\mathsf{Pr}(v \leq x).$$

We may compute $\mathsf{Pr}(v \in I)$ from ρ using **Lebesgue's theory of integration**:

$$\mathsf{Pr}(v \in I) \stackrel{\text{def}}{=} \int_X \mathbf{1}_I(v(\xi))\,\mathsf{Pr}(d\xi) = \int_{\mathbf{R}} \mathbf{1}_I(x)\frac{d}{dx}\mathsf{Pr}(v \leq x)\,dx = \int_I \rho(x)\,dx.$$

Expected values

The *expected value* of a random variable v, namely its average over X, is likewise computable from ρ:

$$E(v) \stackrel{\text{def}}{=} \int_X v(\xi)\,\mathsf{Pr}(d\xi) = \int_{\mathbf{R}} x \frac{d}{dx}\mathsf{Pr}(v \le x)\,dx = \int_{\mathbf{R}} x\rho(x)\,dx.$$

In general, if $f : \mathbf{R} \to \mathbf{R}$ is a continuous function, then the expected value of $f(v)$ is computed by

$$E(f(v)) \stackrel{\text{def}}{=} \int_{\mathbf{R}} f(x)\rho(x)\,dx,$$

whenever the integral exists.

Functions of a single random variable may be added, and the expected value preserves sums:

$$\begin{aligned} E(f(v)+g(v)) &= \int_{\mathbf{R}} [f(x)+g(x)]\rho(x)\,dx = \int_{\mathbf{R}} f(x)\rho(x)\,dx + \int_{\mathbf{R}} g(x)\rho(x)\,dx \\ &= E(f(v)) + E(g(v)). \end{aligned}$$

Noisy measurement

Suppose $u \in L^2(\mathbf{R})$ is adequately approximated by $u_0 \longleftrightarrow \{c_i : i = 1, \ldots, M\}$.

Suppose the coefficients $\{c_i\}$ acquire *additive errors*:

$$z_i(\xi) \stackrel{\text{def}}{=} c_i + n_i(\xi), \qquad \xi \in X; \quad i = 1, \ldots, M,$$

where X is a set encoding, for example, a particular measurement giving the particular errors $n_i(\xi)$. This is a model of actual measurement. The errors are called *noise*.

Given only $\{z_i(\xi) : i = 1, \ldots, M\}$, for one or perhaps many values of ξ, what can we say about $\{c_i : i = 1, \ldots, M\}$?

Idea: suppose there is an underlying probability space X, events topology T, and probability measure Pr such that $n_i : X \to \mathbf{R}$ (and thus $z_i : X \to \mathbf{R}$) is nice enough to be a random variable with density ρ_i for each $i = 1, \ldots, M$.

Then knowledge of $\{\rho_i\}$ yields estimates of $\{c_i\}$ from $\{z_i\}$.

Probabilistic noise model

A common model for the noise error random variables supposes that they are:

independent: For all $i \neq j$ and all intervals $I, J \subset \mathbf{R}$,

$$\Pr(n_i \in I \text{ and } n_j \in J) = \Pr(n_i \in I) \times \Pr(n_j \in J),$$

normal: $n_i = \mathcal{N}(0, \sigma_i)$, namely for any interval $I \subset \mathbf{R}$,

$$\Pr(n_i \in I) = \int_{x \in I} \rho_i(x)\, dx, \qquad \text{for } \rho_i(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \text{ with } \begin{cases} \mu = 0 \\ \sigma = \sigma_i \end{cases}$$

(In general, $\mathcal{N}(\mu, \sigma)$ means "normal with mean μ and standard deviation σ")

Note that $E(n_i) = 0$ and $E(c_i) = c_i$ for all $i = 1, \ldots, M$. Thus,

$$E(z_i) = E(c_i + n_i) = E(c_i) + E(n_i) = c_i.$$

The expected value of a noisy coefficient is the clean coefficient.

De-noising by averaging

Approximate $E(z_i)$ by an average of repeated measurements:

$$E(z_i) \approx E_K(z_i) \stackrel{\text{def}}{=} \frac{1}{K} \sum_{k=1}^{K} z_i(\xi_k),$$

where $\{\xi_k : k = 1, \ldots, K\} \subset X$ indexes the particular measurements.

Assume ξ_k is randomly located in X and $n_i = \mathcal{N}(0, \sigma_i)$. Then the **Central Limit Theorem** applies:

$$E_K(z_i) = \mathcal{N}(c_i, \sigma_i/\sqrt{K}) \quad \Rightarrow E_K(z_i) - c_i = \mathcal{N}(0, \sigma_i/\sqrt{K}),$$

so for any $\epsilon > 0$ we have

$$E(|E_K(z_i) - c_i|^2) = \frac{\sigma_i^2}{K}.$$

We can therefore expect to get c_i within any given absolute error ϵ by averaging $K > (\sigma_i/\epsilon)^2$ repeated measurements.

However, K may be large and we may have $\{z_i(\xi)\}$ for $K' \ll K$ (or $K' = 1$) values.

De-noising from one sample

Suppose $\widehat{c}_i$, $i = 1, \ldots, M$, is an estimator for the coefficient c_i of the form

$$\widehat{c}_i = \widehat{c}_i(t, \xi) = z_i(\xi) + H_t(z_i(\xi)) = c_i + n_i(\xi) + H_t(z_i(\xi)),$$

where

- t is a fixed parameter (think *threshold*), and
- for any t, $H_t(x)$ is some fixed differentiable function of $x \in \mathbf{R}$.

Interpretation: adjust each measurement z_i individually and independently by subtracting a single-point noise estimate $n_i(\xi) \approx -H_t(z_i(\xi))$ to get

$$c_i = z_i(\xi) - n_i(\xi) \approx z_i(\xi) + H_t(z_i(\xi)) = \widehat{c}_i(t, \xi).$$

Try to choose t so as to minimize variance due to the unknowable ξ.

Risk and optimal risk

Define the *risk* of the coefficient estimator by the *mean squared error*:

$$\mathcal{R} = \mathcal{R}(t) \stackrel{\text{def}}{=} E\left(\sum_{i=1}^{M} |\widehat{c}_i(t) - c_i|^2\right) = \sum_{i=1}^{M} E\left(|n_i + H_t(z_i))|^2\right).$$

The *optimal risk* is the minimum value of $\mathcal{R}$, and we seek an *optimal threshold* t^* giving minimal $\mathcal{R}(t^*)$.

If $\{c_i\}$ and $\{\sigma_i\}$ are known, then we may compute explicitly:

$$t^* \stackrel{\text{def}}{=} \arg\min_{t\geq 0} \mathcal{R}(t) = \arg\min_{t\geq 0} E\left(\sum_{i=1}^{M} |n_i + H_t(c_i + n_i)|^2\right)$$

This is called the *oracle method*.

Approximating the risk

Normally $\{c_i\}$ is unknown (we plan to use $c_i \approx \widehat{c}_i$), so approximate $\mathcal{R}$ with

$$\mathcal{R}(t) \approx \widehat{\mathcal{R}}(t) = \widehat{\mathcal{R}}(t,\xi) \stackrel{\text{def}}{=} \sum_{i=1}^{M} R(\sigma_i, z_i(\xi), t),$$

where

$$R(\sigma, x, t) \stackrel{\text{def}}{=} \sigma^2 + 2\sigma^2 H_t'(x) + H_t^2(x),$$

relying on:

Theorem 1 $E(|\widehat{c}_i(t) - c_i|^2) = E(R(\sigma_i, z_i, t))$ *for each* $i = 1, \ldots, M$.

Summing over $i = 1, \ldots, M$ gives

$$E(\widehat{\mathcal{R}}(t)) = \mathcal{R}(t),$$

so $\widehat{\mathcal{R}}(t) = \widehat{\mathcal{R}}(t,\xi)$ is an *unbiased estimator* for $\mathcal{R}(t)$.

Proof of the main theorem

Estimate each of the summands in $E(\widehat{\mathcal{R}})$ as follows:

$$E(|\widehat{c}_i(t) - c_i|^2) \;=\; E(|n_i + H_t(z_i)|^2) \;=\; E(n_i^2 + 2n_i H_t(z_i) + H_t(z_i)^2)$$

$$= \; E(n_i^2) + E(2n_i H_t(z_i)) + E(H_t(z_i)^2).$$

Now $E(n_i^2) = \sigma_i^2 = E(\sigma_i^2)$, and integration by parts gives:

$$E(2n_i H_t(z_i)) \;=\; \int_R 2x H_t(c_i + x)\rho_i(x)\,dx \;=\; \int_R -2\sigma_i^2 \rho_i'(x) H_t(c_i + x)\,dx$$

$$= \; \int_R 2\sigma_i^2 \rho_i(x) H_t'(c_i + x)\,dx \;=\; E(2\sigma_i^2 H_t'(z_i)),$$

since $2x\rho_i(x) = -2\sigma_i^2 \rho_i'(x)$ for the normal density ρ_i of $n_i = \mathcal{N}(0, \sigma_i)$. Thus

$$E(|\widehat{c}_i(t) - c_i|^2) = E(\sigma_i^2 + 2\sigma_i^2 H_t'(z_i) + H_t(z_i)^2) = E(R(\sigma_i, z_i, t)).$$

Summing over $i = 1, \ldots, M$ gives the result. $\square$

Stein's principle

In 1981, C. M. Stein proposed using

$$\widehat{t} \stackrel{\text{def}}{=} \arg\min_{t \geq 0} \sum_{i=1}^{M} R(\sigma_i, z_i, t)$$

as a *data driven* estimator for t^*. This choice is called *Stein's principle*.

The approximation $\widehat{\mathcal{R}}(t)$ to $\mathcal{R}$ is known as *Stein's unbiased risk estimator* (SURE).

It is not immediately obvious that $\widehat{t}$ is close to t^*. An estimate of $|\widehat{t} - t^*|$ was found in 1991 by Donoho and Johnstone, for particular H_t, using the oracle method.

Soft thresholding

Define *soft thresholding*, or *shrinkage*, by

$$S_t(x) \stackrel{\text{def}}{=} \begin{cases} x - t, & \text{if } x > t; \\ x + t, & \text{if } x < -t; \\ 0, & \text{if } |x| \le t; \end{cases} \qquad \Rightarrow S_t'(x) \stackrel{\text{def}}{=} \begin{cases} 1, & \text{if } |x| > t; \\ 0, & \text{if } |x| \le t. \end{cases} .$$

Thus $S_t(x)$ is differentiable everywhere except $x = \pm t$. Using $\widehat{c}_i = S_t(z_i) = z_i + H_t(z_i)$ identifies

$$H_t(x) \stackrel{\text{def}}{=} \begin{cases} -t, & \text{if } x > t; \\ +t, & \text{if } x < -t; \\ -x, & \text{if } |x| \le t; \end{cases} \qquad \Rightarrow H_t'(x) \stackrel{\text{def}}{=} \begin{cases} 0, & \text{if } |x| > t; \\ -1, & \text{if } |x| \le t, \end{cases} = -\mathbf{1}_{(-t,t)}(x).$$

Given $\{\sigma_i\}$, compute SURE for soft thresholding from

$$R(\sigma, x, t) = \sigma^2[1 + 2H_t'(x)] + H_t^2(x) = \begin{cases} \sigma^2 + t^2, & \text{if } |x| > t; \\ x^2 - \sigma^2, & \text{if } |x| \le t; \end{cases}$$

$$\Rightarrow \widehat{\mathcal{R}}(t) = \sum_{|z_i| \le t} (z_i^2 - \sigma_i^2) + \sum_{|z_i| > t} (\sigma_i^2 + t^2).$$

Renumber $\{z_i\}$ so that $\{|z_i|\}$ is nondecreasing and define

$$F(j) = \sum_{i=1}^{j} (z_i^2 - \sigma_i^2) + \sum_{i=j+1}^{M} (\sigma_i^2 + z_j^2).$$

If $\widehat{\jmath} = \arg\min_{1 \le j \le M} F(j)$, then $\widehat{t} = |z_{\widehat{\jmath}}| = \arg\min_{t \ge 0} \widehat{\mathcal{R}}(t)$.

References and links

Stein, Charles M. Estimation of the mean of a multivariate normal distribution. *Annals of Statistics* 9:6(1981),1135–1151.

Donoho, David, Johnstone, Iain, Kerkyacharian, Gerard, and Picard, Dominique. Density estimation by wavelet thresholding, *Annals of Statistics* 24(1996),508-539.

From web archive `http://www-stat.stanford.edu/~donoho/reports.html`:

Donoho, David, and Johnstone, Iain. Minimax estimation via wavelet shrinkage. Technical Report, Stanford University (1991).

Donoho, David, and Johnstone, Iain. Adapting to unknown smoothness via wavelet shrinkage. Technical Report, Stanford University (1994).

Example: `http://www.quantlet.com/mdstat/scripts/wav/html/wavhtmlnode57.html`